Small Tensor Product Distributed Active Space (STP-DAS) Framework for Relativistic and Non-relativistic Multiconfiguration Calculations: Scaling from 10^9 on a Laptop to 10^{12} Determinants on a Supercomputer

Hang Hu,[†,‡] Shiv Upadhyay,[¶,‡] Lixin Lu,[¶] Andrew J. Jenkins,[¶] Tianyuan Zhang,[¶] Agam Shayit,[§] Stefan Knecht,[‖] and Xiaosong Li[*,¶]

†*Molecular Engineering and Sciences Institute, University of Washington, Seattle, Washington 98195*
‡*Authors contributed equally to this work*
¶*Department of Chemistry, University of Washington, Seattle, Washington 98195*
§*Department of Physics, University of Washington, Seattle, Washington 98195*
‖*Algorithmiq Ltd, Kanavakatu 3C, FI-00160 Helsinki, Finland*

E-mail: xsli@uw.edu

Abstract

Despite the power and flexibility of configuration interaction (CI) based methods in computational chemistry, their broader application is limited by an exponential increase in both computational and storage requirements, particularly due to the substantial memory needed for excitation lists that are crucial for scalable parallel computing. The objective of this work is to develop a new CI framework, namely the small tensor product distributed active space (STP-DAS) framework, aimed at drastically reducing memory demands for extensive CI calculations on individual workstations or laptops, while simultaneously enhancing scalability for extensive parallel computing. Moreover, the STP-DAS framework can support various CI-based techniques, such as complete active space (CAS), restricted active space (RAS), generalized active space (GAS), multireference CI (MRCI), and multireference perturbation theory (MRPT2), applicable to both relativistic (2-component and 4-component) and non-relativistic theories, thus extending the utility of CI methods in computational research. We conducted benchmark studies on a supercomputer to evaluate the storage needs, parallel scalability, and communication downtime using a realistic exact-two-component CASCI (X2C-CASCI) approach, covering a range of determinants from 10^9 (billion) to 10^{12} (trillion). Additionally, we performed extensive X2C-CASCI calculations on a single laptop and examined how the STP-DAS partitioning affects performance.

1 Introduction

Full configuration interaction (CI) provides the most accurate electronic structure description of a molecular system in a given basis.[1–6] However, its computational cost grows factorially with the system's size due to the combinatorial increase in the number of configurations, as it includes all possible excited states within the complete orbital space. Significant work has been invested in developing approximate CI techniques, as exemplified by the complete active space (CAS) approach[7,8] along with its numerous variants and enhancements. The CAS approach specifically focuses on configurations that arise from a limited active space, which consists of a selection of orbitals and electrons deemed chemically significant. While the CAS framework effectively addresses static correlation, it often fails to capture a substantial portion of the dynamic correlation arising from excitations beyond the active space. This limitation can be overcome by incorporating multireference configuration interaction (MRCI) to account for the missing dynamic correlation.[6,9–14] and multireference second-order perturbation theory (MRPT2, CASPT2).[15–22]

To minimize the number of configurations in the CI expansion further, implementing limitations on excitation operators is effective, resulting in the development of the restricted active space (RAS)[8,23,24] and occupation restricted multiple active space (ORMAS) methods.[25,26] A more generalized strategy involves dividing the total correlation space into smaller generalized active spaces (GASs).[8,27–33] The primary benefit of GAS lies in its ability to apply a broad framework of excitation constraints, providing a flexible framework for truncated CI types of electronic structure calculations.

Recent advancements in CI-based relativistic methods, which variationally incorporate both scalar relativity and spin-orbit couplings at the molecular orbital level, have spurred a new era in multi-reference electronic structure theory.[14,19,21,22,30,31,34–44] This progress is largely motivated by improvements in core-electron spectroscopies, research in magnetic and spintronic materials, as well as the chemistry of rare-earth and heavy elements. Employing two- and four-component complex-valued orbitals within Kramers' unrestricted framework[45,46] inherently results in increased computational requirements for CI-based relativistic methods, as manifested by a significantly larger configurational space, a less sparse CI vector, and a higher count of floating-point operations compared to their non-relativistic counterparts.

