Advancing Coupled Cluster Methods: Tensor Factorization And Analytic Gradient Implementation

Tingting Zhao
Southern Methodist University, tz.tingting.zhao@gmail.com

Follow this and additional works at: https://scholar.smu.edu/hum_sci_chemistry_etds

Part of the Computational Chemistry Commons

Recommended Citation
Zhao, Tingting, "Advancing Coupled Cluster Methods: Tensor Factorization And Analytic Gradient Implementation" (2024). Chemistry Theses and Dissertations. 45.
https://scholar.smu.edu/hum_sci_chemistry_etds/45

This Dissertation is brought to you for free and open access by the Chemistry at SMU Scholar. It has been accepted for inclusion in Chemistry Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
ADVANCING COUPLED CLUSTER METHODS:
TENSOR FACTORIZATION AND ANALYTIC GRADIENT IMPLEMENTATIONS

Approved by:

Dr. Devin A. Matthews
Associate Professor

Dr. Brian D. Zoltowski
Professor

Dr. Weihua Geng
Associate Professor

Dr. Doran I. G. B. Raccah
Assistant Professor
ADVANCING COUPLED CLUSTER METHODS:
TENSOR FACTORIZATION AND ANALYTIC GRADIENT IMPLEMENTATIONS

A Dissertation Presented to the Graduate Faculty of the
Dedman College
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Chemistry
by
Tingting Zhao

B.S., Pharmaceutical Sciences, Southern Medical University

April 24, 2024
ACKNOWLEDGMENTS

First and foremost, I extend my heartfelt gratitude to my mentor and advisor, Dr. Devin A. Matthews. I am immensely thankful for his belief in me, and for his guidance and support in helping me achieve feats I didn’t know I could. Words cannot adequately convey my appreciation for his mentorship. It is a profound honor to work with and learn from him. He serves as a beacon of inspiration and a role model whom I aspire to emulate.

I want to give special thanks to Dr. Brian D. Zoltowski for his invaluable guidance on protein dynamics, which has sparked my fascination with proteins and will continue to inspire me. I also greatly appreciate that he was there to talk during difficult times before joining Dr. Matthews’s group.

I would also like to thank all the members of Matthews’ group whose camaraderie and kindness have enriched my journey, and their presence will be sorely missed. Thanks, Megan, for working with me on my first-ever QM project. Special acknowledgment to Postdoc Avdhoot and James, whose unwavering support has been invaluable.

I owe my dear friends a debt of gratitude for their companionship and unwavering encouragement, especially during moments of uncertainty.

I want to thank Mandy Graham, the chemistry department secretary, and Dr. Michael Lattman, the graduate program director, for their assistance during the Ph.D. program.

I would like to thank my parents for always supporting me with all my decisions.
Last but not least, to my best friend and life partner, Nischal, thank you for always believing in me and encouraging me. Your presence has been my anchor throughout this journey.
With the rapid advancement of computing capabilities, computational chemistry has become increasingly indispensable in both experiment design and the interpretation of experimental outcomes. Wave-function-based quantum mechanistic methods are highly sought after for their ability to provide high-accuracy data and their potential for systematic improvement. However, to extend their applicability to larger molecules, it is essential to employ rank-reducing approximations to these methods. This dissertation dedicates Chapters Three and Four to the development of rank-reduced methodologies. Additionally, beyond single-point energies, molecular geometries and properties of molecule excited states hold paramount importance across diverse fields of chemistry. Hence, Chapter Five presents the work on the implementation of the analytic gradient for the EOM-CCSD* method. Moreover, molecular mechanisms play an indispensable role, particularly in their scalability to biocomplexes. Chapter Six exemplifies this by demonstrating the advantageous application of molecular dynamics in exploring the protein allostery mechanism.
5.1. Introduction ................................................................. 76
5.2. Theory ................................................................. 78
  5.2.1. Determination of the lagrange multiplier $\lambda$ ......................... 79
  5.2.2. Determination of the Lagrange multiplier $\Upsilon$ ......................... 79
  5.2.3. Determination of the Lagrange multiplier $\Pi$ ......................... 79
  5.2.4. Determination of the Lagrange multiplier $Z$ ......................... 80
5.3. Results ................................................................. 81
  5.3.1. Implementation ................................................................. 81
  5.3.2. Application ................................................................. 83
  5.3.3. $n \rightarrow \pi^*$ state of Tetrazine ........................................... 84
5.4. Conclusion ............................................................. 85

6 ALLOSTERIC REGULATION IN STAT3 INTERDOMAINS ................... 86
  6.1. Introduction ............................................................. 87
  6.2. Results ................................................................. 90
  6.3. Conclusion ............................................................. 93

7 Conclusion ................................................................. 94

APPENDIX

A Publication on the implementation of LS-THC on open-shell MP2/MP3 ........ 96
B Draft on the rank-reduced CCSD(T) in the context of high-accuracy thermo-
chemistry ................................................................. 112
C Draft on the first order analytic gradient of EOM-CCSD* ....................... 151
D Publication on the allosteric regulation mechanism of STAT3 protein ........ 179

BIBLIOGRAPHY ......................................................... 201
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td></td>
<td><strong>Matrix multiplication calculation.</strong> (A) Compute architecture demonstrates the relative spatial position of memory units: solid-state drive (SSD), random access memory (RAM), and cache, which are located with respect to the central processing unit (CPU). The physical placements of these components introduce latency in data transfer with 1-12 ns for cache, 62.9 ns for RAM, and 0.1 ms for an SSD. (B) The flow of data between the memory and compute units only occurs between adjacent units. SSDs are the storage medium that can scale to several TBs affordably but can only transfer data to and from RAM with latencies as high as 0.1 ms. RAM is the intermediate data storage that can hold sub-TB of data and is fairly expensive. The data in the RAM is transferred to the cache with a low latency. The cache is the immediate data storage for the CPU, which feeds the data and instructions directly to the CPU for computation. This latency is very low at the range of 1 to 12 ns since it is installed directly on the CPU. This makes it difficult to scale the cache to high capacity, typically at the scale of MBs per core, and it is very expensive to manufacture. (C) Different representation of Matrix multiplication. Matrix multiplication requires large amounts of data to pass through the compute unit. The data must first be loaded onto RAM from SSD and then to cache. If the data fits entirely in the cache, the computing process is straightforward and only scales with the number of operations. However, the data might need to be loaded onto the cache in chunks for a large matrix. Moreover, for some very large matrices, the data may need to be loaded onto the RAM in chunks, which significantly increases the compute cost as the high latency causes the CPU to wait for data for a longer time period.</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td><strong>Tensor size with different systems.</strong> This figure was prepared by Chao Yin.</td>
</tr>
<tr>
<td>2.1</td>
<td></td>
<td><strong>Diagrammatic representation of each fragment of the Hamiltonian operator,$\hat{F}_N$ and $\hat{V}_N$.</strong> The excitation level of each diagram is indicated beneath it. The interaction line is indicated by the dashed horizontal line. This figure is taken and modified from a review [70].</td>
</tr>
</tbody>
</table>
2.2 Canonical HF-case antisymmetrized Goldstone diagrams diagrams for the first-order and second-order wave function. Figure is taken and modified from reference [71]. ................................................................. 24

2.3 HOOI procedure demonstration diagram ................................................................. 44

3.1 Graphic representation of different approximation. (A) MP2a, (B) MP2b, (C) MP3a, (D) MP3b, (E) MP3c or MP3d. ................................................................. 48

3.2 Absolute energy errors for open-shell linear alkyl radicals, H(CH2)n•. The absolute value of the per-electron error between the THC and corresponding DF-MPn calculation is plotted for (a) the total LS-THC-MP3b correlation energy, which is the sum of (b) the LS-THC-MP3b third-order correction and (c) the LS-THC-MP2b correlation energy, and finally (d) the LS-THC-MP2a correlation energy. ε is the Cholesky decomposition cutoff parameter when pruning the grid, varying logarithmically from 10^{-1} to 10^{-4} in steps of 0.2 log units. THC calculations with a range of ε were performed in order to examine the dependence of the error with grid size (a smaller ε results in a larger grid and should yield a smaller error). Certain narrow ranges of ε significantly depart from this trend, e.g. ε ∼ 10^{-1.2} in Fig. 3.2a or ε ∼ 10^{-3.0} in Fig. 3.2d. These irregularities occur due to a sign change in the error. ................................................................. 53

3.3 Errors in incremental solvation energy for 2H-2-azabicyclo[1.1.1]pentane in an explicit water solvation environment. From 1 to 11 waters are added in the order indicated in appendix A. See text for details. .................. 54

3.4 Bond dissociation energy errors of glutathione. (a) Absolute BDE error due to THC for homolytic bond cleavage (open-shell). The bond index refers to labeled bonds in Figure 6. (b) Absolute BDE error due to THC for heterolytic bond cleavage (closed-shell). (c) Absolute errors in the difference between the homolytic and heterolytic BDEs due to THC. .................. 55

3.5 Errors in the relative energies of 24 conformations of 9-propyl-4,11-tridecadienoic acid due to THC. Conformational energy differences are calculated with respect to the lowest-energy conformation (index 19), and the remaining conformations are ordered by increasing energy (using the DF-MP3 energies). (a) Errors for open-shell radical conformations. (b) Errors for closed-shell cationic conformations ................................. 57
3.6 Timings for DF- and LS-THC-MP3 methods for linear alkyl radicals, (H(CH2)ₙ⁻, n=5–20) with the cc-pVDZ basis set. Both axes are on a logarithmic scale to highlight the polynomial scaling. All calculations were performed using a single node with 2 × Intel Xeon E5-2695v4 CPUs and 256 GiB of memory; OpenMP was used to parallelize the calculation over all 36 cores. The inset shows the measured scaling of DF-MP3 and LS-THC-MP3 with ε = 10⁻².² obtained via linear regression of the timings for n = 10—20.

4.1 Graphical representations of the various decompositions employed. See text for details. (A) Left: ERIs (v_{kd} or v_{kj} \text{ or } v_{cm} \text{ or } v_{mk}), right: DF-approximated ERIs; (B) Left: T₂ amplitudes, right: rank-reduced T₂ amplitudes; (C) Left: z_i^{abc}, right: τ_i^{abc}; (D) Left: compressed z_i^{abc}, right: simplified representation; (E) Left: compressed τ_i^{abc}, right: simplified representation; (F) Alternative τ_i^{abc} formed using orthogonal rotation, right: simplified representation.

4.2 Methods explored in this work. (A) Canonical (T); (B) Canonical (T) with density fitting; (C) \tilde{Z} \tilde{D}Z or \tilde{Z}' \tilde{D}Z; (D) \tilde{T}Z; (E) \tilde{Z}T; (F) \tilde{Z}\tilde{D}\tilde{Z}.

4.3 T₂ convergence and T₃ convergence behavior for different methods and four representative reactions. The absolute reaction energy error with the aug-cc-pV\{T,Q\}Z extrapolation is shown on the y-axis. The x-axis indicates the number of T₂ projectors as a multiple n (N_{T₂} = nN_{bas}). Separate colored lines correspond to increasing (red to blue) numbers of T₃ projectors (N_{T₃} = mN_{bas}).

4.4 Error statistics of various decomposition schemes for the benchmark reaction set. The thick lines indicate the mean absolute or absolute percentage errors across the set, while the top of the shaded region indicates the maximum error. Plots are labeled by the extrapolation scheme used, and “aug-cc-pV\{Q,5\}Z Corrected” indicates an indicates a frozen core aug-cc-pV\{Q,5\}Z valence-only energy plus an aug-cc-pCV\{T,Q\}Z core-valance correction.

4.5 “Violin plots” of the mean absolute percent errors (MAPE) of reaction energies with the aug-cc-pV\{T,Q\}Z basis set extrapolation. The break in the y-axis is used to present the extremal data in detail.

4.6 Absolute relative errors (in percent) of each reaction species at the aug-cc-pV\{T,Q\}Z frozen-core level. In each figure, the right axis gives the absolute error in kJ/mol.
6.1 STAT3 structure. 1BG1 was used as the template; Amino-terminal Domain (NTD) is not shown. (A) STAT3 domain structure (Y705 is shown as spheres, D170 is shown as sticks). See below for details of the initial structure. (B) Secondary structures are labeled according to the UniProt database. α helices are colored blue, β sheets are colored red, and unstructured regions (loops) are colored yellow (transverse view). The assigned secondary structures combine information from multiple x-ray crystal structures, thus there is some mismatch with the specific structures used in this work.

6.2 Conformational analysis of SH2 domain binding pockets. (A) PCA of both the pY and pY+3 pockets in the wild type (blue) and D170A mutant (orange). The contour lines show the density of recorded frames in each region, and the crystal structure (PDB ID: 6NUQ) is marked for reference. (B): pY and pY+3 pockets PCA 2D plane colored by different macro-states. (C): The averaged structures for each macro-state. Residues were shown as sticks.

6.3 Rigid core analysis. (A) The proposed allosteric path from D/A170 to Y657 was obtained from REDAN analysis. (B) Conserved hydrogen bond network between CCD, LD, and DBD. (C) Inter-domain pair Cα distances are shown as dashed lines and colored by the normalized standard deviation across all trajectories. (D) Representative structure macrostate 0 (light cyan) compared with macrostate 4 (salmon). (E) Representative structure macrostate 3 (green) compared with macrostate 4 (salmon). (F) Distribution of global helical tilt of α3 in CCD for each macro-state.
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Molecule set selected from W4-17, %TAE[(T)] values are taken from W4-17 SI, and approximately indicate multi-reference behavior.</td>
<td>69</td>
</tr>
<tr>
<td>5.1 Gradients calculated by analytic derivative and finite-difference. Parameter values are in angstroms and degrees.</td>
<td>84</td>
</tr>
<tr>
<td>5.2 Optimized geometry of $n \rightarrow \pi^*$ excited state of formaldehyde with AUG-PVTZ basis set.</td>
<td>85</td>
</tr>
<tr>
<td>5.3 Optimized geometry of $n \rightarrow \pi^*$ state of Tetrazine with basis set AUG-PVTZ</td>
<td>85</td>
</tr>
</tbody>
</table>
Dedicated to my family and my partner Nischal Karki
CHAPTER 1
INTRODUCTION

Understanding chemical and biological processes necessitates qualitative and quantitative analysis achieved through computational calculations. Such computational studies serve as invaluable tools, facilitating the interpretation of experimental data and providing crucial structural, spectral, and reactivity insights that may be inaccessible through experiments alone. Just as experiments employ specialized equipment tailored to different spatial and temporal resolutions, computational chemistry utilizes a diverse array of tools. Each tool is designed to address varying levels of detail resolution and timescale, all while operating within the constraints of modern computational resources. For instance, molecular mechanics (MM) enables the study of biological macromolecules across timescales ranging from nanoseconds to microseconds, while Quantum mechanics/Molecular mechanics (QM/MM) allows for the examination of chemical reactions within macromolecules over the timescale of picoseconds to nanoseconds. Quantum mechanics (QM), on the other hand, provides a means to explore molecular properties with exceptional accuracy within the timescale of picoseconds. In this dissertation, there will be a pronounced emphasis on the development and application of QM methods. However, chapter VI will also include a brief discussion on the utilization of MM techniques to investigate protein allostery. The introductory section serves to provide a concise overview of pertinent terminology that will be elaborated upon in subsequent chapters. Additionally, it offers both motivation and contextual background for the following chapters.

Beginning with the least complex tool, molecular mechanics allows us to study the dynamics of macromolecules that consist of millions of atoms. Force-field development (molecular
mechanics) can be traced back to the 1970s and is motivated to describe the conformational landscape of biomolecular systems [1–4]. It is an empirical model that applies a classical “ball-spring” model to describe the atomic interaction. The parameters of the “ball-spring” model are entirely parameterized using experimental or ab initio computational data. As a consequence of deliberate simplifications in describing atomic interactions, absolute potential energy lacks direct physical significance. Nonetheless, these parameterized force fields are adequate in describing the relative energy differences of conformational landscapes. Macromolecular systems like proteins typically contain thousands of atoms (including water) and can be as large as an entire viral complex [5]. These intentional simplifications render computational calculations of macromolecule systems economically viable compared to the resource-intensive nature of quantum mechanical methods, which are practically impossible.

However, the traditional force field is powerless to answer questions involving the formation and fracture of chemical bonds in the system. Although there are reactive force fields available that offer an approximate description of reactions [6–9], they are often not accurate enough for quantitative studies or may be limited to specific types of reactions. Machine learning force field [10] is being investigated as a potential alternative tool to predict quantum chemical properties. However, a reliable training model requires thousands to millions of high-quality experimental or ab initio target data points. Furthermore, the ability to interpret the experimental data using QM models is essential for understanding chemical behavior. Therefore, it is important to improve both the accuracy and efficiency of QM methods to aid in both the prediction and understanding of chemical and biological applications.

The field of quantum mechanics leaped into existence owing to the Schrödinger equation providing a quantum mechanical way of describing the motion of nucleus and electrons in molecules [11]. It is the equation to solve to understand atomic phenomena. However, it was soon realized that the mathematics was too complex to be practically computed
for any system with more than two electrons. Born-Oppenheimer approximation simplifies the quantum mechanical treatment of molecules by separating the motion of electrons and nuclei [12]. The instantaneous interaction between each pair of electrons described by the electron repulsion term introduces the inseparability of all the electron coordinates [13]. Thus, many approximations had to be introduced to simplify the calculation. Hartree Fock (HF) theory [14, 15] is one of the simplest approximate theories for solving the many-body Hamiltonian. It is based on a simple approximation to the true many-body wavefunction: the Hartree product wavefunction, a product of one electron function. This wavefunction was later expanded as a single Slater determinant of $N$ spin-orbitals to include the antisymmetric property of the wavefunction.

The HF model accounts for approximately 99% of the total energy of a molecule. However, it assumes that each electron moves in an average field created by all other electrons, neglecting some information originating from the correlation between electrons and giving rise to the 1% error. One may stop and wonder, if the error is only 1%, why is it insufficient to answer all the questions? While a 1% error might seem negligible in terms of the total energy, it can significantly affect crucial properties such as bond dissociation energy, reaction energy, or conformational energy differences, which often fall within the 1% range of the total energy. Therefore, achieving high levels of accuracy in predicting molecular structures, properties, and energy landscapes is imperative. This precision enables the identification of reaction intermediates, transition states, and the overall thermodynamics and kinetics information of reactions, with far-reaching implications across material science, drug discovery, catalysis, green chemistry, and environmental science. To attain such accuracy, it is essential to meticulously account for electron correlation. Consequently, the bulk of quantum chemistry research is dedicated to precisely calculating electron correlation. In the subsequent section, I will delve into the concept of electron correlation, elucidating its significance and the computational challenges it presents. Additionally, I will provide a brief overview of my research endeavors pertaining to this intricate aspect of quantum chemistry.
1.1. What is electron correlation energy

The error in total energies introduced by the mean-field approximation in HF compared with the exact total energies is defined as the electron correlation energies [16].

\[ E_{\text{corr}} = E_{\text{error}} = E - E_{HF} \]

Where \( E \) is the exact eigenvalue of the Hamiltonian \( H \) in post-HF methods since exact energy is not available. And since \( E_{HF} \) is an upper bound to the \( E \), the correlation energy is always negative.

In a multi-electron system, each electron is influenced not only by the positively charged nucleus but also by the positions and spins of other negatively charged electrons. This is why electron density is a more sophisticated picture than a uniform distribution of electrons around the nucleus. These electron-electron interactions encompass various phenomena arising from the mutual repulsion and exchange interactions between electrons within atoms and molecules. The electron-electron repulsion described by \( \frac{1}{r_{12}} \) term in Hamiltonian diverges when the distance of two electrons is close to 0. However, the energy has a finite value, thus the wavefunction is considered to have a hole when the distance of two electrons is close to 0, which is called a Coulomb hole. By excluding close electrons, this Coulomb hole decreases the Coulomb interactions, thus lowering the total energies. In the HF model, this is described in the J Coulomb operator and K exchange operator raised from the antisymmetric of the wavefunction. when the distance between a pair of electrons with parallel spins approaches zero, a similar hole emerges in the wavefunction, termed the Fermi hole or exchange hole [17]. However, this phenomenon is not observed for electron pairs with opposite spins. The tendency of opposite-spin electrons to avoid each other is often disregarded, as reflected in the integration of the exchange operator, which yields zero for
opposite spins. This electron correlation arising from anti-parallel-spin electrons is referred to as the dynamical correlation \[18,19\].

Further, system-global (not operator-specific) correlation is also completely discarded in HF since only one Slater determinant(configuration) is considered. In the chemical reaction process, near-degenerate electron configurations usually appear when breaking a bond. The interaction of the near-degenerate electron configurations is called the non-dynamic correlation since the electron configuration is not transferable.

The dynamic correlation can be accounted for by including the electron-pair relaxation. It refers to the correlation arising from the small contribution of a large number of configurations. Whereas the non-dynamic correlation represents correlations involving large contributions from a few configurations. Most post-HF theories try to incorporate electron correlations efficiently and elegantly.

It is worth noting that besides the inherent error due to the deficiency in correlation technique, the truncation of basis set functions introduces another error, which is usually tackled with basis-set extrapolation \[20,21\] and explicit correlation methods \[22,23\]. Although we mention explicit correlation methods here for completeness, it will not be further discussed. Basis-set extrapolation method is applied in chapter four.

1.2. Different theories to approximate electron correlation energy

As defined above, the correlation energy is a measure of the error of the HF model. In wavefunction-based methods, a lot of correction models have been proposed on the basis of the HF model. The main idea of the majority of approaches is to incorporate electron correlation effects by expanding the wave function as a linear combination of multiple electron configurations. Among them, Møller-Plesset (MP) perturbation theory, Coupled cluster (CC), and Configuration Interaction(CI) are the most applicable ones. Detailed equations are collectively presented in Chapter 2.
The simplicity of the implementation and relatively low computational scaling (scale as the fifth order of the system size) have made MP2 the most widely applicable method [24–26]. Very large MP2 calculation, eg $C_{60}$ [27], has been made possible. However, it was observed in the analysis of the trends in the MP2, MP3, and MP4 energies that higher-order terms are needed to overcome the slow and oscillatory convergence of perturbation theory [28–31]. The higher computational cost of including fourth or fifth-order terms encourages the switch of focus to other replacement methods, such as CI and CC [32]. Nevertheless, it’s worth emphasizing that the connection of MP methods with CI and CC methods has been instrumental in advancing the accuracy of correlation methods. For example, the significance of triple excitations was first recognized subsequent to the establishment of the MP4 method [33,34].

Configuration Interaction (CI) is a straightforward correlation method where a linear combination of different excitation configurations (single, double,...) are included to provide more accurate electronic wavefunctions. When all the possible excitation configurations are included, this full CI (FCI) is valuable in providing the "exact" energy for a given combination of extrapolation techniques of basis sets. Unfortunately, FCI is only practical for systems that have a limited number of valence electrons. Truncation of excited configuration can be done to calculate large systems but is done so at the cost of accuracy. Despite its many advantages, such as the straightforward definition of excited states and the simple analytic evaluation, the CI method has a major deficiency - it is not size-consistent.

The introduction of the exponential form wavefunction in coupled cluster theory ensures the inclusion of higher-order excitation as well as size consistency. Coupled cluster single and double (CCSD), owning its accuracy, size consistency, and computational affordability, has made itself the go-to approach for professionals worldwide. However, the iterative method used in CC to solve the amplitudes limited the application of higher-order excitation due to computational requirements. For example, the CCSDT scales as eighth powers of the system.
Numerous approximations $[35–39]$ to $\hat{T}_3$ were proposed, among them, CCSD(T) $[35,38]$ is the most famously known, also referred to as the “gold standard” of quantum chemistry.

In most cases, the above methods with triple excitation can offer a good approximation to the dynamic correlation of the ground state molecules dominated by a single configuration. However, there are cases, such as bond breaking and photo-excitation, where the wave function is dominated by multi-electron configurations. The above methods fail to offer good approximation unless infinite excitation is included since all methods are still on the basis of a single Slater determinant. The inclusion of higher excitation usually leads to challenges in the convergence of energy. Thus, many multireference methods were introduced to treat multi-electron configurations on the same footing. For instance, MCSCF $[40]$, complete active space (CAS) $[41]$, Multireference configuration-interaction (MR-CI) $[42]$, CASPT2 $[43–45]$, and multi-reference coupled clusters $[46–48]$ have been developed. Owing to the favorable accessibility and predictable accuracy over a wide range of chemical problems, equation of motion coupled cluster singles and doubles (EOM-CCSD) has become a to-go method (in the wavefunction-based domain, DFT(TD-DFT) always the cheaper way to go) for predicting excited state properties.

This dissertation mainly focuses on implementing CC theory, specifically CCSD(T) in Chapter 3, EOM-CCSD* in Chapter 4, and MP2/MP3 in Chapter 2, as they are good testing methods for CC theory.

1.3. Why it is expensive to calculate electron correlation energy

Regardless of the various Quantum Mechanics theories, energy calculation always involves tensor (high-dimensional matrix) contraction. Similar to matrix multiplication (Fig. 1.1), tensor contraction involves element-wise multiplication and summation over corresponding indices. In wavefunction-based quantum chemistry methodologies, two key types of tensors
are prevalent: the four-dimensional two-electron integrals (ERIs) and wavefunction amplitudes.

In methods like Coupled Cluster (CC) theory, first-order excitation amplitudes ($t_{ia}^a$) manifest as matrices, while second-order excitation amplitudes ($t_{ia}^{ab}$) assume the form of four-dimensional tensors. For nth-order excitation amplitudes, the dimension expands to $2^n$. Here, $a$ and $b$ denote virtual molecular orbitals, and $i$ and $j$ denote occupied molecular orbitals. Molecular orbitals, constructed as linear combinations of atomic orbitals, are, in turn, composed of Gaussian functions known as basis sets. While the dimensions of occupied orbitals do not change, the dimensions of virtual orbitals increase as a larger basis set is used. Timing and memory constraints are paramount considerations in computational calculations. Timing encompasses both the requisite floating-point operation count and data transfer rates, while memory concerns primarily revolve around storage requirements, typically on SSDs (refer to Figure 1.1). Thus, the dimensions and sizes of ERIs and amplitude tensors serve as crucial limiting factors in computational chemistry. Regrettably, achieving higher accuracy often demands a larger basis set and higher excitation levels. As depicted in Fig. 1.2, the size of the tensor escalates dramatically with the system’s size. For instance, the CCSD(T) method scales exponentially as $N_{\text{basis}}^7$, restricting its computational feasibility to systems with merely dozens of non-hydrogen atoms.

Over the past few decades, numerous methodologies have emerged to mitigate this scaling issue. Broadly, these approaches fall into three categories. First, there are alternative low-cost methodologies to wavefunction-based theories, such as Density Functional Theory (DFT) and Reduced-Density-Matrix (RDM) methods [49–51]. Additionally, Quantum Monte Carlo (QMC) methods have been explored [52, 53]. A brief overview of these techniques is presented in Chapter 2. These approaches typically circumvent the need to solve the wavefunction directly. Functional theory, for instance, posits that the information within the wavefunction is redundant, thus opting for simpler descriptors like density, which uniquely
Matrix multiplication calculation. (A) Compute architecture demonstrates the relative spatial position of memory units: solid-state drive (SSD), random access memory (RAM), and cache, which are located with respect to the central processing unit (CPU). The physical placements of these components introduce latency in data transfer with 1-12 ns for cache, 62.9 ns for RAM, and 0.1 ms for an SSD. (B) The flow of data between the memory and compute units only occurs between adjacent units. SSDs are the storage medium that can scale to several TBs affordably but can only transfer data to and from RAM with latencies as high as 0.1 ms. RAM is the intermediate data storage that can hold sub-TB of data and is fairly expensive. The data in the RAM is transferred to the cache with a low latency. The cache is the immediate data storage for the CPU, which feeds the data and instructions directly to the CPU for computation. This latency is very low at the range of 1 to 12 ns since it is installed directly on the CPU. This makes it difficult to scale the cache to high capacity, typically at the scale of MBs per core, and it is very expensive to manufacture. (C) Different representation of Matrix multiplication. Matrix multiplication requires large amounts of data to pass through the compute unit. The data must first be loaded onto RAM from SSD and then to cache. If the data fits entirely in the cache, the computing process is straightforward and only scales with the number of operations. However, the data might need to be loaded onto the cache in chunks for a large matrix. Moreover, for some very large matrices, the data may need to be loaded onto the RAM in chunks, which significantly increases the compute cost as the high latency causes the CPU to wait for data for a longer time period.

Tensor size with different systems. This figure was prepared by Chao Yin.
define the prosperity of a many-electron system as demonstrated by the Hohenberg-Kohn theorems [54]. Meanwhile, QMC methods endeavor to statistically sample the wave function to derive solutions. Nonetheless, these methods often entail a trade-off, sacrificing some accuracy, particularly for strongly correlated systems.

The following two categories work within the realm of wavefunction-based methods. The primary advantage of wavefunction-based ab initio methods lies in their systematic approach towards approximating the exact solution of the time-independent many-particle Schrödinger equation.

One group of methods exploits the locality of dynamic electron correlation, exemplified by the domain-based local pair natural orbital (DL-PNO) [55], which has empowered the calculation of (T)-like corrections for systems containing hundreds of atoms [56–60]. While early results suggest that achieving 99% accuracy is feasible, reaching an accuracy level of 99.9% is significantly more challenging [55,61].

On the other hand, tensor decomposition methods seek to break down the formidable large tensor into the product of a few lower-dimensional tensors. This diverse array of approaches includes density fitting (DF), Cholesky decomposition (CD), rank-reduced (RR), single value decomposition [62], Tucker decomposition, Density matrix renormalization group, and Tensor Network Methods [63]. Additionally, grid-based tensor numerical methods [64], such as least-squares tensor hypercontraction (LS-THC) and pseudospectral (PS) method, have demonstrated significant efficacy in quantum chemistry. Notably, the potency of randomized methods [65] has gained recognition. All these methodologies adhere to a common principle. Initially, they truncate the size of each dimension by eliminating linear dependence, retaining only crucial eigenvalues and eigenvectors. Subsequently, they reduce dimensionality by decomposing the high-dimensional tensor into a product of lower-dimensional counterparts. The foremost advantage of tensor approximation methods lies in their inherent ability to modulate accuracy and cost with ease.
1.4. Overview of projects working with correlation energy

This dissertation is split into four chapters, excluding introduction, theory, and conclusion, which represent each of the projects that I have worked on. Of those four chapters, Chapters III and IV use tensor approximation methods to optimize computations. Specifically, Chapter III presents the work of applying Least-squares Tensor Hypercontraction to open-shell MP2 and MP3 methods. In Chapter IV, the combination of DF, Rank-reduced, and Tucker decomposition approximation methods was used to reduce the $N^7$ scaling of CCSD(T) while targeting thermochemical accuracy (1 kJ/mol). On the other hand, Chapter V describes my work in solving and implementing the first-order analytic gradient of the excited state modeling method, EOM-CCSD*. Finally, chapter VI presents the work done on identifying the allostery mechanism of Signal Transducers and Activators of Transcription 3(STAT3) protein using molecular mechanisms(MM) and molecular dynamics(MD). The theories related to this dissertation have been briefly discussed in Chapter II. Further details can be found in the publications resulted from these works and will not be discussed here. Thus, each chapter only provides a brief summary of methods and results and focuses on the motivation of my study and the novel contribution of my work.
The following notation is used throughout this work:

- The letters $\mu\nu\rho\sigma$ represent atomic orbital (AO) basis;
- The letters $ijkl$ represent the occupied molecular orbitals in the molecular orbital (MO) basis;
- The letters $abcd$ represent the virtual orbitals in the MO basis;
- The letters $pqrs$ represent the general orbitals in the MO basis;

In quantum mechanics, the allowed energy levels and the corresponding wave functions for a given quantum system are based on the solution of the non-relativistic, time-independent Schrödinger equation:

$$H|\psi\rangle = \varepsilon|\psi\rangle \quad (2.1)$$

This fundamental approach is attractive since molecular structures, energies, and many associated properties can be calculated on the basis of fundamental physical principles, namely the kinetic energy of the electrons and nuclei (red), the Coulomb attraction between electrons and nuclei (purple), the repulsion (blue) between electrons and between nuclei (Eq.
2.2). $H$ is the Hamiltonian operator in atomic units for a system of $M$ nuclei and $N$ electrons.

$$H = -\sum_{i=1}^{N} \frac{1}{2} \nabla_i^2 - \sum_{i=1}^{N} \frac{1}{2M_A} \nabla_A^2 - \sum_{i=1}^{N} \sum_{A=1}^{M} \frac{Z_A}{r_{iA}}$$
$$+ \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{1}{r_{ij}} + \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}}$$  \hspace{1cm} (2.2)

where $Z_A$ represents the nuclear charge, $R_{AB}$ is the distance between two nuclei, $r_{iA}$ is the distance between nuclei and electron, and $r_{ij}$ is the distance between two electrons. In most situations, the motions of electrons are way faster than the motion of nuclei, and thus, the electronic and nuclear wave functions can be treated independently (Born-Oppenheimer approximation [12]). Thus, in Eq. 2.2, the second term and the last term can be considered as constant. The remaining terms are named the electronic Hamiltonian: this Hamiltonian describes the motion of $N$ electrons in the field of $M$ point charges. In most cases, we solve the Schrödinger equation involving the electronic Hamiltonian and the electronic wave functions,

$$H_{elec} = -\sum_{i=1}^{N} \frac{1}{2} \nabla_i^2 - \sum_{i=1}^{N} \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{1}{r_{ij}}$$  \hspace{1cm} (2.3)

$$H_{elec} \psi_{elec} = \varepsilon_{elec} \psi_{elec}$$  \hspace{1cm} (2.4)

$$\psi_{elec} = \psi_{elec}(r_i; R_A)$$  \hspace{1cm} (2.5)

which describes the electron motion that explicitly depends on the electronic coordinates but parametrically depends on the nuclear coordinates. The total energy for fixed nuclei is represented by electronic energy and nuclear repulsion energy.

$$\varepsilon_{tot} = \varepsilon_{elec} + \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}}.$$  \hspace{1cm} (2.6)

Thus, the nuclei in the Born-Oppenheimer approximation move along the potential energy surface obtained by solving the electronic Hamiltonian equation.
In the Born-Oppenheimer approximation, the electronic Schrödinger equation is the equation to solve to study the quantum behavior in chemical systems. However, it is only solved for the Hydrogen-like atoms exactly. This is because the interaction between electrons through their electron-electron repulsion makes it impossible to find an exact wave function for a multi-electron system, which is a single function that depends on the coordinates of all the electrons simultaneously, $\Psi(r_1, r_2, r_3...r_n)$. Thus, many approximations have to be made to tackle the mathematical complexities of the above equation. The single configuration HF wave function is the most widely applied approximation to the ground state wavefunction. Hartree first came up with Hartree product wavefunction [14], $\Psi(r_1, r_2, r_3...r_n) \approx \Psi_H = \phi_1(r_1)...\phi_n(r_n)$, where each electron is described by its own one-electron-like “wavefunction” called an atomic orbital($\phi_1(r_1)$).

However, the Pauli exclusion principle states that two or more identical fermions (particles with half-integer spin) cannot simultaneously occupy the same quantum state within a quantum system. To fully describe this property of electrons, two spin functions($\phi*\alpha, \phi*\beta$) are introduced. However, because the Hamiltonian operator does not contain spin, simply including spin in the wave function does not lead to a satisfactory theory. To obtain a more satisfactory theory, we require a multi-electron wave function to be antisymmetric with respect to the exchange of the coordinates of any two electrons. While the Hartree product wavefunction does not obey the Pauli exclusion principle, Slater extended the Hartree product as a linear combination of Hartree product wavefunction with exchanging signs [66]:

$$\Psi = \Psi_H - P_{12}\Psi_H - P_{13}\Psi_H + ... \quad (2.7)$$
Which can be written as the determinant of a matrix:

\[ \Phi = \mathcal{N} \begin{vmatrix} \mathcal{X}_1(x_1) & \mathcal{X}_2(x_1) & \cdots & \mathcal{X}_n(x_1) \\ \mathcal{X}_1(x_2) & \mathcal{X}_2(x_2) & \cdots & \mathcal{X}_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{X}_1(x_n) & \mathcal{X}_2(x_n) & \cdots & \mathcal{X}_n(x_n) \end{vmatrix} \]  

(2.8)

Where \( \mathcal{X}_i \) is the molecule spin-orbital function (which is a linear combination of atomic orbitals, see Chapter Preface): \( \mathcal{X}_i = \sum_{\mu=1}^{N} c_{\mu i} \phi_{\mu} \sigma_\alpha \), where \( \phi_{\mu} \) is the symmetry adapted atomic orbital function, \( \sigma_\alpha \) is the spin function, \( c_{\mu i} \) are the molecule orbital coefficients mapping atomic orbitals to molecular orbitals, \( N \) is the number of electrons, and \( \mathcal{N} = N!^{-1/2} \) is the normalization factor. It should be noted that while these approximations allow us to model the Schrödinger equation closely, they are still not an exact solution.

### 2.1. Techniques to manipulate Slater determinant

The necessary permutations introduced into the Slater-determinant wave function greatly complicate the equation. A second quantization formulation was introduced to manipulate the Slater-determinant easily [67–69]. Using the anti-commutation relations of second-quantized operators, an arbitrary string of annihilation and creation operators can be written as a linear combination of simplified normal-ordered strings (with fewer operators) multiplied by Kronecker \( \delta \) functions. The normal-ordered strings make the bookkeeping system much easier. In the Fermi vacuum space (Hartree-Fock wavefunction), the normal order means all the q-annihilation \( (a_i^\dagger, a_q) \) operators are put on the right of q-creation \( (a_i^\dagger, a_q) \). Thus, when operating on the Fermi vacuum, any normally ordered operator string containing second-quantized operators is zero. Wick’s theorem provides a recipe for doing the normal ordering. Detailed rules can be found in numerous documentation [70, 71]. For example, the second-
quantized form of the electronic Hamiltonian

\[ \hat{H} = \sum_{pq} \langle p | h | q \rangle a^+_p a_q + \frac{1}{4} \sum_{pqrs} \langle pq | rs \rangle a^+_p a^+_q a_r a_s \]  

(2.9)

\[ \hat{H} = \sum_{pq} \langle p | h | q \rangle \{ a^+_p a_q \} + \sum_{pq} \langle p | i | q \rangle \{ a^+_p a^+_q \} + \frac{1}{4} \sum_{pqrs} \langle pq | rs \rangle \{ a^+_p a^+_q a_r a_s \} + \sum_i \langle i | h | i \rangle + \frac{1}{2} \sum_{ij} \langle ij | ij \rangle \]  

(2.10)

\[ \hat{H} = \sum_{pq} f_{pq} \{ a^+_i a_q \} + \frac{1}{4} \sum_{pqrs} \langle pq | rs \rangle \{ a^+_p a^+_q a_r a_s \} + \langle \Phi_0 | \hat{H} | \Phi_0 \rangle \]  

(2.11)

\[ \hat{H} = \hat{H}_N + \langle \Phi_0 | \hat{H} | \Phi_0 \rangle \]  

(2.12)

where the string of operators inside the curly brackets are normal-ordered. Notice the first two terms in Eq. (2.10) are simply the spin-orbital Fock operator and that the last term is the Hartree-Fock energy. \( \hat{H}_N \) is called the normal-ordered Hamiltonian and can be regarded as the correlation operator. However, Wick’s theorem is rather tedious and error-prone. The diagram representation that was invented by Feynman and populated by Kucharski and Bartlett offered an easy way to construct the energy equations. The basic principle is that the lines represent the action of creating or destroying particles or holes, and the horizontal lines represent the operators. The lines with the downward arrow represent the particles (occupied orbitals in the Fermi vacuum), and the ones with the upward arrow represent holes (virtual orbitals in the Fermi vacuum). If the lines go toward the operator lines (horizontal lines), it means particles or holes are being destroyed; otherwise, it means they are being created. The diagrammatic representation of the above normal-ordered Hamiltonian is presented in Figure 2.1. For example, the \( F^a_i \) means one particle \( i \) is being destroyed and a hole, \( a \), is being created.
Figure 2.1. Diagrammatic representation of each fragment of the Hamiltonian operator, $\hat{F}_N$ and $\hat{V}_N$. The excitation level of each diagram is indicated beneath it. The interaction line is indicated by the dashed horizontal line. This figure is taken and modified from a review [70].
2.2. Hartree Fock

In HF, the antisymmetric wavefunction is based on the independent-particle approximation, a single Slater determinant. With the approximated wavefunction Eq. 2.8 and electronic Hamiltonian Eq. 2.3, we can solve the electronic Schrödinger equation:

\[ E_{RHF} = \langle \Phi | \hat{h}^p_q + \frac{1}{2} \sum_{pqrs} \hat{g}^pq_{rs} | \Phi \rangle \]  

(2.13)

where \( \hat{h}^p_q = \hat{T}_e + \hat{V}_{en} \) and \( \hat{g}^pq_{rs} \) represents the one electron and two electron operator, \( \langle \Phi | \) is the Slater determinant wavefunction.

The introduction of second quantization simplifies the 2.13 into 2.14:

\[ E_{RHF} = \langle 0 | \sum_{pq} \hat{h}^p_q a^+_p a_q + \frac{1}{2} \sum_{pqrs} \hat{g}^pq_{rs} a^+_p a^+_q a_s a_r | 0 \rangle \]  

(2.14)

where \( | 0 \rangle \) is the Fermi vacuum space, which is the reference Slater-determinant. Using Wick’s theorem, we will get,

\[ E_{RHF} = \sum_i h^i_i + \frac{1}{2} \sum_{ij} g^{ij}_{ij} - \sum_{ij} g^{ij}_{ji}, \]  

(2.15)

where \( N \) is number molecular orbital function, \( 2N \) sum of the spin function \( \alpha \) and \( \beta \). and the factor \( \frac{1}{2} \) comes from the sum of \( i \) and \( j \), where it should be \( i < j \). By replacing the MO
with AO, we will get

$$E_{RHF} = \sum_{i} \sum_{\mu \nu} h_{\mu \nu} c_{\mu i}^* c_{\nu i} + \frac{1}{2} \sum_{ij} \sum_{\mu \nu \rho \sigma} (g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) c_{\mu i}^* c_{\nu j}^* c_{\rho i} c_{\sigma j}$$

$$= 2 \sum_{i} \sum_{\mu \nu} h_{\mu \nu} c_{\mu i}^* c_{\nu i} + \frac{1}{2} \sum_{ij} \sum_{\mu \nu \rho \sigma} (2g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) c_{\mu i}^* c_{\nu j}^* c_{\rho i} c_{\sigma j},$$

(2.16)

$$E_{RHF} = \sum_{\mu \nu} h_{\mu \nu} P_{\mu \nu} + 1/4 \sum_{\mu \nu \rho \sigma} (2g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) P_{\mu \rho} P_{\nu \sigma}$$

$$= \sum_{\mu \nu} h_{\mu \nu} P_{\mu \nu} + 1/4 \sum_{\mu \nu} \sum_{\rho \sigma} (2g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) P_{\mu \rho} P_{\nu \sigma}$$

$$= \frac{1}{2} \sum_{\mu \nu} (h_{\mu \nu} + f_{\mu \nu}) P_{\mu \nu}$$

$$= \sum_{\mu \nu} (h_{\mu \nu} + g_{\mu \nu}) P_{\mu \nu}$$

$$= \langle 0 | \hat{h} + \hat{g}_{HF} | 0 \rangle,$$

(2.17)

where $P_{\mu \nu}$ is the atomic orbital (AO) density matrix, $f_{\mu \nu}$ is the Fock operator, $\hat{g}_{HF}$ is the new effective external potential.

$$P_{\mu \nu} = 2 \sum_{i} c_{\mu i}^* c_{\nu i},$$

(2.18)

$$f_{\mu \nu} = h_{\mu \nu} + 1/2 \sum_{\rho \sigma} (2g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) P_{\rho \sigma},$$

(2.19)

$$\hat{g}_{HF} = \frac{1}{4} \sum_{\rho \sigma} (2g_{\mu \nu \rho \sigma} - g_{\mu \nu \sigma \rho}) P_{\rho \sigma}.$$

(2.20)

Variation of the total energy with respect to the MO coefficients $c_{\mu i}$ Eq. (2.21) leads to a set of algebraic equations that can be written in matrix form, which are also called Self-Consistent Field (SCF) equations Eq. (2.24).

$$\frac{\partial}{\partial c_{\mu i}^*} \left( \sum_{\nu \rho} h_{\nu \rho} P_{\nu \rho} + 1/4 \sum_{\nu \rho \sigma \tau} (2g_{\nu \rho \sigma \tau} - g_{\nu \sigma \rho \tau}) P_{\nu \sigma} P_{\rho \tau} - 2 \sum_{jk} \lambda_k (\langle \psi_j | \psi_k \rangle - \delta_{jk}) \right) = 0$$

(2.21)
where Lagrange multipliers, \( \lambda_{kj} \), are included to make sure the orbitals are orthonormal.

\[
\sum_\nu f_{\mu\nu}c_{\nu i} = \sum_j \sum_\nu S_{\mu\nu}c_{\nu i}\lambda_{ji},
\]

\( S_{\mu\nu} = \int \Phi_\mu \Phi_\nu d\tau, E_{ij} = \varepsilon_i \delta_{ij}, \)

\[
\hat{f} = h_{\mu\nu} + 1/2 \sum_{\rho\sigma} (2g_{\mu\rho\sigma} - g_{\mu\sigma\rho}) P_{\mu\nu}
= \hat{h} + 2\hat{J} - \hat{K} = \hat{h} + 2\hat{g}_{HF}.
\]

where \( J \) is the Coulomb operator, \( K \) is the exchange operator, which is non-classical since it arises from the antisymmetry of the determinant wavefunction.

Upon unitary transformation, we can diagonalize the Fock matrix. We can convert the problem into a general eigenvalue problem, which is solvable by standard linear algebra techniques. The diagonalization leads to the canonical orbital energy but delocalized orbitals. Compared with Eq. 2.17, notice that the sum of the Fock operator eigenvalues (orbital energy) is not the HF energy. The Fock operator counts the electron-electron interaction twice. According to Koopmans’ theorem [72], orbital energy is the Ionization Potential \( (\epsilon_k) \) and Electron Attachment \( (\epsilon_a) \) energy.

\[
FCU^H = SCU^H \lambda
\]
\[
FCU^H U = SCU^H \lambda U
\]
\[
FC = SC\epsilon
\]

As outlined in the introductory chapter, the Hartree-Fock (HF) method fails to incorporate correlation energy, resulting in a 1% error in the total energy estimation. Subsequently, several correlation correction methods will be succinctly elucidated.
2.3. Møller–Plesset perturbation theory

Perturbation theory enhances the Hartree-Fock (HF) method by introducing a small perturbation $\hat{V}$ to the unperturbed Hamiltonian operator $\hat{H}^{(0)}$. The choice of perturbation operator and various types of perturbation expansions yield distinct perturbation theories. For instance, Rayleigh–Schrödinger perturbation theory (RS-PT) necessitates knowledge of the zero-order energy and eigenfunctions, whereas Brillouin–Wigner perturbation theory (BWPT) requires awareness of the exact energy within the “Known” $P$ space, thus necessitating an iterative solution approach, as expounded in the literature [71,73].