In computations using CI or its various approximations, constructing the Hamiltonian matrix demands an algorithm capable of efficiently navigating through all determinants, with constraints on excitation operators and the active space. This process is commonly referred to as the CI addressing scheme. Since the CI addressing scheme is the core engine of any high-performance CI calculation, considerable efforts have been directed toward developing highly vectorized CI algorithms optimized to take advantage of the capabilities of modern computational infrastructures. Handy's string-based addressing scheme[47] stands out as one of the most effective methodologies and is implemented in almost all high-performance CI programs. The string-based addressing scheme can efficiently produce a unique address for each configuration, leading to a precomputed list of non-zero excitations that facilitates a highly vectorized construction of the CI Hamiltonian. However, the storage of the excitation list from calculations involving large active spaces presents a significant computational challenge. For a CI problem with N configurations, the size of each CI vector grows linearly with respect to N while the size of the one-electron excitation list expands quadratically. For example, in a CAS calculation involving 42 Kramers' unrestricted orbitals and 24 electrons, the combinatorial factor $\binom{42}{24}$ implies that the storage requirement for each CI vector is 1 TB (using complex-valued double precision). However, to store the non-zero elements of the one-electron excitation list, an estimated 1.8 petabytes (PB) of memory would be needed even with bit-wise compression. The size of this list is proportional to $n_e \times (n_h + 1) \times N$, where n_e is the number of electrons and n_h is the number of holes (unoccupied orbitals) in the complete active space.

Simply put, the bottleneck for CI methods is the memory requirement arising from the storage of the excitation list, which practical CI-based methods must work to circumvent. The seminal works for direct CI calculations aimed to optimize memory storage and locality[47–50] sometimes at the cost of additional computation. In the modern era of heterogeneous computation, this bottleneck is particularly critical for performance on accelerators such as graphical processing units (GPUs) as the memory on accelerators is limited and data transfer with the accelerator might be constrained by the bandwidth and latency.[51,52] A straightforward solution is to distribute the excitation list among multiple computing nodes.[53] Indeed, several recent efforts have focused on developing a distributed CI framework, albeit at the cost of utilizing several hundreds of high-memory computing nodes.[30,31,54,55] Nonetheless, this approach does not alleviate the demand for large-scale CI applications that can be executed on conventional workstations or small computing clusters.

Minimizing the memory needs for extensive CI calculations on a single workstation or laptop, while simultaneously enhancing scalability for massively parallel computing, may appear as divergent goals. Yet, they form the central objective of this work, and we demonstrate that optimizations for both small-scale and large-scale resources can complement each other. Additionally, we aim to establish a versatile framework capable of supporting various CI methods such as CAS, RAS, GAS, MRCI, and MRPT2, applicable to both relativistic and non-relativistic theories.

In this work, we advance CI methods by introducing a novel small-tensor-product (STP) addressing scheme that dramatically reduces storage requirements while being fully compatible with distributed computing environments. This advancement leverages the proposed distributed active space (DAS) framework, which separates inter-space and intra-space excitations in the loop structure, as well as global and local phase factors in the symbolic matrix elements. Through a carefully designed three-tiered addressing structure, the STP-DAS framework eliminates the necessity of storing a global excitation list. Instead, it adopts a small-tensor-product algorithm using a localized addressing strategy and is tailored for efficient vectorization and parallel processing. The hallmark of this research is the STP-DAS framework's seamless support for large-scale CI calculations, applicable across both single-node systems with constrained memory resources and expansive, distributed supercomputing infrastructures.

2 A Brief Background on Configuration Interaction

In this section, the following notations are used, unless otherwise specified:

- i, j, k, l: occupied molecular orbitals (MOs)
- a, b, c, d: virtual MOs
- p, q, r, s: general MOs
- $I,\ J,\ K,\ L$: Slater determinants

This work primarily focuses on developing a small-tensor-product distributed active space (STP-DAS) framework that is capable of supporting CI-based methods. To ensure comprehensive coverage, we briefly introduce the complete active space configuration interaction (CASCI) in which the wave function, $|\Psi\rangle$, is represented as a CI expansion or a linear combination of Slater determinants, constructed from a selected group of active MOs and a selected group of active electrons. The STP-DAS framework can be applied to both relativistic and nonrelativistic CI problems, however in this work we are primarily focused on two-component relativistic CI so we will present the following equations in a spinor basis.