Møller–Plesset perturbation theory (MPPT) improves the Hartree–Fock method by adding electron correlation effects using RS-PT. In MPPT, the unperturbed Hamiltonian operator is defined as the Fock operator in the HF method. Thus, the MP perturbation operator is given by the following:

$$\hat{H}(\lambda) = \hat{H}^{(0)} + \hat{H}^{(1)} = \sum_p \hat{F}_0 + \lambda \hat{V}, \quad (2.26)$$

$$\hat{V} = \sum_{p \geq q} \frac{1}{r_{ij}} - \sum_{p,i} [(\hat{J}_i(p) + \hat{K}_i(p))] \quad (2.27)$$

where $\lambda$ is used to control the various contributions by their order, $\hat{J}$ is the Coulomb operator, and $\hat{K}$ is the exchange operator. The perturbed wave function and perturbed energy are expressed as a power series of $\lambda$, $\Psi^{(0)} = \Phi_0$:

$$\hat{H}(\lambda)(\Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \cdots) = (E^{(0)} + \lambda E^{(1)} + \lambda^2 E^{(2)} + \cdots)(\Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \cdots) \quad (2.28)$$

According to equation 2.28, we can get 1st:

$$(E_0 - H_0)\Psi^{(1)} = (V - E^{(1)})\Psi^{(0)}. \quad (2.29)$$
To obtain $\Psi^{(1)}$, we take advantage of intermediate normalization [74] that $\langle \Psi^{(0)} | \Psi^{(0)} \rangle = 1, < \Psi^{(0)} | \Psi^{(m)} >= 0$ for $|m\rangle$. Right multiplying $\Psi^{(0)}$, we can get

$$E^{(1)} = \langle \Psi^{(0)} | V | \Psi^{(0)} \rangle = -1/2 \sum_{ij} \langle ij || ij \rangle$$

(2.30)

$$E(\text{MP1}) = E_0 + E^{(1)} = E_{\text{SCF}} - 1/2 \sum_{ij} \langle ij || ij \rangle$$

(2.31)

This reveals that first-order correction is zero.

Left multiplying $\langle \Phi_{ij}^{ab} |$ (instead of $\langle \Phi_i^a |$ since the direct mixing between single excitation and the ground state is prevented by self-consistent optimization of the HF wavefunction), we can get

$$\Psi^{(1)} = \sum_{i>j; a>b} \frac{\langle \Phi_{ij}^{ab} | V | \Psi^{(0)} \rangle}{\varepsilon_{ij}^{ab}} \Phi_{ij}^{ab}$$

(2.32)

where $\langle \Phi_{ij}^{ab} |$ is an eigenfunction of $\hat{H}_0$, $\hat{H}_0 | \Phi_{ij}^{ab} \rangle = \varepsilon_{ij}^{ab} | \Phi_{ij}^{ab} \rangle$, $\varepsilon_{ij}^{ab} = -\varepsilon_i + \varepsilon_a - \varepsilon_j + \varepsilon_b$ is the orbital energy differences. Similarly,

$$E_0 \Psi^{(2)} + E^{(2)} \Psi^{(0)} = H_0 \Psi^{(2)} + V \Psi^{(1)}.$$ 

(2.33)

To obtain the second order energy correction $E^{(2)}$, left multiply $\Psi^{(0)}$:

$$E^{(2)} = \langle \Psi^{(0)} | V | \Psi^{(1)} \rangle$$

(2.34)

$$E^{(2)} = \frac{1}{4} \sum_{abij} \frac{\langle \Phi_0 | V | \Phi_{ij}^{ab} \rangle \langle \Phi_{ij}^{ab} | V | \Phi_0 \rangle}{\varepsilon_{ij}^{ab}}$$

(2.35)

$$= \frac{1}{4} \sum_{abij} \frac{\langle ij | ab \rangle \langle ab | ij \rangle}{\varepsilon_{ij}^{ab}}$$

(2.36)

$$= \langle 0 | V | D \rangle \langle D | V | 0 \rangle$$

(2.37)
Similarly, we can get:

\[ E^{(3)} = \langle 0 | V | D \rangle \langle D | V | D \rangle \langle D | V | 0 \rangle, \]
\[ E^{(4)} = \langle 0 | V | D \rangle \langle D | V | S, D, T, Q \rangle \langle S, D, T, Q | V | D \rangle \langle D | V | 0 \rangle, \]

... 

As we can see, the correlation correction arises from the interaction of excitation states with the ground state. Since the Hamiltonian includes only one- and two-electron terms, double excitations \( |\psi_{ij}^{ab}\rangle \) are the only ones that can directly couple to the ground state in the HF reference. Higher excitations are coupled to the ground state through double excitations indirectly. According to Wigner’s rule [75], knowledge of \( \Psi^{(0)} \) allows us to determine directly the \( E^{2l+1} \). For example, the first-order wavefunction can be used to compute both the second-order and third-order energies, whereas the second-order wavefunction contributes to the fourth- and fifth-order perturbed energies. The perturbed wave function can be viewed as a linear combination of excited determinants. The first-order wavefunction contains only doubly excited determinants (Figure 2.2), while the second-order wavefunction contains contributions from singly (term 4 and 5 in Figure 2.2), doubly (term 1, 2 and 3 in Figure 2.2), triply (term 6 and 7 in Figure 2.2), and quadruply (term 8 in Figure 2.2) excited determinants.

### 2.4. Configuration Interaction

Configuration Interaction (CI) is a straightforward theory to include the correlation energy. In CI, the wavefunction is represented by a linear combination of excited determinants, taking the HF Slater determinant as the reference function,

\[ |\Psi\rangle = c_0 \psi_0 + \sum_{ai} c_i^a \psi_i^a + \frac{1}{4} \sum_{abij} c_{ij}^{ab} \psi_{ij}^{ab} + \ldots \quad (2.38) \]
Figure 2.2. Canonical HF-case antisymmetrized Goldstone diagrams diagrams for the first-order and second-order wave function. Figure is taken and modified from reference [71].
\( \psi_0 \) is the HF Slater determinant, \( \psi_i^a \) is formed by replacing occupied spin-orbital \( i \) in the reference determinant \( \psi_0 \) with virtual spin-orbital \( a \).

### 2.5. Coupled Cluster theory

In coupled cluster theory [70, 71, 76, 77], the correlation feature among electrons is captured by the excitation operators \( \hat{T} \). The excitation operators produce the excited state determinants from \( \Phi_0 \).

\[
\Psi_{CC} = e^{\hat{T}} \Phi_0 = (1 + \hat{T} + \frac{\hat{T}^2}{2!} + \ldots + \frac{\hat{T}^n}{n!}) \Phi_0 \tag{2.39}
\]

\[
\hat{H} |\Psi_{CC}\rangle = E |\Psi_{CC}\rangle \tag{2.40}
\]

\[
\hat{H} e^{\hat{T}} |\Phi_0\rangle = E e^{\hat{T}} |\Phi_0\rangle \tag{2.41}
\]

For CCSD, \( \hat{T} \) was truncated at double excitation:

\[
\hat{T} = \hat{T}_1 + \hat{T}_2
\]

\[
= \sum_i \hat{a}_i^a + \frac{1}{4} \sum_{ij} \hat{a}_{ij}^{ab}
\]

\[
= \sum_{ia} t_i^a a_i^a + \frac{1}{4} \sum_{iab} t_{ij}^{ab} a_i^a a_j^b a_i^a \tag{2.42}
\]

where \( a_i \) deletes the electron from one occupied orbital, and \( a_i^a \) puts the electron in the virtual orbital, creating the one electron excited determinant. \( t_i^a \) and \( t_{ij}^{ab} \) are the amplitudes indicating the weight of that specific determinant. Similar to the MP\(_n\) method, the “Projection” method is used to obtain the energy and cluster amplitudes.

\[
\langle \Phi_0 | \hat{H} e^{\hat{T}} |\Phi_0\rangle = E \langle \Phi_0 | e^{\hat{T}} |\Phi_0\rangle = E \tag{2.43}
\]

\[
\langle \Phi_{ab\ldots} | \hat{H} e^{\hat{T}} |\Phi_0\rangle = E \langle \Phi_{ab\ldots} | e^{\hat{T}} |\Phi_0\rangle \tag{2.44}
\]
where intermediate normalization, \( \langle \Phi_0 | \Phi_{CC} \rangle = 1 \), is assumed. However, Eq. 2.43 and Eq. 2.44 are non-linear (due to the presence of \( e^\hat{T} \)) and amplitude Eq. 2.44 is energy dependent. To solve the problems, Eq. 2.41 is similarity transformed via \( e^{-\hat{T}} \),

\[
\langle \Phi_0 | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Phi_0 \rangle = E \langle \Phi_0 | e^{-\hat{T}} e^{\hat{T}} | \Phi_0 \rangle = E \tag{2.45}
\]

\[
\langle \Phi_{ij...}^{ab...} | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Phi_0 \rangle = E \langle \Phi_{ij...}^{ab...} | e^{-\hat{T}} e^{\hat{T}} | \Phi_0 \rangle = 0 \tag{2.46}
\]

Eq. 2.45 and Eq. 2.46 are the conventional coupled cluster method equations, where \( e^{-\hat{T}} \hat{H} e^{\hat{T}} \) is the similarity-transformed Hamiltonian usually denoted as \( \hat{H} \). The new amplitude equation Eq. 2.46 is decoupled from energy. Moreover, the Baker-Campbell-Hausdorff (BCH) formula of \( e^{-\hat{T}} \hat{H} e^{\hat{T}} \) greatly simplifies the exponential expression into a linear combination expression:

\[
e^{-\hat{T}} \hat{H} e^{\hat{T}} = \hat{H} + [\hat{H}, \hat{T}] + \frac{1}{2!}[[\hat{H}, \hat{T}], \hat{T}] + \frac{1}{3!}[[[\hat{H}, \hat{T}], \hat{T}], \hat{T}] + \frac{1}{4!}[[[[\hat{H}, \hat{T}], \hat{T}], \hat{T}], \hat{T}] + \cdots \tag{2.47}
\]

which terminates at the fourth order since the Hamiltonian only contains up to two electrons and can connect with up to four excitation operators.

Plugging Eq. 2.47 into Eq. 2.46, we get the amplitudes Equation 2.48 [78]. Where \( \hat{H} \) is represented in the normal ordered \( \hat{H}_N = F_N + V_N = \hat{H} - \langle \Phi_0 | \hat{H} | \Phi_0 \rangle \), the subscript c indicates that only those terms in which the Hamiltonian is connected (in the Wick’s theorem sense) to every cluster operator on its right should be included in the algebraic interpretation of
the operator.

\[ \langle \Phi_{ij...} | (F + V + FT + VT + \frac{1}{2} FT^2 + \frac{1}{2} VT^2 + \frac{1}{6} VT^3 + \frac{1}{24} VT^4) | \Phi_0 \rangle = 0 \]  

(2.48)

\[ \langle \Phi_{ij...} | (F + V + F'T + VT + \frac{1}{2} FT^2 + \frac{1}{2} VT^2 + \frac{1}{6} VT^3 + \frac{1}{24} VT^4) | \Phi_0 \rangle = - \langle \Phi_{ij...} | DT | \Phi_0 \rangle \]  

(2.49)

Rewriting \( F \) as \( F = F' + D \), \( D \) is the diagonal element of the fock matrix, we can get the Eq. 2.49. The first guess of \( T \) can be obtained as \( T = - \frac{\langle \Phi_{ab...}|(F+V)|\Phi_0 \rangle}{D} \), and then converged \( T \) will be obtained through iteratively solving Eq. 2.49.

2.6. Connection between MBPT and CC

In perturbation theory, the Hamiltonian was partitioned into unperturbed and perturbed operators, and the HF determinant is taken as the zeroth order wavefunction. The perturbed wavefunctions are represented as a linear combination of excited determinants. Similarly, the \( T \) operator in the CC can be partitioned into different orders according to their contribution to different orders of energy.

\[ T_n = T_n^{(1)} + T_n^{(2)} + T_n^{(3)} + ... T_n^{(n)} \]  

(2.50)

Correspondingly, the effective Hamiltonian can be partitioned into different orders:

\[ \tilde{H} = \tilde{H}^{(0)} + \tilde{H}^{(1)} + \tilde{H}^{(2)} + ... \tilde{H}^{(n)} \]

\[ \tilde{H}^{(0)} = F_N \]

\[ \tilde{H}^{(1)} = V_N + (F_N T_2^{(1)})_c \]

\[ \tilde{H}^{(2)} = (F_N T_1^{(2)} + V_N T_2^{(1)} + \frac{1}{2} F_N (T_2^{(1)})^2)_c \]
Then the second-order energy can be obtained through $E^{(2)} = \langle 0 | \bar{H}^{(2)} | 0 \rangle$. It is easy to recognize that the first term of $\bar{H}^{(2)}$ is zero, and the third term does not contribute since $F_N$ cannot contract with quadruple excitation. Thus $E^{(2)} = \frac{1}{4} \sum_{abij} \langle ij | ab \rangle T_2^{(1)}$, where $T_2^{(1)} = \frac{\langle ab || ij \rangle}{D_{ij}}$. Detailed derivation can be found in paper [70]. As we can see, we arrive at the same equation with MBPT theory. Researchers have been motivated by the relationship between coupled cluster theory and MBPT to develop perturbation-based corrections to the CCSD energy to include higher-order excitation. CCSD(T) is the most successful outcome [32,35].

2.7. Equation of Motion Coupled Cluster

Equation of motion coupled cluster (EOM-CC) is an elegant extension of the CC method to describe a spectrum of electronic states in a single calculation [79,80]. In this method, the CI-like linear excitation operator $R$ is used to generate excited states from the CC ground state. The reference state is described by the $T$ operator, while the orbital relaxation of the excited state is described by the $R$ operator.

\[
R | \Psi_{cc} \rangle = R e^T | \Psi_0 \rangle \tag{2.51}
\]

\[
R = R_0 + R_1 + R_2 + ... \tag{2.52}
\]

EOM methods can describe different excited states by using different excitation operators, EOM-EE (electron excitation), EOM-IP (ionization potential), EOM-EA(electron attachment), and EOM-SF(spin-flip). In the EOM-EE method, the $R$ operator is spin-conserved,$r_0 = 1$ describes the CC the ground state.

\[
R_0 = r_0, \tag{2.53}
\]

\[
R_1 = \sum_{ia} r_i^a \hat{a}_i^a \tag{2.54}
\]

\[
R_2 = \frac{1}{4} \sum_{ijab} r_{ij}^{ab} \hat{a}_i^{a} \hat{b}_j^{b} \tag{2.55}
\]
Similar to the CC theory, by introducing the similarity transformed Hamiltonian $\tilde{H}$, the EOM-CC equation can be treated as an eigenvalue problem.

$$\tilde{H} R |\Psi_0\rangle = E R |\Psi_0\rangle \quad (2.56)$$

The operator $R$ and the energy $E$ can be solved by projecting on excited determinants:

$$\langle \psi_{ij}^{ab} | \tilde{H} R |\Psi_0\rangle = \langle \psi_{ij}^{ab} | E R |\Psi_0\rangle \quad (2.57)$$

To describe the left eigenvector of the non-Hermitian operator, we introduce the $L$ operator:

$$L_0 = l_0 \quad (2.58)$$

$$L_1 = \sum_{ia} l_i^o \hat{a}_i \hat{a}_i^\dagger \quad (2.59)$$

$$L_2 = \frac{1}{4} \sum_{ijab} l_{ij}^{ab} \hat{a}_i \hat{b}_i \hat{a}_j \hat{b}_j \quad (2.60)$$

where $l_0 = 0$ in EOM-CC excited states, $l_0 = 1$ in CC reference state $[79,81]$. And $l$ amplitudes can be determined by:

$$\langle \Psi_0 | L H |\psi_{ij}^{ab}\rangle = \langle \Psi_0 | L E |\psi_{ij}^{ab}\rangle$$

and we usually require the left and right eigenstates to be biorthonormal.

$$\langle \Psi_0 | L_i R_j |\Psi_0\rangle = \delta_{ij}$$
2.8. Alternatives to Wavefunction-based methods

As highlighted in the introduction, wavefunction-based methods are computationally expensive, especially when high excitation levels and large basis sets are needed. Numerous alternative low-cost methodologies to wavefunction-based theories have been developed. The functional theory is that instead of using the wave function, simpler quantities such as density or density matrixes are used to describe the system.

2.8.1. Reduced density matrix methods (RDM)

It was first realized and presented by the mathematician John Coleman that the complex problem of many electrons could potentially be reduced as a linear functional of the two-dimensional RDM since electrons only interact pairwise and are indistinguishable, while many-electron wave function tells us more than we need to know [82,83]. For an N-fermion system, let $\psi(x_1, x_2, ..., x_p, x_q, ..., x_N)$ be the ground state wavefunction in the case of a single Slater determinant. We frequently divide N-indistinguishable particles of our physical system into two sets of p and q particles. The density matrix that describes the interaction among p particles in the existence of the other q set particles can be represented as:

$$D^p(x_1, ..., x_p, x'_1, ..., x'_p) = \int \psi(x_1, ..., x_p, x_q, ...x_N) \psi(x'_1, ..., x'_p, x_q, ...x_N) d(x_q...x_N) \quad (2.61)$$

Since the Hamiltonian of a system of N indistinguishable fermions involves only one- and two-particle interactions, only the first order density matrix ($D^1(x_1, x'_1)$), usually denoted as $D^i_j$ and second density matrix ($D^2(x_1, x_2, x'_1, x'_2)$), usually denoted as $D^{ij}_{kl}$ are needed. In the second quantization expression, these tensors are:

$$D^i_j = \langle \psi | \hat{a}_j^\dagger \hat{a}_i | \psi \rangle \quad (2.62)$$

$$D^{ij}_{kl} = \langle \psi | \hat{a}_k^\dagger \hat{a}_l^\dagger \hat{a}_j \hat{a}_i | \psi \rangle \quad (2.63)$$
and thus, in a reduced density matrix, the diagonal of the first-order density matrix is the electron density. Only two particle coordinates are needed instead of $N$ particle coordinates. Similarly, for 2-RDM only four particle coordinates are needed. Using the 2-RDM as the fundamental variable in electronic structure theory is a cost-effective and practical substitute for wave function-based approaches. However, it requires the use of $N$-representability conditions that ensure the 2-RDM aligns with a physically sound $N$-electron wave function.

2.8.2. Density Functional theory

The basic principle of DFT is that the ground state energy of the system can be expressed as the function of ground state electron density. DFT originates from the Thomas-Fermi model for atoms [84]. Different forms of additional correlation terms were incorporated but still with poor accuracy [85]. Later, Hohenberg and Kohn laid out the universal density functional and the variational theorem, where they proved that the derived energy with the trail density function is always larger than the exact energy [54]. However, DFT was not a practical computational tool until Kohn and Sham reintroduced the orbitals into DFT, which improves the accuracy of the electron’s kinetic energy [86]. The Kohn-Sham equation is in the following form:

$$E_{KS}[\rho] = T_{\text{ref}}[\rho] + V_{eN}[\rho] + J[\rho] + E_{XC}[\rho],$$  

where $E_{T_{\text{ref}}}[\rho]$ is the approximate kinetic energy of electrons calculated from a arbitrary reference system of non-interacting electrons (single Slater determinant), $V_{eN}[\rho]$ is the electron-nuclear repulsion, and electron repulsion term $J[\rho]$, and an additional term $E_{XC}[\rho]$ to correct...
the approximation.

\[ E_{XC}[\rho] = (T[\rho] - T_{ref}[\rho]) + (V_{ee}[\rho] - J[\rho]) \quad (2.66) \]

Different exchange-correlation functionals have been presented to approximate the \( E_{XC}[\rho] \) [87]. One of the most successful and accurate hybrid approximations is the B3LYP exchange-correlation functional:

\[ E_{XC}^{B3LYP} = aE_X^{HF} + bE_X^{B88} + (1 - a - b)E_X^{LDA} + cE_c^{LYP} + (1 - c)E_c^{LDA}, \quad (2.67) \]

where \( a = 0.20, b = 0.72, \) and \( c = 0.81 \) were parameterized by optimizing on a set of experimental atomization energies, ionization potentials, proton affinities, and total atomic energies [88].

2.9. Green’s Function

The time-independent Schrödinger equation, the Hamiltonian eigenvalue problem, is the central equation of quantum mechanics. Since it includes the Laplacian operator, it is also a linear partial differential equation. Green’s function is a powerful tool for many difficult differential equations. Given a general linear differential equation:

\[ \hat{L}\mu(x) = f(x) \quad (2.68) \]

where \( \hat{L} \) is the linear differential operator, \( \mu(x) \) is the function to be solved, \( f(x) \) is the given function named forcing term.

In this case, Green’s function \( G(x, y) \) is the inverse of the linear operator that satisfies \( \hat{L}G(x, y) = \hat{L} \hat{L}^{-1} = \delta(x - y) \) with two variable \( x \) and \( y \), where \( x \) is the given source point, \( y \)
is the reference point, $\hat{L}_{x,y}^{-1}$ is equal to $G(x, y)$, and thus $\hat{L}_{x}^{-1} = \int G(x, y)dy$.

$$\hat{L}G(x, y)\mu(x) = G(x, y)f(x)$$

$$\hat{L}\hat{L}_{x,y}^{-1}\mu(x) = G(x, y)f(x)$$

$$\int y \delta(x - y)\mu(x)dy = \int y G(x, y)f(y)dy$$

$$\mu(x) = \int y G(x, y)f(y)dy$$

(2.69)

As shown above, the method for solving a differential equation that incorporates an arbitrary forcing term $f(x)$ involves a point-by-point integration of the Green’s function against the forcing term. Thus, Green’s functions are not functions but rather distributions. This approach is similar to combining an infinite number of solutions to the equation with the point source to produce the desired arbitrary forcing term. The process is illustrated below:

$$\hat{L}\mu(x) = f(x)$$

$$\hat{L}\int y G(x, y)f(y)dy = \int y \delta(x - y)f(y)dy$$

$$f(x) = \int y \delta(x - y)f(y)dy$$

(2.70)

There are numerous ways of constructing Green’s functions. The above only serves as an example.

Now let’s consider the time-independent Schrödinger equation, $H_0\psi(r) = E\psi(r)$, where $H_0 = -\frac{1}{2}\nabla^2 + V_0(r)$, the Green’s function solution of this homogeneous equation is $G(r, r', E) = (E - H_0(r))^{-1}\delta^3(r, r')$. [89,90]. When a perturbed source $V(x)$ is added to the system, $H(r)\psi(r) = (H_0(r) + V(r))\psi(r) = E\psi(r)$, the forcing term(external source) is $f(r) =$
\[ V(r) \psi(r) \]

\[
(E - H(r)) G(r, r', E) = \delta^3(r, r') \tag{2.71}
\]

\[
G(r, r', E) = (E - H(r))^{-1} \delta^3(r, r')
\]

\[
= (E - H(r))^{-1} \delta^3(r, r')
\]

\[
= (E - H(r))^{-1} [(E - H_0(r))G(r, r', E)_0]
\]

\[
= (E - H(r))^{-1} [(E - H(r) + V(r))G(r, r', E)_0]
\]

\[
= G(r, r', E)_0 + (E - H(r))^{-1} V(r) G(r, r', E)_0
\]

\[
= G(r, r', E)_0 + \int G(r, r_1) V(r_1) G(r_1, r')_0 dr_1 \tag{2.72}
\]

where \( H \) acts on the \( r \)-dependence of \( G \)(the "field" point), while \( r' \)(the "source" point) are \( E \) considered parameters.

### 2.10. Density Matrix Renormalization Group

Density Matrix Renormalization Group methods, also referred to as matrix product state, were coming to researchers’ attention while the fundamental limitations of the density functional approach to “multiconfigurational” electron problems had started to become apparent. This method finds a way to represent larger tensors as a product of small dimensional tensors. Great review papers can be found [91, 92]. Here, the basis principle idea is introduced. Let \( C^{1,2,...,n} \) be the wavefunction amplitudes. The idea is very similar to the resolution of identity, given an \( N \)-dimension tensor, if we insert infinite identity matrix in between \( \sum_i C^{1,2,...,k} |i\rangle \langle i| C^{k,...,n} \), we separate the \( N \)-dimension tensor into two smaller tensors with dimension \( k \) and \( n - k \), respectively, while these two tensors still correlated/entangled by the indices \( i \). One way to represent \( C^{n_1,n_2,...,a_n} \) as a product of tensor \( \sum_k = A[1]_{n_1} A[2]_{n_1,n_2}... A[N]_{n_{n-1}} \) is by a single value decomposition, this process is called infinite DMRG since theoretically infinite site can be generated. However, the dimension of the indices \( N_k \) grows exponentially towards the middle of this tensor product train, thus,
the truncation of the dimension is needed. The truncation of dimension D is often called the bond or virtual dimension. There are several methods to optimize this yielding a variational upper bound for the ground state energy. The simplest way is SVD compression, where we just throw away small single values, which is a local update. A better way is the variation minimization and sweep algorithm (DMRG algorithm).

2.11. Tensor Decomposition

In quantum chemistry calculations, the matrices we encounter often have more than two dimensions. For convenience, matrices with three or more dimensions are named tensors. The following will introduce tensor decomposition methods in the context of the electron repulsion integrals in quantum chemistry calculations. These integrals are defined as:

\[
\langle \mu \nu || \rho \sigma \rangle = \int \phi_\mu(r_1)\phi_\nu(r_1) \frac{1}{r_{12}} \phi_\rho(r_2)\phi_\sigma(r_2) d^3r_1 d^3r_2 \\
- \int \phi_\mu(r_1)\phi_\nu(r_2) \frac{1}{r_{12}} \phi_\rho(r_1)\phi_\sigma(r_2) d^3r_1 d^3r_2
\]  

(2.73)

Owning to the central importance of the two-electron repulsion integral (Eq. 2.73), its approximation has been greatly explored with, for example, density fitting, Cholesky decomposition, pseudospectral, and tensor hypercontraction methods. Note that the wavefunction is represented in the Slater determinant, Eq. 2.73 represents a matrix element of a four-dimensional tensor. The transformation of AOs to MOs scales as \( N^5 \). The general idea of tensor decomposition is to reduce the computational cost in the calculation of energy by either separating the index or truncating the tensor into smaller sizes.

2.11.1. Density fitting (DF)

The general idea of the resolution of identity (RI) [93] is factoring a four-dimensional tensor into two separate three-dimensional tensors by inserting an identity matrix: \( I = \)
\( \sum_m |m\rangle \langle m| \) where \( |m\rangle \) is an orthonormal basis. While Density fitting approximates the original integral in a resolution of the identity (RI) format (the reason it was also called RI approximations), the main idea is expanding a product of basis functions in a new auxiliary basis set \( \{ \alpha_n(r) \} \) (Eq. 2.74) [94–96].

\[
\rho_{pq} = \psi_p(r)\psi_q(r) \approx \bar{\rho}_{pq} = \sum_n C^{pq}_n \alpha_n(r)
\]

(2.74)

Pre-optimized auxiliary basis sets are the most commonly used choice in implementations [97–100]. Various formulations and fitting choices were thoroughly discussed [95], and the Coulomb metric was shown to be the most accurate one.

\[
\Delta_{pq} = \int dr_1 \int dr_2 \frac{[\rho_{pq}(r_1)-\bar{\rho}_{pq}(r_1)][\rho_{pq}(r_2)-\bar{\rho}_{pq}(r_2)]}{r_{12}}
\]

(2.75)

\[
C^M_{pq} = \sum_N (pq|N)[J^{-1}]_{NM}
\]

(2.76)

\[
(pq|N) = \int dr_1 \int dr_2 \psi_p(r)\psi_q(r) \frac{1}{r_{12}} \alpha_N(r_2)
\]

(2.77)

\[
J_{MN} = \int dr_1 \int dr_2 \alpha_N(r_1) \frac{1}{r_{12}} \alpha_M(r_2)
\]

(2.78)

To reduce the cost of constructing the four-index integrals from the three-index integrals, it can be advantageous to use a symmetric expression that splits the matrix \( J^{-1} \) into a product \( J^{-1/2}J^{-1/2} \):

\[
\langle pq|rs \rangle \approx \sum_{MN} (pq|M)(M|N)^{-1}(N|rs)
\]

\[
= \sum_{LMN} (pq|M)(L|N)^{-\frac{1}{2}}(L|N)^{-\frac{1}{2}}(N|rs)
\]

\[
= \sum_L B^L_{pq}B^L_{rs}
\]

(2.79)
\( B_{pq}^L \) is the so-called three-index two-electron integral, detailed implementation algorithm can be found [101]. DF decomposes the four-dimensional ERIs into a product of a three-dimensional tensor.

2.11.2. Cholesky Decomposition (CD)

Similarly, Cholesky decomposition [102,103] exploits the numerical linear dependence by decomposing the Hermitian, positive-definite matrix two-electron repulsion integral matrix \( V = (pq|rs) \) into the product of a lower triangular matrix and its conjugate transpose, the detailed algorithm can be found in paper [103]. The fact that the \( L \) matrix depends on a pair of orbitals rather than four greatly simplifies the transformation of changing orbital basis shown as in 2.82,

\[
\langle \mu \nu | \rho \sigma \rangle = \sum_m L^{m}_{\mu \nu} L^{m}_{\rho \sigma} \\
\psi_p = \sum_{\mu} c_{u}^{p} \phi_{\mu} \\
L^{m}_{pq} = \sum_{\mu \nu} c_{\mu}^{p} L^{m}_{\mu \nu} c_{\nu}^{q}
\]

The advantage of the Density Fitting (DF) and Cholesky Decomposition (CD) methods lies in their ability to transform a four-index integral problem into a three-index integral problem. This transformation notably diminishes the computational burden and storage requirements for intermediate calculations in post-Hartree-Fock (HF) procedures. For instance, it mitigates the expense of transforming Atomic Orbitals (AOs) to Molecular Orbitals (MOs) from \( N^5 \) to \( N^4 \). However, a drawback of DF and CD methods is the inherent coupling of indexes, hindering reduced scaling for exchange-like terms such as \( K_{ps}^{w} = \langle pq|rs \rangle D_{qr}^{w} \). Within DF/CD methods, the indexes \( p \) and \( q \) remain fused, restricting the scaling to \( N^5 \). In contrast, in approaches like PseudoSpectral and LS-THC (as discussed subsequently), where \( p \) and \( q \) are not bound together, the scaling can be diminished to \( N^4 \).
2.11.3. PseudoSpectral(PS)

The basic principle of the pseudospectral (PS) methods [104–106] is to solve the ERIs (partially) in physical space. The ERIs can be represented as follows:

\[
\langle \mu \nu | \rho \sigma \rangle \approx Q^N_{\rho} A^N_{\mu \nu} R^N_{\sigma}, \quad (2.83)
\]

\[
R^N_{\sigma} = \phi_{\sigma}(r_n), \quad (2.84)
\]

\[
A^N_{\mu \nu} = \int \frac{\psi_{\mu}(r)\psi_{\nu}(r)}{r - r_P} dr \quad (2.85)
\]

where R is the collocation matrices with dimension N by M transforming between the spectral space and physical space, N is the number of physical grids, and M is the number of basis functions. To achieve chemical accuracy, it is critical to choose a basis set \( \phi_{m=1,M} \) (spectral space) and a grid \( \{r_1...r_N\} \) representation (physical space) carefully that can transform through a transformation matrix \( R(N \times M) \). If \( N = M \), \( Q^N_{\rho} \) is the inverse of \( R^N_{\rho} \). However, the number of basis functions is often way fewer than the number of grid N. It is computationally infeasible to increase the M to N. Instead, we turn to the least-square fitting of the overlap matrix between the basis function in the spectral space and the basis function represented in the physical space [106, 107].

\[
Q^P_{\mu} = S_{\mu l}[R^Q_{l} w_Q R^Q_{g}]^{-1} R^P_{g} w_P \quad (2.86)
\]

where \( S_{\mu l} \) is the overlap metric between basis functions \( \phi_{\mu} \) and \( \phi_{l} \), and w contains the grid weights. In this factorization, \( K^w_{ps} \) can be calculated with \( N^4 \)

\[
K^w_{ps} = \langle pq|rs \rangle D^w_{qr} \approx Q^P_{P} [A^P_{rs}[R^P_{q} D^w_{qr}]]
\]

38
2.11.4. Least-Squares tensor hypercontraction (LS-THC)

The advantage of LS-THC is it represents a four-dimensional tensor as a product of 5 matrices: $\langle pq|rs \rangle \approx \sum_{RS} X^R_p (X)_r^R V_{RS} (X)_q^S (X)_s^S$. The separation of different MO indexes also reduces the computational cost of the exchange terms. One challenge for this method is the formation of the molecule grid. Several different parent grid has been explored and tested [108]; the Becke family of atom-centered grids are the most widely adopted method [109]. In this method, the grid divides the whole space into a series of atom-centered spherical integrals, each one of which is complete to the whole of 3-dimensional space. The atom-centered spherical integrals consist of a direct product of angular and radial coordinates. For the spherical coordinate, the family of Lebedev-Laikov grids [110] is usually used as the quadrature rule. The Becke-type grid is usually used for the radial coordinates, and the radial nodes(quadrature points) of Treutler-Ahlrichs radial quadrature can be optimized [111]. After constructing the atomic grids, they are then rotated into a standard molecular orientation according to the guidelines established by Gill et al [112, 113]. The final weights for the molecular grid are determined using Becke’s fuzzy Voronoi partitioning scheme, modified by Treutler/Alhrichs [114].

Given a large enough parent grid, the atomic spatial function $\phi(x)$ can be evaluated at each parent grid $X^{RR'}_\mu = \phi_\mu(x_{RR'})$. The fitting can be done in the AO space by optimizing the error between $\langle \mu\nu|\rho\sigma \rangle$ and $\sum_{RS} (X^R_\mu X^R_\nu V_{RS} X^S_\rho X^S_\sigma)$, where $R$ represents the parent grid. However, Parrish et al. found that directly fitting MO integrals improve accuracy over transforming AO integrals after factorization [107, 115]. The molecular orbital collocation matrix is given as $X^{RR'}_p = \sum_{\mu} C_\mu^p \phi_\mu(x_{RR'})$. Besides, the grid can be further pruned to reduce the computational cost [116]. A pivoted Cholesky-decomposition on the metric matrix $S^{RR'}$ can be performed [116], other pruning procedures are possible, e.g. based on domain
decomposition [117,118].

\[
S^{RR'} = \sum_{pq} X^R_{pq} X^{R'}_{pq}, \tag{2.87}
\]

\[
S \approx \Pi R^T R \Pi^{-1}, \tag{2.88}
\]

Where the selected rows of R define the pruned grid \(x_P\) from the original grid \(x_{P'}\) and the pruned collocation matrix \(X_p^P\) from \(X_{p'}^{P'}\). The matrices \(X\) are the collocation matrices, determined \textit{a priori} by evaluation of the (spatial) molecular orbitals \(\psi_p\) at the pruned grid points \(r_R\): \(X^R_p = \psi_p(r_R)\). The pruning process is specific to the occupation of the pair of molecular orbitals. Superscripts such as \((pr)\) differentiate different sets of “pruned” grid points [116], leading to separate occupied–occupied (\(X \equiv X^{(ij)}\)), mixed virtual–occupied (\(\tilde{X} \equiv X^{(ai)} = X^{(ia)}\)), or virtual–virtual (\(\tilde{\tilde{X}} \equiv X^{(ab)}\)) collocation matrices. These three cases may be further classified by the spin of the associated molecular orbitals (differentiated by an overbar on the MO and grid indices), leading to six unique collocation matrices for open-shell LS-THC-MP3 [119].

Once the pruned grid is obtained, the core matrix, \(V_{RS}\), is evaluated by least squares fitting of either the exact ERI tensor Eq. 2.89 with stationary condition Eq. 2.90 or some intermediate approximation. In this work, we utilize density fitting Eq. 2.91 in order to
maintain an overall scaling of $O(N^4)$ [120], and we arrive at the analytical results Eq. 2.92.

$$\Delta \equiv \frac{1}{2} \langle pq|rs \rangle - \sum_{RS} (X^{(pq)})^R_p (X^{(pq)})^R_q V_{RS} (X^{(rs)})^S_r (X^{(rs)})^S_s$$ (2.89)

$$\frac{\partial \Delta}{\partial V_{RS'}} = -X^{R'}_{pq} \Delta_{pqrs} X^{S'}_{rs}$$ (2.90)

$$\tilde{g}_{qs}^{pr} = \sum_{JK} \langle pq|J \rangle \langle J|K \rangle^{-1} \langle K|rs \rangle$$ (2.91)

$$V_{RS} = \sum_{R'S'} (S^{(pr)}_{RR'})^{-1} E_{R'S'} (S^{(qs)}_{S'S})^{-1}$$ (2.92)

$$E_{RS} = \sum_{pqrs} (X^{(pq)})^R_p (X^{(pq)})^R_q \tilde{g}_{qs}^{pr} (X^{(rs)})^S_r (X^{(rs)})^S_s$$ (2.93)

$$S^{(pq)}_{RS} = \sum_{pq} (X^{(pq)})^R_p (X^{(pq)})^R_q (X^{(pq)})^S_p (X^{(pq)})^S_q$$ (2.94)

$$X_{pq}^R = (X^{(pq)})^R_p (X^{(pq)})^R_q$$ (2.95)

where in each case a superscript $\bullet^{-1}$ is understood as an element of the matrix inverse rather than an inverse of the matrix element itself. Note that, as with the collocation matrices, there are as many as six distinct metric matrices $S$. Due to the possible combination of each of these six electron distributions, there are as many as 21 different core ($V$) and fitting ($E$) matrices, of which 13 are utilized in open-shell LS-THC-MP3 (assuming a canonical Hartree–Fock reference for which the single excitation amplitudes can be neglected). In order to avoid notational clutter, we rely on context to determine which of these 13 core matrices is intended, unless explicitly specified.

The LS-THC form of the ERIs [107] is thus,

$$\langle pq|rs \rangle \equiv g_{rs}^{pq} \approx \sum_{RS} (X^{(pq)})^R_p (X^{(pq)})^R_q V_{RS} (X^{(rs)})^S_r (X^{(rs)})^S_s$$ (2.96)
2.11.5. Laplace transform

We must deal with the inseparability of the energy denominators present in Eq. (3.1)–(3.9). A convenient approach is the method of “Laplace denominators”, pioneered by Almlöf and others, \[121,122\]

\[
\frac{1}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} = - \int_0^\infty e^{-(\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j) t} dt \\
\approx - \sum_{g=1}^{N_g} w_g e^{(\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j) t_g} \\
= \sum_{g=1}^{N_g} \tau_a^g \tau_b^g \tau_i^g \tau_j^g 
\]  \hspace{1cm} (2.97)

where \(N_g\) is the number of Laplace quadrature points. In this work we use the quadrature developed by Braess and Hackbusch \[123\]. This Laplace transform technique has been known for some time and works remarkably well with between three and twelve Laplace points depending on the level of accuracy required and whether or not core orbitals are included. \[122,124–126\]

2.11.6. Tucker decomposition

The Tucker decomposition is a form of higher-order singular value decomposition (HOSVD). It decomposes a tensor into a core tensor multiplied by a matrix along each dimension. For example, a 3-D tensor \(T^{IJK}\) can be approximated as a smaller 3-D core tensor \(t^{PQR}\) multiplied by three factor matrix \(A_I^P, B_J^Q\) and \(C^K_R\), where \(P < I, Q < J, R < K\).

\[
T^{IJK} \approx \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} t^{PQR} A_I^P B_J^Q C^K_R 
\]  \hspace{1cm} (2.98)

However, obtaining via a HOSVD of a full, \(T_3\)-like quantity, would be very expensive. Instead, following the work of Lesiuk \[127\] (and the earlier work of Bell, Lambrecht, and Head-Gordon
in the context of MP2 [128]) we employ higher-order orthogonal iteration (HOOI) which constructs the factor matrix in a self-consistent manner. In most of what follows, we talk about this in the context of CCSD(T) in chapter four. The Tucker-3 compression of the $Z_3$ intermediates is defined as,

$$z_{ij}^{abc} \approx \tilde{z}_{ij}^{abc} = \sum_{XYZ} Z_{ai}^X U_{bj}^Y U_{ck}^Z$$

(2.99)

The initial guess for $U_{ai}^X$ is seeded from the eigenvectors of $T_2$. These are then partially contracted against the three-electron property in question, such as $z_{ij}^{abc}$, obtaining for example $Z_{ck}^{XY}$,

$$Z_{ck}^{XY} = \sum_{ai bj} z_{ij}^{abc} U_{ai}^X U_{bj}^Y$$

(2.100)

Because the projectors $U$ are orthogonal but rectangular, they form an approximate resolution of the identity. Thus, $Z_{ck}^{XY}$ contracted with itself to form a matrix $Z_{ck,c'k}$, provides eigenvectors and eigenvalues approximating the right singular vectors and values of $z_{ij}^{abc}$ formatted as a matrix $z_{ai bj,ck}$,

$$Z_{ck,c'k'} = \sum_{XY} Z_{ck}^{XY} Z_{c'k'}^{XY} = \sum_{ai bj a'i'bj'} z_{ij}^{abc} \left( \sum_{X} U_{ai}^X U_{a'i'}^X \right) \left( \sum_{Y} U_{bj}^Y U_{b'j'}^Y \right) z_{i'j'k'}^{a'b'c'}$$

$$= \sum_{Z} U_{ck}^Z \epsilon_{ck}^Z U_{c'k'}^Z$$

(2.101)

$$z_{ij}^{abc} = z_{ai bj,ck} = \sum_{Z} I_{ai bj}^Z \sigma_{ck}^Z R_{ck}^Z$$

(2.102)

$$U_{ck}^Z \approx R_{ck}^Z, \quad \epsilon_{ck}^Z \approx (\sigma_{ck}^Z)^2$$

(2.103)

where we assume a self-consistent solution of $U$. The positive square roots of $\epsilon_{ck}^Z$ are used to determine the number of projectors ($N_{T_3}$) retained in the compressed space. The associated
Higher-Order Orthogonal Iteration (HOOI)

Initial guess of $U_{ai}^{X}$

\[ \left| N_{T3} - \sum_{X} \sigma^{X} \right| < 10^{-5} \]

False

\[ Z_{Al}^{YZ} = U_{bj}^{Y} U_{ck}^{Z} W_{ijk}^{abc} = U^{(n)} U^{(n-1)} W_{ijk}^{abc} \]

EVD

\[ Z_{ij}^{AB} = W \Sigma W^{T}, U^{n} = W \]

\[ M_{XY} = \sum_{ai} U_{ai}^{X(n)} U_{ai}^{Y(n-1)} \]

SVD: $M_{XY} = v \sigma w^{T}$

$U_{ai}^{X}$

Figure 2.3. HOOI procedure demonstration diagram

eigenvectors are then used as the next iteration of guesses for the Tucker-3 compression vectors. Convergence is checked by computing the singular values $\sigma^{X}$ of the overlap matrix $M_{XY} = \sum_{ai} U_{ai}^{X(n)} U_{ai}^{Y(n-1)}$ for projectors from successive iterations $n - 1$ and $n$. If the value $|N_{T3} - \sum_{X} \sigma^{X}|$ falls below some cutoff, chosen here as $10^{-5}$, then the process is stopped. A schematic of this procedure is shown in Fig. 2.3. The final core tensor $Z^{XYZ}$ is then obtained by compressing $z_{ijk}^{abc}$ with the converged $U$ factors,

\[ Z^{XYZ} = \sum_{abjck} z_{ijk}^{abc} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} \]

(2.104)
Chapter 3
IMPLEMENTATION OF LS-THC ON OPEN-SHELL MP2/MP3

3.1. Introduction

As outlined in the introduction chapter, utilizing tensor decomposition methods to approximate wave-function-based techniques offers distinct advantages, notably in terms of easily adjustable accuracy. The fourth-order electron repulsion integral (ERI) tensor, denoted as $g^{pq}_{rs} = \langle pq | rs \rangle$, is an essential component of all electronic structure methods. Consequently, it is an appealing focal point for computational scaling reduction efforts. Various approaches have been developed to reduce the scaling of the ERIs. Among these, the least-squares tensor hypercontraction (LS-THC) combines the desirable attributes of several other factorization techniques. Specifically, it incorporates: 1) The capacity to represent electron distributions via a linear scaling auxiliary basis as in Density Fitting (DF)/Resolution of Identity (RI) and Cholesky Decomposition (CD) methods. 2) The flexibility to factor exchange terms, a desirable characteristic found in pseudospectral method. 3) The use of least-squares fitting as in DF and alternating least-squares-based CP factorization. 4) A grid-based expansion similar to the numerical integration in exchange-correlation functions of density functional theory (DFT).

Owing to its desirable features, numerous implementation works with LS-THC have been explored across various theoretical methods, such as MP2 [107, 129], MP3 [107], CC2 [130, 131], and $T_2$ amplitudes in CCSD [132]. However, all of these implementations and tests have been conducted solely within the realm of closed-shell reference. Closed-shell systems are crucial to the field of chemistry, as both reactants and products typically consist of neutral, stable chemical species that favor a closed-shell configuration. On the other hand,
open-shell species are also of great significance due to the high reactivity and complexity of
electronic structure displayed by radical species. Moreover, the inherently reactive nature
of radical transient intermediates or transition states makes them challenging to study in
the laboratory, requiring more advanced instrumentation or complex spectroscopic analysis.
Therefore, a pressing need exists to develop quantum chemistry methods to aid the experi-
mental investigation of such systems alongside the development of approximation techniques
aimed at scaling down computational demands, particularly for the study of larger systems.
Besides the importance of open-shell species, the nature of open-shell systems also poses
additional challenges to the theoretical model that closed-shell systems may not encounter.
In closed-shell systems, all the electrons are paired, and the spin symmetry can be eas-
ily enforced by making the spatial part of the $\alpha$ and $\beta$ spin-orbital the same. Conversely,
open-shell systems, such as doublet and triplet radicals, biradicals, and open-shell singlets
(excited states), harbor unpaired electrons, leading to differing spatial parts in the $\alpha$ and
$\beta$ spin-orbitals. It is widely recognized that treating electron correlation explicitly in such
systems poses significant challenges [133–135]. These challenges include spin contamination,
instability and near-singularity of the HF-SCF solution, strong (non-dynamical) electron
correlation effects, and so on. For instance, maintaining spin symmetry becomes intricate,
as the spin state of electrons should remain invariant to spin-free operators. Therefore, while
the benefits of LS-THC approximation are evident in closed-shell systems, their translation
to open-shell systems is not straightforward.

MP2/MP3 serves as an ideal precursor to delving into more intricate CC theory, thanks
to the diverse array of interpretations MP3 offers ("MP3a," "MP3b," "MP3c," "MP3d" see
below) and the comprehensive set of foundational elements encapsulated within MP3 energy,
notably the particle-particle ladder, hole-hole ladder, and “ring” terms. Moreover, the im-
plementation and testing of MP2/MP3 are comparatively more straightforward. This study
introduces an extension of the LS-THC approach to open-shell MP2 and MP3 energies. The
derivation of working equations employed a mixed graphical–algebraic technique. Specifi-
cally, our implementation focuses on the MP3b variant of LS-THC-MP3, with the developed techniques readily applicable to other MP3 variants and more intricate theories such as CC theory.

3.2. Method

In the canonical spin-orbital representation, the MP2 and MP3 energies are defined as,

\[
E_{\text{MP2}} = \frac{1}{2} \sum_{abij} (g_{ab} - g_{ba}) g_{ij}^{ab} \frac{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}{(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b)(\varepsilon_k + \varepsilon_l - \varepsilon_a - \varepsilon_b)}
\]  

\[
E_{\text{MP3}} = \frac{1}{2} \sum_{abijkl} (g_{ij}^{ab} - g_{ij}^{ba}) g_{kl}^{ij} g_{ij}^{ab} \frac{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}{(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b)(\varepsilon_i + \varepsilon_j - \varepsilon_c - \varepsilon_d)} + \frac{1}{2} \sum_{abcdij} (g_{ij}^{ab} - g_{ij}^{ba}) g_{cd}^{ij} g_{ij}^{cd} \frac{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}{(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b)(\varepsilon_a + \varepsilon_k - \varepsilon_a - \varepsilon_c)}
\]  

(3.2)

Where \(g_{rs}^{pq}\) is the ERI, and \(\frac{1}{(\varepsilon_i + \varepsilon_k - \varepsilon_a - \varepsilon_c)}\) is the orbital energy denominator. \(a, b, c, d...\) indicates the virtual orbitals, while \(i, j, k, l...\) represents the occupied orbitals.

LS-THC theory and implementation details can be found in the section 2.11.4 and the original paper [119]. Briefly, given a ”pruned” grid, the MOs are evaluated on each grid point and represented by the collocation matrix \(X_p^R\). The core matrix \(V_{RS}\) is evaluated by least squares fitting of either the exact ERI tensor or some intermediate approximation. In this work, we utilize density fitting in order to maintain an overall scaling of \(O(N^4)\) [120],

\[
(pr|qs) \equiv g_{rs}^{pq} \approx \sum_{RS} (X^{(pr)})_p^R (X^{(pr)})_r^R V_{RS} (X^{(qs)})_q^S (X^{(qs)})_s^S
\]  

(3.3)

Superscripts such as \((pr)\) differentiate different sets of ”pruned” grid points [116]. The pruning process is specific to the occupation of the pair of molecular orbitals, leading to separate occupied–occupied \((X \equiv X^{(ij)})\), mixed virtual–occupied \((\tilde{X} \equiv X^{(ai)} = X^{(ia)})\), or
Two different approximation methods can be applied to MP2, namely "MP2a" and "MP2b". Four different approximation methods are possible for MP3, namely, "MP3a", "MP3b", "MP3c," and "MP3d". Only "MP3b" is implemented in this work, but the same principle applies to other variants. A diagrammatic representation is presented in Figure 3.1. "MP2a" (Fig. 3.1 A) and "MP3a" (Fig. 3.1 C) are the same as the Eq. 3.1 and Eq. 3.2 except $g_{rs}^p$ is approximated by LS-THC. In "MP2b" (Fig. 3.1 B) and "MP3b" (Fig. 3.1 D), the first order $T_2$, which is $g_{rs}^p$ with denominators, is fitted instead of $g_{rs}^p$. During the fitting $E_{PQ} = X_a^P X_i^P X_b^Q X_j^Q l_a l_i b_j l_j a_i b_i a_j b_j l_a l_j X_a^R V_{RS} X_b^S X_j^S$ is used. In "MP3c" and "MP3d" (Fig. 3.1 E,F), the second order $T_2$ is fitted instead. In "MP3d", the orange box in Fig. 3.1 F is fitted first, and the Laplace transformed denominator is included in the second round of

---

**Figure 3.1.** Graphic representation of different approximation. (A) MP2a, (B) MP2b, (C) MP3a, (D) MP3b, (E) MP3c or MP3d.
fitting (green box in Fig. 3.1 F).

\[
E_{MP^{2b}} = \frac{1}{2} \sum_{abij} (g_{ij}^{ab} - g_{ba}^{ij}) \tilde{t}^{[1]}_{ij} ab
\]

\[
E_{MP^{3b}} = \frac{1}{2} \sum_{abijkl} (\tilde{t}^{[1]}_{ij} ab - \tilde{t}^{[1]}_{ij} ba) g_{kl}^{ij} \tilde{t}^{[1]}_{kl} cd
\]

\[
+ \sum_{abcijk} (\tilde{t}^{[1]}_{ij} ab - \tilde{t}^{[1]}_{ij} ba) (g_{ik}^{bk} - g_{ik}^{bk}) (\tilde{t}^{[1]}_{ij} ac - \tilde{t}^{[1]}_{ij} ac)
\]

\[
E_{MP^{3c}} = \frac{1}{2} \sum_{abij} (g_{ij}^{ab} - g_{ba}^{ij}) \tilde{t}^{[2]}_{ij} ab
\]

\[
\tilde{t}^{[n]}_{ij} = \tilde{t}^{[n]}_{ij} ab - \tilde{t}^{[n]}_{ij} ba = \tilde{t}^{[n]}_{ij} ab - \tilde{t}^{[n]}_{ij} ba
\]

\[
\tilde{t}^{[1]}_{ij} ab = \frac{g_{ij}^{ab}}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}
\]

\[
\tilde{t}^{[2]}_{ij} ab = \sum_{kl} g_{kl}^{ij} \tilde{t}^{[1]}_{ij} ab + \sum_{cd} g_{cd}^{ij} \tilde{t}^{[1]}_{ij} cd + \sum_{ck} (g_{jc}^{bk} - g_{jc}^{bk}) (\tilde{t}^{[1]}_{ij} ac - \tilde{t}^{[1]}_{ij} ac)
\]

Note that \( g_{pq}^{rs} \) and \( t_{ij}^{ab} \) are non-antisymmetrized as avoiding explicit antisymmetrization allows the same-spin amplitudes and integrals to be factorized in a way that avoids factorization of “exchange-like” terms.

We can justify the validity of such a representation in the open-shell case and use it to easily derive open-shell variants of LS-THC-MPn methods utilizing diagrammatic methods. The justification can refer to the original article [119].

By adhering to the diagrammatic method, wherein the top and bottom Hamiltonian vertices are substituted with the first-order amplitudes, and subsequently applying THC approximation to both the integrals and first-order amplitudes for each spin case, we arrive at equations for open-shell LS-THC-MP3b,

\[
E_{LS-THC-MP3b} = \sum_{abij} \sum_{RS} \left( \tilde{X}_a^R \tilde{X}_i^R T_{RS}^{[1]} \tilde{X}_b^S \tilde{X}_j^S - \tilde{X}_a^R \tilde{X}_b^R T_{RS}^{[1]} \tilde{X}_b^S \tilde{X}_i^S \right)
\]
\[
\times \left[ \frac{1}{2} \sum_{kl} \sum_{TUWY} X_k^T X_l^T V_{TU} X_i^U X_j^U X_a^W X_b^W T_{WY}^{[l]} \tilde{X}_i^Y \tilde{X}_j^Y \right.
\]
\[
+ \frac{1}{2} \sum_{cd} \sum_{TUWY} \tilde{X}_a^T \tilde{X}_c^T V_{TU} \tilde{X}_b^U \tilde{X}_d \tilde{X}_e^W \tilde{X}_f \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_j^Y \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_e^W \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W \tilde{X}_c^T \tilde{X}_e^Y \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{obj} \sum_{RS} \tilde{X}_a^R \tilde{X}_i^R T_{RS}^{[l]} \tilde{X}_b \tilde{X}_j^S
\]
\[
\times \left[ \sum_{kl} \sum_{TUWY} X_k^T X_l^T V_{TU} X_i^U X_j^U X_a^W X_b^W T_{WY}^{[l]} \tilde{X}_i^Y \tilde{X}_j^Y \right.
\]
\[
+ \sum_{cd} \sum_{TUWY} \tilde{X}_a^T \tilde{X}_c^T V_{TU} \tilde{X}_b^U \tilde{X}_d \tilde{X}_e^W \tilde{X}_f \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_j^Y \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_e^W \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W \tilde{X}_c^T \tilde{X}_e^Y \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
- \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
\[
+ \sum_{ck} \sum_{TUWY} \tilde{X}_c^T \tilde{X}_k^T V_{TU} \tilde{X}_b^U \tilde{X}_j \tilde{X}_a \tilde{X}_i^W T_{WY}^{[l]} \tilde{X}_c \tilde{X}_k^Y
\]
3.3. Results

Upon implementing the methods, we conducted tests on four different systems. As LS-THC is a grid-based method, we took into account the following questions:

1. How does the error vary with different system sizes (linear alkyl radicals (n=1 to 20))?
2. How effectively does the method describe various interactions (electrostatic interactions, van der Waals interactions or hydrogen-bond): (2H-2-azabicyclo[1.1.1]-pentane (ABP) solvation)?
3. How well does the method correlate with different bond cleavages and radical stability (glutathione and 9-propyl-4,11-tridecadienoic acid)?

4. Will the method accurately capture conformational differences (9-propyl-4,11-tridecadienoic acid)?

A detailed analysis is presented in the original paper [119](appendix A), but only the key findings are summarized here.

3.3.1. Size Extensivity of the Error

To evaluate the correlation of absolute energy error of THC approximation with respect to the system size, linear alkyl radicals (n=1 to 20) were examined. The absolute value of the per-electron error between the THC and corresponding DF-MP<sub>n</sub> calculation is presented in Fig. 3.2. The error per electron quickly reaches an approximately constant value, whether for the total MP3 correlation energy (Fig. 3.2a), the MP2 correlation energy (Fig. 3.2c,d), or the MP3 correction by itself (Fig. 3.2b). The latter contribution is the most irregular, perhaps due to its smaller magnitude and sensitivity to the virtual-virtual grid represented by $\tilde{X}$. Notably, the error for LS-THC-MP2a is much smaller than for LS-THC-MP2b, as previously observed in the closed-shell case. As the total LS-THC-MP3b correlation energy includes an LS-THC-MP2b contribution, this error is, in fact, dominated by the MP2 part, with the MP3 correction error being a minor part for most reasonable choices of $\epsilon$. Typical density-fitting errors for DF-MP3 seem to be approximately 15 $\mu$E<sub>h</sub>/e$^-$ based on previous experiments. [120]. This indicates that the LS-THC-MP2b errors are comparable for $\epsilon \sim 10^{-3.2}$, or even earlier near $\epsilon \sim 10^{-2.0}$ if the LS-THC-MP2b contribution is replaced by LS-THC-MP2a. These results confirm the size-extensivity of open-shell LS-THC-MP<sub>n</sub> as was observed for the closed-shell variant. Typical per-electron errors are also similar for the closed- and open-shell THC methods; e.g. for C<sub>8</sub>H<sub>18</sub>/C<sub>8</sub>H<sub>17</sub>• at $\epsilon = 10^{-2.4}$ we find errors of 1.0/0.35 $\mu$E<sub>h</sub>/e$^-$ (MP2a), 102/105 $\mu$E<sub>h</sub>/e$^-$ (MP2b), and 6.9/8.7 $\mu$E<sub>h</sub>/e$^-$ (MP3b
correction), respectively. We also observe, similar to the closed-shell THC, a "threshold" effect where the convergence of the incremental error to the asymptotic value is only reached for a sufficiently long chain. This effect diminishes with looser cutoff values (smaller grids), suggesting a saturation of the orbital pair space for smaller systems.

Figure 3.2. Absolute energy errors for open-shell linear alkyl radicals, H(CH₂)ₙ⁺. The absolute value of the per-electron error between the THC and corresponding DF-MPₙ calculation is plotted for (a) the total LS-THC-MP3b correlation energy, which is the sum of (b) the LS-THC-MP3b third-order correction and (c) the LS-THC-MP2b correlation energy, and finally (d) the LS-THC-MP2a correlation energy. $\epsilon$ is the Cholesky decomposition cutoff parameter when pruning the grid, varying logarithmically from $10^{-1}$ to $10^{-4}$ in steps of 0.2 log units. THC calculations with a range of $\epsilon$ were performed in order to examine the dependence of the error with grid size (a smaller $\epsilon$ results in a larger grid and should yield a smaller error). Certain narrow ranges of $\epsilon$ significantly depart from this trend, e.g. $\epsilon \sim 10^{-1.2}$ in Fig. 3.2a or $\epsilon \sim 10^{-3.0}$ in Fig. 3.2d. These irregularities occur due to a sign change in the error.
3.3.2. Radical Microsolvation Energies

Figure 3.3. Errors in incremental solvation energy for 2H-2-azabicyclo[1.1.1]pentane in an explicit water solvation environment. From 1 to 11 waters are added in the order indicated in appendix A. See text for details.

In order to study how an increase in the solvation shell size impacts solvation energy errors, 2H-2-azabicyclo[1.1.1]-pentane (ABP) solvation energy with up to 11 water molecules is examined. As the number of water molecules in the solvation shell increases, various solvation interactions such as hydrogen bonds, weaker electrostatic interactions, or van der Waals interactions appear. However, there is no clear trend in the error due to THC concerning the strength or type of solvent interaction. Rather, the errors are relatively consistent at moderate cutoff values \((0.001 < \epsilon < 0.01)\). At or below \(\epsilon = 10^{-2.6}\), errors are consistently less than 5%. The absence of a trend or significant outliers indicates a relative insensitivity of THC-MP3 to different types of interactions, and critically, no bias of the open-shell THC-MP3 method toward or against interactions involving unpaired electrons.
Figure 3.4. Bond dissociation energy errors of glutathione. (a) Absolute BDE error due to THC for homolytic bond cleavage (open-shell). The bond index refers to labeled bonds in Figure 6. (b) Absolute BDE error due to THC for heterolytic bond cleavage (closed-shell). (c) Absolute errors in the difference between the homolytic and heterolytic BDEs due to THC.
3.3.3. Bond Cleavages and Radical Stability

The THC approximation is further evaluated for its robustness against various bond cleavage points and charge/spin separation. This is accomplished by scrutinizing 13 different bond dissociation energies (BDEs) for the glutathione system and five different hydrogen atom abstraction (HA) energies for the tridecadienoic acid system (Not shown here). Overall, the THC approximation produces an accurate bond cleavage energy (around 1 kcal/mol) for most of the bond-breaking cases at reasonable selections of $\epsilon$ ($\epsilon < 0.01$). Near-saturation of the THC grid for the smaller fragment errors was also observed (bonds A, F, and L). Across the various bonds, the error in the difference between the homolytic and heterolytic BDEs ($\Delta$BDE) is somewhat lower than for the BDEs themselves. This indicates a reliable error cancellation between the open- and closed-shell LS-THC calculations on similar systems (note that the geometries of the products were optimized separately for hetero- and homolytic cleavages). A similar conclusion can be drawn from hydrogen atom abstraction (HA) energies for the tridecadienoic acid system. An extra trend was observed that more stable cationic products tend to result in lower errors for the (substituted alkyl radicals and especially double-bonded sp2 cationic centers), which may result from enhanced error cancellation between more similar geometries where rehybridization is incomplete.

3.3.4. Conformational Energy Ordering

Moreover, the relative energies of 24 distinct conformations of the 9H hydrogen abstraction (open-shell radical) and hydride abstraction (cationic) products were examined. In Figure 3.5, the errors in relative conformational energies due to THC are presented for both types of products. The errors observed in both closed- and open-shell systems show similar ranges across various values of $\epsilon$. Despite examining chemically relevant features of individual conformations, no discernible trend emerges regarding the error size. Consequently, there appears to be a weak correlation between the errors observed in neutral radicals and those
FIGURE 3.5. Errors in the relative energies of 24 conformations of 9-propyl-4,11-tridecadienoic acid due to THC. Conformational energy differences are calculated with respect to the lowest-energy conformation (index 19), and the remaining conformations are ordered by increasing energy (using the DF-MP3 energies). (a) Errors for open-shell radical conformations. (b) Errors for closed-shell cationic conformations in charged closed-shell products. The smallest errors seen here are much lower than those in Figure 3.4, suggesting some error cancellation. However, this cancellation doesn’t happen consistently between the radical and cationic structures. It’s not unexpected because even though we started with the same initial guess, the final structures were optimized separately.

3.3.5. Computational Scaling

Finally, the scaling reduction was demonstrated by reporting the timings of both DF-MP3 and LS-THC-MP3 methods for linear alkyl radicals with varying cutoff parameters ($\epsilon$). The timings include the calculation of the MP2 and MP3 correlation energy only. As
can be seen in Fig. 3.6, it is clear that LS-THC-MP3 achieves reduced scaling compared to DF-MP3 (see inset for measure scaling), and reaches a crossover at around 70 correlated electrons with a cutoff of $\epsilon = 10^{-2.2}$.

**Figure 3.6.** Timings for DF- and LS-THC-MP3 methods for linear alkyl radicals, $(\text{H(CH}_2)_n \cdot$, $n=5–20$) with the cc-pVDZ basis set. Both axes are on a logarithmic scale to highlight the polynomial scaling. All calculations were performed using a single node with 2× Intel Xeon E5-2695v4 CPUs and 256 GiB of memory; OpenMP was used to parallelize the calculation over all 36 cores. The inset shows the measured scaling of DF-MP3 and LS-THC-MP3 with $\epsilon = 10^{-2.2}$ obtained via linear regression of the timings for $n = 10–20$.

### 3.4. Conclusion

We have successfully applied LS-THC on open-shell MP2 and MP3 calculations using diagrammatic techniques and explicit spin-summation. We found that open-shell THC is just as effective as closed-shell THC. Our results indicate that the LS-THC method is insensitive to particular chemical interactions, geometries, or moderate spin contamination. These results lend to promising development and optimization potential for CC-based methods as the decomposed tensors are preserved between the two methods. Furthermore, as additional
excitations are included into the MP$_n$/CC to improve accuracy, LS-THC methods could be used directly to greatly improve performance, as observed in Fig 3.6.

As it stands, this method can be used to compute energies with accuracy, rivaling competing techniques at a faster speed. However, the accuracy must still be improved before it can be used for reaction predictions. The accuracy of the method can reach 1 kcal/mol (4.18 kJ/mol). For thermochemical computations, an accuracy of 1 kJ/mol is required. It has been observed that the error of the exchange term is larger than the column term. Therefore, a detailed investigation of the error behavior is necessary to further improve the accuracy of this method.
Chapter 4
RANK-REDUCED CCSD(T) IN THE CONTEXT OF THERMOCHEMISTRY

4.1. Introduction

Thermodynamic processes and the stability of transient intermediates are crucial for understanding chemical reactions that are important to various fields of chemistry and industry [136]. In other words, it is critical to determine the relative energy (stability) of species in a chemical reaction with great accuracy. An accuracy of one kJ/mol or better is typically required to describe a thermodynamic process. Relative energy can be determined by comparing experimental measurements or based on theoretical calculations. Thanks to computing hardware and algorithm advancements, quantum chemical calculations have greatly facilitated experimental methods [137,138]. The most straightforward way is to calculate the energy of each species via single energy calculations and obtain the relative energy among reactant and product species. However, obtaining thermochemical accuracy reaction energy via single energy calculation is impractical since one needs large basis sets and higher-order excitation theory. The former has slow convergence, and the latter has an exponential increase in computational cost with system size [139,140]. These problems constrain quantum chemical calculation only accessible to small systems (less than five heavy atoms). Fortunately, there’s an observed trend where successive higher cluster expansion terms exhibit a faster convergence with the basis set [141], as they increasingly capture non-dynamical rather than dynamic correlation effects.

Thus, many composite ab initio methods have been developed, where the effects of higher-order cluster operators are treated with monotonically decreasing basis sets to offset the computational cost.
Over the last thirty years, numerous composite \textit{ab initio} methods have been developed. Extensive reviews on this topic can be found in various sources \cite{141–144}. These techniques assume that the solution of the “true” Schrödinger equation for a particular molecule can be represented as a sum over separable contributions to the energy. In order of importance, these typically include the non-relativistic electronic energy, the (ro)vibrational zero point energy, corrections for scalar relativistic and spin-orbit effects, and the diagonal Born-Oppenheimer correction. The non-relativistic electronic energy (NREE), the focus of this work, is generally the largest contributor to a particular observable and is also one of the most difficult to get right. All of these methods aim to produce the exact non-relativistic electronic energy by incorporating different contributions: the HF contribution, valence CCSD contribution, CCSD(T) contribution, and or post-CCSD(T) contribution, in varying ways. CCSD(T), as the felicitous compromise between accuracy and computational cost, contributes to the third-largest correction to the non-relativistic electronic energy, succeeding the Hartree-Fock and valence CCSD correlation.

Thanks to advances in supercomputers, CCSD(T) can now be applied to molecules with over 20 carbons \cite{145–147} using medium-size basis sets like cc-pVTZ and approaching chemical accuracy(1 Kcal/mol) despite its scaling of $N^7$.

Further, the development of localized techniques has empowered the calculation of (T)-like corrections for systems containing hundreds of atoms, and have been shown to approach kcal/mol accuracy. \cite{56,57,59,60,148} Extensive performance studies have shown these methods have approached kcal/mol for small model systems \cite{58,149–152}. However, it should be noted that the error is system size-dependent. For larger systems, the error is sensitive to the basis set. Additionally, extra attention is required for strongly correlated systems \cite{152–155}. Besides, sub-chemical accuracy usually requires (T) contributions calculated with basis sets of at least QZ size (ideally 5Z or higher) \cite{156–158}, and such accuracy places significant strain on localized methods for calculating this quantity. This paper focuses on exploring
how the scaling of this correction can be reduced in the context of high-accuracy computational thermochemistry.

Tensor decomposition avoids these assumptions in local orbital theory, trying to compress the space into a smaller subspace, and has proven to be a great tool in reducing computational cost in quantum chemistry [94, 95, 121, 122, 124, 127, 153, 159–164].

In this work, we are particularly interested in exploring the combinations of various approximations to gain insight into applying these reduced scaling techniques in high-accuracy thermochemistry, specifically in the context of CCSD(T). The performance and accuracy were investigated on a set of molecules selected from the W-17 dataset.

4.2. Theory Throughout this work, the following notation convention is used:

- The letter $g$ represents one of the $N_g$ Laplace transform quadrature points;
- The letters $VW$ represent one of the $N_{T_2}$ rank-reduced $T_2$ projectors;
- The letters $XYZABC$ represent one of the $N_{T_3}$ Tucker-3 compressed $T_3$ projectors;
- The letters $JKL$ represent one of the $N_{DF}$ density-fitting auxiliary basis functions;
- The letters $pqrs$ represent one of the $N = o + v$ general orbitals in the MO basis, which may be fewer than the number of basis functions $N_{bas}$ if core orbitals are frozen.

The approximate treatment of higher excitation in CCSD(T) to the CCSD energy is twofold, the first is the approximation of triple excitations, $E^{[4]}_T$, based on the four-order excitation from perturbation theory [165]; the second, $E^{[5]}_{ST}$, is the fifth and higher order terms in the conventional perturbation expansion to offset the characteristic overestimation.
of \( E_T^{[4]} \) [166].

\[
E_T(E) = E_T^{[4]} + E_T^{[5]}
\]

\[
E_T^{[4]} = \frac{1}{3} \sum_{abcijk} z_{ijk} S_{ijk}^{abc} t_{ijk}^{abc}
\]

\[
E_T^{[5]} = -\frac{1}{3} \sum_{abcijk} V_{ijk}^{abc} S_{ijk}^{abc} t_{ijk}^{abc}
\]

\[
z_{ijk}^{abc} = P_{ijk}^{abc} \left( \sum_d v_{kda}^{ij} t_{ijk}^{(1)} - \sum_l v_{kjl}^{i} t_{ijk}^{(1)} \right)
\]

where \( t_{ijk}^{abc} = z_{ijk}^{abc} D_{ijk}^{abc} \); \( V_{ijk}^{abc} = \frac{1}{2} P_{ijk}^{abc} t_{ijk}^{abc} \); \( D_{ijk}^{abc} \) is the orbital energy denominator, \( P_{ijk}^{abc} \) and \( S_{ijk}^{abc} \) are permutation operators.

\[
D_{ijk}^{abc} = 1/(\epsilon_i + \epsilon_j + \epsilon_k - \epsilon_a - \epsilon_b - \epsilon_c)
\]

\[
P_{ijk}^{abc} = (abc)_{ijk} + (acb)_{ijk} + (bca)_{ijk} + (cab)_{ijk} + (cba)_{ijk} + (bca)_{ijk}
\]

\[
S_{ijk}^{abc} = 4(abc)_{ijk} - 2(acb)_{ijk} - 2(bac)_{ijk} - 2(cba)_{ijk} + (bca)_{ijk} + (cab)_{ijk}
\]

The overall unfavorable \( O(N^7) \) scaling leading term of the CCSD(T) method comes from \( z_{ijk}^{abc} / t_{ijk}^{abc} \). Thus, ranked reduction can be achieved by approximation of this term. The canonical (T) has 88 unique terms [162], and the number of unique diagrams is reduced to six due to the symmetry introduced by the Tucker-3 representation. For simplification, the simplest term (shown in Fig. 4.2 A) contributing \( E^{[4]} \) is used as an example, where the top represents \( z_{ijk}^{abc} \), the bottom with the line representing denominator indicates \( t_{ijk}^{abc} \).

There are four types of terms that can be approximated in the calculation of the CCSD(T): the two-electron repulsion integrals (ERIs), the two- and three-electron amplitudes, and the three-electron orbital eigenvalue denominators. Each of these has been studied individually in the literature. A brief overview of these methods can be found in the introduction of Appendix B.
Figure 4.1. Graphical representations of the various decompositions employed. See text for details. (A) Left: ERIs ($v_{kd}^{cb}$ or $v_{kj}^{cm}$), right: DF-approximated ERIs; (B) Left: $T_2$ amplitudes, right: rank-reduced $T_2$ amplitudes; (C) Left: $z_{ij,k}^{abc}$, right: $t_{ij,k}^{abc}$; (D) Left: compressed $z_{ij,k}^{abc}$, right: simplified representation; (E) Left: compressed $t_{ij,k}^{abc}$, right: simplified representation; (F) Alternative $t_{ij,k}^{abc}$ formed using orthogonal rotation, right: simplified representation.
There are two layers of approximation: the first layer is approximation on the two-electron repulsion integral using Density fitting (DF) (Figure 4.1A) and $t_2$ amplitudes using rank-reduced (RR) approximation (Figure 4.1B) [163]. We denoted the first layer approximation as $z_{ijkl}^{abc}/\tilde{t}_{ijkl}^{abc}$ (Figure 4.1C), however the scaling term of $ov^5$ is unavoidable. Further tucker decomposition is employed on the $z_{ijkl}^{abc}$ or $\tilde{t}_{ijkl}^{abc}$ (Figure 4.1D, E and F). The inseparability of the orbital energy denominator is eliminated by either Laplace Transformation (LT) (Figure 4.1E) or Orthogonal Rotation (OR) (Figure 4.1F). The details of these techniques are discussed in the theory section of Appendix B.

Considering the possible combination of the aforementioned approximation methods, there are many more possible schemes for these decompositions. However, they were quickly screened out upon testing. Only five different approximation schemes are discussed here. A diagrammatic representation of the five different approximation schemes is presented in Figure 4.2. We now use $E^{[4]}$ to demonstrate how the various approximations discussed above can create a number of reduced scaling implementations of CCSD(T). One or both of $z_{ijkl}^{abc}$ and $\tilde{t}_{ijkl}^{abc}$ are cast into Tucker-3 form ($z_{ijkl}^{abc}$ and $\tilde{t}_{ijkl}^{abc}$), creating five distinct reduced scaling methods:

1. $\tilde{Z}\tilde{D}Z$ The top $Z_3$ is Tucker-3 compressed to form $\tilde{Z}_3$, which is then transformed into $\tilde{T}_3$ via orthogonal rotation. The bottom $Z_3$ remains uncompressed. See Figure 4.2C.

2. $\tilde{Z}'\tilde{D}Z$ As in $\tilde{Z}\tilde{D}Z$, except that $T_{ck}^{XY}$ is used during the HOOI instead of $Z_{ck}^{XY}$.

3. $\tilde{T}Z$ The top $Z_3D_3$ is Tucker-3 compressed into $\tilde{T}_3$, while the bottom remains as $Z_3$.

   This is the method previously published by Lesiuk. [127] See figure 4.2D.

4. $\tilde{Z}T$ The top $Z_3$ is Tucker-3 compressed into $\tilde{Z}_3$ and the bottom $Z_3D_3 = T_3$ is left uncompressed. See Figure 4.2E.

5. $\tilde{Z}\tilde{D}\tilde{Z}$ The top and bottom $Z_3$ are Tucker-3 compressed to form $\tilde{Z}_3$. The top $\tilde{Z}_3$ is then transformed into $\tilde{T}_3$ via orthogonal rotation. See Figure 4.2F.

65
Figure 4.2. Methods explored in this work. (A) Canonical (T); (B) Canonical (T) with density fitting; (C) \( \tilde{Z} \tilde{D}Z \) or \( \tilde{Z}' \tilde{D}Z \); (D) \( \tilde{T}Z \); (E) \( \tilde{Z}T \); (F) \( \tilde{Z} \tilde{D} \tilde{Z} \).

4.3. Results

The accuracy of the five different approximation schemes was tested on a molecule set selected from the W4-17 data set [167]. In order to ensure the universality of our results, the selected test set (Table 4.1) covers a wide range of different molecular sizes, organic and inorganic species, as well as references to both dynamic and non-dynamic characters. The geometries are directly obtained from the W4-17 supporting information. Following the W4 protocol, aug-cc-pV\(X\)Z [168,169] and aug-cc-pCV\(X\)Z [168–170] basis sets were used on non-
hydrogen atom valence and all-electron CCSD(T) correlation energies, respectively, while cc-pVXZ basis sets were employed on hydrogen atoms. RI basis sets \cite{171–173} of the same were used to construct the auxiliary basis for density fitting. Basis-set extrapolations of the CCSD(T) correlation energy were obtained using the two-point extrapolation formula using the basis set cardinality $X$ ($X = 3, 4, 5$ for TZ, QZ, 5Z, etc.),

$$E(X) = E_\infty + A(X + a)^{-\alpha} \quad (4.3)$$

where $a$ and $\alpha$ are taken as 0.5 and 4, respectively. \cite{157, 158, 174} Extrapolated values are indicated in the \{X − 1, X\} notation such that aug-cc-pV{T, Q}Z would indicate a reaction energy or correlation energy obtained from the extrapolation using aug-cc-pVTZ/cc-pVTZ (non-hydrogen/hydrogen) and aug-cc-pVQZ/cc-pVQZ (non-hydrogen/hydrogen) basis sets.

The ANL reaction scheme \cite{175} was used to benchmark the accuracy of the above approximate methods against the canonical DF-CCSD(T) reaction energies using the same basis sets. This scheme was extended to include boron, silicon, sulphur, and chlorine-containing species. H$_2$, B$_2$H$_6$, CH$_4$, H$_2$O, NH$_3$, HF, HCl, H$_2$S, and SiH$_4$ were considered as reference species for the H, B, C, O, N, F, Cl, S, and Si elements, respectively, such that the appropriate reaction products of an arbitrary molecule can be determined as

$$H_a B_b C_c N_d O_e F_f Si_g S_h Cl_i = xH_2 + bB_2H_6 + cCH_4 + dNH_3 \quad (4.4)$$

$$+ eH_2O + fHF + gSiH_4 \quad (4.4)$$

$$+ hH_2S + iHCl,$$

where the number of H$_2$ molecules in the products, $x$, needed to balance the equation can be determined algebraically. While this reaction scheme is not designed to universally take advantage of error cancellation (for example, the carbon dioxide reaction scheme would be 4H$_2$ + CO$_2$ $\rightarrow$ 2H$_2$O + CH$_4$, which preserves no bonding character between the reactants.
and products), all the reference species involved are closed shell, well characterized, and free from significant multireference effects.

In this work, we aim to assess the viability of several approximate (T) methods in the context of sub-thermochemical accuracy (1 kJ/mol) of the reaction energies. We will investigate rank convergence and identify the scope of its application. A detailed analysis is presented in Appendix B, but the key findings are highlighted here.

4.3.1. Separability of parameter

Each of the above five approximation schemes is constructed from individual compressions or approximations of the following: 1) the density fitting on the two-electron integral, 2) the $T_2$ amplitudes (rank-reduced compression with a user-defined $N_{T_2}$), and 3) Tucker-3 compression of the $T_3$ amplitude or $Z_3$ intermediates (with a user-defined $N_{T_3}$ and convergence threshold of the HOOI procedure). The error profile of each approximation is investigated to identify the separability of error control parameters in each approximation method.

DF fitting error was found to be negligible (less than 0.1% of the (T) correction) for the reaction energies studied here, and thus DF is used in the reference method of this work (Fig. 4.2B). Further, the separability of the $T_2$ and $T_3$ approximations was investigated. The result shows that the $T_2$ amplitudes converge fast and basically run parallel with only minor differences with different numbers of $T_3$ projectors. Which indicates the $T_3$ and $T_2$ convergence are independent. Thus, three times $N_{bas} T_2$ projectors are used while investigating the behavior of $T_3$ approximation in this work.

A practical application of the T3 approximation is the use of HOOI to obtain the U factor matrix, replacing the unaffordable HOSVD. It is important to understand its convergence behavior. The convergence tolerance ranging from $10^{-1}$ to $10^{-7}$ was investigated with $N_{T_2} = 3 \times N_{bas}$ and $N_{T_3} = 3 \times N_{bas}$. It takes around ten or fewer cycles to converge even for $10^{-7}$. 

68
<table>
<thead>
<tr>
<th>Number of Heavy atoms</th>
<th>single reference (&lt;4)</th>
<th>minor Multi-reference (4-10)</th>
<th>Moderate Multi-reference (&gt;10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>B2H6 0.74 N2 4.16</td>
<td>B2 14.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SiH3F 0.81 O2 7.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BF2H 1.49 HCNO 4.77</td>
<td>O3 17.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CH3CHO 1.72 HN3 5.64</td>
<td>F2O 14.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C3H4 1.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C2H5F 1.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>BF3 1.77 SO3 5.87</td>
<td>F2O2 16.92</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SiF4 1.75 HClO4 7.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>C6H6 1.96 N2O4 9.11</td>
<td>ClF5 14.78</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>SF6 4.29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1. Molecule set selected from W4-17, %TAE[(T)] values are taken from W4-17 SI, and approximately indicate multi-reference behavior.

Figure 4.3. $T_2$ convergence and $T_3$ convergence behavior for different methods and four representative reactions. The absolute reaction energy error with the aug-cc-pV{T,Q}Z extrapolation is shown on the y-axis. The x-axis indicates the number of $T_2$ projectors as a multiple $n$ ($N_{T_2} = nN_{bas}$). Separate colored lines correspond to increasing (red to blue) numbers of $T_3$ projectors ($N_{T_3} = mN_{bas}$).
The results show that the error is well converged even with $10^{-1}$ as tolerance, which indicates fewer iteration cycles (3 or 4 cycles) are needed. Since the code can be further improved, $10^{-5}$ as the tolerance in the HOOI procedure is used in this work to be on the safe side. Detailed parameter suggestions will be investigated in the future.

4.3.2. Comparison of Approximation Schemes

The performance of the five approximation schemes is evaluated by different error statistics. The Mean Absolute Error (MAE) (Fig. 4.4) values are used to determine the average reaction energy error of the entire molecule set, and the MAE over the whole test set was plotted against different approximation schemes and different numbers of projectors in approximating $T_3$ amplitudes, along with the maximum error shown as the error shade. While MAE can be misleading sometimes, such as when the reaction energy is small, the error will be small. To truly represent the accuracy of each approximation scheme and investigate how these approximation schemes may extrapolate to larger molecule systems, we have also studied the mean absolute percent error (MAPE) (Fig. 4.4) of these various reactions. To illuminate this error distribution, we employed a violin plot (Fig. 4.5) for visual representation and analysis. Finally, to show the error behavior of the different chemical environments, the error of each reaction species was also presented (Fig. 4.6).

For the thermochemistry model, we aim for an accuracy of 1 kJ/mol and even several wavenumbers. As we only measured the error of the contribution of (T) correlation energy to reaction energy, we set a target of 0.1 kJ/mol MAE errors and 1% percent error.

All methods are capable of achieving a sub-0.1 kJ/mol and/or sub-1% average accuracy given a large enough set of $Z_3/t_3$ projectors, though with varying error profiles.
Figure 4.4. Error statistics of various decomposition schemes for the benchmark reaction set. The thick lines indicate the mean absolute or absolute percentage errors across the set, while the top of the shaded region indicates the maximum error. Plots are labeled by the extrapolation scheme used, and “aug-cc-pVQ5Z Corrected” indicates an indicates a frozen core aug-cc-pVQ5Z valance-only energy plus an aug-cc-pCVTQZ core-valance correction.

\( \tilde{Z}T \) seems to perform best or second-best for the MAE and MAPE measures (and for the maximum absolute/percentage errors)(Fig. 4.4), and also is the most resistant to outliers with a tight distribution of errors compared to the other methods.(Fig. 4.5).

\( \tilde{T}Z \) and \( \tilde{Z}'\tilde{D}Z \) (green and red lines, respectively) performed quite similarly, not just for the statistical errors, but also in the individual reaction energies, though \( \tilde{Z}'\tilde{D}Z \) (red line) is slightly more accurate. Consistent with the average and max error plots, \( \tilde{T}Z, \tilde{Z}'\tilde{D}Z \) are also resistant to outliers.

Then \( \tilde{Z}\tilde{D}Z \) follows, although the relative performance of \( \tilde{Z}\tilde{D}Z \) changes depending on the basis sets involved. \( \tilde{Z}\tilde{D}Z \) performs well with the exception of a few outliers that skew the average statistics. These outliers correspond to the reactions : \[ 2 \text{BF}_3 + 6 \text{H}_2 \rightarrow \text{B}_2\text{H}_6 + 6 \text{HF}, \text{SiF}_4 + 4 \text{H}_2 \rightarrow \text{SiH}_4 + 4 \text{HF}, \text{and } 2 \text{BHF}_2 + 3 \text{H}_2 \rightarrow \text{B}_2\text{H}_6 + 2 \text{HF}. \] Interestingly, these reactions all display extremely small (T) contributions to reaction energy.
Figure 4.5. “Violin plots” of the mean absolute percent errors (MAPE) of reaction energies with the aug-cc-pV\{T,Q\}Z basis set extrapolation. The break in the y-axis is used to present the extremal data in detail.
Figure 4.6. Absolute relative errors (in percent) of each reaction species at the aug-cc-pV(T,Q)Z frozen-core level. In each figure, the right axis gives the absolute error in kJ/mol.
Thus, even large percentage errors still meet the desired 0.1 kJ mol\(^{-1}\) threshold. However, some applications may require small relative errors, even in the case of small total reaction energies.

The $\tilde{Z}\tilde{D}\tilde{Z}$ scheme (purple line) requires the largest number of projectors to reach target MAE accuracy, which is perhaps unsurprising as this scheme employs Tucker-3 decompositions of both the top and bottom half of the $E^{[4]}$ diagrams. The same is true in the villon plot, $\tilde{Z}\tilde{D}\tilde{Z}$, the double-Tucker-3 method performs worst.

Most important in the potential application of these methods in model chemistries is the number of $Z_3/T_3$ projectors required to achieve some target accuracy, and hence the overall computation cost required. For example, a relatively expensive approximation scheme such as $\tilde{T}Z$ (green) may out-perform a less expensive scheme such as $\tilde{Z}T$ (orange) if the former requires a sufficiently smaller number of projectors to obtain equivalent confidence intervals. To explore this, we examine the number of projectors required to obtain an MAE of 0.1 kJmol and estimate the relative expense of the five methods using the equations contained within the Supplementary Information. From this perspective, the two most promising methods are $\tilde{Z}T$ (orange) and $\tilde{Z}\tilde{D}Z$ (blue). $\tilde{Z}\tilde{D}Z$ (purple) falls behind as it requires a large number of $T_3$ projectors to obtain target accuracy (and scales cubically with $N_{T_3}$ while the other methods scale quadratically), while $\tilde{T}Z$ (green) and $\tilde{Z}'\tilde{D}Z$ (red), although relatively stable, do not out-compete $\tilde{Z}T$’s (orange) ratio of cost to accuracy.

4.4. Conclusion

We have examined five different approximation schemes for the perturbative triples correction in CCSD(T), four of which are completely new. Our analysis showed that, of the approximations studied, only the $Z_3/t_3$ Tucker-3 decomposition was a limiting factor in the accuracy and cost of the resulting approximation schemes. Among the five schemes we evaluated, $\tilde{Z}T$, novel to this work, demonstrates a favorable combination of accuracy and
computational efficiency. However, to determine exactly how these components should be included in such model chemistries, future work is needed, with particular emphasis on selecting the number of $Z_3/t_3$ projectors on a reaction-by-reaction basis. Nonetheless, it is evident that these techniques have a place in some of our most accurate models of gas-phase thermochemistry and may yield order-of-magnitude improvements in computational cost with minimal and well-controlled error.
5.1. Introduction

When a molecule absorbs photons, electrons from the occupied orbitals will be promoted to unoccupied orbitals. We say the molecule is promoted from the ground state to the excited state, and the process is called photoexcitation. Conversely, the molecular excited states can relax to their ground state electronic configuration through either the radiative or non-radiative decay processes. These processes can be collectively described in the well-known Perrin-Jablonski diagram. Molecular excited states are of great scientific interest as they are crucial in many fields of science and technology: bioluminescence [176], mutagenesis and carcinogenesis [177], and photovoltaics and light-emitting diodes [178, 179], to name a few. Experimentally, photoexcitation can be observed through UV/Vis spectra. The emission from the excited state to the ground state can be observed through fluorescence for the single state and phosphorescence for the triplet state. However, analyzing excited states in spectroscopy can be challenging. First, the experiment is difficult to conduct due to the short time and low population of the excited state, instrumental limitation, and so on. Second, the experimental interpretations are not always straightforward. Theoretical quantum chemistry calculations can provide crucial support in gaining a comprehensive understanding of the excited states.

Consequently, numerous methods have emerged to model excited states. The inherent multiconfigurational nature of excited states poses significant challenges to traditional single-reference methods, which are typically optimized for systems dominated by a single reference in the ground state [135]. Thus, the natural inclination is towards multiconfigura-
tional methods. Among these, complete active space self-consistent field (CASSCF) [180], its second-order perturbation-corrected variant (CASPT2) [181], and the second-order n-electron valence state perturbation theory (NEVPT2) [182–184] have emerged as successful approaches to address this challenge. However, these methods require constructing a suitable zero-order wave function encompassing all relevant configurations, which require advanced knowledge and experience with the system to be treated correctly. In contrast, Equation-of-motion (EOM) methods adopt a distinct approach, encompassing the entire set of multiconfigurational target states within a single-reference framework. This black box theory has gained considerable popularity. Equation of motion coupled cluster singles and doubles (EOM-CCSD) [79], owing to its favorable accessibility and its predictable accuracy over a wide range of chemical problems [185], has become the golden standard method for the prediction of one-electron excited state properties. However, it is poor for double excitation states [186–188]. Efforts have been made in either iterative or non-iterative schemes to include triple excitation to improve its accuracy [189], EOM-CCSD* [190], EOM-CCSD(\tilde{T}) [188], EOM-CCSDT-3 [188], EOM-CC3 [191,192], EOM-CCSDT [193].

In order to explain and predict the spectroscopy, energy differences, and transition probabilities are the key quantities the theoretical method needs to capture [194]. These quantities must be computed at significant points in the potential energy surface (PES). For example, to capture the emission band, the energy difference has to be at the optimized geometry of the excited state. The PES has been the pursuit of quantum chemists. Over the years, analytic-derivative techniques have been well developed [195]. The Z vector method offers the analytical derivative method the advantage of efficiency, besides its numerical stability. And great success has been achieved in the exploration of potential energy surfaces (PESs) of molecules in their electronic ground state [196–205]. Owning the popularity of EOM-CCSD, its analytic gradients of EOM-CCSD have already been implemented [79]. However, the analytic gradient for EOM theory, including triple excitation, is still not available. In what
follows, we present detailed formulas and implementation details for the efficient computation of analytic closed-shell EOM-CCSD* energy gradients.

5.2. Theory

Here, we pursue a different strategy of derivation presented in the Appendix C. The Lagrangian Z-vector approach [206–208] way of the analytic gradient is shown. These two different ways of derivation form a mutual validation.

The perturbative third-order correction [209] to EOM-CCSD energy can be written as follows, and

$$\omega = E_{{\text{EOM-CCSD}}} - E_{{\text{CCSD}}}.$$

$$\Delta E = \langle 0 | L \hat{H} | T \rangle \langle T | (\omega - F)^{-1} | T \rangle \langle T | V R_2 + (V T R_1)_c | 0 \rangle$$

The total EOM-CCSD* energy can be written as:

$$E_{{\text{EOM-CCSD*}}} = E_{{\text{EOM-CCSD}}} + \Delta E = \langle 0 | L \hat{H} R | 0 \rangle + \langle 0 | \hat{L}_3 (\omega - F) \hat{R}_3 | 0 \rangle$$

where

$$\langle 0 | \hat{L}_3 | T \rangle = \langle 0 | L \hat{H} (\omega - F)^{-1} | T \rangle \quad \text{and} \quad \langle T | \hat{R}_3 | 0 \rangle = \langle T | (\omega - F)^{-1} (V R_2 + (V T R_1)_c) | 0 \rangle.$$ 

Then, the fully variational expression for the excited state energy function can be written as follows:

$$L(t, R, L, Z, \Pi, \Upsilon, \lambda) = \langle 0 | L \hat{H} R | 0 \rangle$$

$$+ \langle 0 | L \hat{H} | T \rangle \langle T | (\omega - F)^{-1} | T \rangle \langle T | V R_2 + (V T R_1)_c | 0 \rangle$$

$$+ \langle 0 | Z \hat{H} | 0 \rangle + \langle 0 | \Pi (\hat{H} - E_{\text{EOM}}) R | 0 \rangle + \langle 0 | L (\hat{H} - E_{\text{EOM}}) \Upsilon | 0 \rangle$$

$$+ \lambda (1 - \langle 0 | L R | 0 \rangle)$$

The lagrangian L is required to be stationary with respect to all its parameters $t, R, L, Z, \Pi, \Upsilon, \lambda$, which are considered as variables in Eq. 5.3. $Z, \Pi, \Upsilon$ is introduced to enforce $\langle \psi^{ab}_{ij} | \bar{H} | 0 \rangle = 0$. 

78
\( \langle 0 | \mathcal{L} \hat{H} | P \rangle = E_{EOM} \langle 0 | \mathcal{L} | P \rangle \) and \( \langle P | \mathcal{R} \hat{H} | 0 \rangle = E_{EOM} \langle P | \mathcal{R} | 0 \rangle \), \( \lambda \) is introduced to enforce \( L \) and \( R \) biorthonomalized. One \( Z, \Pi, \Upsilon, \lambda \) has been determined from the stationarity conditions, and derivatives of the excitation energy follow.

### 5.2.1. Determination of the lagrange multiplier \( \lambda \)

\[
\frac{\partial L(t, R, L, Z, \Pi, \Upsilon, \lambda)}{\partial \langle L | R \rangle} = E_{eomccsd} - \langle 0 | L_3 R_3 | 0 \rangle E_{eomccsd} - \langle 0 | \Pi R | 0 \rangle E_{eomccsd} - \langle 0 | L \Upsilon | 0 \rangle E_{eomccsd} - \lambda = 0 \tag{5.4}
\]

where \( \langle 0 | \Pi R | 0 \rangle = 0 \) and \( \langle 0 | L \Upsilon | 0 \rangle = 0 \) is enforced during the convergence problem (see results section).