2.1 Complete Active Space Configuration Interaction

For a CAS wave function, the active electrons are distributed into the active orbitals in *all* combinations that preserve the symmetry of the system.

$$|\Psi\rangle_{\text{CI}} = \sum_{K=1}^{N_{\text{CAS}}} C_K |K\rangle \tag{1}$$
$$|K\rangle \in \{|0\rangle, |0_i^a\rangle, |0_{ij}^{ab}\rangle, ...\}$$

where N_{CAS} is the total number of determinants in the expansion. $|0\rangle$ is the reference configuration. $\{|0_i^a\rangle, |0_{ij}^{ab}\rangle, ...\}$ are the singly-excited and doubly-excited electron configurations generated from the reference determinant $|0\rangle$ by applying the excitation operators:

$$|0_i^a\rangle = \hat{E}_{ai}\,|0\rangle\,; \quad \left|0_{ij}^{ab}\right\rangle = \hat{e}_{aibj}\,|0\rangle\,; \tag{2}$$

where the excitation operators are defined in terms of creation and annihilation operators:

$$\hat{E}_{pq} = a_p^\dagger a_q; \quad \hat{e}_{pqrs} = a_p^\dagger a_r^\dagger a_s a_q \tag{3}$$

2.2 Iterative Solver

The total CI state energy is defined as

$$E_{\mathrm{CI}} = \langle\Psi_{\mathrm{CI}}|\,\hat{H}\,|\Psi_{\mathrm{CI}}\rangle + E^{\mathrm{FC}} \tag{4}$$

$$E^{\mathrm{FC}} = \sum_{i'} h_{i'i'} + \frac{1}{2}\sum_{i'j'}(g_{i'i'j'j'} - g_{i'j'j'i'}) \tag{5}$$

where $\{i', j'\}$ belong to the frozen core space and E^{FC}, computed using reference (HF or CASSCF) orbitals, is the energetic contribution from the frozen core space. Orbitals within the frozen core space are excluded from the CI expansion. h includes contributions from both electron-nuclear repulsion and electron kinetic energy.

The electronic Hamiltonian for the correlated space has the form,

$$\hat{H} = \sum_{pq} h_{pq}^{\mathrm{FC}}\hat{E}_{pq} + \frac{1}{2}\sum_{pqrs} g_{pqrs}\hat{e}_{pqrs} \tag{6}$$

$$h_{pq}^{\mathrm{FC}} = h_{pq} + \sum_{i'}(g_{pqi'i'} - g_{pi'i'q}) \tag{7}$$

where again i' belong to the frozen core space and $\{p, q, r, s\}$ are in the correlated space.

Solving the CI problem (Eq. (4)) in a large determinantal space usually requires the use of iterative diagonalization approaches, such as the Davidson algorithm.[56] The most computationally expensive step in the Davidson algorithm is the $\boldsymbol{\sigma}$ vector formation, *i.e.*, matrix-vector product,

$$\begin{aligned}\sigma_K &= \sum_{pq}\sum_{L} h_{pq}^{\mathrm{FC}}\,\langle K|\,\hat{E}_{pq}\,|L\rangle\, C_L \\ &\quad + \frac{1}{2}\sum_{pqrs}\sum_{L} g_{pqrs}\,\langle K|\,\hat{e}_{pqrs}\,|L\rangle\, C_L \\ &= \sum_{pq}\sum_{L} h'_{pq}\,\langle K|\,\hat{E}_{pq}\,|L\rangle\, C_L \\ &\quad + \frac{1}{2}\sum_{pqrs}\sum_{L,J} g_{pqrs}\,\langle K|\,\hat{E}_{pq}\,|J\rangle\,\langle J|\,\hat{E}_{rs}\,|L\rangle\, C_L\end{aligned} \tag{8}$$

where we used

$$\hat{e}_{pqrs} = \hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps} \tag{9}$$

and

$$h'_{pq} = h_{pq}^{\mathrm{FC}} - \frac{1}{2}\sum_{r} g_{prrq} \tag{10}$$

Since two-electron excitations cannot be stored in memory except for very small systems, we used the N-resolution method[47,49] in Eq. (8) to evaluate the two-electron contribution based on one-electron excitation lists.