### 5.2.2. Determination of the Lagrange multiplier \( \Upsilon \)

\[
\frac{\partial L(t, R, L, Z, \Pi, \Upsilon, \lambda)}{\partial L} = \langle P | \bar{H} R | 0 \rangle + \langle P | H R_3 | 0 \rangle - \langle P | \bar{H} R | 0 \rangle \langle 0 | L_3 R_3 | 0 \rangle \\
- \langle P | \bar{H} R | 0 \rangle \langle 0 | \Pi R | 0 \rangle + \langle P | (\bar{H} - E_{eom}) \Upsilon | 0 \rangle \\
- \langle P | \bar{H} R | 0 \rangle \langle 0 | L \Upsilon | 0 \rangle - \lambda \langle P | R | 0 \rangle \tag{5.5}
\]

\[
= \langle P | \bar{H} R | 0 \rangle + \langle P | H R_3 | 0 \rangle - \langle P | \bar{H} R | 0 \rangle \langle 0 | L_3 R_3 | 0 \rangle \\
+ \langle P | (\bar{H} - E_{eom}) \Upsilon | 0 \rangle - \lambda \langle P | R | 0 \rangle \tag{5.6}
\]

Replacing \( \lambda \) with \( \lambda = E_{eomccsd} - \langle 0 | L_3 R_3 | 0 \rangle E_{eomccsd} \) in above, we can get

\[
\langle P | (\bar{H} - E_{eom}) \Upsilon | 0 \rangle = - \langle P | H R_3 | 0 \rangle
\]

Which can be solved through the linear solution.
5.2.3. Determination of the Lagrange multiplier $\Pi$

\[
\frac{\partial L(t, R, L, Z, \Pi, \Upsilon, \lambda)}{\partial R} = \langle 0 | L\bar{H} | P \rangle + \langle 0 | L_3V | D \rangle + \langle 0 | L_3VT_2 | S \rangle - \langle 0 | L_3R_3 | 0 \rangle \langle 0 | L\bar{H} | P \rangle - \langle 0 | L\bar{H} | P \rangle \langle 0 | \Pi R | 0 \rangle - \lambda \langle 0 | L | P \rangle
\]
\[
= \langle 0 | L\bar{H} | P \rangle + \langle 0 | L_3V | D \rangle + \langle 0 | L_3VT_2 | S \rangle - \langle 0 | L_3R_3 | 0 \rangle \langle 0 | L\bar{H} | P \rangle + \langle 0 | \Pi(\bar{H} - E_{eom}) | P \rangle
\]
\[
= \langle 0 | L\bar{H} | P \rangle + \langle 0 | L_3V | D \rangle + \langle 0 | L_3VT_2 | S \rangle - \langle 0 | L_3R_3 | 0 \rangle \langle 0 | L\bar{H} | P \rangle + \langle 0 | \Pi(\bar{H} - E_{eom}) | P \rangle - \lambda \langle 0 | L | P \rangle \tag{5.7}
\]

Replacing $\lambda$ with $\lambda = E_{eomccsd} - \langle 0 | L_3R_3 | 0 \rangle E_{eomccsd}$ in above, it is clear that

\[
\langle 0 | \Pi(\bar{H} - E_{eom}) | P \rangle = -\langle 0 | L_3V | D \rangle + \langle 0 | L_3VT_2 | S \rangle
\]

which can be solved through the linear solve.

5.2.4. Determination of the Lagrange multiplier $Z$

\[
\frac{\partial L(t, R, L, Z, \Pi, \Upsilon, \lambda)}{\partial t} = \langle 1 - \langle 0 | L_3R_3 | 0 \rangle - \langle 0 | \Pi R | 0 \rangle - \langle 0 | L\bar{Y} | 0 \rangle \rangle *
\]
\[
- \langle 0 | L | P \rangle \langle P | \bar{H}R | 0 \rangle + \langle 0 | L\bar{H}R | P \rangle + \langle 0 | L_3VR_1 | D \rangle
\]
\[
+ \langle 0 | L_3R_3 | 0 \rangle \langle 0 | \bar{H} | P \rangle + \langle 0 | Z\bar{H} | P \rangle
\]
\[
- E_{eom} \langle 0 | \Pi R | 0 \rangle + \langle 0 | \Pi\bar{H}R | P \rangle
\]
\[
+ E_{eom} \langle 0 | L\bar{Y} | P \rangle - \langle 0 | L | P \rangle \langle P | \bar{H}\bar{Y} | 0 \rangle
\]
\[
+ \langle 0 | L\bar{H} | Q \rangle \langle Q | Y | P \rangle \tag{5.9}
\]
\[
= \langle 1 - \langle 0 | L_3R_3 | 0 \rangle \rangle * \langle 0 | L\bar{H} | Q \rangle \langle Q | R | P \rangle + \langle 0 | L_3VR_1 | D \rangle
\]
\[
+ \langle 0 | L_3R_3 | 0 \rangle \langle 0 | \bar{H} | P \rangle + \langle 0 | Z\bar{H} | P \rangle
\]
\[
+ E_{eom} \langle 0 | L\bar{Y} | P \rangle - \langle 0 | L | P \rangle \langle P | \bar{H}\bar{Y} | 0 \rangle + \langle 0 | \Pi\bar{H}R | P \rangle
\]
\[
+ \langle 0 | L\bar{H} | Q \rangle \langle Q | Y | P \rangle \tag{5.10}
\]

80
The final expression is as follows:

\[
L = \langle 0 | H \hat{\mathcal{H}}^{(x)} | 0 \rangle + \langle 0 | \mathcal{L} \hat{\mathcal{H}}^{(x)} \mathcal{Y} | 0 \rangle \\
+ (1 - \Delta - \varepsilon) \langle 0 | \mathcal{L} \hat{\mathcal{H}}^{(x)} | 0 \rangle \\
+ \langle 0 | Z \hat{\mathcal{H}}^{(x)} | 0 \rangle \\
+ \langle 0 | (\mathcal{L}_1 + \mathcal{L}_2) \mathcal{H}^\chi \hat{R}_3 | 0 \rangle \\
+ \langle 0 | \hat{L}_3 (V^\chi \mathcal{B}_2 + V^\chi T_2 \mathcal{B}_1) | 0 \rangle \\
+ \Delta \langle 0 | F^\chi | 0 \rangle + \Delta \langle 0 | \hat{H}^\chi | 0 \rangle + \lambda (1 - \langle 0 | LR | 0 \rangle) \tag{5.11}
\]

5.3. Results

5.3.1. Implementation

Eq. 5.11 represent the gradients in the molecule orbitals, which can be summarized as the product of the effective one and two-particle density matrices and the one- and two-electron derivative of molecular orbitals integral:

\[
E^\chi_{\text{EOM-CCSD}^*} = \sum_{pq} (f^p_q)^x D^p_q + \sum_{pqrs} \langle V^p_{r}q^s \rangle x \Gamma^p_{rs} \tag{5.12}
\]

where the \(D^p_q\) and \(\Gamma^p_{rs}\) are the effective one and two-particle density matrices. The derivatives of molecular orbital integrals will be mapped into the derivatives of atomic orbital integrals for the calculation. Further details are omitted here since this is well-documented and implemented in most of the software [196–205].

The effective one- and two-particle density matrices are implemented in the development version of CFOUR. Given the structure of the CFOUR ncc program, code recycling can be achieved, as shown below. Notably, the gradient of EOM-CCSD is of the format in Eq. 5.13.
Thus, the highlighted terms in Eq. 5.11 can be refactored as in Eq. 5.14 and Eq. 5.15. Each of them is a function call of the code for Eq. 5.13, in the Eq. 5.15 \( Z = 0 \). This refactorization reduces the computational cost by consolidating three function calls into two.

\[
E_{EOM-CCSD}^X = \langle 0 | \mathcal{L} \hat{H}^{(x)} \mathcal{R} | 0 \rangle + \langle 0 | Z \hat{H}^{(x)} | 0 \rangle \quad (5.13)
\]

Highlighted – terms = \( \langle 0 | [(1 - \Delta - \varepsilon) \mathcal{L} + \Pi] \hat{H}^{(x)} [\mathcal{R} + \Upsilon] + \langle 0 | Z \hat{H}^{(x)} | 0 \rangle \) \quad (5.14)

\[+ \langle 0 | [(\Delta + \varepsilon) \mathcal{L} - \Pi] \hat{H}^{(x)} \Upsilon | 0 \rangle \quad (5.15)\]

One challenge we face during the implementation is the convergence problem of the linear equation for Eq. 5.2.2 and 5.2.3.

In the gradients equation, \( \langle 0 | \Omega \mathcal{R} | 0 \rangle \) and \( \langle 0 | \mathcal{L} \mathcal{R} \Sigma | 0 \rangle \) appears. Given a complete basis set, the \( \langle 0 | \Omega | P \rangle \) can be represented as Eq. 5.16. \( \langle 0 | \mathcal{L} | P \rangle \) and \( \langle P | \mathcal{R} | 0 \rangle \) form a biorthogonal set (Eq. 5.17), we can require \( \langle 0 | \mathcal{L} \mathcal{R} | 0 \rangle = 0 \) and \( \langle 0 | \mathcal{R} \mathcal{L} | 0 \rangle = 0 \).

\[
\langle 0 | \Omega | P \rangle = \sum_i c_i \langle 0 | L_i | P \rangle \quad (5.16)
\]

\[
\langle 0 | \mathcal{L} \mathcal{R} | 0 \rangle + \langle 0 | \mathcal{R} \mathcal{L} | 0 \rangle = 0 \quad (5.17)
\]

\[
I = \sum_k \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle \quad (5.18)
\]

\[
\langle 0 | \Omega | P \rangle = \sum_i c_i \langle 0 | L_i | P \rangle \sum_k \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle \quad (5.19)
\]

Inserting the identity(Eq. 5.18), we will get Eq. 5.19. In the gradient function, the \( \langle 0 | \Omega | P \rangle \) is multiplied with the derivative of \( R(\langle P | R^n | 0 \rangle) \), thus only the orthogonal part of \( \langle \omega | P \rangle \) with respect to \( \langle P | R \ker 0 \) will survive as shown below.
\[
\langle 0 | \Omega | P \rangle = \sum_i c_i \langle 0 | L_i | P \rangle \sum_{k \neq i} \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle \\
+ \sum_i c_i \langle 0 | L_i | P \rangle \sum_{k=i} \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle 
\] 
(5.20)

\[
\langle 0 | \Omega R^x | P \rangle = \sum_i c_i \langle 0 | L_i | P \rangle \sum_{k \neq i} \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle \langle P | R^x_k | 0 \rangle \\
+ \sum_i c_i \langle 0 | L_i | P \rangle \sum_{k=i} \langle P | R_k | 0 \rangle \langle 0 | L_k | P \rangle \langle P | R^x_k | 0 \rangle \\
= \langle 0 | \Omega | P \rangle^\perp \langle P | R^x | 0 \rangle + 0 
\] 
(5.21)

\[
\langle 0 | \Pi | P \rangle \langle P | (\bar{H} - E_k)^{-1} | P \rangle = \langle 0 | \Omega^\perp | P \rangle 
\] 
(5.22)

\[
\bar{H} - E_k = \sum_{i=0} \langle P | R_i | 0 \rangle \langle 0 | L_i | P \rangle (E_i - E_k) 
\] 
(5.23)

\[
\langle 0 | \Pi | P \rangle \langle P | (\bar{H} - E_k)^+ | P \rangle = \langle 0 | \Omega | P \rangle^\perp 
\] 
(5.24)

Thus, during the linear solve \( \langle 0 | \Pi | P \rangle = \langle 0 | \Pi | P \rangle - \langle 0 | \Pi | P \rangle \langle P | R | 0 \rangle \langle 0 | L | P \rangle \) is ensured and \( \langle 0 | \Pi R | 0 \rangle \) is enforced. Similar treatment was applied on \( \langle 0 | L \Upsilon | 0 \rangle \).

5.3.2. Application

The analytic gradient of EOM-CCSD* is implemented in the development version of CFOUR, and all the calculations are also performed in the CFOUR program package. In this section, we discuss the validation of our implementation and demonstrate the illustrative application of analytic EOM-CCSD* gradients to a set of excited states. In the geometry optimization, the following convergence thresholds were used: SCF\_CONV = \( 10^{-10} \);
We validated the correctness of our implementation using formaldehyde as a test system. Formaldehyde has been extensively investigated both experimentally \cite{210} and theoretically \cite{211–216}. Its small system size makes it a great candidate for validating the correctness of our implementation against numerical differentiation. The first excited state $1^1A_2(n \rightarrow \pi^*)$ valence state is investigated here. It is well known that $1^1A_2$ state has a non-planar equilibrium geometry \cite{210,211,216}. Thus, $Cs$ symmetry is used during optimization. The ground state geometry is obtained from the QUEST2 dataset \cite{217}, which has been optimized at the CC3/aug-cc-pVTZ level. We computed the EOM-CCSD* gradients (with the 6-31G** basis) using both the analytic gradient we implemented and the finite-difference differentiation. The differences between the numerical and analytic values for the individual gradient components were, in all cases, less than $10^{-7}$ a.u. (Table 5.3), which confirmed the correctness of our implementation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RCO</th>
<th>RCH</th>
<th>Angle(COH)</th>
<th>Dihedral</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG</td>
<td>-0.1699816224</td>
<td>0.0028813050</td>
<td>0.0069008662</td>
<td>0</td>
</tr>
<tr>
<td>FD</td>
<td>-0.1699819929</td>
<td>0.0028810153</td>
<td>0.0069011922</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.1. Gradients calculated by analytic derivative and finite-difference. Parameter values are in angstroms and degrees.

Further, the geometry is optimized at different theory levels, ranging from EOM-CCSD, EOM-CCSD* to EOM-CCSDT with an AUG-PVTZ basis set. The optimized geometries are summarized in Table 5.2.

5.3.3. $n \rightarrow \pi^*$ state of Tetrazine

Excited states of tetrazine are of great interest in the scientific community, as greatly summarized in the introduction of the work by Angeli \cite{218}. The ground state geometry...
Table 5.2. Optimized geometry of $n \rightarrow \pi^*$ excited state of formaldehyde with AUG-PVTZ basis set.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>EOM-CCSD</th>
<th>EOM-CCSD*</th>
<th>EOM-CCSDT</th>
<th>Expt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCO</td>
<td>1.305955429</td>
<td>1.343091596</td>
<td>1.331052151</td>
<td>1.323</td>
</tr>
<tr>
<td>RCH</td>
<td>1.09130959</td>
<td>1.088900852</td>
<td>1.093015002</td>
<td>1.103</td>
</tr>
<tr>
<td>Angle(COH)</td>
<td>119.0449</td>
<td>120.9835</td>
<td>118.4577</td>
<td>118.1</td>
</tr>
<tr>
<td>Out of plane dihedral</td>
<td>149.6198434</td>
<td>152.25129</td>
<td>143.8952518</td>
<td>146.0</td>
</tr>
</tbody>
</table>

is obtained from the QUEST2 dataset [217], which has been optimized at the CC3/aug-cc-pVTZ level. $n \rightarrow \pi^*$ excited state promoting charge from a $b_{3g}$ to a $a_u$ orbital is studied here, which has shown to be that the first excited state is of D2h symmetry [219]. The optimized geometries are summarized in table 5.3; as this is a pure single excited state, there are not many differences besides the angle.

Table 5.3. Optimized geometry of $n \rightarrow \pi^*$ state of Tetrazine with basis set AUG-PVTZ

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RCH</th>
<th>RCN</th>
<th>RNN</th>
<th>AHCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOM-CCSD</td>
<td>1.0768</td>
<td>1.3359</td>
<td>1.2503</td>
<td>121.2169</td>
</tr>
<tr>
<td>EOM-CCSD*</td>
<td>1.0777</td>
<td>1.3365</td>
<td>1.2541</td>
<td>121.1106</td>
</tr>
</tbody>
</table>

5.4. Conclusion

This chapter elucidates the theoretical framework and practical application of the first-order analytic gradient within the EOM-CCSD* method. Our analytic gradient formulation is consistent with numerical gradients. This work empowers the study of excited states in medium-sized molecules with enhanced theoretical precision.
Chapter 6
ALLOSTERIC REGULATION IN STAT3 INTERDOMAINS

While quantum mechanics offers accurate total energy, its computational complication limits its application to systems no larger than hundreds of atoms (with DFT) [220, 221]. Which makes molecular mechanics have its irreplaceable role in chemistry, biology, material science, and many other areas. Molecular mechanics describes molecular motions in terms of ball and spring models, thus avoiding the expensive computational costs of accurately describing electronic motions. This allows us to approximate the motions of macromolecules with thousands of atoms at $\mu s$ timescales. Briefly, the atoms are assumed to be perfect spheres, and the interactions between atoms are described by the classical model of spring, where different terms are added to account for different physical natures. The potential energy of the whole system is approximated as the combination of contributions from each atom. The potential energy $U(x, y, z)$ is calculated as a function of the atomic coordinates:

$$U(x, y, z) = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{6.1}$$

$$E_{\text{bonded}} = \sum_{\text{bonds}} K_b (b - b_0)^2$$

$$+ \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2$$

$$+ \sum_{\text{dihedrals}} K_\phi (1 + \cos(n\Phi - \delta)) \tag{6.2}$$

$$E_{\text{non-bonded}} = \sum_{\text{nbpair}(ij)} \left( \frac{A_{i,j}}{|< x_i, y_i, z_i > - < x_j, y_j, z_j >|^{12}} \right)$$

$$- \frac{B_{i,j}}{|< x_i, y_i, z_i > - < x_j, y_j, z_j >|^{6}} \right)$$

$$+ \sum_{\text{nbpair}(ij)} K_e \frac{q_i q_j}{|< x_i, y_i, z_i > - < x_j, y_j, z_j >|^2} \tag{6.3}$$
where $b$ is the bond length ($b_0$ is the equilibrium bond length); $\theta$ is the atomic angle ($\theta_0$ is the equilibrium bond angle); $\Phi$ is the dihedral angle ($\delta$ is the phase angle); integer $n$ is the periodicity; $x_i, y_i, z_i$ is the coordinate of atom $i$, $K_b, K_{\theta}, K_{\Phi}$ are corresponding force constant. Van der Waals interactions were modeled using the Lennard-Jones potential, $A_{ij}$ and $B_{ij}$ are adjustable parameters responsible for short-range repulsion interactions and dispersion (London) attraction, respectively. The charged (electrostatic) interactions were modeled using Coulomb’s law, where $q_i$ is the effective atom charge for atom $i$; $K_e$ is the Coulomb’s constant.

Given the atomic position of a bio-molecular system and the above force field function, the potential energy can be calculated. The force exerted on each atom can then be calculated by the differentiation of potential energy in terms of atomic coordinates. Using Newtonian laws of motion, the position of each atom of the next time step can be predicted. The repeating procedure of calculating the atomic position and velocity of each atom is Molecular Dynamics simulations [222–224]. The resulting trajectory allows us to study the conformation spaces of a bio-molecular system.

This chapter presents a case study using molecular mechanics to investigate the signal propagation between different domains in Signal Transducers and Activators of Transcription 3 (STAT3) protein.

6.1. Introduction

STAT3 functions as both signal transducers in the cytoplasm and transcription factor upon nuclear translocation. Its ability to seamlessly switch between these functions makes it an indispensable component of cellular signaling pathways. However, constitutive activation of STAT3 has been shown to play a crucial role in cancer progression [225]. The direct therapeutic inhibition of STAT3 is highly desirable but remains challenging, as evidenced by the lack of FDA–approved drugs.
Figure 6.1. STAT3 structure. 1BG1 was used as the template; Amino-terminal Domain (NTD) is not shown. (A) STAT3 domain structure (Y705 is shown as spheres, D170 is shown as sticks). See below for details of the initial structure. (B) Secondary structures are labeled according to the UniProt database. α helices are colored blue, β sheets are colored red, and unstructured regions (loops) are colored yellow (transverse view). The assigned secondary structures combine information from multiple x-ray crystal structures, thus there is some mismatch with the specific structures used in this work.

STAT3 consists of six domains (Figure 6.1A): amino-terminal domain (NTD), coiled-coil domain (CCD), DNA-binding domain (DBD), linker domain (LD), SRC-homology 2 domain (SH2), and transactivation domain (TAD) [226, 227]. It binds, via the SH2 domain, to cell-surface receptors and is activated by various receptor-associated kinases. Upon binding, the recruited kinases activate STAT3 through phosphorylation within the TAD (at Y705), followed by dissociation from the receptor to form homodimers through reciprocal interactions between the SH2 domain and the phosphotyrosine (pY705) residue. These activated homodimers are then translocated to the nucleus, where the DBD binds to target genes, and TAD activates the expression of proteins crucial for cell growth and survival.
A significant amount of effort has been employed in the development of molecules targeting the SH2 domain [228–230]. Most of these compounds have yet to be explored in clinical studies, or further development of these compounds was limited due to concerns of lacking potency and selectivity. Surprisingly, several studies [231–233] have determined that effector (small molecule and polypeptide) binding to CCD interferes with SH2 domain binding or precludes STAT3 nuclear translocation, which suggests potential targets for further drug design. The discovery of several diverse inhibitory agents that bind to CCD rather than SH2 is a fascinating development. However, the rational design of allosteric effectors requires a more detailed, mechanistic knowledge of how CCD binding affects SH2 structure and activity. In this work, molecular dynamics (MD) was applied to the wild-type system and D170A (in CCD) variant, which showed the diminishment of both pY-peptide binding and tyrosine phosphorylation (SH2 binding ability), to study the allosteric mechanism.

In total, 3 µs (6 replicas, 500ns/replica) simulation was obtained for both the wild-type and the D170A variant. Then, the sampled configurational space from these simulation trajectories was used to carry out statistical analysis for the allostery mechanism study.
6.2. Results

Figure 6.2. Conformational analysis of SH2 domain binding pockets. (A) PCA of both the pY and pY+3 pockets in the wild type (blue) and D170A mutant (orange). The contour lines show the density of recorded frames in each region, and the crystal structure (PDB ID: 6NUQ) is marked for reference. (B): pY and pY+3 pockets PCA 2D plane colored by different macro-states. (C): The averaged structures for each macro-state. Residues were shown as sticks.

The regulation of STAT3 behavior crucially depends on the conformations of binding pockets in the SH2 domain. To identify the primary conformational difference of the binding pockets, the SH2 binding pocket was characterized using the pair center of mass (COM) distance of key residues for both pY and pY+3 pockets. The distance matrix was projected into a 2D plane to visualize the conformational differences of these key residues (Figure 6.2, A). A combination of Markov State Modeling and Perron Cluster-Cluster Analysis was employed to cluster transient conformations into kinetically meta-stable macro-states (Figure 6.2, B). These analyses highlight the unique conformations (macro-states 0, 1, and 3) explored by the D170A variant. A representative structure from each macro-state shows distinct conformations of Q644 and Y657 found in Q644 and Y657 residues, which is also highlighted by the explained variance of the PCA (data not shown; refer to appendix D). The pY+3 pocket is blocked in macro-state 2 by Y657 and Y640, with both of them pointing towards the pocket. Conversely, in macro-states 0, 1, and 3, Y657 points away from the pY+3 pocket, leading to an open conformation. The shared conformational states between the wild type and the D170A variant consist of viable functional states of STAT3. However, the D170A
variant has reduced occupancy at those conformational states, leading to a differentiated function relative to the wild type. These demonstrate the correlation between conformational differences in the pY+3 binding pocket and the decreased binding affinity of the D170A variant comparing the wild type.

Figure 6.3. Rigid core analysis. (A) The proposed allosteric path from D/A170 to Y657 was obtained from REDAN analysis. (B) Conserved hydrogen bond network between CCD, LD, and DBD. (C) Inter-domain pair \( C_\alpha \) distances are shown as dashed lines and colored by the normalized standard deviation across all trajectories. (D) Representative structure macrostate 0 (light cyan) compared with macrostate 4 (salmon). (E) Representative structure macrostate 3 (green) compared with macrostate 4 (salmon). (F) Distribution of global helical tilt of \( \alpha_3 \) in CCD for each macro-state.

Owing to the rigidity of the \( \alpha \) helices, the pair CA distances give a poor characterization of allosteric signal transduction. To explore the correlation between SH2 domain conformational changes and CCD conformations, the \( \alpha \)-helices of CCD were characterized by the global tilt.
Significant differences in global helical tilt for each of the macro-states of the SH2 domain are observed in α3, showing functional correlative differences among macro-states (Figure 6.3 F). The inter-domain pair Cα distances standard deviation analysis (shown in Figure 6.3 C) shows that this motion from CCD is transmitted allosterically to SH2 via a “rigid core”. A rigid core, in STAT3, can be defined as an interlocking sequence of conserved interactions that function as a sort of molecular machine. α3, α20, α21 compose the first section of the rigid core between CCD, DBD, and LD (Figure 6.3 C), which conveys the dynamics of CCD into this highly rigid region. Upon close inspection, the three helices α3, α20, and α21 are locked via a hydrogen bond network between I252, Q511, and W474 (Figure 6.3 B). Such that any rotation of these residues results in a corresponding reorientation of the helices to preserve the hydrogen bond network. To further elucidate the allostery pathway and show dynamical correlation, we employ both a Relative entropy-based dynamical allosteric network (REDAN) analysis and an analysis of differences in the global hydrogen bond network among macro-states. REDAN analysis was employed as a means to identify residue pairs that are responsive to allosteric perturbation, followed by shortest path analysis using Dijkstra’s algorithm. The most structurally relevant pathway from effector residue (D/A170) to regulatory site (Y657) was identified by REDAN and is shown in Figure 6.3 A. The pathway originates from CCD, through LD, and to the SH2 domain, bypassing DBD, despite passing through a nearby α20, which was identified as a component of the rigid core.

The specific allosteric pathway was further elucidated via differential hydrogen bonding analysis. These analyses both point to a very specific mechanism (structural details shown in 6.3 D and E): 1) stronger interaction between α5 and α2 causes a tilt in the α2/α3 helix, 2) α3 tilt interferes the interaction between β22/α23 and α26/α27. In macrostates 0 and 1, the interference breaks the salt bridge between D566 and R335 and causes α26 to shift away β22. While in macrostate 3, α26 shifts toward β22, which causes a steric clash between Lys 573 and β22. 3) in turn, the movement of α26 and α26/α27 causes breakage of the hydrogen bonds between α33 and the α32-α33 loop, 4) α33 extends significantly and alters
the conformation of the pY+3 pocket. We also identified conserved interactions between α26 and β29 in SH2, which may provide further coupling.

6.3. Conclusion

To summarize, we identify a distinct network of highly inflexible network hydrogen bonds that transmits any perturbation in CCD to affect the conformation of the SH2 domain. This network can be leveraged in drug discovery projects to identify small molecules that bind the CCD domain as an effective allosteric inhibitor of the SH2 domain. Such a small molecule would be highly desirable, as the STAT3 signal transduction relies on a highly sensitive positive feedback loop. Effective inhibition of STAT3 homodimerization can be used to inhibit the signal transduction rapidly and does not rely on inhibition of kinases, which risk high degrees of cross-reactivity.

This project demonstrates the feasibility of large-scale molecular studies using computational techniques, whereas QM methods are impractical.
CHAPTER 7
Conclusion

The development of quantum chemistry methods tailored for medium to large molecules holds significant importance. This dissertation is primarily focused on contributing to this area. Specifically, Chapters Three and Four delve into the exploration of rank-reduced approximations of Coupled-cluster theory for ground states, aiming to broaden its utility for large molecules.

In Chapter Three, the behavior of LS-THC methods within open-shell systems is investigated in the context of MP2/MP3 theory. The study reveals that open-shell LS-THC behaves similarly to closed-shell systems. Moreover, the presented diagrammatic derivation method facilitates the implementation of open-shell THC methods with minimal additional complexity compared to their closed-shell counterparts. LS-THC demonstrates the capability to achieve an accuracy of 1 kcal/mol, though further methodological enhancements are required for improved precision.

In Chapter Four, five distinct approximation schemes of CCSD(T) are explored, incorporating various tensor decomposition techniques aimed at thermochemical applications with a target error margin of 0.1 kJ/mol. Among these schemes, two novel approaches show promise as viable candidates for approximating thermochemistry calculations effectively.

Chapter Five represents a continued endeavor towards understanding excited states. While EOM-CC serves as a reliable "black-box" model for the excited state, the gradient of EOM-CCSD was introduced approximately 30 years ago. Despite the development of numerous iterative and non-iterative triple corrections to EOM-CCSD over time, the ana-
lytic gradient remained unexplored. In this chapter, the theoretical framework and practical application of the first-order analytic gradient within the EOM-CCSD* method are elucidated. This advancement empowers the study of excited states in medium-sized molecules with enhanced theoretical precision.

In Chapter Six, the critical significance of molecular mechanisms in elucidating protein dynamics and conformational landscapes is emphasized through an examination of the allosteric mechanism of the STAT3 protein. This investigation offers profound insights into the allosteric mechanism, furnishing atomistic details regarding the regulation of the CCD to SH2 domain in the STAT3 protein. Such insights not only deepen our understanding of protein dynamics but also pave the way for informed drug design strategies targeting the CCD to evoke desired downstream effects on the SH2 domain and, consequently, modulate the overall function of STAT3.
APPENDIX A

Publication on the implementation of LS-THC on open-shell MP2/MP3

Tingting Zhao, Megan Simons, and Devin A. Matthews Journal of Chemical Theory and Computation 2023 19 (13), 3996-4010 DOI: 10.1021/acs.jctc.3c00392
Open-Shell Tensor Hypercontraction
Tingting Zhao, Megan Simons, and Devin A. Matthews*

ABSTRACT: The extension of least-squares tensor hypercontracted second- and third-order Møller–Plesset perturbation theory (LS-THC-MP2 and LS-THC-MP3) to open-shell systems is an important development due to the scaling reduction afforded by THC and the ubiquity of molecular ions, radicals, and other open-shell reactive species. The complexity of wavefunction-based quantum-chemical methods such as Møller–Plesset and coupled cluster theory is reflected in the steep scaling of the computational costs with the molecular size. The least-squares tensor hypercontraction (LS-THC) method is an efficient, single-step factorization for the two-electron integral tensor but can also be used to factorize the double excitation amplitudes, leading to significant scaling reduction. Here we extend this promising method to open-shell variants of LS-THC-MP2 and MP3 by using diagrammatic techniques and explicit spin summation. The accuracy of the resulting methods for open-shell species is benchmarked on test set systems such as regular alkanes as well as realistic systems involving bond breaking, radical stabilization, and other effects. We find that open-shell LS-THC-MP methods exhibit errors highly comparable to those produced by closed-shell LS-THC-MP methods and are highly insensitive to particular chemical interactions, geometries, or even moderate spin contamination.

INTRODUCTION

Complexity arising due to the presence of high-order tensors is well-known as a bottleneck in wavefunction-based quantum-chemical methods, leading to steep scaling of computational cost with system size. As the essential component of all electronic structure methods, the fourth-order electron repulsion integral (ERI) tensor, \( \mathcal{T} \), is the most obvious target for cost and scaling reduction. Many approaches have been developed to tackle the complexity of the ERIs, such as the density fitting (DF) approximation,\(^7\) resolution of the identity (RI),\(^1\) Cholesky decomposition (CD),\(^11\) the pseudospectral (PS) method,\(^15\) and more recently tensor hypercontraction (THC).\(^16\)–\(^21\) Approximations of the wavefunction itself have received less attention, but due to the presence of the fourth-order double excitation amplitudes (\( C_4 \) or \( P_4 \)), which appear in virtually any wavefunction theory,\(^12\) some approximation is necessary to achieve significant reductions in computational scaling.\(^19\),\(^25\),\(^30\)

While a wide range of approximations to the ERIs and/or doubles amplitudes have proven to be highly successful, the vast majority of these methods have been implemented, tested on, and subsequently applied to closed-shell systems. While closed-shell systems are of vital importance to chemistry as a whole—neutral stable chemical species, both reactants and products, tend to adopt a closed-shell configuration—open-shell species are of equal importance due to the high reactivity and richness of electronic structure exhibited by radical species.\(^31\) Such species are central to a number of chemical fields, such as organic catalysis,\(^35\) environmental and health studies,\(^33\) combustion and alternative fuel chemistry,\(^34\) and photochemistry.\(^35\),\(^36\) Additionally, due to the reactive nature of radical transient intermediates or transition states, they are often difficult to study in the laboratory, requiring more elaborate instrumentation or complicated spectroscopic analysis.\(^38\) Thus, it is necessary to develop quantum-chemical methods to assist in the experimental studies of these systems, along with approximation techniques to reduce computational scaling for the study of larger systems.

We are especially interested in the least-squares tensor hypercontraction (LS-THC) method\(^18\),\(^20\) as an efficient approximation of both the ERI and doubles amplitude tensors. We and others have previously demonstrated the potential for high accuracy and low scaling (\( O(N^4) \), where \( N \) is a measure of system size) for closed-shell third-order Møller–Plesset perturbation theory (MP3).\(^21\),\(^29\) In LS-THC, the fourth-order ERI tensor is approximated as a (hyper)product of five matrices: four collocation matrices, determined by evaluation of the molecular orbitals at a set of grid points, and a core

Received: April 6, 2023
Published: June 23, 2023
matrix evaluated by least-squares fitting of the canonical ERIs (or ERIs approximated by a second method such as density fitting). In the "MP3a" variant, this approximation is combined with the Laplace transform quadrature technique in order to compute the energy without referencing the doubles amplitudes explicitly. Alternatively, the "MP3b" method invokes an LS-THC factorization of the first-order double amplitudes and instead avoids the computational complexity of the Laplace transform quadrature (and a quadratic scaling with the number of quadrature points). Lastly, the "MP3c" and "MP3d" methods directly approximate the second-order doubles amplitudes using an energy functional form identical to that used for coupled cluster and other more complex methods.

This range of interpretations of MP3, as well as the basic building blocks of the MP3 energy (especially the well-known particle–particle, hole–hole, and “ring” terms), makes MP3 an ideal stepping stone to more complex theories such as coupled cluster with single and double excitations (CCSD).

In this work, an extension of the LS-THC approach to open-shell MP2 and MP3 energies was developed by using a mixed graphical–algebraic technique for deriving the working equations. We have implemented the MP3b variant of LS-THC-MP3, but the techniques developed are immediately applicable to the other MP3 variants as well as more complex LS-THC methods.

**THEORY**

The following notational conventions are used throughout this work:

- The letters $pqrs$ denote arbitrary molecular orbitals (MOs).
- The letters $abdef$ $(ijklnn)$ denote virtual (occupied) MOs.
- The letters $RSTUVWXYZ$ denote grid points.
- The letters $JK$ denote auxiliary (density fitting) basis functions.
- Where applicable, $\alpha (\beta)$ molecular orbitals, pruned grid points, and other quantities are denoted by an overbar (no overbar).

**Least-Squares Tensor Hypercontraction.** LS-THC is a method that combines the desirable features of several other factorizations: the representation of electron distributions over a linear-scaling auxiliary basis as in DF/BF/CD methods, the pseudospectral method’s flexibility of factoring exchange terms, the use of least-squares fitting as in DF and alternating least-squares-based CP factorizations (although the LS-THC factorization is linear and noniterative), and finally the least-squares form a grid-based expansion as well as the numerical integration of exchange–correlation functionals in density functional theory. The LS-THC form of the ERIs is

\[
\langle pr | qs \rangle \equiv g_{pq}^{rs} \approx \sum_{RS} (X^{(pr)})_p (X^{(qs)})_q V_{RS} (X^{(qs)})_q (X^{(qs)})_q
\]

The matrices $X$ are the collocation matrices, determined $\text{a priori}$ by evaluation of the (spatial) molecular orbitals $y_{\alpha}$ at a set of grid points $r_{\alpha}$: $X_{\alpha} = y_{\alpha} (r_{\alpha})$. Superscripts such as $\langle pr \rangle$ in eq 1 differentiate different sets of “pruned” grid points. The pruning process is specific to the occupation of the pair of molecular orbitals, leading to separate occupied–occupied (X $\equiv X^{(\uparrow\uparrow)}$), mixed virtual–occupied (X $\equiv X^{(\uparrow\downarrow)} \equiv X^{(\downarrow\uparrow)}$), or virtual–virtual (X $\equiv X^{(\downarrow\downarrow)}$) collocation matrices. The pruning procedure here is based on the Cholesky decomposition, but other pruning procedures are possible, e.g., based on domain decomposition. These three cases may be further classified by the spin of the associated molecular orbitals (differentiated by the absence or presence of an overbar on the MO and grid indices), leading to six unique collocation matrices for open-shell LS-THC-MP3.

The final matrix $V$ is the core matrix and is evaluated by least-squares fitting of either the exact ERI tensor or some intermediate approximation. In this work we utilize density fitting in order to maintain an overall scaling of $O(N^4)$:

\[
V_{pq} = \arg\min_{v_{qs}} \frac{1}{2} \sum_{prs} \left( g_{pq}^{rs} - \sum_{RS} (X^{(pr)})_p (X^{(qs)})_q V_{RS} (X^{(qs)})_q (X^{(qs)})_q \right)^2
\]

\[
E_{RSP} = \sum_{pqr} (X^{(pr)})_p (X^{(qs)})_q g_{pq}^{rs} (X^{(qs)})_q (X^{(qs)})_q
\]

\[
S_{pq}^{rs} = \sum_{pq} (X^{(pr)})_p (X^{(qs)})_q g_{pq}^{rs} (X^{(qs)})_q (X^{(qs)})_q
\]

where in each case a superscript $\beta^{-1}$ is understood to denote an element of the matrix inverse rather than an inverse of the matrix element itself. Note that as with the collocation matrices, there are as many as six distinct metric matrices $S$. Due to the possible combination of each of these six electron distributions, there are as many as 21 different core ($V$) and fitting ($E$) matrices, of which 13 are utilized in open-shell LS-THC-MP3 (assuming a canonical Hartree–Fock reference, for which the single excitation amplitudes can be neglected). In order to avoid notational clutter, we rely on context to determine which of these 13 core matrices is intended unless explicitly specified.

**Canonical Formulation of MP2 and MP3.** In the canonical spin–orbit representation, the MP2 and MP3 energies are defined as

\[
E_{\text{MP2}} = E_{\text{MP2a}} = \sum_{\alpha \beta} \frac{(g_{\alpha \beta}^{\alpha \beta} - g_{\alpha \beta}^{\beta \alpha}) g_{\alpha \beta}^{\beta \alpha}}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\alpha} - \epsilon_{\beta}}
\]

\[
E_{\text{MP3}} = E_{\text{MP3a}} = \sum_{\alpha \beta \gamma \delta} \frac{(g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\delta \gamma}) g_{\alpha \beta}^{\delta \gamma}}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\gamma} - \epsilon_{\delta}} + \frac{1}{2} \sum_{\alpha \beta \gamma \delta} \frac{(g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\delta \gamma}) (g_{\alpha \beta}^{\delta \gamma} - g_{\alpha \beta}^{\gamma \delta}) (g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\beta \gamma}) (g_{\alpha \beta}^{\beta \gamma} - (g_{\alpha \beta}^{\delta \gamma}) (g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\beta \gamma}) (g_{\alpha \beta}^{\beta \gamma} - g_{\alpha \beta}^{\gamma \delta})}
\]

\[
E_{\text{MP3}} = E_{\text{MP3a}} = \sum_{\alpha \beta \gamma \delta} \frac{(g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\delta \gamma}) g_{\alpha \beta}^{\delta \gamma}}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\gamma} - \epsilon_{\delta}} + \frac{1}{2} \sum_{\alpha \beta \gamma \delta} \frac{(g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\delta \gamma}) (g_{\alpha \beta}^{\delta \gamma} - g_{\alpha \beta}^{\gamma \delta}) (g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\beta \gamma}) (g_{\alpha \beta}^{\beta \gamma} - (g_{\alpha \beta}^{\delta \gamma}) (g_{\alpha \beta}^{\gamma \delta} - g_{\alpha \beta}^{\beta \gamma}) (g_{\alpha \beta}^{\beta \gamma} - g_{\alpha \beta}^{\gamma \delta})}
\]
Note that the more common formulation of these energies uses the fully second-quantized representation of the Hamiltonian:

$$\hat{H} = \sum_{pq} \epsilon_{pq}^{\text{ij}} a_{pq}^{\dagger} a_{pq} \| + \frac{1}{4} \sum_{pqrs} \epsilon_{pqrs}^{\text{ij}} a_{pq}^{\dagger} a_{rs}^{\dagger} a_{rs} a_{pq} \|$$

where \(\{|i\}\) denotes normal ordering. The two representations are connected by the simple identity \(\epsilon_{pq}^{\text{ij}} = \epsilon_{pq}^{\text{ij}} - \epsilon_{qj}^{\text{ij}}\). The first-order and second-order perturbed double excitation amplitudes are most commonly defined by

$$t^{[1]}_{\text{ab}} = t^{[1]}_{\text{ab}} - t^{[1]}_{\text{ab}} \| = \frac{\int_{0}^{\infty} \epsilon_{\text{ab}}^{\text{ij}} e^{-t_{\text{ab}}^{\text{ij}}} \| dt}{\int_{0}^{\infty} e^{-t_{\text{ab}}^{\text{ij}}} \| dt}$$

$$t^{[2]}_{\text{ab}} = \frac{1}{\epsilon_{\text{ab}}^{\text{ij}}} \sum_{l} (g^{\text{ab}}_{l} - g^{\text{ab}}_{l}) (t^{[1]}_{\text{ab}}(l) - t^{[1]}_{\text{ab}}(l)) + \sum_{g^{\text{ab}}_{l} - g^{\text{ab}}_{l}) (t^{[1]}_{\text{ab}}(l) - t^{[1]}_{\text{ab}}(l)) + \sum_{g^{\text{ab}}_{l} - g^{\text{ab}}_{l}) (t^{[1]}_{\text{ab}}(l) - t^{[1]}_{\text{ab}}(l))$$

The representation of \(t^{[n]} \) in a non-antisymmetric form is critical in the following derivation, where we also show that such a representation is always well-defined by following a physically motivated graphical derivation. With these definitions we can write several additional variant forms of the MP2 and MP3 energies, although all variants are mathematically identical when exact ERIs and perturbed amplitudes are used.

$$E_{\text{MP2a}} = \frac{1}{2} \sum_{ab} (g_{ab}^{\text{ij}} - g_{ab}^{\text{ij}}) t^{[1]}_{\text{ab}}$$

$$E_{\text{MP3a}} = \frac{1}{2} \sum_{abkl} (t^{[1]}_{\text{abkl}} - t^{[1]}_{\text{abkl}}(g_{ab}^{\text{kl}} - g_{ab}^{\text{kl}})(t^{[1]}_{\text{abkl}} - t^{[1]}_{\text{abkl}}))$$

$$E_{\text{MP3c}} = \frac{1}{2} \sum_{ab} (g_{ab}^{\text{ij}} - g_{ab}^{\text{ij}}) t^{[2]}_{\text{ab}}$$

A further variant, MP3d, is obtained by separating the formation of \(t^{[3]} \) into two parts: evaluation of the residual starting with the first-order amplitudes and division of the residual by the orbital energy differences to obtain the second-order amplitudes. In ref 21 we showed that this approach results in a distinct LS-THC-MP3 method with lower cost and nearly identical numerical error.

Before moving on to the derivation of the LS-THC-MPn approximations, we must deal with the inseparability of the energy denominators present in eqs 6–10. A convenient approach is the method of “Laplace denominators” pioneered by Almlöf and others.

$$\frac{1}{\epsilon_{\text{ab}}^{\text{ij}}} \approx \frac{1}{\epsilon_{\text{ab}}^{\text{ij}}} = \int_{0}^{\infty} e^{-t_{\text{ab}}^{\text{ij}}} \| dt$$

where \(L \) is the number of Laplace quadrature points. In this work we use the quadrature developed by Braess and Hackbusch and a sufficient number of quadrature points to evaluate \(1/x \) to a relative accuracy of \(10^{-8} \) (approximately nine points).

**Closed-Shell LS-THC-MPn.** The MP2a and MP3a formulations of MP2 and MP3, which are defined solely in terms of integrals and orbital energy denominators, are easily represented in compressed form by using tensor hyper-contractive. Spin integration for closed-shell systems is straightforward (see, e.g., ref 1), and we may exploit the equivalence of the different spin cases of the integrals:

$$g_{ab}^{\text{ij}} = g_{ab}^{\text{ij}}$$

We may then expand the equations in terms of, say, \(g_{ab}^{\text{ij}} \) alone and then approximate these integrals using the form of eq 1 while also approximating the energy denominators using eq 14. This results in the LS-THC-MP2a and LS-THC-MP3a methods:

$$E_{\text{LS-THC-MP2a}} = -\sum_{ab} \sum_{RSTU} \sum_{ijkl} t^{(2)}_{abkl} t^{(2)}_{abkl} (2X^{R}_{ab} X^{R}_{ab} X^{S}_{ab} X^{S}_{ab})$$

$$E_{\text{LS-THC-MP3a}} = \sum_{ab} \sum_{RSTU} t^{(2)}_{abkl} t^{(2)}_{abkl} (2X^{R}_{ab} X^{R}_{ab} X^{S}_{ab} X^{S}_{ab})$$

$$E_{\text{LS-THC-MP3c}} = \sum_{ab} \sum_{RSTU} t^{(2)}_{abkl} t^{(2)}_{abkl} (2X^{R}_{ab} X^{R}_{ab} X^{S}_{ab} X^{S}_{ab})$$

The "b" variants are simply obtained by directly approximating the first-order (orbital) doubles amplitudes using THC:

$$t^{[1]}_{\text{ab}} \| = \int_{0}^{\infty} e^{-t_{\text{ab}}^{\text{ij}}} \| dt$$


This factorization enables each step to be completed in at most $O(N^4)$ time, given a linear number of occupied and virtual orbitals as well as grid points. The full working equations for the closed-shell case are given in ref 21, and those for the open-shell case are given in the Supporting Information.

**Open-Shell LS-THC-MPn.** In the preceding section, we used the non-antisymmetrized doubles amplitudes $\tilde{t}^{[4]}$. We can justify the validity of such a representation in the open-shell case and use it to easily derive open-shell variants of LS-THC-MPn methods by utilizing diagrammatic methods.

First, we may recognize each term in eqs 6, 7, and 10, after distribution of the parentheses, as a unique Goldstone diagram. Here we use (non-antisymmetrized) Goldstone diagrams instead of the usual antisymmetrized Goldstone diagrams (ASGs), also called Brandow diagrams. As will be seen below, avoiding explicit antisymmetrization allows the same-spin amplitudes and integrals to be factorized in a way which avoids factorization of “exchange-like” terms. The unique Goldstone diagrams are reproduced in Figures 1, 2, and 3, respectively. In particular, the diagrams in Figure 3 are easily recognized as the necessary contributions to the closed-shell second-order $\tilde{T}$ amplitudes or, with replacement of the bottom integral vertex by doubles amplitudes, the iterative $\tilde{T}_2 \rightarrow \tilde{T}_2$ contributions in coupled cluster theory. In the closed-

Figure 1. Goldstone diagrams for the MP2 energy, omitting denominator lines for clarity.

Figure 2. Goldstone diagrams for the MP3 energy, omitting denominator lines for clarity.
shell interpretation, closed loops contribute a factor of 2 and each diagram is symmetrized, leading to the well-known expression:

$$
(\varepsilon_i + \varepsilon_b - \varepsilon_j - \varepsilon_a) f^{[2]}_{\text{ij}} = \sum_{\sigma_f} \phi_{\sigma_f} \Phi_{\sigma_f}^{\text{ij}} + \sum_{\text{mn}} \phi_{\text{mn}} \Phi_{\text{mn}}^{\text{ij}}
$$

$$
+ \sum_{\text{em}} \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}} + (1 + p_{\text{ij}}) \left( \sum_{\text{em}} 2 \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}} \right)
$$

$$
- \sum_{\text{em}} \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}} - \sum_{\text{em}} \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}}
$$

(25)

where the permutation operator $P_{\text{ij}}$ exchanges the top labels with those on the bottom in the following expression. Note that in the last term on the right-hand side the $ij$ orbitals are ordered differently than in the remaining terms. This highlights the rule for Goldstone diagrams that, for orbitals sharing the same "column" in the external vertex (e.g., $ai$ or $bj$), we must be able to follow a continuous loop through the diagram from one label to the other. Thus, the "PHCC" diagram of Figure 3 results in such modified labeling. This choice is discussed further and theoretically motivated below.

For the closed-shell case we may use $\tilde{t}^{[1]}_{\text{ij}}$ without ambiguity since it is precisely defined as the mixed-spin amplitudes $t^{[1]}_{\text{ij}}$ due to the relationship between the various spin cases for a closed-shell reference. However, in the open-shell case, we could instead replace the factor of 2 for closed loops by an explicit summation over spin. In any case, each contraction line along a loop must carry the same spin since contraction implies a spin integral over orthonormal spin functions. Then we can derive equations for the three distinct spin cases in the spin-unrestricted formalism:

$$
(\varepsilon_i + \varepsilon_b - \varepsilon_j - \varepsilon_a) f^{[2]}_{\text{ij}} = \sum_{\sigma_f} \phi_{\sigma_f} \Phi_{\sigma_f}^{\text{ij}} + \sum_{\text{mn}} \phi_{\text{mn}} \Phi_{\text{mn}}^{\text{ij}}
$$

$$
+ (1 + p_{\text{ij}}) \left( \sum_{\text{em}} 2 \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}} \right)
$$

$$
- \sum_{\text{em}} \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}} - \sum_{\text{em}} \phi_{\text{em}} \Phi_{\text{em}}^{\text{ij}}
$$

(26)

Antisymmetrization of eqs 26 and 28 in accordance with our definition of $f^{[1]}_{\text{ij}}$ arrives precisely at the standard equations for the second-order amplitudes in a (canonical) unrestricted Hartree–Fock reference. The application of Goldstone diagrams along with explicit spin summation then gives us a rather simple route to derive the open-shell working equations. The use of Goldstone diagrams also offers a straightforward way to implement THC for spin-component-scaled Möller–Plesset (SCS-MP) perturbation theories by simply including a scaling coefficient in each spin-labeled Goldstone diagram. The necessary equivalence of these equations with the standard spin-integrated form after antisymmetrization and the fact that these equations trivially reduce to the one given above in the closed-shell case are what we mean by claiming that $f^{[2]}$ is well-defined. As we discuss below, however, these amplitudes and their factorized form are not numerically well-defined, but we can make a consistent choice based on theoretical arguments.

A recursive application of eqs 26–28, with additional terms accounting for single excitation amplitudes, provides a route to define non-antisymmetrized amplitudes for methods such as coupled cluster with single and double excitations (CCSD). In the present work, we focus on LS-THC-MP3b, where in fact we only require $f^{[1]}$. These amplitudes are trivially defined in terms of the orbital two-electron integrals:

$$
\begin{align*}
\tilde{t}^{[1]}_{\text{ij}} &= \sum_{\text{RS}} X_{X} X_{Y} t^{2\text{RS}}_{\text{ij}} X_{X}^{S} X_{Y}^{S} \\
\tilde{t}^{[2]}_{\text{ij}} &= \sum_{\text{RS}} X_{X} X_{Y} t^{2\text{RS}}_{\text{ij}} X_{X}^{S} X_{Y}^{S} \\
\tilde{t}^{[3]}_{\text{ij}} &= \sum_{\text{RS}} X_{X} X_{Y} t^{2\text{RS}}_{\text{ij}} X_{X}^{S} X_{Y}^{S} \\
\end{align*}
$$

(29–31)
We note that these chosen definitions, while perhaps obvious given the close relationship of the first-order amplitudes to the integrals, are also derivable using the diagrammatic technique outlined above.

Returning to the issue of relabeling the \( ij \) indices in the ring terms, we do note that for the same-spin amplitudes, every term may in fact be written in one of two ways due to the \( P(\{ij\}) \) [or equivalently \( P(ab) \)] factor which relates the antisymmetric and non-antisymmetric amplitudes. Thus, we may technically define a large number of equivalent non-antisymmetric amplitudes. However, once a tensor factorization such as THC is applied, these different definitions are no longer equivalent and may incur errors of differing magnitude. In this regard, our specific choice of eqs 26 and 28 is motivated by the physical compressibility (in an information-theoretic sense) of interelectron interactions. The "same column, same loop" rule espoused above allows us to interpret a tensor factorization of the entire diagram as a "recompression" step. In Figure 4 we conceptually identify regions of each Goldstone diagram with the resulting factors in the THC compressed form: electron pairs are represented by a pair of collocation matrices, e.g., \( X^a X^b \), which effect a transformation from molecular orbital to grid space, and pair correlation is represented by the core matrix \( T^{[2]} \). The orbital labeling in Figure 4 demands that the collocation matrices capture some correlation effects leading to "dressed" particle–hole pairs (which contributes to the numerical errors discussed previously by us\(^2\)), while the interaction must capture bare, dressed, and higher-dimensional interactions via a linear-scaling extraction of important features, in a similar spirit to the singular value decomposition and other low-rank decompositions. Switching either the \( ij \) or \( ab \) labels destroys this clear separation in terms of local (raw or dressed Coulombic) interactions of dressed particle–hole pairs and instead introduces a picture with more exchange-like interactions. The THC decomposition, like most low-rank decompositions, is not able to effectively capture such nonlocal interactions.

A final numerical issue for the \( i_{ij}^{ab} \) and \( i_{ij}^{EP} \) amplitudes is the exclusion-principle-violating (EPV) amplitudes, which occur when \( i = j \) or \( a = b \). In the canonical equations, these amplitudes cancel after antisymmetrization and do not affect the energy or properties in any way. Thus, we could assign these amplitudes to any numerical value. However, when applying the tensor hypercontraction approximation, information from all doubles amplitudes is mixed together to determine the elements of the core matrix \( T^{[2]} \) via least-squares fitting. Thus, the EPV amplitudes potentially contribute to the LS-THC-MP3 energy. We argue that in order to minimize the impact of EPV amplitudes, we should choose a value for these amplitudes that produces the most accurate THC decomposition of the doubles amplitudes as a whole. This guideline is based on the fact that as the THC decomposition approaches exactness (e.g., as the grid size is increased), the cancellation of the EPV terms, regardless of their numerical value, becomes more complete. This approach is quite different from setting the EPV amplitudes themselves to the smallest possible value (namely, zero). Thus, we suggest not to modify the EPV terms from their definition as obtained using the above diagrammatic approach, e.g., eqs 26–28. Because the THC approximation captures the global mathematical structure of the amplitudes (i.e., it is an interpolation), a consistent choice of EPV and non-EPV terms should provide the most compressible amplitudes. For \( i_{ij}^{(1)} \), which are the only amplitudes used in LS-THC-MP3b, the proposed choice of the EPV amplitudes also corresponds directly to the "correct" Coulomb self-interaction.

Following the diagrammatic method using Figure 2 (where we replace the top and bottom Hamiltonian vertices by the first-order amplitudes), followed by THC approximation of the integrals and first-order amplitudes for each spin case, we arrive at the equations for open-shell LS-THC-MP3b:
These equations bear a striking similarity to those in the closed-shell case, where each open-shell term corresponds exactly to one of the closed-shell terms, except for numerical prefactor and the spin of each electron (loop). We leverage this similarity in our implementation by adding loops over spins to the closed-shell code, resulting in a highly efficient and maintainable implementation.

In summary, the procedure for deriving open-shell THC methods is quite straightforward:

1. Enumerate all distinct Goldstone diagrams, for example by expanding the Brandom diagrams/ASGs via permutation of each vertex’s indices.

2. Algebraically evaluate each diagram as usual, except that the factor of 2 for closed loops is replaced by an unrestricted sum over \( \alpha \) and \( \beta \) spin for each loop (with the same spin for all indices in the loop). For open diagrams (amplitude equations), the spin of external loops is fixed and determined by the spin of the corresponding amplitude to be determined.

3. Replace each two-electron integral and double excitation amplitude by its THC-factorized form. For open diagrams, additional collocation matrices are applied to the external lines in order to form the fitting matrix which will be used to determine the core matrix by least-squares fitting as in eqs 3 and 4.

### COMPUTATIONAL DETAILS

Open-shell LS-THC-MP2a, -MP2b, and -MP3b were implemented in a development version of CFOUR. We tested the accuracy of these methods on four types of test systems:

1. Linear alkyl radicals (H(CH\(_2\))\(_n\)*, \( n = 1-20 \)): Geometries were optimized at the B3LYP/def2-TZVP level with D3 dispersion correction, \(^{1-14}\) starting with synthetic
structures with $R_{CC} = 1.54$ Å, $R_{CH} = 1.1$ Å, and tetrahedral angles.

2. Microsolvated 2H-2-azabicyclo[1.1.1]pentane radical (Figure 5): The aqueous microsolvation environment and solute radical geometry were determined from a short QM/MM simulation (see the Supporting Information for details). Up to the 11 closest water molecules (based on the distance of the water oxygen atom to the closest solute atom) were retained in the THC calculations. Solvation energies were computed without geometry relaxation or counterpoise corrections.

3. Glutathione (Figure 6): A gas-phase structure for glutathione was optimized at the B3LYP/def2-TZVP level with D3 dispersion correction. Heterolytic and homolytic bond cleavage energies were calculated for each bond indicated in Figure 6. Isolated bond cleavage fragments were reoptimized at the same level of theory. The fragment charges after heterolytic cleavage were assigned based on the lowest-energy configuration.

4. 9-Propyl-4,11-tridecadienoic acid (Figure 7): An initial structure was obtained at the same B3LYP/def2-TZVP level with D3 dispersion correction, as well as radical and ionic structures produced by removing in turn each hydrogen atom indicated in Figure 7 followed by reoptimization. We also generated 24 distinct conformations using the FRee Online druG conformation generation (FROG) tool. We then removed the tertiary hydrogen (#5) from each conformation and reoptimized using B3LYP/def2-TZVP.

All B3LYP geometry optimizations were performed with Q-Chem. For all THC calculations, we used the cc-pVDZ basis set, density fitting with the cc-pVdz-R1 auxiliary basis set, and SGO as the parent grid. The parent grid was pruned as in ref 20, based on a numerical cutoff $\varepsilon$ which was varied from $10^{-7}$ to $10^{-4}$ in each experiment. An unrestricted Hartree–Fock (UHF) reference was used in all cases.

In most cases, spin contamination of the UHF reference was negligible (<5%), although certain systems (e.g., some radical bond-breaking fragments of glutathione and the vinlyc tridecanoic acid radical) exhibit moderate spin contamination, with values of $(S^2)$ as high as 0.99. We specifically avoid the use of a restricted open-shell Hartree–Fock (ROHF) reference, as these systems provide a test case for how the THC approximation is affected by spin contamination.

### RESULTS

#### Size Extensivity of the Error

We first examine the error of the THC approximations compared with their density-fitted counterparts for linear alkyl radicals in order to correlate absolute energy errors with system size. Figure 8 gives the THC error per correlated electron. THC calculations with a range of grid cutoff parameters, $\varepsilon$, were performed in order to examine the dependence of the error with grid size (a smaller $\varepsilon$ results in a larger grid and should yield a smaller error). The error per electron quickly reaches an approximately constant value, whether for the total MP3 correlation energy (Figure 8a), the MP2 correlation energy (Figure 8c,d), or the MP3 correction by itself (Figure 8b). The latter contribution is the most irregular, perhaps due to its smaller magnitude and sensitivity to the virtual–virtual grid represented by $X$.

Notably, the error for LS-THC-MP2a is much smaller than that for LS-THC-MP2b, as previously observed in the closed-shell case. As the total LS-THC-MP3b correlation energy includes an LS-THC-MP2b contribution, this error is in fact dominated by the MP2 part, with the MP3 correction error being a minor part for most reasonable choices of $\varepsilon$. In each case, the asymptotic error per electron decreases roughly linearly with respect to $\varepsilon$ (note that the $\varepsilon$ values are chosen on a log scale and the y axis is also logarithmic). Certain narrow ranges of $\varepsilon$ significantly depart from this trend, e.g., $\varepsilon \sim 10^{-12}$ in Figure 8a or $\varepsilon \sim 10^{-33}$ in Figure 8d. These irregularities occur due to a sign change in the error.

Typical density-fitting errors for DF-MP3 seem to be approximately 15 $\mu$Eh/e$^2$ based on previous experiments. This indicates that the LS-THC-MP2b errors are comparable...
for $\varepsilon \sim 10^{-3.2}$, or even earlier near $\varepsilon \sim 10^{-2.0}$ if the LS-THC-MP2b contribution is replaced by LS-THC-MP2a. These results confirm the size extensivity of open-shell LS-THC-MPs, as was observed for the closed-shell variant. Typical per-electron errors are also similar for the closed- and open-shell THC methods; e.g., for $C_8H_{18}/C_8H_{17}^\bullet$ at $\varepsilon = 10^{-2.4}$, we find errors of $1.0/0.35 \mu E_h/e^-$ (MP2a), $102/105 \mu E_h/e^-$ (MP2b), and $6.9/8.7 \mu E_h/e^-$ (MP3b correction), respectively. We also observe, as for closed-shell THC, a “threshold” effect where convergence of the incremental error to the asymptotic value is only reached for a sufficiently long chain. This effect diminishes with looser cutoff values (smaller grids), suggesting a saturation of the orbital pair space for smaller systems.

**Radical Microsolvation Energies.** We next examine the error of the THC approximations for 2H-2-azabicyclo[1.1.1]pentane (ABP) in order to study how an increase in the solvation shell size impacts solvation energy errors. This is an important test, given that subsequent solvent waters will contribute very different physical interactions to the total solvation (interaction) energy. For example, the fifth and seventh waters added (see Figure 5) interact directly with the radical center, and the seventh water at least forms a hydrogen bond with the amino nitrogen. Other waters interact via weaker electrostatic interactions or van der Waals interactions and instead hydrogen-bond with other solvent molecules (e.g., the sixth, eighth, and ninth waters), and some more distant

![Figure 8. Absolute energy errors for open-shell linear alkyl radicals, H(CH$_2$)$_n$$^\bullet$. The absolute value of the per-electron error between the THC and corresponding DF-MP$_n$ calculation is plotted for (a) the total LS-THC-MP3b correlation energy, which is the sum of (b) the LS-THC-MP3b third-order correction and (c) the LS-THC-MP2b correlation energy, and finally (d) the LS-THC-MP2a correlation energy. $\varepsilon$ is the Cholesky decomposition cutoff parameter when pruning the grid, varying logarithmically from $1 \times 10^{-1}$ to $1 \times 10^{-4}$ in steps of 0.2 log units.](image1)

![Figure 9. Errors in incremental solvation energy for 2H-2-azabicyclo[1.1.1]pentane in an explicit water solvation environment. One to 11 waters are added in the order indicated in Figure 5. See the text for details.](image2)
The waters do not seem to form any strong interactions, at least with other fragments included in the present calculations (e.g., the 10th and 11th waters). However, the errors due to THC, presented as percentages of the incremental solvation energy,

Figure 10. Bond dissociation energy errors of glutathione. (a) Absolute BDE error due to THC for homolytic bond cleavage (open-shell). The bond index refers to labeled bonds in Figure 6. (b) Absolute BDE error due to THC for heterolytic bond cleavage (closed-shell). (c) Absolute errors in the difference between the homolytic and heterolytic BDEs due to THC.
do not show any clear trend with the strength or type of solvent interaction (Figure 9). Rather, the errors are relatively consistent at moderate cutoff values (0.001 < \( \varepsilon \) < 0.01). At or below \( \varepsilon = 10^{-2.6} \), errors are consistently less than 5%. The absence of a trend or significant outliers indicates a relative insensitivity of THC-MP3 to different types of interactions and, critically, no bias of the open-shell THC-MP3 method toward or against interactions involving unpaired electrons.

**Bond Cleavages and Radical Stability.** We evaluate the robustness of the THC approximation with respect to various bond cleavage points and charge/spin separation by examining 13 different bond dissociation energies (BDEs) for the glutathione system (Figure 6) and five different hydrogen...
atom abstraction (HA) energies for the tridecadienoic acid system (Figure 7).

We only consider bond breakages in glutathione between backbone C−C or C−N bonds as well as the C−S bond in the cysteine peptide. For each bond cleavage, we calculate both a homolytic (AB → A• + B•) and a heterolytic (AB → A− + B+ or AB → A• + B•, whichever results in more stable products) bond dissociation energy.

The homolytic and heterolytic bond dissociation energy errors with respect to standard density fitting calculations are listed in Figure 10a,b. Overall, the THC approximation produces an accurate bond cleavage energy (around 1 kcal/mol) for most of the bond-breaking cases at reasonable selections of ε (ε < 0.01). Additionally, the error differences between the homolytic (solid lines) and heterolytic (dashed lines) bond cleavage energies are of similar magnitude for each value of ε, and at tighter thresholds follow the same trend toward lower errors for bonds A, F, and L. These bonds cleave the thiol or other terminal functional groups—because of the “threshold” effect noted above, these bond dissociations result in the lowest total error due to near-saturation of the THC grid for the smaller fragment. The closed-shell calculations result in smaller errors for these three cases, perhaps indicating a slight difference in how quickly the grid saturates in the open- and closed-shell cases. However, “moderate” cutoff values do not show such a trend and instead provide a rather consistent magnitude of error irrespective of the bond being broken.

The error in the difference between the homolytic and heterolytic BDEs (ΔBDE) was also computed and is depicted in Figure 10c. Across the various bonds, errors for this relative measure are somewhat lower than for the BDEs themselves. This indicates a reliable error cancellation between the open- and closed-shell LS-THC calculations on similar systems (note that the geometries of the products were optimized separately for hetero- and homolytic cleavages).

In Figure 11a,b, bond dissociation energy errors are shown for H\(^{+},\cdot,\ast\) abstraction from 9-propyl-4,11-tridecadienoic acid. For the closed-shell products (dashed lines; these are all cationic with the exception of the carboxylate), there does seem to be a trend that tends to result in lower errors for the more stable cationic products (substituted alkyl radicals and especially double-bonded sp\(^2\) cationic centers). This may result from enhanced error cancellation between more similar geometries where rehybridization is incomplete. The errors for radical open-shell products (solid lines) are more consistent, perhaps again due to reduced rehybridization even in the primary and secondary carbon radicals. Figure 11c gives the error in the relative energy between the charged and the neutral abstraction products. As for glutathione, there is some cancellation of errors, in particular for the carboxyl and vinyl abstractions, where the geometric changes are more similar.

Conformational Energy Ordering. Again focusing on 9-propyl-4,11-tridecadienoic acid, we examine the relative energies of 24 distinct conformations of the 9H hydrogen abstraction (open-shell radical) and hydride abstraction (cationic) products. In Figure 12, the errors in relative conformational energies due to THC are presented for both
types of products. While it is clear that there is not a strong correlation of the errors between the neutral radical and charged closed-shell products (cf. the lack of any distinct trend in the open-shell errors with respect to the closed-shell errors ordered from smallest to largest at $\varepsilon = 10^{-4}$), both the closed- and open-shell errors seem to span almost identical ranges for each value of $\varepsilon$. The smallest and largest errors, even at a tight tolerance of $\varepsilon = 10^{-22}$, seem to span at least an order of magnitude, although there does not seem to be a trend in the size of the error with any chemically relevant features of the individual conformations. The fact that the smallest errors are significantly lower than observed in, for example, Figure 10 may then simply indicate fortuitous error cancellation, which is not reproducible between the radical and cationic structures. Since the final structures in these cases were optimized independently (despite starting with the same algorithmically generated guess structure), this is perhaps not surprising.

Computational Scaling. In the above sections, we have shown that LS-THC-MP3 can achieve errors below 1 kcal/mol with moderate cutoff values ($\varepsilon \lesssim 10^{-22}$). It is equally important to demonstrate the scaling reduction that can be achieved by the LS-THC method. To that end, we report the timings of both DF-MP3 and LS-THC-MP3 methods for linear alkyl radicals with varying cutoff parameters ($\varepsilon$), analyzed in Size Extensivity of the Error. The timings include the calculation of the MP2 and MP3 correlation energy only. As can be seen in Figure 13, it is clear that LS-THC-MP3 achieves reduced scaling compared to DF-MP3 (see the inset for measured scaling) and reaches a crossover at around 70 correlated electrons with a cutoff of $\varepsilon = 10^{-22}$. Note that the DF-MP3 scaling in this regime is actually dominated by the formation of the $(abcd)$ integrals, which scales as $O(n^8)$, and other lower-scaling operations.

**CONCLUSIONS**

The LS-THC method has proven to be an efficient approximation of both the ERI and doubles amplitude tensors with high accuracy and low scaling for closed-shell systems. Here, LS-THC was implemented for MP2 and MP3 calculations on open-shell systems and was evaluated with various test systems: linear alkyl radicals, a microsolvated amine radical, heterolytic and homolytic bond cleavages in glutathione, and proton/hydrogen/hydride abstractions as well as conformational differences of 9-propyl-4,11-tridecadienoic acid. A number of trends were observed:

1. Similar to the RHF implementation, the LS-THC-MPs errors scale linearly with system size after a critical “threshold” molecular size is reached (this effect is diminished for looser cutoff values). Errors for LS-THC-MP2a are essentially negligible, while errors in LS-THC-MP3b are dominated by the MP2b contribution. Remaining errors in only the LS-THC-MP3b contribution are reasonably small with a double-$\zeta$ basis set.

2. Reliable error cancellation is evident in almost all calculations of energy differences. The greatest degree of error cancellation occurs when the systems are most chemically similar, resulting fragments are unequally distributed in size, or potentially when more stable products are produced. Errors in relative energies are typically below 1 kcal/mol for moderate cutoff values in the range $\varepsilon < 10^{-22}$.

3. The errors produced for open-shell systems are highly similar to the errors produced for chemically similar closed-shell systems. In some cases, error cancellation can also be exploited between open- and closed-shell processes, such as in the relative BDEs of heterolytic and homolytic bond cleavages.

Figure 13. Timings for DF-MP3 and LS-THC-MP3 methods for linear alkyl radicals (H(CH$_2$)$_n^\bullet$, $n = 5$–20) with the cc-pVDZ basis set. Both axes are on a logarithmic scale to highlight the polynomial scaling. All calculations were performed using a single node with 2x Intel Xeon ES-2695v4 CPUs and 256 GB of memory; OpenMP was used to parallelize the calculation over all 36 cores. The inset shows the measured scaling of DF-MP3 and LS-THC-MP3 with $\varepsilon = 10^{-22}$, obtained via linear regression of the timings for $n = 10$–20.
4. The error of the open-shell LS-THC-MPn methods is highly insensitive to the specific nature of the chemical structure, type of interactions, and even moderately severe spin contamination of the reference wavefunction.

In summary, open-shell THC seems to be equally as applicable as closed-shell THC and multireference THC. The diagrammatic method of derivation presented also enables the implementation of open-shell THC methods with little additional effort compared to the closed-shell version and a highly similar code structure which should enable maintainable, efficient codes.

ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00392.

Factorized LS-THC MP3b equations and QM/MM simulation details (PDF)
All calculated DF- and LS-THC-MPn correlation energies (XLSX)
Molecular geometries used for all calculations (XLSX)

AUTHOR INFORMATION

Corresponding Author
Devin A. Matthews – Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; orcid.org/0000-0003-2795-5483; Email: damatthews@smu.edu

Authors
Tingting Zhao – Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; orcid.org/0000-0001-8429-2239
Megan Simons – Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; orcid.org/0000-0001-9457-7019

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00392

Author Contributions
‡T.Z. and M.S. contributed equally to this work.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under Grant DOE-SC0022893 and in part by the U.S. National Science Foundation under Grants OAC-2003931 and CHE-2143725. M.S. was supported by a SMU Center for Research Computing Graduate Fellowship. All calculations were performed on the ManeFrame II computing system at SMU.

REFERENCES


APPENDIX B

Draft on the rank-reduced CCSD(T) in the context of high-accuracy thermochemistry

Tingting Zhao, James Thorpe, and Devin A. Matthews.

Almost ready to submit.
Prospects for rank-reduced CCSD(T) in the context of high-accuracy thermochemistry

Tingting Zhao, James H. Thorpe, and Devin A. Matthews*

Department of Chemistry, Southern Methodist University, Dallas, TX

(*Electronic mail: dmatthews@smu.edu)

(Dated: 10 April 2024)

Obtaining sub-chemical accuracy (1 kJ mol$^{-1}$) reaction energies for medium-sized gas-phase molecules is a longstanding challenge in the field of thermochemical modeling. The perturbative triples correction to CCSD, CCSD(T), constitutes an important component of all high-accuracy composite model chemistries that obtain this accuracy, but can be a roadblock in the calculation of medium to large systems due to its $O(N^7)$ scaling, particularly in HEAT-like model chemistries that eschew separation of core and valance correlation. This study extends the work of Lesiuk, and assesses the accuracy of five different approximations of (T) in the context of a subset of molecules selected from the W4-17 dataset. It is demonstrated that all of these approximate methods can achieve sub-0.1 kJ mol$^{-1}$ accuracy with respect to canonical, density-fitted (T) contributions with a modest number of projectors. The approximation labelled $\tilde{Z}T$ appears to offer the best trade-off between cost and accuracy, and shows significant promise in an order-of-magnitude reduction in computational cost of the CCSD(T) component of high-accuracy model chemistries.
I. INTRODUCTION

Accurately determining reaction energies to sub-chemical accuracy (1 kJ mol$^{-1}$) is crucial to various fields of chemistry and industry,$^1$ such as materials science,$^2$ drug discovery,$^3$ catalysis,$^4$ green chemistry,$^5$ environmental chemistry,$^6$ and many more. Beyond traditional roles, such as transient species investigation,$^7$ thermochemical and kinetics predictions,$^8$ and interpretation of experiment results,$^9$ the rise of machine learning has also accelerated the demand for a large quantity of high-accuracy data as training sets.$^{10}$

Wavefunction-based composite model chemistries are the leading methods of choice for obtaining reaction energies and molecular enthalpies of formation. These techniques assume that the solution of the “true” Schrödinger equation for a particular molecule can be represented as a sum over separable contributions to the energy. These typically include, in order of importance, the non-relativistic electronic energy, the (ro)vibrational zero point energy, corrections for scalar relativistic and spin-orbit effects, and the so-called diagonal Born-Oppenheimer correction. The focus of this work, the non-relativistic electronic energy (NREE), is generally the largest contributor for a particular observable.$^{11-13}$ This term is usually obtained via a further series of additive coupled-cluster calculations, where the effects of higher-order cluster operators are treated with monotonically decreasing basis sets to offset the computational cost. There are few families of such high-accuracy model chemistries that fall into one of two camps: “fixed” recipes such as Wn,$^{11,14,15}$ HEAT,$^{13,16-19}$ and ANL-n,$^{20}$ which prescribe a fixed protocol and obtain Type-A uncertainties,$^{21}$ and “free” recipes such as FPA$^{22-25}$ and FPD$^{12,26-29}$ which provide guidelines for constructing a model chemistry for a molecule of interest and obtain Type-B uncertainties.$^{21}$

While the particulars of these calculations differ from family to family, it is of particular note that every single one of them employs CCSD(T)$^{30}$ as an additivity point in their recipes, often as a component of a $T_3$ correction to the $T_1 + T_2$ correlation obtained by CCSD$^{31}$ (the exception to this statement is the HEAT family of methods, which avoid separation of CCSD and CCSD(T) correlation). CCSD(T) provides a felicitous compromise between accuracy and computational cost and goes a long way towards recovering post-CCSD correlation with basis sets large enough to provide the accuracy required of these model chemistries. This last point is important: CCSD(T) scales as $O(N^7)$ and can be one of the most expensive parts of these model chemistries for larger systems when large basis sets are required, particularly in HEAT-like methods which do not separate the basis-set dependence of CCSD from its perturbative (T) corrections.
Thanks to advances in hardware and software of supercomputers CCSD(T) can now be applied to molecules with over twenty carbons using medium-size basis sets like cc-pVTZ. Further, the development of localized techniques has empowered the calculation of (T)-like corrections for systems containing hundreds of atoms, and have been shown to approach kcal mol\(^{-1}\) accuracy. However, sub-chemical accuracy usually requires (T) contributions calculated with basis sets of at least QZ size (ideally 5Z or higher), and such accuracy places significant strain on localized methods for calculating this quantity. The focus of this paper is to explore how the scaling of this correction can be reduced in the context of high-accuracy computational thermochemistry.

There are four types of terms that can be usefully approximated in the calculation of the perturbative corrections in CCSD(T): the two-electron repulsion integrals (ERIs), the two- and three-electron cluster amplitudes (\(T_2\) and \(T_3\)), and the three-electron orbital eigenvalue denominators (\(D_3\)). Each of these has been studied individually in the literature, and we provide a brief overview of the relevant work here. Density fitting via the Coulomb metric has long been used to approximate the two-electron ERIs. Of the methods discussed here, this is probably the best understood and most well characterized, and for most applications, including this one, results in little-to-no appreciable error in predicted properties. Approximation of the orbital eigenvalue denominator in CCSD(T) via an (inverse) Laplace transform followed numerical quadrature was first studied by Constans et al., who demonstrated that the canonical \(O(N^7)\) scaling could be reduced to \(O(N^6)\), although the leading term still scales unfavorably with the fifth power of the number of basis functions. The Laplace transform can be made essentially exact with a sufficient number of quadrature points, though this results in an increase in the numerical prefactor of the computational cost. In the approximation of \(T_2\), Kinoshita, Hino, and Bartlett demonstrated that the CCD equations could be represented in an SVD subspace, effectively “compressing” the two-electron amplitudes of the CC equations. Parrish et al. later presented a “rank-reduced” coupled cluster (RR-CC) method for compression of \(T_2\), where the compressed doubles amplitudes are solved directly via a Lagrangian formulation of the CCSD equations. These works demonstrated that the \(T_2\) amplitudes were of significantly lower rank than their formal sizes; in fact it appears that the meaningful parts of \(T_2\) scale only linearly with system size. Further, as the compression vectors used in these rank-reduced methods formed an orthogonal basis, orbital-rotation could be used to simplify expressions for the three-electron orbital energy denominator. Similar attempts to approximate the \(T_3\) amplitudes were explored by Hino, Kinoshita, and Bartlett in the context of CCSDT-1, but the computational cost of decomposing \(T_3\) remained prohibitively expen-
sive until developments utilizing Golub-Kahan bidiagonalization and/or higher order orthogonal iteration (HOOI) reduced the cost of decomposing $T_3$-like quantities to $O(N^6)$. Of particular relevance is the recent work of Lesiuk, who used HOOI to reduce the cost of the entire CCSD(T) correction to an overall scaling the same as CCSD. Importantly, that manuscript reported less than 1% errors in the compressed CCSD(T) isomerization energies, which makes it a prime candidate for application in model chemistries aiming for sub-chemical accuracy.

Combinations of these various approximations are explored in this work to attempt to gain some insight into the application of these reduced scaling techniques in high-accuracy thermochemistry. It should be noted that other decompositions of the electron repulsion integrals and CC amplitudes are possible – of particular note are Tensor Hypercontraction methods – but as the focus of this work is on high-accuracy thermochemistry we have elected to focus only on the set of techniques discussed above.

II. THEORY

- The letters $ijkl$ represent one of the $o$ occupied orbitals in the MO basis;

- The letters $abcd$ represent one of the $v$ virtual orbitals in the MO basis;

- The letters $pqrs$ represent one of the $N = o + v$ general orbitals in the MO basis, which may be fewer than the number of basis functions $N_{bas}$ if core orbitals are frozen;

- The letter $g$ represents one of the $N_g$ Laplace transform quadrature points;

- The letters $VW$ represent one of the $N_{T_2}$ rank-reduced $T_2$ projectors;

- The letters $XYZABC$ represent one of the $N_{T_3}$ Tucker-3 compressed $T_3$ projectors;

- The letters $JKL$ represent one of the $N_{DF}$ density-fitting auxiliary basis functions.
A. Perturbative triples correction to CCSD

As a starting point, we take the coupled cluster singles and doubles (CCSD) energy \( E_{CC} \) and amplitudes \( t_n \),

\[
E_{CC} = \langle 0 | \hat{H} | 0 \rangle = \langle 0 | e^{-\hat{F}} \hat{H}_N e^{\hat{F}} | 0 \rangle = \langle 0 | (\hat{H}_N e^{\hat{F}})_C | 0 \rangle \tag{1}
\]

\( \hat{H}_N = \hat{F}_N + \hat{V}_N \)

\[
\hat{F}_N = \sum_{pq} f_p^q \{ a_p^\dagger a_q \} + \frac{1}{2} \sum_{pqrs} v_{pq}^{rs} \{ a_p^\dagger a_q^\dagger a_s a_r \} \tag{2}
\]

\[
\hat{T} = \hat{T}_1 + \hat{T}_2
\]

\[
= \sum_{ai} t_{ai} a_i^\dagger a_i + \frac{1}{4} \sum_{abij} t_{abij} a_i^\dagger a_j^\dagger a_i a_j \tag{3}
\]

where we use the non-antisymmetrized two-electron integrals \( v_{pq}^{rs} = \langle pq | rs \rangle = \langle pr | qs \rangle \). From this point, we will assume the use of a restricted Hartree–Fock (RHF) reference and refer only to spatial orbitals. In CCSD(T), \( E_{CC} \) is supplemented by two terms arising from triple excitations: the so-called fourth-order correction, \( E^{[4]}_T \), and the so-called fifth-order correction, \( E^{[5]}_{ST} \),

\[
E_{(T)} = E^{[4]}_T + E^{[5]}_{ST} \tag{4}
\]

\[
E^{[4]}_T = \frac{1}{3} \sum_{abcij} z_{abc}^{ijk} s_{abc}^{ijk} \tag{5}
\]

\[
E^{[5]}_{ST} = \frac{1}{6} \sum_{abcijk} \left( P_{ijk}^{abc} v_{ijk}^{a} \right) s_{ijk}^{abc} \tag{6}
\]

\[
z_{abc}^{ijk} = P_{ijk}^{abc} \left( \sum_d v_{id}^{ab} v_{id}^{ad} - \sum_l v_{jd}^{cm} v_{td}^{cm} \right) \tag{7}
\]

\[
i_{abc}^{ijk} = z_{abc}^{ijk} D_{ijk}^{abc} \tag{8}
\]

\( D_{ijk}^{abc} \) are the orbital eigenvalue denominators and \( P_{ijk}^{abc} \) and \( s_{ijk}^{abc} \) are permutation operators,

\[
D_{ijk}^{abc} = \frac{1}{(\varepsilon_i + \varepsilon_j + \varepsilon_k - \varepsilon_a - \varepsilon_b - \varepsilon_c)} \tag{9}
\]

\[
P_{ijk}^{abc} = \left( abc \right)_{ijk} + \frac{1}{2} \left( abc \right)_{ikj} + \frac{1}{2} \left( abc \right)_{kij} + \left( bca \right)_{jki} + \left( cab \right)_{kji} \tag{10}
\]

\[
s_{ijk}^{abc} = 4 \left( abc \right)_{ijk} - 2 \left( abc \right)_{ij} - 2 \left( abc \right)_{jk} - 2 \left( abc \right)_{ki} + \left( bca \right)_{ij} + \left( cab \right)_{ij} \tag{11}
\]
FIG. 1: Graphical representations of the various decompositions employed. See text for details.

(A) Left: ERIs ($v_{cb}^{kd}$ or $v_{cm}^{kj}$), right: DF-approximated ERIs; (B) Left: $T_2$ amplitudes, right: rank-reduced $T_2$ amplitudes; (C) Left: $z_{ijk}^{abc}$, right: $t_{ijk}^{abc}$; (D) Left: compressed $z_{ijk}^{abc}$, right: simplified representation; (E) Left: compressed $t_{ijk}^{abc}$, right: simplified representation; (F) Alternative $t_{ijk}^{abc}$ formed using orthogonal rotation, right: simplified representation.

The unfavorable $O(N^7)$ scaling comes from the construction of $z_{ijk}^{abc}$ from the two-electron integral and amplitudes, and approximating this term is particularly important in reducing the cost of CCSD(T). We will use the $E_T^{[4]}$ term (diagrammatically represented in Fig. 3A) to demonstrate the various approximate methods explored here. Scaling reduction of $E_T^{[5]}$ is achieved using the same principles as $E_T^{[4]}$.
B. Density fitting of the ERIs

Density fitting factorizes the four-dimensional ERI tensor into a contraction between two, three-dimensional tensors via a resolution of the identity (RI) of an auxiliary basis set:\textsuperscript{43–45,61}

\[ v_{pr}^{pq} \approx \sum_{KL} (pq|K)(K|L)^{-1}(L|rs) \equiv \sum_{JL} B_{pq}^{L} B_{rs}^{J} \]

where \((pq|K)\) and \((K|L)\) are three- and two-center ERIs. While the basis can be constructed on the fly, pre-optimized auxiliary basis sets are generally the option of choice.\textsuperscript{62–65} \(B_{pq}^{L}\) are the density fitting factors, shown as yellow-green triangles in Fig. 1A. Testing on all the species in this work found that the use of DF-approximated ERIs introduced no more than 0.12 kJ mol\(^{-1}\) of error in any of the \((T)\) contributions to the reaction energies (see Supplemental Information), and thus is used as the “reference” calculation (Figure 3B) for the analysis of the other approximate methods. A further reduction in the density fitting error can be achieved by using a larger auxiliary basis set, which has a relatively minor impact on the cost of the \((T)\) correction, \textit{vide infra}.

C. Rank-reduced \(T_2\) amplitudes

In rank-reduced CCSD,\textsuperscript{50,51} the \(T_2\) amplitudes are approximated as,

\[ t_{ij}^{ab} \approx \sum_{VV} V_{ai}^{V} T^{VV} V_{bj}^{W} \]

where \(V_{ai}^{W}\) is an eigenvector of a reference set of \(T_2\) amplitudes, obtained by folding into a \(\nu_o \times \nu_o\) matrix and then diagonalizing. Typically these amplitudes are taken from an MP2 or MP3 calculation so that they may be fixed throughout the CCSD iterations. \(T^{VV}\) is a core matrix that represents the compressed CCSD \(T_2\) amplitudes. Alternatively, if one does not wish to accelerate CCSD, but rather to approximate \((T)\), it is sufficient to perform an eigenvalue decomposition (EVD) of the converged CCSD \(T_2\) amplitudes,

\[ t_{ij}^{ab} \approx \sum_{V} V_{ai}^{V} \lambda_{V}^{V} V_{bj}^{V} \]

where \(V_{ai}^{V}\) are the eigenvectors of the converged \(T_2\) and \(\lambda_{V}^{V}\) are the associated eigenvalues (diagrammatic representation in Figure 1B). For this work, we adopt the more general formula in
order to retain the option of using MP2 or MP3 $T_2$ compression vectors (or some other choice), in which case $T^{VW}$ is not diagonal. The decomposition is truncated using the absolute value of the $T_2$ eigenvalues as the metric of interest, and results in $N_{T_2} = O(N)$ “projectors” retained in the compressed space. This decomposition incurs $O(\sigma^3 v^3)$ cost, and scales favorably compared to the $O(\sigma^2 v^4)$ of later steps in the process for most basis sets of interest.

D. Tucker-3 factorization of the $T_3$ amplitudes

The Tucker-3 compression of the $Z_3$ intermediates or $T_3$ amplitudes is defined as,

$$z_{abc}^{i,j,k} \approx \tilde{z}_{abc}^{i,j,k} = \sum_{XYZ} Z_{XYZ} U_{ai} X_{Y} U_{ck} Z_{ck}$$  \hspace{1cm} (15)

$$t_{abc}^{i,j,k} \approx \tilde{t}_{abc}^{i,j,k} = \sum_{XYZ} T_{XYZ} U_{ai} Y_{b} U_{ck} Z_{ck}$$  \hspace{1cm} (16)

where the graphical representations of each quantity are depicted in Fig. 1C–E. $Z_{XYZ}$ is the core tensor in the compressed basis and $U_{ai}^X$ are the compression vectors or projectors.\textsuperscript{54,55} This representation can be obtained via a higher-order SVD (HOSVD) of a full, $T_3$-like quantity, but this process would be very expensive. Instead, following the work of Lesiuk\textsuperscript{56} (and the earlier work of Bell, Lambrecht, and Head-Gordon in the context of MP2\textsuperscript{66}) we employ higher-order orthogonal iteration (HOOI) which constructs $U_{ai}^X$ in a self-consistent manner. In most of what follows, the initial guess for $U_{ai}^X$ is seeded from the eigenvectors of $T_2$. These are then partially contracted against the three-electron property in question, such as $z_{ij,k}^{abc}$, obtaining for example $Z_{ck}^{XY}$,

$$Z_{ck}^{XY} = \sum_{abj} z_{ij,k}^{abc} U_{ai} X_{abj} Y_{b} U_{ck} Z_{ck}$$  \hspace{1cm} (17)

Because the projectors $U$ are orthogonal but rectangular, they form an approximate resolution of the identity. Thus, $Z_{ck}^{XY}$ contracted with itself to form a matrix $Z_{ck,c'}^{c'}$, provides eigenvectors and eigenvalues approximating the right singular vectors and values of $z_{ij,k}^{abc}$ formatted as a matrix.
where we assume a self-consistent solution of \( U \). The positive square roots of \( \epsilon^Z \) are used to determine the number of projectors (\( N_{T_3} \)) retained in the compressed space. The associated eigenvectors are then used as the next iteration of guesses for the Tucker-3 compression vectors. Convergence is checked by computing the singular values \( \sigma^X \) of the overlap matrix \( M_{XY} = \sum_{ai} U_{ai}^X U_{ai}^Y (n-1) \) for projectors from successive iterations \( n-1 \) and \( n \). If the value \( |N_{T_3} - \sum_X \sigma^X| \) falls below some cutoff, chosen here as \( 10^{-5} \), then the process is stopped. A schematic of this procedure is shown in Fig. 2. The final core tensor \( Z^{XYZ} \) is then obtained by compressing \( z_{ijk}^{abc} \) with the converged \( U \) factors,

\[
Z^{XYZ} = \sum_{aibjck} z_{ijk}^{abc} U_{ai}^X U_{bj}^Y U_{ck}^Z \tag{21}
\]

E. Laplace denominators and orthogonal rotation

In order to avoid the construction of any full, three-electron tensor in our decomposition of \( Z_3 \) or \( T_3 \), we employ two approximations of the three-electron orbital energy denominators (\( D_3 \)): Laplace transform (LT)\( ^{47,49,67,68} \) and orthogonal rotation (OR)\( ^{52,56} \). In LT, \( D_{ij}^{abc} \) is represented as a sum over Laplace transform quadrature points,

\[
\frac{1}{\varepsilon_i + \varepsilon_j + \varepsilon_k - \varepsilon_a - \varepsilon_b - \varepsilon_c} = -\int_0^\infty e^{-(\varepsilon_a + \varepsilon_b + \varepsilon_c - \varepsilon_j - \varepsilon_k)t} dt \\
\approx -\sum g \tau_g^a \tau_g^b \tau_g^c \tau_g^d \tau_g^e \tau_g^f \tau_g^i \tau_g^j \tau_g^k \tag{22}
\]

where \( \tau_g^s = \omega_g^{1/6} e^{it_g} \) and \( \tau_g^s = \omega_g^{1/6} e^{-it_g} \). \( \omega_g \) and \( t_g \) are the weights and abscissas of a suitable quadrature.\( ^{48} \) When used in the HOOI process described above, the partially-projected \( T_{ck}^{XY} \) term
is then constructed from a sum over these points,

\[ T_{XY}^{ck} = -\sum_g T_{XY}^{ck(g)} \]  

This Laplace transform technique has been known for some time and works remarkably well with between three and twelve Laplace points depending on the level of accuracy required and whether or not core orbitals are included.\(^{47,49,67,68}\)

Alternatively, one can perform an orthogonal rotation of the \( U_{ai}^{X} \) projectors in order to simplify \( D_3 \) without the Laplace transform. When the denominators are applied to the compressed \( z_{ijk}^{abc} \) intermediates, and then the result is immediately recompressed, the projectors can be rotated to a “diagonal” representation such that equivalently a set of orbital denominators can be applied directly to the core tensor \( Z_{XYZ} \).\(^{52}\)

\[ T^{XYZ} = D^{XYZ} Z^{XYZ} \]  

\[ D^{XYZ} = \frac{1}{\varepsilon^X + \varepsilon^Y + \varepsilon^Z} \]  

\[ U^{XY} = \sum_{ai} U_{ai}^{X} (\varepsilon_i - \varepsilon_a) U_{ai}^{Y} = \delta_{XY} \varepsilon^X \]
### Table I: Scaling of the leading term contributing to distinct phases of the approximations discussed. \( N_{\text{it}} \) is the number of iteration cycles in the HOOI procedure.

<table>
<thead>
<tr>
<th>Method</th>
<th>HOOI</th>
<th>( Z_{XYZ} )</th>
<th>( T_{ABC} )</th>
<th>Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{Z}T )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>0</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
</tr>
<tr>
<td>( \tilde{T}Z )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>0</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
</tr>
<tr>
<td>( \tilde{Z}'DZ )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>0</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
</tr>
<tr>
<td>( \tilde{Z}'DZ )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
<td>( N_{\text{DF}} N_{T_3}^{2} )</td>
</tr>
</tbody>
</table>

The required form can be achieved by diagonalizing the matrix \( U_{XY} \) formed from the original projectors, and then rotating the projectors by the matrix of eigenvectors. Except where noted, a Tucker-3 compression of the \( T_3 \) amplitudes, denoted \( \tilde{T} \) in Fig. 1, also implies the inclusion of \( D_3 \) during the HOOI process (that is, \( T_{XY}^{ck} \) is used in place of \( Z_{XY}^{ck} \)), while compression of \( Z_3 \) uses \( Z_{XY}^{ck} \). In the \( T_{XY}^{ck} \) case, the Laplace denominators must be used during the HOOI.

### F. Approximation schemes

We now use \( E[4] \) to demonstrate how the various approximations discussed above can create a number of reduced scaling implementations of CCSD(T). The most important term to approximate is either \( z_{ij}^{abc} \) or its denominator weighted version, \( t_{ij}^{abc} \). These terms constitute the two halves of the \( E[4] \) diagram in Fig. 3A. In order to guarantee \( O(N^6) \) scaling during the HOOI and in the calculation of the energy for all methods, density fitting is applied to the Hamiltonian term and the rank-reduced compression is applied to \( T_2 \) (Figure 1C). Then, one or both of \( z_{ij}^{abc} \) and \( t_{ij}^{abc} \) are cast into Tucker-3 form (\( \tilde{z}_{ij}^{abc} \) and \( \tilde{t}_{ij}^{abc} \)), creating five distinct reduced scaling methods:

- **\( \tilde{Z}DZ \)** The top \( Z_3 \) is Tucker-3 compressed to form \( \tilde{Z}_3 \), which is then transformed into \( \tilde{T}_3 \) via orthogonal rotation. The bottom \( Z_3 \) remains uncompressed. See Figure 3C.

- **\( \tilde{Z}'DZ \)** As in \( \tilde{Z}DZ \), except that \( T_{XY}^{ck} \) is used during the HOOI instead of \( Z_{XY}^{ck} \).

- **\( \tilde{T}Z \)** The top \( Z_3 D_3 \) is Tucker-3 compressed into \( \tilde{T}_3 \), while the bottom remains as \( Z_3 \). This is the method previously published by Lesiuk.\(^5^6\) See figure 3D.
- $\tilde{Z}T$ The top $Z_3$ is Tucker-3 compressed into $\tilde{Z}_3$ and the bottom $Z_3D_3 = T_3$ is left uncompressed. See Figure 3E.