3 Distributed Active Space Configuration Interaction

3.1 Statement of the Problem

As stated above the $\boldsymbol{\sigma}$ vector formation is the most expensive part of a CI calculation. Eq. (8) can be readily partitioned into one-electron and two-electron contributions,

$$\begin{aligned}\sigma_K &= {}^{1\mathrm{e}}\sigma_K + {}^{2\mathrm{e}}\sigma_K \\ {}^{1\mathrm{e}}\sigma_K &= \sum_{pq}\sum_{L} h'_{pq}\,\langle K|\,\hat{E}_{pq}\,|L\rangle\, C_L \\ {}^{2\mathrm{e}}\sigma_K &= \frac{1}{2}\sum_{pqrs}\sum_{L,J} g_{pqrs}\,\langle K|\,\hat{E}_{pq}\,|J\rangle\,\langle J|\,\hat{E}_{rs}\,|L\rangle\, C_L\end{aligned} \tag{11}$$

The bulk of the computational effort comes from the two-electron part of Eq. (11). As such the efficient construction of the two-electron contribution to the $\boldsymbol{\sigma}$ vector formation has been the focus of several studies over the years primarily for nonrelativistic Hamiltonians.[47–50] The emergent state of the art methods for large-scale CI problems from the decades of effort can be characterized as direct methods which generate matrix elements on-the-fly from one-electron excitation lists $\langle K| \hat{E}_{pq} |J\rangle$ with differing orders of contraction in Eq. (11). However, for extremely large CI calculations distributed over many nodes the storage of the excitation lists becomes unmanageable.

In this work, we aim to reformulate the structure of the loops in Eq. (11) by factorization into many small tensor products and considering only nonzero combinations – which is henceforth referred to as a tensor-looping algorithm – facilitated by a distributed active space partitioning scheme. The goal is to identify a partitioning of the orbital space which allows us to circumvent the storage of extremely large excitation lists by writing the two-electron contribution solely in terms of small tensor products local to the partitioned space. If such a partitioning is possible, one would be able to reuse excitation lists between spaces and additionally store much smaller excitation lists. Crucially, we seek a partitioning which is *exact* and does not introduce any approximations relative to the full space when all excitations between the partitioned space are allowed.

In the CI community, ORMAS[25,26] and GAS are common orbital partitioning schemes.[8,27–33] These partitioning schemes differ based on their occupation schemes. ORMAS bounds the minimum and maximum occupation for each space, and GAS places bounds on the accumulated electron occupation number for each successive orbital space. These methods have been successful when forming approximate full CI spaces, however neither partitioning scheme provides a computational advantage when exactly partitioning a full CI space, *i.e.*, when there is no limit on the interspace excitaitons.

As we will demonstrate in the following sections, the STP-DAS framework can provide a computational advantage even in the limit of full interspace excitations, by breaking the sigma build into small tensor products. The following sections are technical but lay out important foundations of the STP-DAS framework for CI-based calculations. Readers who are primarily interested in the final STP-DAS $\boldsymbol{\sigma}$ build expressions may skip directly to Section 3.7.

3.2 The Distributed Active Space Framework

In the DAS framework, the total correlation space is defined with a collection of active orbitals $\{\phi\}$ whose cardinality is $M = |\{\phi\}|$ (*i.e.*, the total number of orbitals), a total number of active electrons n_e, and the determinants $|K\rangle$ generated either by the CAS or MRCI method.

Analogous to the generalized active space (GAS) approach, the total correlation space can be partitioned into an arbitrary number of distributed active spaces (DASs). In the absence of excitation constraints between these spaces, DAS becomes equivalent to CAS. On the other hand, imposing limits on the number of electrons or holes within each DAS creates a situation reminiscent of RAS, ORMAS, and GAS.

In the CI $\boldsymbol{\sigma}$ build, the global address of each $|K\rangle$ is uniquely defined. Addressing each determinant in CAS can be efficiently done with a string-based method, leading to a highly vectorized algorithm. A similar string-based addressing scheme has been extended to RAS with excitation restrictions.[14] However, for large CAS calculations, it is not feasible to save the complete global address with the excitation list. An efficient addressing scheme is critical for the large-scale parallelization of CI codes, and reducing the memory demands for the excitation list is essential for their broad practical use. Our strategy to accomplish these objectives involves fully separating global and local addresses to enable small tensor products in the $\boldsymbol{\sigma}$ build.

Figure 1 illustrates the STP-DAS mapping scheme, using CAS as example. In this illustration, a CAS problem is mapped onto multiple DASs. In contrast to the RAS addressing scheme where excitation restrictions are enforced in the map that generates strings,[14] the first step in the STP-DAS framework involves differentiating between excitations within a space (intra-space) and those between different spaces (inter-space). This separation is achieved through a process known as categorical excitations, which creates various configuration categories by promoting electrons from one DAS to another, following specific excitation constraints. This process is shown schematically in Figure 1.