- $\tilde{Z}D\tilde{Z}$ The top and bottom $Z_3$ are Tucker-3 compressed to form $\tilde{Z}_3$. The top $\tilde{Z}_3$ is then transformed into $\tilde{T}_3$ via orthogonal rotation. See Figure 3F.

It should be noted that there are additional possible schemes utilizing the same approximations, however initial screening clearly singled out these five methods as the most accurate; see Section IV for more discussion on this point.

The scaling of the leading order terms for each of these five reduced scaling methods are displayed in Table I, and the full analysis is provided in the Supplementary Information. Note especially that the scaling of the HOOI term can in general be lowered to $O(N^5)$ if the rank-reduced $T_2$ amplitudes are fully leveraged. In order to test un-approximated $T_2$ as well, we keep a slightly lower-performing factorization. Additionally, the use of randomized projection techniques\(^{69}\) can further accelerate this portion of the calculation. For each method, we separately report the leading order term in a) the HOOI, b) the compression of $Z_3$ (if done), c) the compression of $T_3$ (if done), and d) the evaluation of the approximate (T) energy correction. All five methods achieve $O(N^6)$ scaling, allowing that $N_{\text{H}}$ and $N_{\text{g}}$ are essentially constant. There are three distinct types of methods proposed here: those that employ the orthogonal rotation technique in order to incorporate the orbital energy denominators (although the Laplace denominators are required in the HOOI step of $\tilde{Z}/\tilde{D}Z$), those that directly employ the Laplace denominators in either the compression of $T_3$ or the evaluation of the energy, and, by itself, the $\tilde{Z}D\tilde{Z}$ method which employs a double compression. Note that the similar method $T\tilde{Z}$ is not presented here, for reasons which will become clear in Section IV.

III. COMPUTATIONAL DETAILS

The accuracy of the above approximation schemes was tested on a subset of the W4-17 data set\(^{70}\) detailed in Table II. The species were selected to represent a range of bonding environments, number of heavy atoms, and multi-reference character. The geometries are directly obtained from the W4-17 supporting information. Following the W4 protocol, aug-cc-pVXZ\(^{71,72}\) and aug-cc-pCVXZ\(^{71-73}\) basis sets were used on non-hydrogen atom valance and all-electron CCSD(T) correlation energies, respectively, while cc-pVXZ basis sets were employed on hydrogen atoms. RI
basis sets\textsuperscript{74–76} of the same were used to construct the auxiliary basis for density fitting. Basis-set extrapolations of the CCSD(T) correlation energy were obtained using the two-point extrapolation formula using the basis set cardinality $X$ ($X = 3, 4, 5$ for TZ, QZ, 5Z, etc.),

\begin{equation}
E(X) = E_\infty + A(X + a)^{-\alpha}
\end{equation}

where $a$ and $\alpha$ are taken as 0.5 and 4, respectively.\textsuperscript{41,42,77} There are other extrapolation schemes that may be chosen, but this will not significantly change the relative performance of the methods explored here. Extrapolated values are indicated in the \{X − 1, X\} notation such that \text{aug-cc-pV} [T, Q] Z would indicate a reaction energy or correlation energy obtained from the extrapolation using \text{aug-cc-pVTZ/cc-pVTZ} (non-hydrogen/hydrogen) and \text{aug-cc-pVQZ/cc-pVQZ} (non-hydrogen/hydrogen) basis sets.

Aside from density fitting (controlled by the auxiliary basis set used as noted above) and the Laplace denominators (controlled by the number of quadrature points, which is 12 in all presented calculations), the cost and accuracy of the approximate methods are controlled by two adjustable parameters, $N_{T2}$ and $N_{T3}$, which determine the size of the rank-reduced and Tucker-3 compressions.
of $T_2$ and $T_3$. We report these values as multiples of the number of basis functions $N_{bas}$. The HOOI procedure was terminated when $|N_{T_3} - \sum X^{\sigma^X}| < 10^{-5}$ which usually occurs in six iterations. All calculations were performed in the development version of cFOUR.\textsuperscript{78}

The ANL reaction scheme\textsuperscript{20} was used to benchmark the accuracy of the above approximate methods against the canonical DF-CCSD(T) reaction energies using the same basis sets. This scheme was extended to include boron, silicon, sulphur, and chlorine-containing species. $H_2$, $B_2H_6$, $CH_4$, $H_2O$, $NH_3$, $HF$, $HCl$, $H_2S$, and $SiH_4$ were considered as reference species for the H, B, C, O, N, F, Cl, S, and Si elements, respectively, such that the appropriate reaction products of an arbitrary molecule can be determined as

\begin{equation}
H_d B_b C_c N_d O_e F_f Si_g S_h Cl_i = xH_2 + bB_2H_6 + cCH_4 + dNH_3 + eH_2O + fHF + gSiH_4 + hH_2S + iHCl,
\end{equation}

where the number of $H_2$ molecules in the products, $x$, needed to balance the equation can be determined algebraically. While this reaction scheme is not designed to universally take advantage of error cancellation (for example, the carbon dioxide reaction scheme would be $4H_2 + CO_2 \rightarrow 2H_2O + CH_4$, which preserves no bonding character between the reactants and products), all the reference species involved are closed shell, well characterized, and free from significant multireference effects.

IV. RESULTS AND DISCUSSION

A. Separability of the approximations

Each of the above approximate schemes is constructed from individual compressions or approximations of the two-electron ERIs (only density fitting is used here), the $T_2$ amplitudes (rank-reduced compression with a user-defined $N_{T_2}$), and Tucker-3 compression of the $T_3$ amplitude or $Z_3$ intermediates (with a user-defined $N_{T_3}$ and convergence threshold of the HOOI procedure). As might be expected, density fitting of the ERIs introduces exceptionally small errors (less than 0.5% of the (T) correction) for the reaction energies studied here and is applied to all ERI terms in the approximate (T) schemes and in the CCSD calculation which determines $T_2$. More information on the effect of density fitting in the present calculations can be found in the Supporting Information.
FIG. 4: $T_2$ convergence and $T_3$ convergence behavior for different methods and four representative reactions. The absolute reaction energy error with the aug-cc-pV(T,Q)Z extrapolation is shown on the y-axis. The x-axis indicates the number of $T_2$ projectors as a multiple $n$ ($N_{T_2} = nN_{bas}$). Separate colored lines correspond to increasing (red to blue) numbers of $T_3$ projectors ($N_{T_3} = mN_{bas}$).
<table>
<thead>
<tr>
<th># of Heavy Atoms</th>
<th>Single Reference</th>
<th>Minor MR</th>
<th>Moderate MR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(%TAE[(T)] &lt; 4)</td>
<td>(4 &lt; %TAE[(T)] &lt; 10)</td>
<td>(%TAE[(T)] &gt; 10)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₂H₆</td>
<td>0.74</td>
<td>N₂</td>
<td>4.16</td>
</tr>
<tr>
<td>SiH₃F</td>
<td>0.81</td>
<td>O₂</td>
<td>7.67</td>
</tr>
<tr>
<td>BF₂H</td>
<td>1.49</td>
<td>HCNO</td>
<td>4.77</td>
</tr>
<tr>
<td>CH₃CHO</td>
<td>1.72</td>
<td>HN₃</td>
<td>5.64</td>
</tr>
<tr>
<td>C₃H₄</td>
<td>1.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₂H₃F</td>
<td>1.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF₃</td>
<td>1.77</td>
<td>SO₃</td>
<td>5.87</td>
</tr>
<tr>
<td>SiF₄</td>
<td>1.75</td>
<td>HClO₄</td>
<td>7.92</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₆H₆</td>
<td>1.96</td>
<td>N₂O₄</td>
<td>9.11</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II: Molecule set selected from W4-17, along with %TAE[(T)] values from Ref. 70 which approximately indicate multi-reference (MR) behavior.

The HOOI procedure, the only iterative portion of these approximation schemes, converges very rapidly, requiring only 6 cycles to meet a very tight tolerance of $10^{-5}$. Further, it appears that the predicted reaction energies depend only minimally upon the number of HOOI iterations; even a loose convergence threshold (or potentially, a non-iterative determination of $U$) appears to be sufficient to achieve convergence in the reaction energies (see Fig. S1). Thus, consideration of the DF and HOOI treatments can be separated from the $T₂$ rank-reduction and $Z₃/T₃$ compression. These findings are consistent with those observed in previous literature.⁵⁶

At first glance, it is not obvious that this separability should extend to the compression of the $T₂$ and $Z₃/T₃$ space, especially considering that the compressed $T₂$ amplitudes are used in the construction of the compressed three-electron quantities. Fig. 4 displays the dependence of a representative subset of reaction energies on the number of $T₂$ projectors, with separate curves accounting for increasing numbers of $T₃$ projectors. Increasing the number of $T₂$ projectors past roughly $2.2N_{bas}$ leads to no significant improvement in the predicted reaction energies for all numbers of $T₃$ projectors. Additionally, the curves for separate numbers of $T₃$ projectors run essentially parallel across the $N_{T₂}$ space, indicating that these two approximations are decidedly separable.
FIG. 5: Error statistics of various decomposition schemes for the benchmark reaction set. The thick lines indicate the mean absolute or absolute percentage errors across the set, while the top of the shaded region indicates the maximum error. Plots are labeled by the extrapolation scheme used, and “aug-cc-pV\{Q,5\}Z Corrected” indicates a frozen core aug-cc-pV\{Q,5\}Z valence-only energy plus an aug-cc-pCV\{T,Q\}Z core-valence correction.

Thus, we fix the number of $T_2$ projectors to a “safe” value of $3N_{\text{bas}}$ for all reactions, effectively making the number of $T_3$ projectors the sole variable that controls both cost and accuracy of the five approximate methods studied. Note that $N_{T_2}$ has minimal impact on the cost of the overall energy evaluation.

B. Comparison of approximation schemes in frozen core and core-valence corrected calculations

To recapitulate, the purpose of this work is to assess the viability of several approximate (T) methods in the context of sub-chemical (1 kJ mol$^{-1}$) accuracy reaction energies. Given that the perturbative triples are but one component in most model chemistries, we have arbitrarily selected 0.1 kJ mol$^{-1}$ as the target mean absolute error (MAE) and 1% as the target mean absolute percent error (MAPE) for these approximate schemes applied to the test-suite of reaction energies. The number of $Z_3/T_3$ projectors required to hit these goals then determines the relative cost of these methods, though these targets may be changed to suit the needs of the individual model chemistry.
FIG. 6: “Violin plots” of the mean absolute percent errors (MAPE) of reaction energies with the aug-cc-pV(T,Q)Z basis set extrapolation. The break in the y-axis is used to present the extremal data in detail.
that these methods would be employed.

Figure 5 displays the mean absolute error (MAE) and mean absolute percent error (MAPE) for the five approximate schemes discussed above for valence aug-cc-pV{T,Q}Z, valence aug-cc-pV{Q,5}Z, and core-valence corrected aug-cc-pV{Q,5}Z calculations. In this figure, the solid line indicates the mean value for the respective error statistic, while the upper edge of the shaded region represents the maximum error across the test set. All methods are capable of achieving a sub-0.1 kJ mol$^{-1}$ and/or sub-1% average accuracy given a large enough set of $Z_3/T_3$ projectors, though with varying error profiles. The $\tilde{Z}DZ$ scheme (purple line) requires the largest number of projectors to reach target MAE accuracy, which is perhaps unsurprising as this scheme employs Tucker-3 decompositions of both the top and bottom half of the $E^{[4]}$ diagrams. $\tilde{T}Z$ and $\tilde{Z}'\tilde{D}Z$ (green and red lines, respectively) performed quite similarly, not just for the statistical errors, but also in the individual reaction energies, though $\tilde{Z}'\tilde{D}Z$ (red line) is slightly more accurate. These two methods both create a Tucker-3 representation of $T_3$, but in $\tilde{Z}'\tilde{D}Z$ the effect of the denominators is included in the HOOI iterations. Between those potential approximation schemes reported and not reported here, the main differentiating factor is the handling of the denominators in the determination of $U$ via HOOI. For example, $\tilde{T}Z$ includes the denominators in the HOOI since it is the $T_3$ amplitudes which are being directly approximated. The similar method $\tilde{T}'Z$ which differs only in that it does not include the denominators during the HOOI is much less accurate. Similar comparisons can be made for almost every method tested. This indicates a strong dependence of the projectors on the structure of $Z_3$ vs. $T_3$. In this light, the $\tilde{Z}'\tilde{D}Z$ method (and to some extent, $\tilde{Z}\tilde{D}\tilde{Z}$) is then an outlier because it only slightly underperforms the “correct” version $\tilde{Z}'\tilde{D}Z$ (note that $\tilde{Z}\tilde{D} \sim \tilde{T}$ indicating that including denominators in the HOOI should perform better). There is also an inconsistency for $\tilde{Z}'\tilde{D}Z$ compared to $\tilde{T}Z$—the former method is identical to the latter if the initial compression of $Z_3$ is removed (orthogonal rotation is an exact representation of the denominators in the compressed space so these would revert to the similarly “exact” Laplace denominators). Instead, $\tilde{Z}\tilde{D}\tilde{Z}$ often outperforms $\tilde{T}Z$, particularly when the core-valence correlation is included. This may indicate a partial cancellation of errors between the compression of $Z_3$ and later $T_3$. In the future, we will explore the use of multiple sets of projectors to further improve accuracy. $\tilde{Z}T$ seems to perform best or second-best for the MAE and MAPE measures (and for the maximum absolute/percentage errors), while the relative performance of $\tilde{Z}DZ$ changes depending on the basis sets involved. This in turn may indicate a slightly better compressibility of $Z_3$ compared to $T_3$.

Figure 6 displays violin plots of the MAPE for each of these methods. Consistent with the
\begin{align*}
\text{SiH}_2\text{F} + \text{H}_2 &\rightarrow \text{SiH}_3 + \text{HF} (0.81) \\
\text{N}_2 + 3\text{H}_2 &\rightarrow 2\text{NH}_3 (4.16) \\
2\text{O}_2 + \text{H}_2 &\rightarrow 2\text{H}_2\text{O} (7.67) \\
\text{B}_2 + 3\text{H}_2 &\rightarrow \text{B}_3\text{H}_6 (14.66) \\
\text{C}_2 + 2\text{H}_2 &\rightarrow \text{C}_2\text{H}_4 (13.26) \\
\text{C}_3\text{H}_5\text{F} + 2\text{H}_2 &\rightarrow 2\text{CH}_4 + \text{HF} (1.25) \\
2\text{BH}_2 + 3\text{H}_2 &\rightarrow \text{B}_3\text{H}_6 + 2\text{HF} (1.49) \\
\text{CH}_2\text{CHO} + \text{H}_2 &\rightarrow \text{CH}_4 + \text{H}_2\text{O} (1.72) \\
\text{C}_6\text{H}_5 + 4\text{H}_2 &\rightarrow 3\text{CH}_4 (1.75) \\
\text{H}_2\text{NO} + 4\text{H}_2 &\rightarrow \text{CH}_4 + \text{NH}_3 + \text{H}_2\text{O} (4.77) \\
\text{HN}_3 + 4\text{H}_2 &\rightarrow 3\text{NH}_3 (5.64) \\
\text{F}_2\text{O} + 2\text{H}_2 &\rightarrow \text{H}_2\text{O} + 2\text{HF} (14.6) \\
\text{O}_3 + 3\text{H}_2 &\rightarrow 3\text{H}_2\text{O} (17.39) \\
2\text{BF}_3 + 6\text{H}_2 &\rightarrow \text{B}_2\text{H}_6 + 6\text{HF} (1.77) \\
\text{SO}_3 + 4\text{H}_2 &\rightarrow \text{H}_2\text{S} + 3\text{H}_2\text{O} (5.87) \\
\text{FOOF} + 3\text{H}_2 &\rightarrow 2\text{H}_2\text{O} + 2\text{HF} (16.92) \\
\text{SiF}_4 + 4\text{H}_2 &\rightarrow \text{SiH}_4 + 4\text{HF} (1.75) \\
\text{HClO}_4 + 4\text{H}_2 &\rightarrow \text{HCl} + 4\text{H}_2\text{O} (7.92) \\
\text{C}_2\text{H}_2 + 9\text{H}_2 &\rightarrow 6\text{CH}_4 (1.96) \\
\text{N}_2\text{O}_3 + 7\text{H}_2 &\rightarrow 2\text{NH}_3 + 4\text{H}_2\text{O} (9.11) \\
\text{CIF}_3 + 3\text{H}_2 &\rightarrow 5\text{HF} + \text{HCl} (14.78) \\
\end{align*}
FIG. 7: Absolute relative errors (in percent) of each reaction species at the aug-cc-pV{T,Q}Z frozen-core level. In each figure, the right axis gives the absolute error in kJ mol$^{-1}$.

average and max error plots, $\tilde{T}Z$, $\tilde{Z}'\tilde{D}Z$, and $\tilde{Z}T$ are most resistant to outliers. $\tilde{Z}T$ in particular obtains a tight distribution of errors compared to the other methods. $\tilde{Z}DZ$, the double-Tucker-3 method, performs worst, and $\tilde{Z}DZ$ performs well with a exception of a few outliers that skew the average statistics. These outliers correspond to the reactions: $2\text{BF}_3 + 6\text{H}_2 \rightarrow \text{B}_2\text{H}_6 + 6\text{HF}$, $\text{SiF}_4 + 4\text{H}_2 \rightarrow \text{SiH}_44\text{HF}$, and $2\text{BHF}_2 + 3\text{H}_2 \rightarrow \text{B}_2\text{H}_6 + 2\text{HF}$. Interestingly, these reactions all display extremely small (T) contributions to reaction energy. Thus, even large percentage errors still meet the desired 0.1 kJ mol$^{-1}$ threshold. However, some applications may required small relative errors even in the case of small total reaction energies. Thus, the family of methods proposed can allow practitioners to adopt a method most suited to their specific application.

Most important in the potential application of these methods in model chemistries is the number of $Z_3/T_3$ projectors required to achieve some target accuracy, and hence the overall computation cost required. For example, a relatively expensive approximation scheme such as $\tilde{T}Z$ (green) may out-perform a less expensive scheme such as $\tilde{Z}T$ (orange) if the former requires a sufficiently smaller number of projectors to obtain equivalent confidence intervals. To explore this, we examine the number of projectors required to obtain an MAE of 0.1 kJ mol$^{-1}$ and estimate the relative expense of the five methods using the equations contained within the Supplementary Information. From this perspective, the two most promising methods are $\tilde{Z}T$ (orange) and $\tilde{Z}DZ$ (blue). $\tilde{Z}DZ$ (purple) falls behind as it requires a large number of $T_3$ projectors to obtain target accuracy (and scales cubically with $N_{T_3}$ while the other methods scale quadratically), while $\tilde{T}Z$ (green) and $\tilde{Z}'\tilde{D}Z$ (red), although relatively stable, do not out-compete $\tilde{Z}T$’s (orange) ratio of cost to accuracy. Note that, in our present factorization of the working equations and implementation, the cost of the HOOI is a major factor in the overall wall time. Improvements to the HOOI (smaller number of iterations, random projection, etc.) could change the relative ranking of the methods. While a full analysis of performance data and empirical scaling is beyond the scope of this work, we do note already significant speedups compared to DF-CCSD(T) even for the present test set of small molecules.

It should be noted that while the accumulated statistics displayed in Fig. 5 are used in the above discussion (and demonstrate a roughly linear profile in the log space of the respective errors), the errors in each individual reaction (see Fig. IV B) are frequently non-monotonic, and the conclu-
FIG. 8: Mean absolute percent errors (MAPE) of extrapolated reaction energies across the entire test set.

Sions above should therefore be taken as general commentary on possible applicability rather than guidance on constructing a model chemistry using these techniques. We are currently working on further characterizing the errors in these methods and how to best predict the number of $T_3$ projectors required to obtain a target accuracy, but these are the subject of future manuscripts.

C. Comparison of approximation schemes in all-electron calculations

All high-accuracy model chemistries employ some variety of core-valance correlation corrections, typically calculated with aug-cc-pCVXZ variants of smaller canonical size than the aug-cc-pVXZ basis sets used to capture valance correlation. The exception to this rule is HEAT-like model chemistries, $^{13,16,17,19}$ which avoid potential additivity problems with this differing basis set treatment by correlating all electrons through CCSD(T). Fig. 8 displays the MAE statistics for the five above approximate schemes for aug-cc-pV{T,Q}Z frozen-core, aug-cc-pCV{T,Q}Z frozen-core, and aug-cc-pCV{T,Q}Z all-electron calculations. As expected, all the methods require a
smallest ratio of $Z_3/T_3$ projectors to total basis set size when increasing the size of the basis sets used in valence calculations from aug-cc-pV[T,Q]Z to aug-cc-pCV{Q,5}Z. It has been previously shown that there are only a linear number of projectors that are important in describing molecular systems,\textsuperscript{51} and thus these methods become comparatively less expensive with larger basis sets where they can project onto the parts of correlation that “matter” from a larger space of basis functions. This is very desirable property from the perspective of model chemistry development: the larger the basis set the better, relatively, these rank-reduced techniques perform.

However, while all methods require comparatively more projectors when treating all-electron correlation energies (which causes the upward shift in the aug-cc-pV{Q,5}Z core-valence corrected plots in Figure 5), not all of the five approximate methods discussed here respond equivalently to the inclusion of core-electrons in the correlated wavefunction. In particular the $\tilde{Z}T$ method, which generally performs best in valence-only aug-cc-pV{T,Q}Z, performs only a little better than its competitors for all-electron aug-cc-pCV{T,Q}Z. The underlying cause for this change is unclear, and warrants further investigation in future work. However, there seems to be no reason to believe that these techniques cannot be used to address core-valance correlation in model chemistries.

D. Discussion

There are some interesting conclusions one can draw by making comparisons of the relative performance of these five approximate (T) methods, that may provide insight for future schemes that use these techniques.

\(\tilde{Z}D\tilde{Z}\) is unequivocally the least accurate method studied here, likely due to the use of two Tucker-3 compressions. As shown above, the number of $T_3$ projectors is the primary determinant of the accuracy of these methods, and $\tilde{Z}D\tilde{Z}$ trades off decreased scaling with the number of density-fitting auxiliary basis functions for an increased scaling with $N_{T_3}$. Given that it requires more $T_3$ projectors than DF auxiliary functions to obtain the accuracy goals desired here, this trade off simply doesn’t net any gain.

As stated above, $\tilde{Z}'D\tilde{Z}$ and $\tilde{T}Z$ differ only in that the former performs an orthogonal rotation of $\tilde{T}_3$ to compress the denominator terms while the latter uses a Laplace transform and quadrature. As the orthogonal rotation is exact within the compressed subspace (and the additional compression of $Z_3$ does not seem to lead to additional error) and the Laplace transform is a very precise
approximation, there are only slight measurable differences between these methods even on a reaction-by-reaction basis, and the only discrepancy is the cost of performing the orthogonal rotation vs the cost of utilizing the Laplace transform. Both methods perform a Tucker-3 compression of the amplitudes rather than the residuals, and, in this case, both are outperformed by the $\tilde{Z}T$ method on a cost basis due to an increased cost of the HOOI step.

Likewise, $\tilde{Z}\tilde{D}Z$ and $\tilde{T}$ both perform a Tucker-3 compression of the $Z_3$ intermediates and differ only in how they compress the denominator, as with $Z'\tilde{D}Z$ and $\tilde{T}Z$. Unlike the above case, however, this results in significant differences between the two methods, especially when considered on a reaction-by-reaction basis. Generally it seems that $\tilde{T}$ is a bit more stable: it performs on par or better than $Z'\tilde{D}Z$ and $\tilde{T}Z$ with a significantly reduced cost in the HOOI iterations, and only performs significantly worse than these in the $2BF_3 + 6H_2 \rightarrow B_2H_6 + 6HF$ reaction, for which the reaction energy is very small. $\tilde{Z}\tilde{D}Z$, on the other hand, is somewhat more erratic—an undesirable trait in an approximation to be used in high-accuracy thermochemistry.

The final comparison we make is between $Z'\tilde{D}Z$ and $\tilde{Z}\tilde{D}Z$. The only difference between these methods is whether or not the Laplace denominators are included in the HOOI, which determines the projectors $U$. Both then undergo orthogonal rotation to create the $\tilde{T}_3$ used in the final energy calculations. However, while $Z'\tilde{D}Z$ performs extremely similarly to its Laplace transform counterpart ($\tilde{T}Z$), $\tilde{Z}\tilde{D}Z$ features qualitatively different behavior to $\tilde{T}Z$, especially on a species-by-species comparison. In the case of $Z'\tilde{D}Z$, we noted that the additional compression of $Z_3$ does not seem to introduce additional error, and may even exhibit partial error cancellation. Even if the same were true in the case of $\tilde{Z}\tilde{D}Z$, the “incorrect” compression subspace in which the orthogonal rotation is performed is the most likely cause of this discrepancy, and overwhelms any other considerations of the improved (due to the more appropriate choice of $U$) initial compression of $Z_3$.

V. CONCLUSIONS

In this work, we have examined five different approximation schemes for the perturbative triples correction in CCSD(T). Four of these five are entirely novel. For a subset of species taken from the W4-17 dataset, we have constructed a series of reaction energies using an extended version of the ANL-$n$ reaction scheme. It was demonstrated that of the approximations studied here, including density-fitting of ERIs, rank-reduced compression of the $T_2$ amplitudes, Tucker-3 compression of the $Z_3/T_3$ intermediates in the (T) correction, and either Laplace transform or orthogonal rotation
of the orbital eigenvalue denominators, only the $Z_3/T_3$ Tucker-3 decomposition was a limiting factor in the accuracy and cost of the resulting approximation schemes, and the number of projectors $U$ can be used to control the accuracy of the calculation. These five schemes, $\tilde{Z}T$, $\tilde{T}Z$, $\tilde{Z}\tilde{D}Z$, $\tilde{Z}'\tilde{D}Z$, and $\tilde{Z}\tilde{D}Z$, are all capable of achieving sub-0.1 kJ mol$^{-1}$ mean absolute errors with modest numbers of projectors when compared against canonical, density-fitted DF-CCSD(T) calculations. In particular $\tilde{Z}T$, novel to this work, demonstrates a favorable combination of accuracy and computational efficiency. Future work is needed to determine exactly how these components should be included in high-accuracy model chemistries, with particular emphasis on selecting the number of $Z_3/T_3$ projectors on a reaction-by-reaction basis, but the work here illustrates that these techniques have a place in some of our most accurate models of gas-phase thermochemistry and may yield order-of-magnitude improvements in computational cost of CCSD(T) with minimal and well-controlled error.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation under grant CHE-2143725. JHT acknowledges support from the SMU Moody School of Graduate and Advanced Studies. All calculations were performed on the ManeFrame III computing system at Southern Methodist University.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available within the article and its supplementary material. Available supplemental information files:

- df_error.xlsx: absolute and relative errors of DF-CCSD(T) vs. CCSD(T) (aug-cc-pVTZ and aug-cc-pVQZ, frozen core).

- T2_convergence_error.xlsx and T2_convergence_raw_data.xlsx: rank-reduced $T_2$ amplitude convergence data.

- H00I_convergence_error.xlsx and H00I_convergence_raw_data.xlsx: H00I convergence data.
• T3_PVNZ_FC_raw_data.xlsx, T3_PCVNZ_FC_raw_data.xlsx, and T3_PCVNZ_AE_raw_data.xlsx: raw and extrapolated data for all reported results.

• supplemental.pdf: factorized formulae for the implemented approximate CCSD(T) methods and additional supporting figures.

REFERENCES


21 B. Ruscic, “Uncertainty quantification in thermochemistry, benchmarking electronic structure


33 A. Karton, P. R. Schreiner, and J. M. L. Martin, “Heats of formation of platonic hydrocarbon


45 A. P. Rendell and T. J. Lee, “Coupled-cluster theory employing approximate integrals: An approach to avoid the input/output and storage bottlenecks,” The Journal of Chemical Physics 101,
400–408 (1994).


A. Karton, N. Sylvestsky, and J. M. L. Martin, “W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods,” Journal of


Prospects for rank-reduced CCSD(T) in the context of high-accuracy thermochemistry: Supplemental Information

Tingting Zhao, James H. Thorpe, and Devin A. Matthews*

Department of Chemistry, Southern Methodist University, Dallas, TX

(*Electronic mail: dmatthews@smu.edu)
I. HOOI CONVERGENCE

FIG. S1. Absolute errors of AUG-PV[T,Q]Z extrapolated (T) correction energies with respect to the HOOI convergence tolerance. \( N_{T_3} = N_{T_3} = 3N_{bas} \).

The HOOI procedure was found to converge very rapidly, and the approximate CCSD(T) energies are highly insensitive to the chosen HOOI convergence criterion (Fig. S1). In some calculations, difficulties were encountered in achieving the requested convergence, typically due to switching of singular vectors between the selected and excluded subsets. In future work, it may be sufficient to simply use a non-iterative estimate of \( U \).

II. COST COMPAIRATION

III. FACTORIZATION OF THE WORKING EQUATIONS

Factorization of the equations for the HOOI procedure, compression of \( Z_3/T_3 \), and the approximate energy for each approximation scheme were obtained by an in-house contraction search
<table>
<thead>
<tr>
<th>Methods</th>
<th>ŽDŽ</th>
<th>ŽT</th>
<th>Ž</th>
<th>ŽDŽ</th>
<th>ŽDŽ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiH₃F + H₂ → SiH₄ + HF</td>
<td>2.0</td>
<td>2.0</td>
<td>1.5</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>C₂H₅F + 2H₂ → 2CH₄ + HF</td>
<td>2.0</td>
<td>3.0</td>
<td>3.5</td>
<td>3.0</td>
<td>4.5</td>
</tr>
<tr>
<td>2BH₃ + 3H₂ → B₂H₆ + 2HF</td>
<td>4.5</td>
<td>3.0</td>
<td>1.5</td>
<td>1.5</td>
<td>5.5</td>
</tr>
<tr>
<td>CH₃CHO + H₂ → CH₄ + H₂O</td>
<td>2.0</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>4.5</td>
</tr>
<tr>
<td>C₃H₄ + 4H₂ → 3CH₄</td>
<td>2.0</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>2BF₃ + 6H₂ → B₂H₆ + 6HF</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>2.5</td>
<td>&gt;6</td>
</tr>
<tr>
<td>SiF₄ + 4H₂ → SiH₄ + 4HF</td>
<td>2.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.0</td>
<td>&gt;6</td>
</tr>
<tr>
<td>C₆H₆ + 9H₂ → 6CH₄</td>
<td>5.5</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
</tr>
<tr>
<td>Mean(Single Reference)</td>
<td>2.75</td>
<td>2.79</td>
<td>2.64</td>
<td>2.5</td>
<td>3.9</td>
</tr>
<tr>
<td>N₂ + 3H₂ → 2NH₃</td>
<td>1.5</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>2O₂ + H₂ → 2H₂O</td>
<td>2.0</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>HCNO + 4H₂ → CH₄ + NH₃ + H₂O</td>
<td>1.0</td>
<td>2.5</td>
<td>1.5</td>
<td>1.5</td>
<td>5.5</td>
</tr>
<tr>
<td>HNNN + 4H₂ → 3NH₃</td>
<td>3.5</td>
<td>3.5</td>
<td>4.0</td>
<td>3.5</td>
<td>5.5</td>
</tr>
<tr>
<td>SO₃ + 4H₂ → H₂S + 3H₂O</td>
<td>6.0</td>
<td>3.5</td>
<td>5.0</td>
<td>4.5</td>
<td>&gt;6</td>
</tr>
<tr>
<td>HClO₄ + 4H₂ → HCl + 4H₂O</td>
<td>&gt;6</td>
<td>4.5</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
</tr>
<tr>
<td>N₂O₄ + 7H₂ → 2NH₃ + 4H₂O</td>
<td>&gt;6</td>
<td>6.0</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
</tr>
<tr>
<td>SF₆ + 4H₂ → 6HF + H₂S</td>
<td>&gt;6</td>
<td>5.0</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
</tr>
<tr>
<td>Mean(Minor MR)</td>
<td>2.8</td>
<td>3.8</td>
<td>3.2</td>
<td>3.0</td>
<td>4.5</td>
</tr>
<tr>
<td>B₂ + 3H₂ → B₂H₆</td>
<td>1.5</td>
<td>1.5</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>C₂ + 2H₂ → C₂H₄</td>
<td>2.5</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>F₂ + 2H₂ → H₂O + 2HF</td>
<td>6.0</td>
<td>4.5</td>
<td>6.0</td>
<td>6.0</td>
<td>&gt;6</td>
</tr>
<tr>
<td>O₃ + 3H₂ → 3H₂O</td>
<td>4.5</td>
<td>4.0</td>
<td>5.0</td>
<td>5.0</td>
<td>&gt;6</td>
</tr>
<tr>
<td>FOOF + 3H₂ → 2H₂O + 2HF</td>
<td>5.5</td>
<td>4.5</td>
<td>5.5</td>
<td>5.5</td>
<td>&gt;6</td>
</tr>
<tr>
<td>ClF₅ + 3H₂ → 5HF + HCl</td>
<td>&gt;6</td>
<td>5.5</td>
<td>&gt;6</td>
<td>&gt;6</td>
<td>&gt;6</td>
</tr>
<tr>
<td>Mean(Moderate MR)</td>
<td>4.0</td>
<td>3.6</td>
<td>4.1</td>
<td>4.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>

TABLE I. N₇ in n*N₂max needed to achieve 0.1 kJ/mol for each reaction species and method.

3
147
code, with some hand modification as noted. The Einstein summation convention is assumed throughout, and parentheses are used to denote the order of operations.

A. Explicit formulae and factorization for $Z_{ck}^{XY}$ and $T_{ck}^{XY}$

The factorized equations for the $Z_{ck}^{XY}$ intermediates are given below, along with the associated number of multiplications for each step. Note that a lower-scaling factorization is possible when using rank-reduced amplitudes; the factorization here was chosen so that un-approximated amplitudes $t_{ij}^{ab}$ can also be used easily.

$$Z_{ck}^{XY} = \frac{1}{6} P(XY) \{(C_{ai}^{X}(E1_{ai}^{Y} + E2_{ai}^{Y}))B_{ck}^{J} + C_{ab}^{X}D_{ca}^{Y} - C_{ci}^{X}D_{ik}^{Y} + i_{ik}^{bc}(U_{ai}^{Y}D_{ab}^{1Z} - U_{bj}^{Y}D_{ij}^{2Z})\}$$

$$C_{ai}^{X} = V_{ai}^{W}(T_{VW}^{VW}(V_{b}^{V}U_{bi}^{X}))$$

$$D_{ab}^{1X} = B_{ab}^{J}(B_{ck}^{J}U_{ck}^{X})$$

$$D_{ij}^{2X} = B_{ij}^{J}(B_{ck}^{J}U_{ck}^{X})$$

$$E_{ai}^{1X} = B_{ab}^{J}U_{bi}^{X}$$

$$E_{ai}^{2X} = -B_{ij}^{J}U_{aj}^{X}$$

$$i_{ij}^{ab} = V_{ai}^{V}T_{VW}^{VW}V_{b}^{W}$$

where the permutation operator $P(XY) = (XY) + (YX)$. The construction of $T_{ck}^{XY}$ uses an identical factorization apart from the insertion of the Laplace factors $\tau_{a}^{g}$ and $\tau_{i}^{g}$ and of course the summation of all terms over $g$. Laplace factors must be inserted for every molecular orbital index of $U$, and also for the external indices $ck$ in (1).

B. Explicit formulae and factorization for $Z^{XYZ}$

The explicit formulae and factorization for the “fourth-order” E4T portion of $Z^{XYZ}$ are shown as follows, including the number of multiplications required for each step,

$$(Z^{XYZ})^{[4]} = P(XYZ) \{(C_{ai}^{X}(E1_{ai}^{Y} + E2_{ai}^{Y}))F_{ai}^{JX} \}$$

$$F_{ai}^{JX} = B_{ai}^{J}U_{ai}^{X}$$

4

148
where the permutation operator \( P(XYZ) = (XY) + (XZ) + (YX) + (YZ) + (ZX) + (ZY) \).

To construct \( T^{XYZ} \) via Laplace denominators, the Laplace factors \( \tau^g_x \) and \( \tau^g_y \) must be again included for each molecular orbital index of \( U \), followed by a summation over \( g \). The “fifth-order” term E5ST is always included through the “top” Z3 or T3 fragment in our calculations, and is thus always included in the Tucker-3 compression. We do not include this term in the HOOI procedure, and limited testing shows a negligible effect on the computed projectors. The explicit formula and factorization, including number of multiplications, for this term is,

\[
(Z^{X_{1}Y_{1}Z_{1}})^5 = \frac{1}{2} P(XYZ) \{ G^X F^{JY} F^{JZ} \} \quad (N_{DF} + N_{T3})N_{T3}^2
\]

\[
G^X = U^X_{ai} U^A_{ai} \quad N_{T3} \quad 1
\]

C. Explicit formulae and factorization for approximate (T) energy: \( \tilde{T} \tilde{D} \tilde{Z} \)

The explicit formulae and factorization for \( E_{(T)}^{\tilde{T} \tilde{D} \tilde{Z}} \) are given below. E5ST is included in the top \( \tilde{T}_3 \tilde{D}_3 = \tilde{T}_3 \) fragment in this approximation scheme.

\[
E_{(T)}^{\tilde{T} \tilde{D} \tilde{Z}} = \frac{1}{3} \left\{ 4 T^{X_{1}Y_{1}Z_{1}} U^X_{ai} U^Y_{bj} U^{Z}_{ck} U^A_{ai} U^B_{bj} U^C_{ck} Z^{ABC} 
\right.
\]

\[
- 6 T^{X_{1}Y_{1}Z_{1}} U^X_{ai} U^Y_{bj} U^{Z}_{ck} U^A_{ak} U^B_{bj} U^C_{ci} Z^{ABC} 
\]

\[
+ 2 T^{X_{1}Y_{1}Z_{1}} U^X_{ai} U^Y_{bj} U^{Z}_{ck} U^A_{ak} U^B_{bi} U^C_{ci} Z^{ABC} \right\}
\]

\[
= \frac{1}{3} \left\{ 4 T^{X_{1}Y_{1}Z_{1}} Z^{X_{1}Y_{1}Z_{1}} 
\right.
\]

\[
- J_{ci}^{Y_{1}} (6 J_{oi}^{A_{1}B_{1}} \delta_{i} - 2 J_{ci}^{Y_{1}}) \right\} \quad N_{T3} \quad 1
\]

\[
H_{ij}^{X_A} = U^X_{ai} U^Y_{aj} \quad N_{T3} \quad \delta_{X_A} \quad (14)
\]

\[
I_{ck}^{X_{1}Y_{1}} = U^Z_{ck} T^{X_{1}Y_{1}Z_{1}} \quad N_{T3} \quad 1
\]

\[
J_{ci}^{A_{1}B_{1}} = U^C_{ci} Z^{A_{1}B_{1}} \quad N_{T3} \quad 1
\]

\[
J_{ci}^{A_{1}B_{1}} = H_{ik}^{X_{1}A_{1}} I_{ck}^{X_{1}Y_{1}} \quad N_{T3} \quad 1
\]

\[
J_{ci}^{A_{1}B_{1}} = H_{ji}^{Y_{1}B_{1}} I_{cj}^{A_{1}B_{1}} \quad N_{T3} \quad 1
\]

where \( U^X_{ai} U^Y_{ai} = \delta_{X_A} \).

D. Explicit formulae and factorization for approximate (T) energy: other schemes

The explicit formulae and factorization for the remaining approximation schemes: \( \tilde{T}Z \), \( \tilde{Z} \tilde{D} \tilde{Z} \), and \( \tilde{Z}' \tilde{D} \tilde{Z} \) share a common structure, and the equations for \( \tilde{T}Z \) are given below as an exemplar.
For $\tilde{Z}T$, there is in addition a sum over the Laplace quadrature points $g$, and Laplace factors $\tau^g_a$ and $\tau^g_i$ must be inserted for each MO $a b c i j k$ (which can be easily accomplished by pre-applying the factors to the $U$ projectors). Note that here also, the equations are factored in such a way as to allow exact $T_2$ to be easily used if desired, although for the (T) energy little to no cost reduction is possible by assuming rank-reduced $T_2$ amplitudes.

$$E^{\tilde{Z}}_{(T)} = \{ 4 T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J}$$

$$- 2 T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J}$$

$$- 2 T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J}$$

$$+ T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J}$$

$$- 2 T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J}$$

$$+ T^{XYZ} U_{ai}^{X} U_{bj}^{Y} U_{ck}^{Z} (\tilde{r}^{ad} B_{bd}^{J} - \tilde{r}^{ab} B_{bi}^{J}) B_{ck}^{J} \}$$

$$= \{ 2 (\tilde{C}_{X}^{d l} (E^{1 Y}_{d l} + E^{2 J Y}_{d l}) T^{XYZ}) F^{J Z}$$

$$- (2 \tilde{r}^{ad} - \tilde{r}^{ab} (E^{1 Y}_{d l} + E^{2 J Y}_{d l}) (E^{3 J Z}_{d l} I^{Y Z}_{a i})$$

$$- \tilde{C}_{X}^{d l} K_{X}^{d l} \}$$

$$\tilde{C}_{a i}^{X} = (2 \tilde{r}^{ab} - \tilde{r}^{ab} U_{b j}^{X})$$

$$E^{3 J Z}_{d l} = B_{c l}^{J} U_{c k}^{Z}$$

$$K_{d l}^{X} = (D^{1 Y}_{d c} U_{c l}^{Z} - D^{2 Y}_{l k} U_{d k}^{Z}) T^{XYZ}$$

$$N_{DF N_{T_{3}}}^{2} v \gamma + N_{DF N_{T_{3}}}^{3} + N_{DF N_{T_{3}}}^{3}$$

$$N_{DF N_{T_{3}}}^{2} v \gamma + N_{DF N_{T_{3}}}^{3} v^{2} \gamma^{2} + v^{2} o^{2}$$

$$N_{T_{3}} v \gamma$$

$$N_{T_{3}} v^{2} \gamma^{2}$$

$$N_{DF N_{T_{3}}}^{2} v \gamma + N_{DF N_{T_{3}}}^{3} v^{2} \gamma^{2} + v^{2} o^{2}$$

$$N_{T_{3}} v \gamma$$

$$N_{T_{3}} v^{2} \gamma^{2}$$

$$N_{DF N_{T_{3}}}^{2} v \gamma + N_{DF N_{T_{3}}}^{3} v^{2} \gamma^{2} + v^{2} o^{2}$$
APPENDIX C

Draft on the first order analytic gradient of EOM-CCSD*

Tingting Zhao, Devin A. Matthews. In preparation.
Analytic gradients for equation-of-motion coupled cluster with single, double, and perturbative triple excitations

Tingting Zhao and Devin. A. Matthews

Department of Chemistry, Southern Methodist University, Dallas, TX

E-mail: dmatthews@smu.edu

Abstract

Understanding the process of molecular photoexcitation is crucial in various fields, including drug development, materials science, photovoltaics, and more. The electronic vertical excitation energy is a critical property, for example in determining the singlet-triplet gap of chromophores. However, a full understanding of excited state processes requires in addition an exploration of the excited state potential energy surface and electronic properties, which is greatly aided by the availability of analytic energy gradients. Owing to its robust high accuracy over a wide range of chemical problems, equation-of-motion coupled cluster with single and double excitations (EOM-CCSD) is a powerful method for predicting excited state properties, and the implementation of analytic gradients of all many of EOM-CCSD (excitation energies, ionization potentials, electron attachment energies, etc.) along with numerous successful applications highlights the flexibility of the method. In specific cases where a higher level of accuracy is needed, or in more complex electronic structures, the inclusion of triple excitations becomes essential, for example in the EOM-CCSD* approach of Saeh and Stanton. In this work, we derive and implement for the first time the analytic gradients.
of EOMEE-CCSD*, which also provides a template for analytic gradients of related excited state methods with perturbative triple excitations. The capabilities of analytic EOMEE-CCSD* gradients are illustrated by several representative examples.

Introduction

Molecular excited states provide a rich potential energy landscape and varied electronic properties which are accessible for example through optical or ultraviolet (UV) absorption. Relaxation and decay to the ground state via either radiative or non-radiative processes provides additional functionality and spectroscopic information. These molecular excited states have unique properties that are of great interest to many fields of science and technology: bioluminescence,\textsuperscript{1} mutagenesis and carcinogenesis,\textsuperscript{2} and photovoltaics and light-emitting diodes,\textsuperscript{3,4} to name a few. Experimentally, excited states can be investigated through UV spectroscopy, fluorescence, phosphorescence, and so on. The interpretation of this experimental spectroscopy data can be aided by theoretical quantum chemistry calculations in order to gain a comprehensive understanding of the excited states.

Despite the inherent complexity of excited states (for example, open-shell and multi-reference character), numerous models have been developed to calculate excited state energies and properties. Generally, these models fall into two categories. The first category comprises explicitly multi-reference methods which deal with solutions spanning a space of determinants or other suitable basis functions, which aim to provide either state-selective, state-averaged, or state-universal descriptions of electronic states. A comprehensive review of these methods is available elsewhere.\textsuperscript{5} While these methods, such as complete active space self-consistent field (CASSCF) method,\textsuperscript{6} its second perturbation-corrected variant (CASPT2),\textsuperscript{7} and the second-order $n$-electron valence state perturbation theory (NEVPT2)$^{8-10}$ have been highly successful, they demand a high level of expertise and familiarity with the system under investigation in order to tune the active orbital space and other calculation parameters. Additionally, interpreting the results can be a formidable challenge, as it often involves analyz-
ing many interacting configurations and orbitals.\textsuperscript{5,11} On the other hand, equation-of-motion coupled cluster (EOM-CC) methods adopt a distinct approach, encompassing the entire set of multi-configurational target states within a single-reference framework. This black box theory has gained considerable popularity, owing to its favorable computational accessibility and its robust high accuracy over a wide range of chemical problems.\textsuperscript{12} For excited states, equation-of-motion coupled cluster with single and double excitations (EOMEE-CCSD)\textsuperscript{13} has become the gold standard method for the prediction of singly excited state energies and properties. The development of variants for ionization potentials (EOMIP-CCSD), electron attachment energies (EOMEA-CCSD), and spin-flip excitations (EOMSF-CCSD) has greatly extended its capability to describe doublet radicals, diradicals, triradicals, and bond breaking.\textsuperscript{14,15}

Energy differences and transition probabilities are the key quantities that are desirable to explain and predict the spectroscopic.\textsuperscript{16} Importantly, these quantities have to be computed at significant points in the potential energy surface (PES). The development of analytic-derivative techniques has significantly facilitated the exploration of potential energy surfaces (PESs) of molecules in their electronic ground state, essentially for all available quantum-chemical schemes.\textsuperscript{17–27} Gradients are accessible for various multireference methods, including MCSCF,\textsuperscript{28} MR-CI,\textsuperscript{29,30} and fully internally contracted CASPT2.\textsuperscript{31} Nevertheless, the efficiency of the EOM-CCSD gradient notably distinguishes itself. Analytic gradients for all the variants of EOM-CCSD methods have been documented.\textsuperscript{13,32–37}

While EOM-CCSD offers great accuracy for predominantly one-electron excitations, it is poor for double electron excitations.\textsuperscript{38–40} Efforts have been made in either iterative or non-iterative schemes to include triple excitation to improve accuracy, particularly for double-excited states,\textsuperscript{41} EOM-CCSD*,\textsuperscript{42} EOM-CCSD(\textdegree{}),\textsuperscript{40} EOM-CCSDT-3,\textsuperscript{40} EOM-CC3 (equivalent to LR-CC3 for energies),\textsuperscript{43,44} and EOM-CCSDT.\textsuperscript{45} However, no analytic gradients have been published for these methods, although gradients for full EOMEE-CCSDT are available in the cfour program package\textsuperscript{46} and transition dipole moments for EOM-CCSDT were re-
ported by Hirata. \(^{47}\) Counter-intuitively, the analytic gradients for the non-iterative triples methods are more complicated than for iterative models, as discussed below. In this paper, we present, for the first time, detailed formulas and implementation details of analytic closed-shell EOMEE-CCSD* energy gradients. Several demonstrative applications are also discussed.

**Theory**

For convenience, in the following context \(|H\rangle\) represents the set of all possible \(n\)-electron Slater determinants within a given spin-orbital basis. \(|H\rangle\) can be further split into \(|H\rangle = |P\rangle + |Q\rangle = |0\rangle + |G\rangle + |Q\rangle = |0\rangle + |S\rangle + |D\rangle + |Q\rangle\), where \(|0\rangle\) is the reference determinant (Fermi vacuum), \(|S\rangle\) are the singly-excited determinants, and \(|D\rangle = |\vec{e}^b_{ij}\rangle\) are the doubly-excited determinants. \(|Q\rangle = |\vec{e}^{abc}_{ijk}\rangle + |\vec{e}^{abcd}_{ijkl}\rangle + \ldots \), represents any determinant of excitation rank higher than two, out of which we specifically identify the triply-excited determinants \(|T\rangle = |\vec{e}^{abc}_{ijk}\rangle\).

Briefly, the ground CCSD and excited state EOMEE-CCSD energies are given by,

\[
E_{\text{CCSD}} = \langle 0 | e^{-\hat{T}} \hat{H} e^{\hat{T}} | 0 \rangle = \langle 0 \rangle \langle \hat{H} e^{\hat{T}} | | 0 \rangle = \langle 0 | \hat{H} | 0 \rangle \tag{1}
\]

\[
0 = \langle G | \hat{H} | 0 \rangle \tag{2}
\]

\[
\hat{H} = \sum_{pq} F_{pq}^a \{ a_p^\dagger a_q \} + \frac{1}{4} \sum_{pqrst} V_{pqrst}^a \{ a_p^\dagger a_q^\dagger a_t a_r \} = \hat{F} + \hat{V} \tag{3}
\]

\[
\hat{T} = \sum_{ai} t_{ai}^a a_i^\dagger a_i + \frac{1}{4} \sum_{abij} t_{abij}^a a_b^\dagger a_i^\dagger a_j a_i = \hat{T}_1 + \hat{T}_2 \tag{4}
\]

\[
E_{\text{EOM-CCSD}} = \langle 0 | \hat{L} \hat{H} \hat{R} | 0 \rangle \tag{5}
\]

\[
0 = \langle G | (\hat{H} - E_{\text{EOM-CCSD}}) \hat{R} | 0 \rangle \tag{6}
\]

\[
0 = \langle 0 \rangle \langle \hat{H} - E_{\text{EOM-CCSD}} | | G \rangle \tag{7}
\]

\[
\hat{R} = r_0 + \sum_{ai} r_{ai}^a a_i a_i + \frac{1}{4} \sum_{abij} r_{abij}^a a_b^\dagger a_i^\dagger a_j a_i = \hat{R}_0 + \hat{R}_1 + \hat{R}_2 \tag{8}
\]

\[
\hat{L} = \sum_{ai} l_{ai}^a a_i a_i + \frac{1}{4} \sum_{abij} l_{abij}^a a_b^\dagger a_i^\dagger a_j a_i = \hat{L}_1 + \hat{L}_2 \tag{9}
\]
where {...} indicates normal ordering and (...)_c indicates a connected operator product. The energy \( E_{\text{EOM-CCSD}} \) and amplitudes \( \hat{R} \) and \( \hat{L} \) correspond to a single unspecified excited state. When properties of a specific excited state \( \mu \) are required, we will employ the notation \( E_{\text{EOM-CCSD}}(\mu), \hat{R}(\mu), \hat{L}(\mu), \) etc.

The perturbative third-order correction\(^{48}\) to the EOM-EE-CCSD energy can be written as follows, with \( \omega = E_{\text{EOM-CCSD}} - E_{\text{CCSD}} \) being the EOM-EE-CCSD vertical excitation energy,

\[
\Delta E = \langle 0 | \hat{L} \hat{H} | T \rangle \langle T | (\omega - \hat{F})^{-1} | T \rangle \langle T | \hat{V} \hat{R}_2 + (\hat{V} \hat{T}_2 \hat{R}_1)_c | 0 \rangle
\]

The total EOM-CCSD\(^*\) energy can be written as,

\[
E_{\text{EOM-CCSD\(^*\)}} = E_{\text{EOM-CCSD}} + \Delta E = \langle 0 | \hat{L} \hat{H} \hat{R} | 0 \rangle + \langle 0 | \hat{L}_3 (\omega - \hat{F}) \hat{R}_3 | 0 \rangle
\]

\[
\langle 0 | \hat{L}_3 | T \rangle = \langle 0 | \hat{L} \hat{V} (\omega - \hat{F})^{-1} | T \rangle
\]

\[
\langle T | \hat{R}_3 | 0 \rangle = \langle T | (\omega - \hat{F})^{-1} (\hat{V} \hat{R}_2 + (\hat{V} \hat{T}_2 \hat{R}_1)_c) | 0 \rangle
\]

assuming a canonical reference determinant. Straightforward differentiation with respect to an arbitrary perturbation \( \chi \) gives,

\[
E_{\text{EOM-CCSD\(^*\)}}^\chi = E_{\text{EOM-CCSD\(^*\)}}^\chi + \langle 0 | \hat{L} \hat{V} \hat{R}_3 | 0 \rangle + \langle 0 | \hat{L} \hat{V}^\chi \hat{R}_3 | 0 \rangle
\]

\[
- \langle 0 | \hat{L}_3 (\omega - \hat{F})^\chi \hat{R}_3 | 0 \rangle + \langle 0 | \hat{L}_3 (\hat{V} \hat{R}_2 + \hat{V} \hat{T}_2 \hat{R}_1)_c^\chi | 0 \rangle
\]

\[
+ \langle 0 | \hat{L}_3 \hat{V} \hat{T}_2 \hat{R}_1^\chi | 0 \rangle + \langle 0 | \hat{L}_3 \hat{V}^\chi (\hat{R}_2 + \hat{T}_2 \hat{R}_1)_c | 0 \rangle
\]

Note there is a negative sign in the term \( \langle 0 | \hat{L}_3 (\omega - \hat{F})^\chi \hat{R}_3 | 0 \rangle \) as,

\[
((\omega - \hat{F})^{-1})^\chi = -((\omega - \hat{F})^{-1})^\chi ((\omega - \hat{F})^{-1})
\]
This term can be further expanded as follows, where for convenience we set \( \delta = \langle 0 | \hat{L}_3 \hat{R}_3 | 0 \rangle \),

\[
- \langle 0 | \hat{L}_3 (\omega - \hat{F}) \hat{R}_3 | 0 \rangle = \langle 0 | \hat{L}_3 (\hat{F}^x + E_{\text{CCSD}}^x - E_{\text{EOM-CCSD}}^x) \hat{R}_3 | 0 \rangle \\
= \langle 0 | \hat{L}_3 \hat{F}^x \hat{R}_3 | 0 \rangle + \delta (E_{\text{CCSD}}^x - E_{\text{EOM-CCSD}}^x) 
\]

(16)

\( E_{\text{CCSD}}^x \) and \( E_{\text{EOM-CCSD}}^x \) can be obtained by directly differentiating the energy equations (1) and (5), respectively. Noting the amplitude equations (2), (6), and (7) and that \( \langle 0 | \hat{L} \hat{R} | 0 \rangle^x = 0 \),

\[
E_{\text{CCSD}}^x = \langle 0 | \hat{H} \hat{F}^x | 0 \rangle + \langle 0 | \hat{H}^{(x)} | 0 \rangle \\
E_{\text{EOM-CCSD}}^x = \langle 0 | \hat{L} \hat{R}^{(x)} | 0 \rangle + \langle 0 | \hat{L} \hat{Q} | \hat{Q} \rangle + \langle 0 | \hat{R} \hat{T}^x | 0 \rangle 
\]

(17)

(18)

where \( \hat{H}^{(x)} = (\hat{H} e^\hat{T} x) \).

Combining (14), (16), (17), and (18) and grouping all terms into four parts we obtain,

\[
E_{\text{EOM-CCSD}}^x = \text{I} + \text{II} + \text{III} + \text{IV} \\
\text{I} = \langle 0 | \hat{L} \hat{F}^x \hat{R}_3 | 0 \rangle \\
\text{II} = \langle 0 | \hat{L}_3 (\hat{V} \hat{R}_2^x + \hat{T}_2 \hat{R}_1^x) | 0 \rangle \\
\text{III} = \langle 0 | \hat{L}_3 (\hat{V} \hat{T}_2^x \hat{R}_1) | 0 \rangle + \langle 0 | (1 - \delta) \hat{L} \hat{H} | \hat{Q} \rangle + \langle 0 | \hat{R} \hat{T}^x | 0 \rangle + \delta \langle 0 | \hat{H} \hat{T}^x | 0 \rangle \\
\text{IV} = \langle 0 | (1 - \delta) \hat{L} \hat{H}^{(x)} \hat{R} | 0 \rangle + \langle 0 | \hat{L}_3 (\hat{V} \hat{R}_2^x + \hat{V} \hat{T}_2 \hat{R}_1) | 0 \rangle \\
+ \langle 0 | \hat{L} \hat{F}^x \hat{R}_3 | 0 \rangle + \langle 0 | \hat{L}_3 \hat{F}^x \hat{R}_3 | 0 \rangle + \delta \langle 0 | \hat{H}^{(x)} | 0 \rangle 
\]

(19)

(20)

(21)

(22)

The terms are grouped such that \( \text{I} \) depends (directly) only on \( \hat{L}^x \), \( \text{II} \) depends only on \( \hat{R}^x \), \( \text{III} \) depends only on \( \hat{T}^x \), and \( \text{IV} \) depends only on the derivatives of integrals.

\( \hat{L}^x \) and \( \hat{R}^x \) of course depend in turn on \( \hat{T}^x \) due to the presence of \( \hat{H} \) in their definite amplitude equations. Thus, we should tackle \( \text{I} \) and \( \text{II} \) before moving on to \( \text{III} \). To analyze
we first define a convenient intermediate, noting that $\hat{L}^\chi$ spans only the $\langle G |$ space,

$$\mathbf{I} = \langle 0 | \hat{L}^\chi \hat{V} \hat{R}_3 | 0 \rangle = \langle 0 | \hat{L}^\chi \hat{\Sigma} | 0 \rangle$$  \hspace{1cm} (23) \\
$$\langle G | \hat{\Sigma} | 0 \rangle = \langle G | \hat{V} \hat{R}_3 | 0 \rangle$$  \hspace{1cm} (24)

We then introduce a definition of $\hat{L}^\chi$ by differentiating (7),

$$\langle 0 | \hat{L}^\chi (\hat{H} - E_{\text{EOM-CCSD}}) | G \rangle = \langle 0 | \hat{L} \hat{\gamma} (\hat{H} - E_{\text{EOM-CCSD}}) | G \rangle - \langle 0 | \hat{L} \hat{H}^{(\chi)} | G \rangle$$

$$+ E_{\text{EOM-CCSD}}^\chi \langle 0 | \hat{L} | G \rangle - \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{T}^\gamma | G \rangle$$  \hspace{1cm} (25)

Notionally, one would utilize this definition by applying the inverse of $\langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle$ to both sides of the equation, yielding $\langle 0 | \hat{L}^\chi | G \rangle = \langle 0 | \hat{X} | G \rangle \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^{-1}$ with $\hat{X}$ collecting all terms from the R.H.S., which then could be substituted in $\mathbf{I}$ giving $\langle 0 | \hat{X} | G \rangle \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^{-1} \langle G | \hat{\Sigma} | 0 \rangle$. Finally, one would then define a new set of amplitudes $\langle G | \hat{T} | 0 \rangle$ defined by the system of equations $\langle G | (\hat{H} - E_{\text{EOM-CCSD}}) \hat{T} | 0 \rangle = \langle G | \hat{\Sigma} | 0 \rangle$.

However, this plan is hindered by the fact that $E_{\text{EOM-CCSD}}^\chi$ is an exact eigenvalue of $\langle G | \hat{H} | G \rangle$, and thus the shifted matrix is exactly singular. This problem can be avoided by imposing an additional condition on $\hat{L}^\chi$ (and $\hat{R}^\chi$): while the normalization condition $\langle 0 | \hat{L} \hat{R} | 0 \rangle = 1$ ensures that $\langle 0 | \hat{L}^\chi \hat{R} | 0 \rangle + \langle 0 | \hat{L} \hat{R}^\chi | 0 \rangle = 0$, we can additionally require a biorthogonal form of intermediate normalization such that $\langle 0 | \hat{L}^\chi \hat{R} | 0 \rangle = \langle 0 | \hat{L} \hat{R}^\chi | 0 \rangle = 0$. Thus, the part of $\hat{\Sigma}$ parallel to $\hat{R}$ is not required and can be projected out. Similarly, the fact that $\hat{L}$ lies in the null space of $\langle G | (\hat{H} - E_{\text{EOM-CCSD}}) | G \rangle$ implies that each term on the R.H.S. of (25) is perpendicular to $\hat{L}$. Thus, it is sufficient and safe to employ the pseudo-inverse,

$$\langle 0 | \hat{L}^\chi | G \rangle = \langle 0 | \hat{L} \hat{T}^\chi | G \rangle (\mathbf{I} - \langle G | \hat{R} | 0 \rangle \langle 0 | \hat{L} | G \rangle)$$

$$- \left[ \langle 0 | \hat{L} \hat{H}^{(\chi)} | G \rangle - E_{\text{EOM-CCSD}}^\chi \langle 0 | \hat{L} | G \rangle \right]$$

$$+ \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{T}^\chi | G \rangle \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+$$  \hspace{1cm} (26)
\[
\langle G | \hat{H} - E_{\text{EOM-CCSD}}(\mu) | G \rangle^+ = \sum_{\nu \neq \mu} \frac{\langle G | \hat{R}(\nu) | 0 \rangle \langle 0 | \hat{L}(\nu) | G \rangle}{E_{\text{EOM-CCSD}}(\nu) - E_{\text{EOM-CCSD}}(\mu)}
\]

(27)

where the pseudo-inverse satisfies,

\[
\langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+ = \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+ \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle = \hat{I} - \langle G | \hat{R} | 0 \rangle \langle 0 | \hat{L} | G \rangle
\]

(28)

The introduction of the pseudo-inverse now allows \( \hat{L}^\chi \) to be substituted into \( \mathbf{I} \),

\[
\mathbf{I} = \langle 0 | \hat{L} \hat{T}^\chi \hat{\Sigma}^\perp | 0 \rangle - \left[ \langle 0 | \hat{L} \hat{H}^{(\chi)} | G \rangle - E_{\text{EOM-CCSD}}^{\chi} \langle 0 | \hat{L} | G \rangle \right.
\]
\[
+ \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{T}^\chi | G \rangle \right] \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+ \langle G | \hat{\Sigma} | 0 \rangle
\]
\[
= \langle 0 | \hat{L} \hat{\Sigma}^\perp \hat{T}^\chi | 0 \rangle + \langle 0 | \hat{L} \hat{H}^{(\chi)} \hat{T} | 0 \rangle - E_{\text{EOM-CCSD}}^{\chi} \langle 0 | \hat{L} \hat{T} | 0 \rangle
\]
\[
+ \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{T}^\chi \hat{T}^\chi | 0 \rangle
\]
\[
\langle G | \hat{\Sigma}^\perp | 0 \rangle = \langle G | \hat{\Sigma} | 0 \rangle - \langle G | \hat{R} | 0 \rangle \langle 0 | \hat{L} \hat{\Sigma} | 0 \rangle
\]

(29)

(30)

where we introduce an excitation operator \( \hat{T} \) spanning \( | G \rangle \) as,

\[
\langle G | (\hat{H} - E_{\text{EOM-CCSD}}) \hat{T} | 0 \rangle = - \langle G | \hat{\Sigma}^\perp | 0 \rangle
\]

(31)

The portion of \( \hat{T} \) parallel to \( \hat{R} \) is undefined, but it is convenient and numerically stable to require \( \hat{T} \) be perpendicular to \( \hat{R} \). In particular, this means that \( \langle 0 | \hat{L} \hat{T} | 0 \rangle = 0 \). Finally we note that biorthogonal intermediate normalization is only one choice of uniquely determining \( \langle 0 | \hat{L} \hat{T} | 0 \rangle \) and \( \langle 0 | \hat{L} \hat{R}^\chi | 0 \rangle \). For example, one could require that \( \langle 0 | \hat{R}^\dagger \hat{R} | 0 \rangle = 1 \) which would still allow the definition of a consistent set of \( \hat{T} \) amplitudes with a modified set of equations (involving an operator more complicated than the pseudo-inverse). Thus to avoid a loss of generality we do not assume that \( \langle 0 | \hat{L} \hat{T} | 0 \rangle = 0 \) or other similar consequences of biorthogonal intermediate normalization.
For II, we employ a similar strategy,

\[
\Pi = \langle 0 | \hat{L}_3 (\hat{V} \hat{R}^x + \hat{V} \hat{T}_2 \hat{R}^x) | 0 \rangle = \langle 0 | \hat{\Omega} \hat{R}^x | 0 \rangle \tag{32}
\]

\[
\langle 0 | \hat{\Omega} | S \rangle = \langle 0 | \hat{L}_3 \hat{V} \hat{T}_2 | S \rangle \tag{33}
\]

\[
\langle 0 | \hat{\Omega} | D \rangle = \langle 0 | \hat{L}_3 \hat{V} | D \rangle \tag{34}
\]

\[
\langle G | \hat{R}^x | 0 \rangle = - (\hat{I} - \langle G | \hat{R} | 0 \rangle \langle 0 | \hat{L} | G \rangle) \langle G | \hat{R} \hat{T}^x | 0 \rangle - \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+ \left[ \langle G | \hat{H}^{(x)} \hat{R} | 0 \rangle - E_{\text{EOM-CCSD}}^{\chi} \langle G | \hat{R}^x | 0 \rangle \right]
\]

\[
- \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+ \left[ \langle G | \hat{H}^{(x)} \hat{R} | 0 \rangle - E_{\text{EOM-CCSD}}^{\chi} \langle G | \hat{R}^x | 0 \rangle \right] - \langle G | \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}^x | 0 \rangle
\]

Substitution followed by the definition of another set of de-excitation amplitudes \(\Pi\) gives,

\[
\Pi = - \langle 0 | \hat{\Omega}^\dagger \hat{T}^x R^x | 0 \rangle + \langle 0 | \hat{\Omega} | G \rangle \langle G | \hat{H} - E_{\text{EOM-CCSD}} | G \rangle^+
\]

\[
\times \left[ \langle G | \hat{H}^{(x)} \hat{R} | 0 \rangle - E_{\text{EOM}}^{\chi} \langle G | \hat{R} | 0 \rangle + \langle G | \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}^x | 0 \rangle \right]
\]

\[
= - \langle 0 | \hat{\Omega}^\dagger \hat{T}^x R^x | 0 \rangle + \langle 0 | \hat{\Pi} \hat{H}^{(x)} \hat{R} | 0 \rangle - E_{\text{EOM-CCSD}}^{\chi} \langle 0 | \hat{\Pi} \hat{R} | 0 \rangle
\]

\[
+ \langle 0 | \hat{\Pi} \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}^x | 0 \rangle
\]

(36)

with \(\Pi\) defined as,

\[
\langle 0 | \hat{\Pi} (\hat{H} - E_{\text{EOM-CCSD}}) | G \rangle = - \langle 0 | \hat{\Omega}^\dagger | G \rangle \tag{37}
\]

Combining I and II and expanding \(E_{\text{EOM-CCSD}}^{\chi}\) using (18), we arrive at an expression where the direct dependence on \(\hat{R}^x\) and \(\hat{L}^x\) has been replaced by dependence on \(\hat{T}^x\) and \(\hat{H}^x\) (via \(\hat{H}^{(x)}\)),

\[
I + II = \left[ \langle 0 | (\hat{L} \hat{\Sigma}^\perp - \hat{\Omega}^\dagger \hat{R}) \hat{T}^x | 0 \rangle + \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{T}^x \hat{T}^x | 0 \rangle \right]
\]

\[
+ \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}^x | 0 \rangle - \epsilon \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}^x | 0 \rangle
\]

\[
+ \left[ \langle 0 | \hat{L} \hat{H}^{(x)} \hat{T} | 0 \rangle + \langle 0 | \hat{L} \hat{H}^{(x)} \hat{R} | 0 \rangle - \epsilon \langle 0 | \hat{L} \hat{H}^{(x)} \hat{R} | 0 \rangle \right]
\]

9

160
\[
\begin{align*}
\text{III}' + \text{IV}' &= \text{(38)} \\
\varepsilon &= \langle 0 | (\hat{\Pi} \hat{R} + \hat{L} \hat{Y}) | 0 \rangle \quad \text{(39)}
\end{align*}
\]

Note that based on assumptions above, \( \varepsilon = 0 \). However, we keep \( \varepsilon \) in the following derivation for completeness.

The additional terms in \( \text{III}' \) can now be added to \( \text{III}, \)

\[
\begin{align*}
\text{III} + \text{III}' &= \langle 0 | \hat{L}_3 \hat{V} \hat{R}_1 \hat{T}_2^x | 0 \rangle + \langle 0 | (1 - \delta - \varepsilon) \hat{L} \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}_2^x | 0 \rangle + \delta \langle 0 | \hat{H} \hat{T}_2^x | 0 \rangle \\
&\quad + \langle 0 | (\hat{L} \hat{\Sigma}^\perp - \hat{\Omega}^\perp \hat{R}) \hat{T}_2^x | 0 \rangle + \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{\Upsilon} \hat{T}_2^x | 0 \rangle \\
&\quad + \langle 0 | \hat{\Pi} \hat{H} | Q \rangle \langle Q | \hat{R} \hat{T}_2^x | 0 \rangle \\
&= \langle 0 | \hat{\Xi} \hat{T}_2^x | 0 \rangle \quad \text{(40)}
\end{align*}
\]

\[
\langle 0 | \hat{\Xi} | G \rangle = \langle 0 | \hat{L}_3 \hat{V} \hat{R}_1 | D \rangle + \langle 0 | (1 - \delta - \varepsilon) \hat{L} \hat{H} | Q \rangle \langle Q | \hat{R} | G \rangle + \delta \langle 0 | \hat{H} | G \rangle \\
&\quad + \langle 0 | (\hat{L} \hat{\Sigma}^\perp - \hat{\Omega}^\perp \hat{R}) | G \rangle + \langle 0 | \hat{L} \hat{H} | Q \rangle \langle Q | \hat{\Upsilon} | G \rangle \\
&\quad + \langle 0 | \hat{\Pi} \hat{H} | Q \rangle \langle Q | \hat{R} | G \rangle \quad \text{(41)}
\]

Differentiation of the \( \hat{T} \) amplitude equations (2) yields an expression for \( \langle P | \hat{T}_2^x | 0 \rangle \), where a pseudo-inverse is no longer required due to the fact that while \( E_{\text{CCSD}} \) is an eigenvalue of \( \langle P | \hat{H} | P \rangle \), removal of the reference determinant in \( \langle G | \hat{H} - E_{\text{CCSD}} | G \rangle \) prevents singularity,

\[
\langle G | \hat{T}_2^x | 0 \rangle = - \langle G | \hat{H} - E_{\text{CCSD}} | G \rangle^{-1} \langle G | \hat{H}^{(\chi)} | 0 \rangle \quad \text{(42)}
\]

Using this definition we introduce a final set of de-excitation amplitudes \( \hat{Z} \),

\[
\begin{align*}
\text{III} + \text{III}' &= - \langle 0 | \hat{\Xi} | G \rangle \langle G | \hat{H} - E_{\text{CCSD}} | G \rangle^{-1} \langle G | \hat{H}^{(\chi)} | 0 \rangle \\
&= \langle 0 | \hat{Z} \hat{H}^{(\chi)} | 0 \rangle \\
&= \text{IV}'' \quad \text{(43)}
\end{align*}
\]

\[
\langle 0 | \hat{Z} (\hat{H} - E_{\text{CCSD}}) | G \rangle = - \langle 0 | \hat{\Xi} | G \rangle \quad \text{(44)}
\]

10

161
Finally, the remaining terms \( \text{IV}' \) and \( \text{IV}'' \) from above can be combined with \( \text{IV} \) to yield the total gradient expression,

\[
E_{\text{EOM-CCSD*}}^x = \text{IV} + \text{IV}' + \text{IV}''
\]

\[
= \langle 0 | \hat{\Pi} H^{(x)} \hat{R} | 0 \rangle + \langle 0 | \hat{\Pi} H^{(x)} \hat{T} | 0 \rangle
+ \langle 0 | \hat{L} H^{(x)} \hat{R} | 0 \rangle + \langle 0 | \hat{L} H^{(x)} \hat{T} | 0 \rangle
+ \langle 0 | \hat{L} \hat{V} x \hat{R}_3 | 0 \rangle + \langle 0 | \hat{L} \hat{V} x \hat{R}_2 + \hat{V} x \hat{T}_3 | 0 \rangle
\]

\[
+ \langle 0 | \hat{L}_3 \hat{F} x \hat{R}_3 | 0 \rangle
\]

\[
= \sum_p D_p^p (\epsilon_p)^x + \sum_{pqrs} \Gamma_{rs}^{pq} (v_{pq}^{rs})^x
\]  

(45)

(46)

where \( \epsilon_p = f_p^p \). The one- and two-particle density matrices \( D_p^q \) and \( \Gamma_{rs}^{pq} \) can be then be constructed from the various amplitudes and then processed using standard techniques such as in the computation of the CCSD(T) gradient.\(^{26,49}\)

**Results**

The analytic gradients of EOMEE-CCSD\(^*\) have implemented in the development version of the cfour program package.\(^{46}\) In this section, we discuss the validation of our implementation and demonstrate the illustrative application of analytic EOMEE-CCSD\(^*\) gradients to a set of excited states. In the geometry optimization, the following convergence thresholds were used, denoted using the relevant cfour keywords: \( \text{SCF\_CONV} = 10^{-10}, \text{CC\_CONV} = 10^{-9}, \text{LINEQ\_CONV} = 10^{-8}, \text{ESTATE\_CONV} = 10^{-8}, \text{GEO\_MAX\_STEP} = 50 \text{ a.u.}, \) and \( \text{GEO\_CONV} = 10^{-7} \), except where indicated.

**Validation**

Formaldehyde has been extensively investigated both experimentally\(^{50}\) and theoretically.\(^{51–56}\)

In addition, its small system size makes it a great candidate for validating the correctness
of our implementation against numerical differentiation. The $1^1A_2(n \rightarrow \pi^*)$ valence excited state is investigated here. The ground state geometry is obtained from the QUEST2 dataset,$^{57}$ which has been optimized at the CC3/aug-cc-pVTZ level. The computed EOM-CCSD* gradients (with the 6-31G** basis) using both the analytic gradient we implemented and finite differences of energies agree well with each other. The differences between the numerical and analytic values for the individual gradient components were, in all cases, less than $10^{-7}$ a.u. (see SI), which confirmed the correctness of our implementation.

It’s established that the $1^1A_2$ state exhibits a non-planar equilibrium geometry.$^{50,51,56}$ Consequently, $C_s$ symmetry is employed during optimization. The optimized geometries and adiabatic excitation energies are summarized in Table 1. The optimized geometries are compared to lower-level EOM-CCSD and higher-level EOM-CCSDT. To facilitate comparison, experimental data are also included.$^{58,60}$ As expected, optimized geometries with EOM-CCSDT are in better agreement with the experimental data. One finds $\leq 0.01 \text{ Å}$ difference in the C–O bond and C–H bonds, $0.4^\circ$ and $2.1^\circ$ difference in the angle and out-of-plane dihedral, respectively. No consistent trend in the theoretical methods’ order is observed for these parameters, except for the consistent elongation of bond distances in EOM-CCSDT calculations compared to EOM-CCSD results. Although the optimized CO bond distance is closer between EOM-CCSD* results and EOM-CCSDT results than with EOM-CCSD results, the other optimized geometric parameters with EOM-CCSD* deviated in the wrong direction, leading to a larger disparity from the experimental data. When examining the computed adiabatic excitation energies, a notable observation arises when comparing the energy calculated by EOM-CCSD with EOM-CCSD* and EOM-CCSDT. The incorporation of triple excitations in the wave function markedly decreases the adiabatic excitation energies. It’s worth noting that while this enhancement occurs, our results deviate further from experimental data. However, conducting a rigorous comparison with experimental findings presents challenges due to the absence of zero-point vibrational energy (ZPVE) corrections in our results.
Table 1. Optimized geometry of $n \rightarrow \pi^*$ excited state of formaldehyde with cc-aug-pVTZ basis set.

<table>
<thead>
<tr>
<th></th>
<th>EOM-CCSD</th>
<th>EOM-CCSD*</th>
<th>EOM-CCSDT</th>
<th>Expt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{CO}$ (Å)</td>
<td>1.306</td>
<td>1.343</td>
<td>1.331</td>
<td>1.323</td>
</tr>
<tr>
<td>$r_{CH}$ (Å)</td>
<td>1.091</td>
<td>1.089</td>
<td>1.093</td>
<td>1.103</td>
</tr>
<tr>
<td>$\angle_{COH}$ (°)</td>
<td>119.0</td>
<td>121.0</td>
<td>118.5</td>
<td>118.1</td>
</tr>
<tr>
<td>Out of plane dihedral (°)</td>
<td>149.6</td>
<td>152.2</td>
<td>143.9</td>
<td>146.0</td>
</tr>
<tr>
<td>$T_e$ (eV)</td>
<td>3.72</td>
<td>3.12</td>
<td>3.06</td>
<td>3.50</td>
</tr>
</tbody>
</table>

Harmonic vibrational frequencies and infrared intensities have been obtained by numerical differentiation of analytic energy derivatives and dipole moments at the optimized geometries and presented in Table 2. From this, we can confirm that the optimized geometry corresponds to a minimum. The frequencies computed using EOM-CCSD* align closely with those from EOM-CCSDT, displaying a consistent trend converging towards experimental data.

Table 2. Harmonic vibrational frequencies ($\omega$), infrared intensities ($I$) for the $n \rightarrow \pi^*$ excited state of formaldehyde with cc-aug-pVDZ.

<table>
<thead>
<tr>
<th>Mode $\omega$</th>
<th>EOM-CCSD $\omega$</th>
<th>EOM-CCSD* $\omega$</th>
<th>EOM-CCSDT $\omega$</th>
<th>Expt.$^a$ $\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_1(A')$</td>
<td>3143.1</td>
<td>3054.4</td>
<td>3011.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$\omega_2(A')$</td>
<td>1362.0</td>
<td>1364.7</td>
<td>1342.8</td>
<td>8.2</td>
</tr>
<tr>
<td>$\omega_3(A')$</td>
<td>1291.7</td>
<td>1131.7</td>
<td>1198.1</td>
<td>21.7</td>
</tr>
<tr>
<td>$\omega_4(A')$</td>
<td>427.0</td>
<td>548.7</td>
<td>655.8</td>
<td>40.2</td>
</tr>
<tr>
<td>$\omega_5(A'')$</td>
<td>3260.1</td>
<td>3176.7</td>
<td>3121.3</td>
<td>0.0</td>
</tr>
<tr>
<td>$\omega_6(A'')$</td>
<td>864.6</td>
<td>921.6</td>
<td>924.9</td>
<td>6.1</td>
</tr>
</tbody>
</table>

(1) C–H stretching; (2) C–O stretching; (3) CH$_2$ bending; (4) out-of-plane bending; (5) C–H stretching; (6) CH$_2$ rocking.

$^a$ Ref. 60

$n \rightarrow \pi^*$ state of s-tetrazine

Excited states of s-tetrazine are of great interest in the scientific community, as greatly summarized in the introduction of the work by Angeli. The ground state geometry is obtained
from the QUEST2 dataset,\textsuperscript{57} which has been optimized at the CC3/aug-cc-pVTZ level. In this study, we focus on the two lowest excited singlet states of $B_{3u}$ and $B_{2g}$ symmetry. The $B_{3u} \ n \rightarrow \pi^*$ excited state is predominantly characterized by the configuration where a single electron is promoted from the highest occupied molecular orbital $b_{3g}$ to the lowest unoccupied molecular orbital $a_u$. Previous research has demonstrated that the first excited state exhibits $D_{2h}$ symmetry,\textsuperscript{62} thus guiding the optimization process with $D_{2h}$ symmetry. The optimized geometries and adiabatic excitation energies are succinctly presented in Table 3.

Consistent with previous studies,\textsuperscript{62–64} the change in geometry upon electronic excitation is small. While minimal disparities are noted in the optimized geometries between the EOM-CCSD and EOM-CCSD* calculations, a consistent elongation in bond length emerges in the EOM-CCSD* results than EOM-CCSD. As previously observed, the inclusion of triple excitations substantially reduces the adiabatic excitation energy by 0.75 eV. For comparison, results from additional theoretical studies and experimental findings are included in Table 3. The absence of zero-point vibrational energy (ZPVE) in our calculation, coupled with experimental data not obtained at 0 K,\textsuperscript{65} hampered the straightforward comparisons.

**Table 3.** Optimized geometry of $s$-tetrazine $1^1B_{3u} \ n \rightarrow \pi^*$ state with the basis set cc-pVTZ.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$r_{CH}$ (Å)</th>
<th>$r_{CN}$ (Å)</th>
<th>$r_{NN}$ (Å)</th>
<th>$\angle_{HCN}$ (°)</th>
<th>$T_e$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground state</td>
<td>1.078</td>
<td>1.336</td>
<td>1.323</td>
<td>116.6</td>
<td></td>
</tr>
<tr>
<td>EOM-CCSD</td>
<td>1.077</td>
<td>1.328</td>
<td>1.312</td>
<td>119.4</td>
<td>2.513</td>
</tr>
<tr>
<td>EOM-CCSD*</td>
<td>1.077</td>
<td>1.329</td>
<td>1.319</td>
<td>119.2</td>
<td>1.760</td>
</tr>
<tr>
<td>CASSCF\textsuperscript{a}</td>
<td>1.067</td>
<td>1.329</td>
<td>1.305</td>
<td>121.4</td>
<td></td>
</tr>
<tr>
<td>CASPT2\textsuperscript{a}</td>
<td>1.073</td>
<td>1.333</td>
<td>1.321</td>
<td>121.5</td>
<td></td>
</tr>
<tr>
<td>Mk-MRCCSD\textsuperscript{b}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expt.\textsuperscript{c}</td>
<td>1.063</td>
<td>1.324</td>
<td>1.349</td>
<td>123.2</td>
<td>2.248\textsuperscript{d}</td>
</tr>
<tr>
<td>Expt.\textsuperscript{e}</td>
<td>1.358 ± 0.010</td>
<td>1.280 ± 0.02</td>
<td></td>
<td>118.5 ± 1.4</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a) Ref. 62} \textsuperscript{b) Ref. 64} \textsuperscript{c) Ref. 66} \textsuperscript{d) Ref. 65} \textsuperscript{e) Ref. 67}

The $B_{2g} \ n \rightarrow \pi^*$ excited state is characterized by a mixed configuration of single excitation.
The optimized geometries are summarised in Table 4. Different from the B$_{3u}$ excited state that small contractions of all bonds upon excitation are predicted, the B$_{2g}$ excited state relaxes to the geometry with $a \geq 0.13$ Å elongation of the N–N bond length and $a \geq 0.015$ Å contraction of the C–N bond length. The elongation of the N–N bond can be partially explained by the excitation from bonding 1$b_1g$ π orbital to the anti-bonding 1$a_u$ π orbital. Similar to previous observations, EOM-CCSD* results in a larger bond length than EOM-CCSD. The triples correction reduced the adiabatic excitation energy by 0.917 eV.

**Table 4.** Optimized geometry of s-tetrazine $1^1B_{2g}$ $n \rightarrow \pi^*$ state mixed with double excitation with basis set PVTZ

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$r_{CH}$ (Å)</th>
<th>$r_{CN}$ (Å)</th>
<th>$r_{NN}$ (Å)</th>
<th>$\angle_{HCN}$ (°)</th>
<th>$T_e$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground state</td>
<td>1.078</td>
<td>1.336</td>
<td>1.323</td>
<td>116.6</td>
<td></td>
</tr>
<tr>
<td>EOM-CCSD</td>
<td>1.081</td>
<td>1.319</td>
<td>1.453</td>
<td>117.2</td>
<td>5.162</td>
</tr>
<tr>
<td>EOM-CCSD*</td>
<td>1.081</td>
<td>1.321</td>
<td>1.456</td>
<td>117.1</td>
<td>4.245</td>
</tr>
</tbody>
</table>

$1^1A'$ state of cytosine

Cytosine, as one of the building blocks of life, has been evoking interest in multiple areas. It has been the motivation of experimental and theoretical efforts to characterize the nature and the properties of the lowest electronically excited states. The lowest bright single-excited state, is of interest in this work. It is recognized that the optimization of excited states frequently leads to relaxed non-planar geometry. Thus $C_1$ symmetry is used during optimization. It is a challenge to numerical differentiation in terms of computational cost. The ground state geometry is optimized at the MP2/cc-pVTZ level. The excited state is optimized with the smaller cc-pVDZ basis set. The optimized geometries are summarized in Table 5, and the atomic labels of cytosine are shown in Figure 1. For EOM-CCSD, only geometries with RMS force convergence below $10^{-5}$ are presented. Continuous optimization results in significant distortion of the ring structure, making convergence challenging, particularly due to mixing with $n \rightarrow \pi^*$. For EOM-CCSD*, geometries with
RMS force convergence below $10^{-7}$ are presented. The easier convergence of EOM-CCSD* could be owing to its superiority at describing the pseudo-Jahn-Teller (PJT) effect.\textsuperscript{42} EOM-CCSD and EOM-CCSD* results show consistent trends in bond length changes. While the $C_6-N_1$ and $C_6-N_8$ bonds contract, all other bonds exhibit elongation. Aligning with prior research,\textsuperscript{70,73,76} the decay of the ($\pi \rightarrow \pi^*$) state involves the carbonyl bond stretching and the pyramidalization of $C_{10}$ and $N_8$. Further frequency calculation confirmed it as a minimum (see SI).

**Figure 1.** Cytosine structures. (A) Cytosine structure labeled with atom indices. (B) The geometric comparison between the ground state (blue) and optimized excited state with EOM-CCSD* (orange), molecules are in the same orientation as shown in A.

**Conclusion**

It has been shown that the effect of triple excitations is important to achieving $\sim 0.1$ eV accuracy of transition energies.\textsuperscript{72} Despite numerous variants that integrate triples effects into EOM-CCSD, emerging over the years, the absence of analytic gradients remains a challenge. In response, this study introduces the formulation and implementation of analytic nuclear gradients for EOMEE-CCSD*.

We showcase the applicability and efficiency of this advancement through demonstrative applications on formaldehyde, $s$-tetrazine, and cytosine systems.

By streamlining the development and implementation of analytic gradient theory for this
Table 5. Optimized bond lengths ($r$, Å), bond angles ($\angle$, °), and dihedral angles ($\phi$, °) of the cytosine $\pi \rightarrow \pi^*$ state with basis set cc-pVDZ. EOM-CCSD RMS force converges at $10^{-5}$, angle and dihedral angle parameter may not converge, see SI; EOM-CCSD* RMS force converges at $10^{-7}$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Ground state</th>
<th>EOM-CCSD</th>
<th>EOM-CCSD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{C_2-N_1}$</td>
<td>1.314</td>
<td>1.417</td>
<td>1.440</td>
</tr>
<tr>
<td>$r_{N_3-C_2}$</td>
<td>1.354</td>
<td>1.376</td>
<td>1.401</td>
</tr>
<tr>
<td>$r_{H_4-N_3}$</td>
<td>1.000</td>
<td>1.008</td>
<td>1.020</td>
</tr>
<tr>
<td>$r_{H_2-N_3}$</td>
<td>1.004</td>
<td>1.009</td>
<td>1.021</td>
</tr>
<tr>
<td>$r_{C_6-N_1}$</td>
<td>1.373</td>
<td>1.323</td>
<td>1.289</td>
</tr>
<tr>
<td>$r_{O_7-C_6}$</td>
<td>1.218</td>
<td>1.269</td>
<td>1.331</td>
</tr>
<tr>
<td>$r_{N_8-C_6}$</td>
<td>1.414</td>
<td>1.410</td>
<td>1.383</td>
</tr>
<tr>
<td>$r_{H_9-N_8}$</td>
<td>1.008</td>
<td>1.015</td>
<td>1.014</td>
</tr>
<tr>
<td>$r_{C_{10}-N_8}$</td>
<td>1.351</td>
<td>1.384</td>
<td>1.411</td>
</tr>
<tr>
<td>$r_{H_{11}-C_{10}}$</td>
<td>1.080</td>
<td>1.089</td>
<td>1.091</td>
</tr>
<tr>
<td>$r_{C_{12}-C_{10}}$</td>
<td>1.355</td>
<td>1.436</td>
<td>1.420</td>
</tr>
<tr>
<td>$r_{H_{13}-C_{12}}$</td>
<td>1.077</td>
<td>1.094</td>
<td>1.096</td>
</tr>
<tr>
<td>$\angle_{N_1-C_2-N_3}$</td>
<td>117.0</td>
<td>112.2</td>
<td>112.5</td>
</tr>
<tr>
<td>$\angle_{C_2-N_3-H_4}$</td>
<td>121.8</td>
<td>122.0</td>
<td>112.7</td>
</tr>
<tr>
<td>$\angle_{C_2-N_3-H_5}$</td>
<td>118.0</td>
<td>117.8</td>
<td>110.1</td>
</tr>
<tr>
<td>$\angle_{C_2-N_1-C_6}$</td>
<td>120.0</td>
<td>116.1</td>
<td>114.5</td>
</tr>
<tr>
<td>$\angle_{N_1-C_6-O_7}$</td>
<td>125.2</td>
<td>123.5</td>
<td>121.9</td>
</tr>
<tr>
<td>$\angle_{N_1-C_6-N_8}$</td>
<td>116.1</td>
<td>123.6</td>
<td>125.8</td>
</tr>
<tr>
<td>$\angle_{C_6-N_8-H_9}$</td>
<td>115.0</td>
<td>115.5</td>
<td>117.0</td>
</tr>
<tr>
<td>$\angle_{C_6-N_8-C_{10}}$</td>
<td>123.8</td>
<td>121.8</td>
<td>118.4</td>
</tr>
<tr>
<td>$\angle_{N_8-C_{10}-H_{11}}$</td>
<td>117.0</td>
<td>118.7</td>
<td>118.6</td>
</tr>
<tr>
<td>$\angle_{N_8-C_{10}-C_{12}}$</td>
<td>119.8</td>
<td>115.9</td>
<td>115.2</td>
</tr>
<tr>
<td>$\angle_{C_{10}-C_{12}-H_{13}}$</td>
<td>121.4</td>
<td>119.1</td>
<td>119.6</td>
</tr>
<tr>
<td>$\phi_{N_1-C_2-N_3-H_4}$</td>
<td>180.0</td>
<td>180.0</td>
<td>155.0</td>
</tr>
<tr>
<td>$\phi_{N_1-C_2-N_3-H_5}$</td>
<td>0.0</td>
<td>0.0</td>
<td>31.7</td>
</tr>
<tr>
<td>$\phi_{C_6-N_1-C_2-N_3}$</td>
<td>-180.0</td>
<td>180.0</td>
<td>163.1</td>
</tr>
<tr>
<td>$\phi_{C_2-N_1-C_6-O_7}$</td>
<td>180.0</td>
<td>-180.0</td>
<td>177.8</td>
</tr>
<tr>
<td>$\phi_{C_2-N_1-C_6-N_7}$</td>
<td>0.0</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td>$\phi_{N_1-C_6-N_8-H_9}$</td>
<td>180.0</td>
<td>-180.0</td>
<td>178.7</td>
</tr>
<tr>
<td>$\phi_{N_1-C_6-N_8-C_{10}}$</td>
<td>0.0</td>
<td>0.0</td>
<td>18.1</td>
</tr>
<tr>
<td>$\phi_{C_6-N_8-C_{10}-H_{11}}$</td>
<td>180.0</td>
<td>180.0</td>
<td>155.3</td>
</tr>
<tr>
<td>$\phi_{C_6-N_8-C_{10}-C_{12}}$</td>
<td>0.0</td>
<td>0.0</td>
<td>-22.6</td>
</tr>
<tr>
<td>$\phi_{N_8-C_{10}-C_{12}-H_{13}}$</td>
<td>180.0</td>
<td>180.0</td>
<td>171.5</td>
</tr>
</tbody>
</table>
method, our work not only facilitates in-depth investigations with better accuracy but also substantially broadens the horizons for addressing complex problems in the field.

Acknowledgement

This work was supported by the US Department of Energy under grant DOE-SC0022893, and in part by the US National Science Foundation under grant OAC-2003931 and CHE-2143725. All calculations were performed on the ManeFrame II computing system at SMU.

Supporting Information

The following electronic supplementary information files are available from the publisher’s website:

References


(23) Lee, T. J.; Rendell, A. P. Analytic gradients for coupled-cluster energies that include noniterative connected triple excitations: Application to ic/i ii/i is/i- and it/i ir/i ia/i in/i is/i-HONO. The Journal of Chemical Physics 1991, 94, 6229-6236.


APPENDIX D

Publication on the allosteric regulation mechanism of STAT3 protein

Allosteric regulation in STAT3 interdomains is mediated by a rigid core: SH2 domain regulation by CCD in D170A variant

Tingting Zhao, Nischal Karki, Brian D. Zoltowski, Devin A. Matthews*

Department of Chemistry, Southern Methodist University, Dallas, Texas, United States of America

* dmatthews@smu.edu

Abstract

Signal Transducer and Activator of Transcription 3 (STAT3) plays a crucial role in cancer development and thus is a viable target for cancer treatment. STAT3 functions as a dimer mediated by phosphorylation of the SRC-homology 2 (SH2) domain, a key target for therapeutic drugs. While great efforts have been employed towards the development of compounds that directly target the SH2 domain, no compound has yet been approved by the FDA due to a lack of specificity and pharmacologic efficacy. Studies have shown that allosteric regulation of SH2 via the coiled-coil domain (CCD) is an alternative drug design strategy. Several CCD effectors have been shown to modulate SH2 binding and affinity, and at the time of writing at least one drug candidate has entered phase I clinical trials. However, the mechanism for SH2 regulation via CCD is poorly understood. Here, we investigate structural and dynamic features of STAT3 and compare the wild type to the reduced function variant D170A in order to delineate mechanistic differences and propose allosteric pathways.

Molecular dynamics simulations were employed to explore conformational space of STAT3 and the variant, followed by structural, conformation, and dynamic analysis. The trajectories explored show distinctive conformational changes in the SH2 domain for the D170A variant, indicating long range allosteric effects. Multiple analyses provide evidence for long range communication pathways between the two STAT3 domains, which seem to be mediated by a rigid core which connects the CCD and SH2 domains via the linker domain (LD) and transmits conformational changes through a network of short-range interactions. The proposed allosteric mechanism provides new insight into the understanding of intramolecular signaling in STAT3 and potential pharmaceutical control of STAT3 specificity and activity.

Author summary

In all living organisms, the proliferation and survival of cells are regulated by various proteins. Signal Transducers and Activators of Transcription 3 (STAT3) protein is one of these important proteins. However, the abnormal regulation of these proteins will contribute to the proliferation of cancer. The constitutive activation of STAT3 has been linked to several types of solid tumors, leukemia, and lymphomas. Consequently, STAT3...
proteins have been a key target for cancer therapy. SH2 (SRC-homology 2) domain is the key interaction site, great efforts have been made to target SH2 domain. However, specificity has been a major challenge in drug discovery. Research showing regulation of SH2 domain via CCD (coiled-coil domain) has opened a new path for drug discovery, but progress is challenged by poor understanding of the allosteric mechanism. Here, we show that CCD regulates SH2 conformation via a rigid backbone. The perturbations in CCD are transmitted through an α-helix to the rigid core that orchestrate the movement of CCD and LD (link domain), leading to structural changes in the SH2 domain. The present findings provide an allosteric mechanism with atomistic details underlying the regulation of CCD to SH2 domain in STAT3 protein. A detailed allosteric pathway allows informed drug design targeting CCD for desired downstream effect on SH2 domain and the overall STAT3 function.

Introduction

Proteins within the Signal Transducers and Activators of Transcription (STAT) family function as both signal transducers in the cytoplasm and transcription factors upon nuclear translocation. All members of STAT family consists of six domains (Fig 1A): amino-terminal domain (NTD), coiled-coil domain (CCD), DNA-binding domain (DBD), linker domain (LD), SRC-homology 2 domain (SH2), and transactivation domain (TAD) which is also named the C terminal domain [1]. STAT proteins are regulated by Janus Kinases (JAKs) and pyMOL files are available at [1BG1](https://osf.io/dvzq7/).

Funding: The authors acknowledge funding sources, including NSF research grant No. (1753167) [https://www.nsf.gov/]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

![Fig 1. STAT3 structure.](https://doi.org/10.1371/journal.pcbi.1010794.g001)
where they play a crucial role in immune response, cell division, and apoptosis, as a gene expression regulatory arm of JAK-STAT signaling pathway [2]. However, each member is activated via different types of cytokines and have unique function in the pathway [3–5].

Constitutive activation of STAT3 has been shown to play a crucial role in cancer progression [7, 8]. STAT3 binds, via the SH2 domain, to cell-surface receptors upon activation and recruitment of receptor-associated kinases. Upon binding, the recruited kinases activate STAT3 through phosphorylation within the TAD (at Y705), followed by dissociation from the receptor to form homodimers through reciprocal interactions between the SH2 domain and the phosphotyrosine (pY705) residue. These activated homodimers are then translocated to the nucleus where the DBD binds to target genes, and TAD activates the expression of proteins crucial for cell growth and survival. In normal cells, the signaling pathway is well-regulated. However, the abnormal activation of this signaling pathway promotes the development of cancer: misregulation of STAT3 in cancer cells promotes pro-oncogenic inflammation and suppresses anti-tumor immunity [9].

The direct therapeutic inhibition of STAT3 is highly desirable but remains challenging as evident from the lack of FDA–approved drugs. Specifically, significant amounts of effort have been employed to develop molecules targeting the SH2 domain of STAT3 [10–12]. The SH2 domain is a structurally conserved protein domain, which appears in many intracellular signal transducing proteins, offering a binding site for phosphorylated tyrosine residues. The SH2 domain in STAT3 contains two regions with specialized functions: the pY pocket, into which the phosphotyrosine of the target inserts, is the binding region, while residues of the pY+3 pocket interact with the three C-terminal residues of the phosphotyrosine in the target, forming a specificity–determining region [13, 14]. The inhibitors targeting the SH2 domain: phosphotyrosine motifs (pY-peptide) or phosphotyrosine-based peptidomimetic inhibitors which mimic the pTyr-Xaa-Yaa-Gln motif, have been previously investigated, as well as the associated binding mode [15–17]. In the pY pocket, R609 is the principal binding partner, along with K591, S636 and S611 which directly interact with pY705. The relative conformation and position of these residues will have a direct effect on STAT3 binding activity. In the pY+3 pocket, V637 in β31 controls accessibility to this pocket, while Y657, Q644, Y640, and E638 facilitate the hydrogen bond interaction with its target, as well as I659, W623 and F621 assist in binding of target peptide by forming hydrophobic environment [17]. However, most of these compounds have yet to be explored in clinical studies or further development of these compounds was limited due to concerns with their relative lack of potency and selectivity [18].

Several studies [20–22] have determined that effectors (small molecule and polypeptide) binding to CCD interfere with SH2 domain binding or preclude STAT3 nuclear translocation, which suggests CCD as a potential target for further drug design. Zhang et al. found that the coiled-coil domain is essential for STAT3 recruitment to the receptor: systematic deletion analysis of the N-domain and α helices of CCD, as well as mutagenesis of conserved residues (D170A) in the CCD of STAT3 were carried out and showed the diminishment of both pY-peptide binding and tyrosine phosphorylation [19]. Furthermore, the small molecule MM-206 was identified as an inhibitor of STAT3 phosphorylation, and Minus et al. surprisingly found the binding site was at α1 of CCD (around F174). In addition, the compound K116, found to bind to CCD by AlloFinder, was shown to be able to inhibit receptor binding, validated by mutagenesis and functional experiments [21]. Recently, a small polypeptide MS3–6 was found to bind to CCD, which caused significant helical tilt in CCD domain compared to the apo conformation, which further diminished DNA binding and nuclear translocation [22]. These observations are summarized in Table 1, and highlight CCD correlation to SH2 domain binding affinity as well as specificity.
The discovery of several and diverse inhibitory agents, which bind to the CCD rather than SH2, but regulate SH2 domain function, is a fascinating development. However, rational design of allosteric effectors requires a more detailed, mechanistic knowledge of how CCD binding affects SH2 structure and activity. There are no crystal structures with MM-206 or K116 binding with STAT3. Moreover, computational modeling of large fragments and optimizing the accurate binding configuration is challenging. D170A mutation, which has shown altered activity in SH2 domain, can thus be a useful avenue of investigation towards understanding the allosteric mechanism. This has the further advantage in that a single point mutation can easily be performed computationally without significantly altering the structural environment. We hypothesize that point mutation and effector binding result in a similar allosteric pathway outside of the CCD domain (where local interactions dominate). While we do not explicitly test this hypothesis here, unraveling the D170A allosteric mechanism will at least help to guide the search for allosteric mechanisms of alternate effectors.

In this work, we study the allosteric mechanism of D170A mutation on the inhibition of STAT3 activity, as predicted by the dynamic structures of the SH2 domain, known to be essential for pY-peptide binding and ensuing Y705 phosphorylation. Specifically, we investigate the structural properties within the SH2 domain upon CCD mutation over a number of molecular dynamics simulations, as well as the dynamical correlations between SH2 and CCD and the associated networks of allosteric residues and interactions within various structural motifs.

Results

Conformational differences in the pY+3 binding pocket correlate to the decreased binding affinity from wild type to D170A variant

The SH2 domain mediates binding of kinase-complexes to unphosphorylated STAT3, directing the phosphorylation of Y705 at TAD. Furthermore, the SH2 domain also provides the interface for dimerization of the phosphorylated TAD to form a functional pSTAT3 homodimer. Thus, any changes to the conformation of the specificity-determining region (pY+3 pocket, Fig 2D) as well as the binding region of phosphorylated TAD (pY pocket, Fig 2D) provides a key regulatory modification to STAT3 behavior. To describe the binding pocket conformations of SH2 domain, the pair residue center of mass (COM) distance matrix of key residues (Fig 2D, see Methods section for details) for both the pY and pY+3 pockets were calculated, separately. Principal component analysis (PCA) was employed to project the high dimensionality of pair residue COM distances into a 2D plane. The conformational space of both pockets for wild type and D170A variant is shown in Fig 2A (note that the principal components are determined from the combined wild type and D170A variant trajectories). The first two principal components contribute 62% of the total variation (S1(A) Fig), encapsulating the majority of the conformational space in just two dimensions.
In the combined PCA plot, the conformational space of the D170A variant partially overlaps with the wild type, while also exploring distinct novel conformations (Fig 2A). The additional conformational space explored by the D170A variant occurs predominately along the PC1 axis. The ten largest coefficients of the first two PCs highlight the pair of residues with largest relative motions across all the trajectories (S1(C) Fig). The motion of Y657 and Q644 yields the largest coefficients across PC1 and PC2 (Fig 2A), respectively, underlining their

Fig 2. Conformational analysis of SH2 domain binding pockets. (A) PCA of both the pY and pY+3 pockets in the wild type (blue) and D170A mutant (orange). The contour lines show the density of recorded frames in each region, and the crystal structure (PDB ID: 6NUQ) is marked for reference. 6NUQ is STAT3 with a ligand bound to the SH2 domain, shown as a reference for the ligand binding conformation. (D) SH2 domain bound to SI109 (light pink) to demonstrate binding mode of ligand to the pY and pY+3 pocket (6NUQ [11]). The pY pocket (yellow) and pY+3 pocket (orange) are shown with key residues included in the COM pair distances are shown as sticks. (B,C) PCAs for wild type and D170A variant, respectively, colored by the COM distance from Q644 to Y657. (E,F) PCAs for wild type and D170A variant, respectively, colored by the COM distance from Q644 to E638. The distance between 644—657 and 638–644 can be used as an approximate representation of PC1 and PC2 (S1(C) Fig). The plot of these two distances distribution in the 12 simulations were shown in (S2 Fig). (G) The averaged structures for each macro-state with key residues show as sticks. (H) pY and pY+3 pockets PCA 2D plane colored by different macro-states. (I) Nested pie charts showing the degree to which each system (apo and D170A) occupy each macro-state. The outer circle is colored by macro-state as in (H), and the inner circle is colored by system.

https://doi.org/10.1371/journal.pcbi.1010794.g002
The Q644–Y657 pair has the largest variance along PC1 (Fig 2B and 2C) while the E638–Q644 pair has the largest variance along PC2 (Fig 2E and 2F).

Comparing Fig 2B, 2C, 2E and 2F with Fig 2A, we see that the E638–Q644 pair has a slightly higher variance in the wild type whereas the Q644–Y657 motion, particularly above 15 Å, is predominately a feature of the D170A variant. Additionally, the D170A variant explores somewhat shorter Q644–Y657 distances than in the wild type. Most notably, all of the residues with highest variance occur in pY+3 pocket rather than the pY pocket. This result is consistent with observation by Zhang et. al., where the authors show that D170A reduces binding affinity of ligands targeting the SH2 domain. The binding mode analysis performed by Dhanik et. al. shows that the binding affinity of a ligand is directly correlated to additional interactions in the pY+3 pocket [15]. In biomolecular recognition, the two aspects binding affinity and binding specificity are coupled to each other, i.e. strong binding affinity is indicative of high substrate specificity and vice versa. Furthermore, strong coupling among flexibility and binding affinity has been shown in different systems [23, 24]. Thus, higher flexibility would lead to exploration of extended conformational landscape and reduced occupancy in substrate binding conformation. The changes in flexibility of pY+3 pocket could potentially explain the reduction of binding affinity observed by Zhang et. al.

To further investigate key differences in the overall structure, a combination of Markov State Modeling and Perron Cluster Cluster Analysis was applied to cluster transient conformations into kinetically meta-stable macro-states. Clustering into five macro-states (0–4) was applied to the combined wild type and mutant conformations (Fig 2H).

Both the D170A variant and wild type are well-represented within macro-states 2 and 4, while macro-states 0, 1 and 3 were uniquely explored by the D170A variant (Fig 2I). The average structures of each macro-state were calculated and are shown in Fig 2G. Distinct conformations of Q644 and Y657 are observed for each of the macro-states, in agreement with the PCA data. The pY+3 pocket is blocked in macro-state 2 by Y657 and Y640, both of them pointing towards the pocket. Conversely, in macro-state 0, 1, and 3, Y657 points away from the pY+3 pocket leading to an open conformation. The shared conformational states between wild type and D170A variant consists of a viable functional state of STAT3, however, D170A variant has reduced occupancy at those conformational states (Fig 2I), thus leading to a differentiated function relative to the wild type.

Allosteric regulation of SH2 pY+3 occurs via translation of motion through a rigid core

To explore the correlation between SH2 domain conformational changes and CCD conformations, CCD was first characterized via PCA analysis of all pair Cα distances. However, as the CCD domain consists of rigid helices, pair Cα distances were not able to characterize the differences between the determined macro-states, as indicated by the low variance contribution of each pair Cα distance (S3(A), S3(C) and S3(D) Fig). Only macro-state 3 showed significant differences, which originate from kinked conformations of α1 (S3(B) Fig).

The rigidity of the α helices diminishes the utility of PCA within the CCD alone to study intradomain conformational differences, however the communication between CCD and SH2 domains has been observed in the literature. To investigate this correlation, we hypothesize that the long helical arms of CCD may act as rigid levers, where any perturbation of the helical arm causes significant changes in inter-domain interaction sites. A role in the helical arm affecting the SH2 domain was in part motivated by prior studies demonstrating that monobodies targeting STAT3 interact at the helical arm and impact function through subtle bending and or rotation of helices in the CCD [22]. The argument for the subtle bending and or
rotation of the helical arm impacting signal transduction is based on crystal structures where there are packing interactions that impact the helices. Thus, whether the movements of the helical arm are a result of the monobodies or crystal packing interactions remains an open point, and the MD simulation is of great value to follow up the crystallography studies. The hypothesis was tested by comparing corresponding helical tilt at the CCD of different macrostates identified in section 'Conformational differences in the pY+3 binding pocket correlate to the decreased binding affinity from wild type to D170A variant for the pY and pY+3 pockets'. Differences in global helical tilt for each of the macro-states of the SH2 domain is observed in α3 (Fig 3F). Although a subtle rotation, these are consistent with the helical movements found in the CCD due to MS3–6 binding that impact biological function. Further, given that α3 helix interfaces with most of the other domains through its C-terminal helical turn, it lies in a central position that may be able to transmit slight movements within the helical arm through subtle motions of interacting residues between α3 and nearby domains. Thus, although the helical tilt is small, and may not be a primary signaling mechanism, it may reflect movements of key residues at the interface between α3 and nearby domains that propagate conformational information from the CCD to the SH2 domain.

Conserved Cα pair distances. The rigidity transmission, where a perturbation of rigidity at one binding site can be allosterically transmitted to a second distant site, was observed in other proteins [25, 26]. Thus we further hypothesized that the motion of α3 is transmitted allosterically to SH2 via a "rigid core", that is, an interlocking sequence of conserved interactions which function as a sort of molecular machine. The existence of such an interaction network is demonstrated in Fig 3B, where the inter-domain pair Cα distances (See Methods) are plotted and colored according to standard deviation values computed across all trajectories. There is a rigid backbone through the protein from CCD, LD, DBD, and finally to SH2 (Fig 3B) which is...
highly conserved during dynamical motion of the protein before and after mutation at D170. α3, α20, α21 compose the first section of the rigid core between CCD, DBD, and LD (Fig 3C), which could convey the dynamics of CCD into this highly rigid region. Upon close inspection, the three helices α3, α20, and α21 are locked via hydrogen bond network between I252, Q511 and W474 (Fig 3A and 3C), such that any rotation of these residues results in a corresponding reorientation of the helices to preserve the hydrogen bond network. We additionally find strongly conserved inter-domain interactions between the DBD and LD (Fig 3D), as well as between the LD and SH2 (Fig 3E), which complete the rigid core. Upon PCA analysis of residues of the rigid core, we see a minimal variance within the rigid core (S4(A) and S4(B) Fig). The PCA 2D plot shows two distinct conformational states, however, this distinction is primarily attributed to the pair C distances between residue 562 and residues from α3 helix (S4(C) Fig). Exclusion of the residue 562 from the PCA analysis of the rigid core shows an indistinguishable conformation landscape (S5(A) and S5(B) Fig). These factors allow subtle changes in the CCD configurations to convey movement from α3 through α21 (in LD) and α20 (in DBD) to α26 and α24 (in LD), and finally leading to allosteric modification of SH2 via β29.

The conformation analysis presented above demonstrates that D170A mutation leads to changes in the orientation of the CCD α helices, which could then lead to allosteric regulation of SH2 domain conformations though a network of conserved interactions. Rigid body analysis shows that these interactions consist of hydrogen bond, π-π, and hydrophobic networks that strongly correlate motions of different domains. However, this analysis does not highlight the allosteric path that differs between the wild type and D170A variant nor provide evidence of a dynamical correlation between CCD and SH2 conformation through this pathway. To further elucidate the allostery pathway and show dynamical correlation, we employ both a REDAN analysis and analysis of differences in the global hydrogen bond network between kinetic macro-states.

REDAN. While a rigid backbone provides a potential pathway for signal propagation, the cumulative long-range allosteric effect is realized through short range interactions and subtle allosteric changes which must occur in concert. To identify a detailed sequence of short range interactions, REDAN analysis was employed as a means to identify residue pairs that are responsive to allosteric perturbation, followed by shortest path analysis using Dijkstra’s algorithm. Using REDAN, subtle yet highly correlated differences in the allosteric network between D170A and the wild type can be resolved allowing us to propose a concrete signal transduction network from CCD to SH2.

From the conformational analysis, Y657 was identified as the residue with highest average difference between different macro-states explored by the SH2 domain. Thus, D/A170 was selected as the starting point and Y657 was selected as the end point for the analysis. The most structurally-relevant pathway from effector residue (D/A170) to regulatory site (Y657) was identified by REDAN and is shown in Fig 4A and 4B. The pathway originates from the CCD, through the LD and to the SH2 domain, bypassing DBD (although it passes nearby α20 which was identified as a component of the rigid core). The residue pairs that connect domains are of most interest, and their distance distributions are shown in S7(C)–S7(E) Fig. The average structures of each macro-state were calculated to structurally verify correlated motions in secondary structure identified by the REDAN (Fig 4C).

The secondary structure designated β22 (Fig 1B) can be observed to dynamically shift between a β sheet and α helix in macro-states 2, 3, and 4, while in macro-state 0 and 1 the α helix is stabilized (S7(F) Fig). β22 extends the α21 helix and reorganizes the loop between β22-α23. It is worth noting that residues 514 to 517 are annotated as a β sheet in the UniProt database (S1 Table), while these residues form an α helix in the reference structure used in this study (PDB ID: 6TLC). This result is not necessarily incongruous with the database
designation as multiple configurations of the secondary structure are observed experimentally. Stabilization of \( \beta_{22} \) as an extension of the \( \alpha_{21} \) helix in macro-states 0 and 1 stabilizes extension and a shift of the \( \alpha_{26} \) helix. The REDAN analysis suggests that further interaction between \( \alpha_{26} \) and \( \alpha_{32} \) then positions the \( \alpha_{32}-\alpha_{33} \) loop such that \( \alpha_{33} \) adopts an extended conformation without close contact to the \( \alpha_{32}-\alpha_{33} \) loop. On the other hand, the structural conformation analysis above suggests that \( \alpha_{26} \) may more indirectly affect \( \alpha_{32} \) via hydrogen bonding and hydrophobic interaction with \( \beta_{29} \).

**Hydrogen bond network.** The formation and breaking of hydrogen bonds plays an important role in stability of secondary structures and conformational variability of tertiary structures of a protein. To complement structural insights and REDAN analysis, the differential rate (preponderance) of hydrogen bond occurrence between different macro-states was compared (Fig 5). Hydrogen bonds were identified using Baker-Hubbard hydrogen bonding analysis, and then the differential hydrogen bond rate was calculated between macro-states 0, 1, 2, and 3, and the wild type-dominant macro-state 4. Macro-state 4 was considered as the primary active state as the wild type and D170A share high occupancy rates at this state. Note that Baker-Hubbard hydrogen bonding analysis algorithm classifies salt-bridge formation between amine and carboxylic acid as a hydrogen bond as well, thus the formation and breaking of salt bridges were also considered in the analysis.

Fig 4. Proposed allosteric path from D/A170 to Y657 obtained from REDAN analysis and its structural details. (A) Residues involved in the allosteric path from CCD to LD and to SH2 domain. The raw path can be found in (S7 Fig). (B) A close-up view of signal transduction from CCD domain to SH2 domain. The residues identified by REDAN show a path through \( \alpha_{3} \) in CCD to \( \beta_{22} \) and \( \alpha_{26} \) in the LD domain and finally to \( \alpha_{32} \) and \( \alpha_{33} \) in the SH2 domain, bypassing DBD. (C) Average structures from all the macro-states show significant reorganization in this interface (colors as in Fig 2H). \( \beta \)-sheet configuration (adjusted in PyMol) of macro-states 2, 3, and 4 are shown to highlight rearrangement of \( \beta_{22} \) based on Ramachandran Dihedral for \( \beta \)-sheets of residues 514 to 517.

https://doi.org/10.1371/journal.pcbi.1010794.g004
Consistent with the observation from REDAN analysis, $\beta_{22}$ has large changes in the hydrogen bond between the native (macro-state 4) and allosteric (macro-states 0 and 1) configurations. In macro-states 0 and 1, hydrogen bonds formation in $\beta_{22}$ stabilizes the alpha helix, while the other macro-states alternate between an $\alpha$ helix and a $\beta$ sheet. Moreover, in macro-state 3, rearrangement causes new a hydrogen bond to form between $\beta_{22}$ and $\alpha_{3}$. These changes highlight key differences between the wild type and D170A variant as only the D170A variant occupies macro-states 0, 1 and 3 (Fig 5). These macro-states also show decoupling of LD to DBD along with alteration of interactions between LD and SH2 (Fig 6). First, a salt bridge between D566 (LD) and R335 (DBD) is lost in macro-states 0 and 1 (Fig 6A and 6B). The loss of this interaction allows for the shift of $\alpha_{26}$ observed in the macro-states (Fig 4C). Second, a new hydrogen bond is formed between I576 (LD, $\alpha_{26}$–$\alpha_{27}$ loop) and N646 (SH2, $\alpha_{32}$) in macro-state 0 (Fig 6A) and a salt bridge is formed between D570 (LD, $\alpha_{26}$) and K642 (SH2, $\alpha_{32}$) in macro-state 3 (Fig 6C). Third, loss of hydrogen bonds of E652 and/or I653 (LD, $\alpha_{33}$) with S649 (LD, $\alpha_{32}$–$\alpha_{33}$ loop) in macro-states 0, 1, and 3 (Fig 6A–6C) contributes to increased flexibility of the $\alpha_{33}$ helix, also as observed in Fig 4C.

Focusing on the upper part of CCD (Fig 6D–6F), it is clear that $\alpha_{1}$ interacts directly with $\alpha_{2}$ and $\alpha_{4}$ through a large interaction surface. $\alpha_{2}$ and $\alpha_{3}$ form a contiguous helix interrupted by a kink in the helix at residue 278, thus perturbation of $\alpha_{2}$ is structurally coupled with $\alpha_{3}$. A salt bridge between E229 ($\alpha_{2}$) and R306 ($\alpha_{5}$) has a high appearance rate in macro-states 0, 1, and 3.
This salt bridge provides a means to strongly correlate the motion of α2 to α5 and mitigate perturbations by α1. These interactions perturb α3 which are translated to the rigid core interface. The residues of α3 domain can thus be used to influence the rigid core to elicit interdomain response (Fig 3C). Q511 also forms a strong hydrogen bond with W474 (DBD). This network of hydrogen bonds allows a strong correlated motion between α3 in CCD and α20/α21 in the LD domain. I252 is also adjacent to E253 which was identified by REDAN analysis as part of the allosteric pathway thus providing a chemical basis for signal transduction from CCD to SH2 via the LD domain.
Discussion
Significant differences in conformational space of the SH2 domain binding pocket (pY+3) between the wild type and D170A mutant were observed. The D170A variant explores and extends conformational space of SH2 domain, specifically with significant changes in opening of the pY+3 pocket. The six independent trajectories obtained for each variant explore fairly distinct areas of conformational space (i.e. different macro-states as identified in the clustering analysis, see Fig 2H). The long transition timescale between these states validates our kinetic clustering model, but also necessitates a “global” view of the trajectories in order to explain the full conformational dynamics of the protein. Based on the coverage of the 2D PCA space of the SH2 domain (Fig 2), we find that the six trajectories obtained are sufficient to sample the dynamics relevant to ligand binding affinity. On the other hand, the long transition timescale necessarily leads to high variability between trajectories for certain computed properties such as RMSF (S8 Fig). While not an indication of insufficiency in the global conformational view obtained from combining all trajectories, this variability instead serves as a valuable additional measure of conformational differences between kinetic macro-states. For transparency, the plots of replicates analysis are available in supplement(S2 and S6 Figs). Since the present MD simulation cannot guarantee sampling of the full conformational space of the protein, other possible allosteric pathways cannot be excluded. The mechanism proposed here is drawn with caution and the results agree with experimental observations. Thus, this work presents a theoretical foundation upon which to draw inspiration for drug design as well as mechanistic studies on the protein.

The hydrophobic environment formed by I659, W623 and F621 in pY+3 were previously shown to assist in binding of target peptide. Changes in the hydrophobic environment by increasing either hydrophobicity or aromaticity leads to hyper-activation, while introduction of polarity and reduction of hydrophobicity and aromaticity diminishes STAT3 function [27]. Furthermore, studies suggest that the side-chain Y657 interaction is important for stabilizing the ligand-protein complex [15–17]. Thus, the diminished binding affinity can be attributed to increased motions in the structures surrounding the pY+3 pocket in D170A. While in principle, translation of motion through the rigid core could affect the conformation of the primary pY binding pocket, only the pY+3 pocket shows differential motions in D170A compared to the wild type. Thus, our simulations support the conclusion that mutation of the D170 residue affects an inactivation of the protein via an allosteric mechanism resulting in conformational changes primarily in the specificity determining pY+3 pocket.

The observed differences in the SH2 conformational space allow us to further characterize the mechanism of this allosteric effect. The α3 helix of the CCD domain correlates functionally with SH2 conformations, as evidenced by strong correlated motions within the rigid core. Given its positioning, it likely communicates changes in the CCD to the LD. Using rigid body analysis, a potential pathway from CCD through LD, and finally to SH2 was identified, wherein conserved hydrogen bond networks and other strong interactions firmly link a series of secondary structures (primarily α3, α20, α21, and α24). While further analysis supports this proposed mechanism (vide infra), controlled mutagenesis of these key residues could also provide experimental evidence for the importance of these interactions to the D170A allosteric pathway. The lack of a significant increase (or decrease) in dynamic motion and overall flexibility in D170A compared to the wild type (as seen in the RMSF analysis) also supports a sequence of interactions between rigid bodies as the main allosteric mechanism.

The specific allosteric pathway was further elucidated via REDAN and differential hydrogen bonding analysis. These analyses both point to a very specific mechanism (structural details shown in S9 Fig): 1) stronger interaction between α5 and α2 causes a tilt in the α2/α3 helix, 2)
α3 tilt interfere the interaction between β22/α23 and α26/α27. In macro-state 0 and 1, the interference breaks the salt bridge between D566 and R335 and caused α26 shift away β22. While in macro-state 3, α26 shift toward β22, which cause a steric clash between Lys 573 and β22. 3) in turn, the movement of α26 and α26/α27 causes breakage of the hydrogen bonds between α33 and the α32-α33 loop, 4) α33 extends significantly and alters the conformation of the pY+3 pocket. We also identified conserved interactions between α26 and β29 in SH2 which may provide further coupling.

This study independently confirms that alterations of the alpha-helical rigid core impacts the SH2 domain in the absence of crystal packing interactions. This supporting prior crystallographic observations indicating that small molecules, mutations, or monobodies can drive long-range allosteric changes to the SH2 domain through subtle alteration of rigid core. We have specifically avoided an elucidation of the allosteric pathway within the CCD domain (i.e. in the immediate vicinity of the mutation site). Although fully exploring comparisons to the allosteric pathways of other effectors (e.g. MS3–6, K116, MM-206) is beyond the scope of the present work, it seems highly likely that the observed structure of the interdomain interactions and the rigid core mechanisms should result in highly similar allosteric mechanisms within the LD, DBD, and SH2 domains. Conversely, the diverse nature of these effectors likely rules out any significant commonality in the initial few steps of the allosteric pathway. Additionally, we have observed that even fairly significant alterations of the CCD structure, such as the kinked conformation of α1 explored by the D170A variant, do not correlate with changes to SH2 structure except via the rigid core. Thus, we consider this feature of the CCD structure as the “trigger” for the overall allosteric pathway.

Previously, the role of α26 in allosteric communication has been identified: nuclear magnetic resonance (NMR) studies showed that mutation of I568F was able to induce a chemical shift perturbation in SH2, DBD, and CCD [28]. Besides, D566A, D570A and D570K mutants showed profound negative effects on transcription, and also unexpectedly tyrosine phosphorylation even before interleukin (IL) 6 induction [29, 30]. Additionally, previous study shows similar allostery pathways in STAT5, however rigid core was not explored [31]. STAT family of proteins are highly similar in primary, secondary and tertiary structures and the similarity in allosteric pathways of STAT3 and STAT5 can be used to posit that STAT family of protein have a rigid core that couples allostery between their domains.

These results present us with novel ways of regulating the CCD domain whereby ligand or peptide interaction with CCD can significantly alter the helical tilt of α3 which is transmitted to the SH2 domain via the identified allosteric pathway. It is unlikely that different effectors will have identical mechanisms within CCD, as the local perturbation caused by point mutation, small molecule binding, peptide binding, etc. is radically different. However, our analysis supports the conclusion that any perturbation resulting in a change of α3 tilt should result in a similar outcome due to the strong and highly concerted motions along the proposed allosteric pathway. Potentially, alteration of helical tilt could also result in tighter binding of target peptides to the SH2 domain while, as observed here, helix-helix interactions in CCD can also promote structural changes in SH2 domain leading to reduced affinity of SH2 binding. In short, the mechanism for signal transduction has been identified through analysis of dynamic correlated motions between CCD and SH2 domain, the exact outcome of this allosteric effect is not clear. Distinguishing between two potential mechanisms: changes in substrate specificity to form heterodimer required for phosphorylation of Y705 and reduction in homodimerization due to reduced affinity towards pSTAT3, needs to be investigated. Drugs designed to specifically alter the motions of the CCD helices would yield valuable insight towards validation of proposed mechanism. Since the pY+3 pocket regulates the affinity of peptide binding, rather than being the catalytic active site, assays developed to probe allosteric regulation of SH2 via
CCD should consider the identity of peptides used to target the SH2 domain in addition to overall activity.

Finally, the methods used here to identify the rigid core mechanism, specifically a combination of structural and conformation analysis (dimensionality reduction, functional clustering, and conserved $\text{Ca}$ pair distances) with dynamical and correlative analyses (REDAN and differential hydrogen bond analysis) should allow for identification of other potential effectors targeting CCD, and more widely, in identifying similar allosteric pathways in a number of (semi-)rigid proteins.

**Methods**

**Initial structure**

The monomer STAT3 with peptide MS3–6 complex (PDB ID: 6TLC) structure was used as the template for this study. In chain A, residue 372–381 and 418–428 in DBD domain were missing, here these residues were modeled using Chimera [32]. The apo structure was created by directly deleting the MS3–6 peptide, and then the apo type was subjected to mutation using the PyMol Mutagenesis Wizard [33] to generate the D170A variant. To show DNA binding (Fig 1), the 1BG1 structure was used which contains a single STAT3 bound to DNA. However, STAT3 binds to DNA in its dimeric state; the dimeric structure was generated in PyMol to show DNA binding of STAT3 dimer. Hydrogen atoms were added to the crystal structures using PyMol. The protonation states for the histidine residues were assigned using the H++ program [34].

**Molecular dynamics simulation**

For each system, a rectangular periodic water box with 84736 TIP3P waters was used with a minimum distance of 10 Å between the box boundary and the protein to avoid image interactions(S12 Fig). To balance charge and provide realistic salinity, 0.15M sodium and chloride ions were added. NAMD 2.13 [35] with the CHARMM 36 force field [36] was used for energy minimization and molecular dynamics (MD) simulations. Initially, the simulation systems were subjected to 5000 steps of energy minimization to remove bad contacts and clashes. Then, systems were heated from 0 K to 300 K, heating 50 K every 200 ps, and then from 300 K to 310 K in 200 ps, with 10 ns isothermal-isobaric ensemble (NPT) short equilibration. Subsequently, six replicas of 600 ns canonical ensemble (NVT) MD simulations at 310 K were conducted. The first 100 ns simulations were discarded as equilibration and the following 500 ns for each replica, 3μs in total, was used for further analysis. The SHAKE algorithm was applied to all bonds containing hydrogen atoms. The electrostatic interaction was evaluated by the particle-mesh Ewald method, and Lennard-Jones interactions were evaluated using 10 Å as a cutoff. The NPT simulations were performed using a Nosé-Hoover Langevin piston pressure. The NVT simulations were performed using the Langevin integrator. For the integrator, a friction coefficient of 1 ps$^{-1}$ was implemented. A step size of 2 fs was used.

NVT was used for the main simulation due to its propensity for greater stability over longer timescales. To reconcile the change in system from NPT to NVT, total energy, kinetic energy, potential energy, temperature and pressure of the system were examined for consistency and smoothness. Both systems were well equilibrated (S10 Fig). Sufficient sampling of the MD simulation for both wild type and D170A variants was evaluated by the pair Root Mean Square Root (RMSD) (S11 Fig), which shows that the simulations have either reached a stationary shape or there is transition between different stationary shapes. Detailed RMSD and RMSF analysis can be found in the Supplemental Information (S1 Text).
Feature characterization

Various feature characterization relevant to trajectory analysis were conducted using two open-source packages: MDTraj 1.9.3.40 [37] and MDAnalysis [38, 39].

**Pair residues center of mass distance.** Pair residues center of mass distance was used to characterize the SH2 domain PY pocket and PY+3 pocket. First SH2 domain was extracted using `atom_slice` function from MDTraj; then the residue center of mass was calculated using the `center_of_mass` function from MDAnalysis, and finally the Euclidean distance of pair residues center of mass was calculated.

**Inter-domain pair Cα distances.** The inter-domain pair Cα distances were used to carry out the rigid core analysis. Take CCD-DBD domain pair Cα distances for example, first, the neighbors Cα atoms from DBD that are within 1 nm of CCD were found using compute_neighbors function from MDTraj, from which we can find the pair residues between CCD and DBD that are within interaction ranges. then the common pair residues between wild type and D170A variant were kept, and the pair ca distance were calculated using `compute_distances` from MDTraj.

**Alpha helix global tilt.** The α helix global tilt angle was used to characterize the geometry of helices according to the procedure of Sugeta and Miyazawa [40]. After alignment the trajectories to the crystal structure, the `helanal.helanal_trajectory` function from MDAnalysis software was used to characterize the alpha helices. [0, 0, 1] was used as the reference axis [41].

RMSD analysis

Root-mean-square deviation (RMSD) analysis shows the conformational dynamics over the trajectory and provides an insight into the variation within the conformational space of a reference structure. After alignment to the reference structure, the RMSD values were measured using MDAnalysis.rms function.

RMSF analysis

Root-mean-square fluctuation (RMSF) allows us to probe average positional changes of each residue. RMSF measures the average deviation of a particle (and individual residue) over time from a reference position (typically the time-averaged position of the particle). The trajectories were first superposed to the backbone of first frame of each trajectory, then the RMSF values for each trajectory were measured using `mdtraj.rmsf` using MDTraj. Finally, the mean RMSF values and the standard error of each residue over the six replicas were calculated and plotted.

Principal components analysis

Linear Principal Components Analysis (PCA [42]) is used to transform high dimensional and often linearly-dependent data points into a low-dimensional space spanned by uncorrelated principle components. The first two principal components (PCs) were used in this work, yielding a two-dimensional reduction of the original data. Given high dimensional data represented in n (sample size) by m (variable size) matrix, the covariance of any two variables X and Y was calculated by,

\[
\text{cov}(X_i, X_j) = \frac{N}{N-1} \langle (X_i - \langle X_i \rangle) \cdot (X_j - \langle X_j \rangle) \rangle,
\]

where \( X_{ai} \) is the \( \alpha \) Cartesian component of the coordinate vector for atom \( i \). The covariance
matrix $C$ is constructed as the pairwise covariance between all variables. The eigenvectors of $C$ are the components of PCA, while the eigenvalues measure the contribution of each PC in the dataset. The eigenvectors also provide a mapping from the high-dimensional dataset to the low-dimensional PC space: the (PC1,PC2) coordinates for each input frame are given by multiplication of the original data set by the first two important eigenvectors. For a given PC, the importance of a feature (variable) is reflected by the absolute magnitude of the corresponding entry in the eigenvector. The PCA analysis was performed by Scikit-learn [43] implemented in Python.

Markov state modeling and Perron cluster cluster analysis

Markov state models (MSMs [44, 45]) have shown great utility in modeling the transitions among functional states. It was used in this study to cluster conformations into kinetically meaningful macro-states. The conformational space was first discretized into $n$ micro-states. Here, agglomerative clustering [46] was applied to divide the sampled conformations into 300 micro-states in the two-dimensional PCA coordinate system. Then, $C_{ij}(\tau)$, the number of observed transitions from micro-state $i$ to micro-state $j$ at a lag time $\tau$ is calculated. Then the transition probability, $P_{ij}$, from micro-state $i$ to micro-state $j$ can be estimated as $P_{ij} \approx (C_{ij} + C_{ji})/\sum_k(C_{ik} + C_{ki})$. According to estimated relaxation timescale (S1(B) Fig), the count matrix and MSM transition probabilities converge beyond $\tau = 5$ ns, which was chosen as the lag time.

Finally, Perron Cluster Cluster analysis (PCCA), implemented in the PyEmma package [47], was used to coarse grain micro-states to "kinetically relevant" macro-states based on the well-sampled micro-state transition matrix. Structures that interconvert frequently were assumed to belong to the same functional metastable state (macro-state)[48]. Five macro-states were determined based on the band gap in the estimated relaxation timescale plot.

Relative entropy-based dynamical allosteric network

The relative entropy-based dynamical allosteric network (REDAN) model [49] was used to quantitatively characterize protein allosteric effects upon mutation. The difference between the distributions of the pair alpha carbon ($C_{\alpha}$) distances of two residues upon perturbation is quantified by the perturbation relative entropy (PRE), which is the average relative entropy,

$$\text{PRE}(P||Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2},$$

$$D_{KL}(P||Q) = \int_0^\infty p(x)\ln \frac{p(x)}{q(x)} dx,$$

where $p(x)$ is the distribution density for system $P$ (before perturbation), and $q(x)$ is the distribution density for system $Q$ (after perturbation). High perturbation relative entropy values indicate that substantial allosteric effects are implied by the significantly different distance distribution of the residue pair.

Then, a weighted graph can be built based on the PRE matrix. Each node is represented by a $C_{\alpha}$ atom, and two nodes will be connected by an edge if the longest possible distance between them is less than 10 Å. Each edge is weighted as $1/\text{PRE}$. Since high PRE values indicate importance in the propagation of the structural changes in the protein, the pathway with the smallest overall weight implies the most structurally relevant route and hence a possible allosteric pathway. Dijkstra’s algorithm was used to identify the shortest pathway.
Supporting information

S1 Text. RMSD and RMSF analysis. D170A mutation induces large structural but minor dynamical changes in STAT3. Table A. Statistical differences for the α3 global tilt among different macrostates (Fig 3F). Kolmogorov-Smirnov test and T-test were done using scipy.stats. ks_2samp and scipy.stats.ttest_ind function respectively.

Table A. Statistical differences for the α3 global tilt among different macrostates (Fig 3F). Kolmogorov-Smirnov test and T-test were done using scipy.stats. ks_2samp and scipy.stats.ttest_ind function respectively.

S1 Table. Secondary structure assigned by the UniProt database by compiling structure information from multiple x-ray crystal structures.

S1 Fig. Additional information of PCA and MSM analysis. (A) PCA scree plot: dot shows the cumulative explained variance of the principal components; the bar chart represents the explained values per component. (B) Relaxation timescales of MSM for SH2 domain conformational space at different lag times. (C) The first ten features that contribute the first two PC the most. The absolute value of PC1 and PC2, and square root of sum of squared PC1 and PC2 values are shown here.

S2 Fig. Pair Cα distance distribution between Q644 and Y657, Q644 and E638 of 6 replicas for wild type and D170A variant.

S3 Fig. Characterization of CCD using pair Cα distances. (A) PCA 2D plane of CCD pair Cα distances colored by macro-state from SH2 domain results. (B) Represent structure of CCD corresponding to Figure A. (C) The coefficients of first 100 features that contribute the first two PC the most.

S4 Fig. The PCA analysis of pair distances of rigid core (residues: 240–252, 474, 479, 511, 546, 549, 550, 562, 564, 568, 610, 611). (A) PCA 2D plot colored by different macrostates; (B) PCA 2D plot colored by systems; (C) The first ten features that contribute the first two PC the most. The absolute value of PC1 and PC2, and square root of sum of squared PC1 and PC2 values are shown here.

S5 Fig. The PCA analysis of pair distances of rigid core without residue 562 (residues: 240–252, 474, 479, 511, 546, 549, 550, 564, 568, 610, 611). (A) PCA 2D plot colored by different macrostates; (B) PCA 2D plot colored by systems; (C) The first ten features that contribute the first two PC the most. The absolute value of PC1 and PC2, and square root of sum of squared PC1 and PC2 values are shown here.

S6 Fig. CCD α3 global tilt angle distribution. (A,B) CCD α3 global tilt angle distribution of different macrostates, plotted separately for the wild type and D170A variant. Average helix tilt angle within each macro-state is illustrated by a vertical dashed line. (C,D) CCD α3 global tilt angle distribution of different replicas for wild type and D170A variant.

S7 Fig. Additional REDAN analysis results. (A) Proposed pathway from 170 to 640 shown in the protein structure; (B) Proposed pathway from 170 to 644; (C,D,E) Key pair residue
distance; (F) Ramachandran Dihedral for residue 513, 514, 515 and 517; (G) Summary of proposed pathways from source residue 170 to target residue 640, 644 and 657.

**S8 Fig.** The mean RMSF of six replicates was plotted, with the RMSF values for each residue separated by domain. The wild type is plotted in blue and the D170A variant in orange, with error bar indicating the stand error among the six replicates. Structures of each domain colored by wild type RMSF values are shown (low RMSF values in white to high RMSF values in red).

**S9 Fig.** Proposed specific mechanism. (A) Representative structure macro-state 0 (light cyan) compare with macro-state 4 (salmon); (B) Representative structure macro-state 3 (green) compare with macro-state 4 (salmon) (C-G) key residue pair CA distance or contact distance (closest heavy atom distance). ILE-252 to LYS-573 distance distribution and PHE-512 to LYS-573 distance distribution show \(\alpha_26\) shift away \(\beta_{22}\) in macro-state 0 and 1, while toward in macro-state 3; PHE-512 to LEU577 and PHE-512 to SER-649 distance distribution show \(\alpha_{26}/\alpha_{27}\) and \(\alpha_{32}/\alpha_{33}\) loops move away \(\beta_{22}\) in macro-state 0, 1 and 3, while in macro-state 3, \(\alpha_{32}\) moves close to \(\alpha_{26}\) as indicated by ASP-570 to LYS-642 distance distribution.

**S10 Fig.** System parameters (pressure, temperature) and system energy (potential energy, kinetic energy and total energy) of wild system and D170A variant system, each system has six replicas: cp1, cp2, cp3, cp4, cp5 and cp6.

**S11 Fig.** First row: Root Mean Square Deviation (RMSD) of wild system and D170A variant system, each system has six replicas: cp1, cp2, cp3, cp4, cp5 and cp6. Note: rolling average of every 100 frames was plotted here for better visualization. Second row: The pair RMSD (frames every 1 ns were extracted and used for pair RMSD calculation) of both systems for each replica.

**S12 Fig.** Figure depicting the simulation system, generated by VMD. Water is shown as lines, ions are shown as vdw, protein is shown as cartoon.

**S13 Fig.** Root Mean Square Deviation (RMSD) analysis. (A, B) Cross-correlation (Pearson correlation) of RMSD values of each domain for wild type and D170A variant. Here RMSD values were calculated with the first frame of reference since the correlation of dynamic changes of each domain is of interested. The cross-correlation was done by merging the rmsd of 6 copies trajectory together, and then the Pearson correlation among different domains were calculated. While worth being noted, the correlation value does not suggest the functional correlation between domains, since RMSD is an overall measurement of conformational changes with regarding to the reference structure, distinctive conformations may have same RMSD value. (C) Violin plot of RMSD values for the whole protein (core full length protein, not including NTD) and each domain. Here crystal structure was using as reference since the conformational changes difference between wild type and D170A was of interested. (D) The RMSD distribution of SH2 domain in the six replicas for D170A variant.
Acknowledgments
TZ was supported by a fellowship from the Department of Chemistry of Southern Methodist University. All calculations were performed on the ManeFrame II supercomputing system at Southern Methodist University.

Author Contributions
Conceptualization: Tingting Zhao, Nischal Karki, Brian D. Zoltowski, Devin A. Matthews.
Data curation: Tingting Zhao.
Formal analysis: Tingting Zhao.
Funding acquisition: Devin A. Matthews.
Investigation: Tingting Zhao.
Methodology: Tingting Zhao, Nischal Karki, Brian D. Zoltowski, Devin A. Matthews.
Project administration: Devin A. Matthews.
Supervision: Brian D. Zoltowski, Devin A. Matthews.
Visualization: Tingting Zhao.
Writing – original draft: Tingting Zhao.
Writing – review & editing: Tingting Zhao, Nischal Karki, Brian D. Zoltowski, Devin A. Matthews.

References


BIBLIOGRAPHY


[16] P.-O. Löwdin, Quantum theory of many-particle systems. iii. extension of the hartree-fock scheme to include degenerate systems and correlation effects, Physical Review 97 (Mar., 1955) 1509–1520. 4


[23] Explicitly Correlated Wave Functions in Chemistry and Physics. Springer Netherlands, 2003, 10.1007/978-94-017-0313-0. 5


[34] M. J. Frisch, R. Krishnan and J. A. Pople, A systematic study of the effect of triple substitutions on the electron correlation energy of small molecules, Chemical Physics Letters 75 (Oct., 1980) 66–68. 6


[67] F. E. Harris, H. K. Monkhorst and D. L. Freeman, *Algebraic and diagrammatic methods in many-fermion theory*, . 15


[74] W. Kutzelnigg, *Quantum chemistry in fock space. i. the universal wave and energy operators*, *The Journal of Chemical Physics* **77** (Sept., 1982) 3081–3097. 22


[83] D. A. Mazziotti, *Two-electron reduced density matrix as the basic variable in many-electron quantum chemistry and physics*, *Chemical Reviews* 112 (Aug., 2011) 244–262. 30


206


207


208


[137] V. Barone, R. Improta and N. Rega, *Quantum mechanical computations and spectroscopy: From small rigid molecules in the gas phase to large flexible molecules in solution*, *Accounts of Chemical Research* **41** (Feb., 2008) 605–616. 60


210
[152] S. Mallick, B. Roy and P. Kumar, A comparison of DLPNO-CCSD(t) and CCSD(t) method for the determination of the energetics of hydrogen atom transfer reactions, *Computational and Theoretical Chemistry* **1187** (Oct., 2020) 112934. 61


212


[210] D. C. Moule and A. D. Walsh, Ultraviolet spectra and excited states of formaldehyde, *Chemical Reviews* **75** (Feb., 1975) 67–84. 84


[234] T. Zhao, N. Karki, B. D. Zoltowski and D. A. Matthews, Allosteric regulation in stat3 interdomains is mediated by a rigid core: Sh2 domain regulation by ccd in d170a variant, PLOS Computational Biology 18 (Dec., 2022) e1010794.