For CAS, there are no limitations on inter-space excitations. Imposing constraints on these excitations transforms the approach into RAS, ORMAS, GAS, or MRCI methods, thus rendering STP-DAS a versatile structure suitable for a broad spectrum of multiconfigurational techniques. Following the separation of inter-space and intra-space excitations, each DAS is treated as an intra-space complete active space. This approach incorporates every possible excitation, making it compatible with the efficient string-based addressing scheme for computational processing, allowing for small tensor products embedded in a tensor looping algorithm.

The STP-DAS framework does not result in decreased accuracy as the total number of excitations, and thus the total number of determinants, remains unchanged. However, the advantage of STP-DAS comes from the fact that it is only necessary to store strings for intra-space excitations used in small tensor products and tensor looping. This means that when two DASs share the same configuration, even if they belong to different categories or consist of different orbitals, they can utilize the same intra-space excitation list and small tensor products. For instance, in CAS calculations involving 40 spinor orbitals and 20 electrons, $\binom{40}{20}$, 637 TB of memory would be required to store the excitation list. By adopting the DAS approach illustrated in Figure 1, where the correlation space is segmented into four DASs each containing 10 orbitals, the memory requirement is dramatically reduced. In this case, one only needs to store intra- and inter-space excitation lists for $\binom{10}{1}$, $\binom{10}{2}$, $\cdots$, $\binom{10}{9}$ CAS strings, which collectively only require 200 MB of memory. Note that any type of space partition will reduce the total number of strings generated from the smaller combinatorics. However, we advocate for equal space partitioning to maximize the reduction and take advantage of the reuse of the excitation list (see Section 3.4 for a detailed discussion).

The discussion above highlights the efficiency of the STP-DAS framework in managing computational resources through the reuse of intra-space excitation lists in small tensor products. Nonetheless, the process of identifying and reutilizing these lists is a challenging task. The complexity arises because, within the excitation list, the symbolic matrix elements vary across different DASs, even when they utilize identical local addressing strings. This complexity presents a significant challenge and restricts the application of the RAS/GAS approach to merely imposing excitation constraints, rather than achieving a reduction in computational demands.

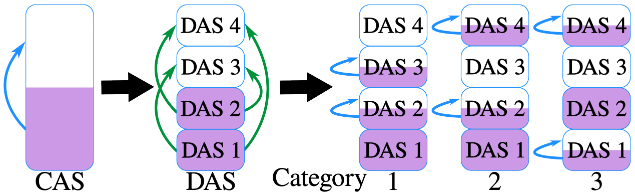

Figure 1. Space Partitioning in STP-DAS. Boxes symbolize active regions. Within these boxes, shaded portions indicate electron occupations, while white areas represent empty orbitals. A CAS problem is mapped onto multiple DASs. Distinct categories of configurations are produced by categorical excitations, depicted by green arrows. In every category, each DAS is treated as a complete active space, including all possible excitations, illustrated by blue arrows.

3.3 Space Partitioning and Organization of Excitations

In this section, the following notations are used, unless otherwise specified:

- $\mu, \nu, \kappa, \lambda$: DAS indices
- $\mathcal{A}, \mathcal{B}, \mathcal{C}$: configuration category
- $|K\rangle, |L\rangle$: determinants in the full space
- $|\mathbb{K}\rangle, |\mathbb{L}\rangle$: sub-determinants in DAS
- p, q, r, s: general MOs

Distributed Active Space

A DAS, $\mathbb{X}_\mu \equiv \{\{\phi\}_\mu, n_\mu, \mathbb{K}_\mu\}$, is defined by a set of orbitals $\{\phi\}_\mu$, an electron occupation number n_μ in the space, and sub-determinants $\mathbb{K}_\mu$. Each DAS is a complete active space where the total number of determinants is given by

$$N_\mu = \binom{M_\mu}{n_\mu} \tag{12}$$

where $M_\mu = |\{\phi\}_\mu|$ is the cardinality of the orbital set in $\mathbb{X}_\mu$, *i.e.*, the number of orbitals. Here, we assume the Kramers' unrestricted two-component or four-component condition where each orbital is a singly occupied spinor.

It's important to emphasize that within the STP-DAS framework, intra-space excitations are not subject to any constraints, as the excitation types typically associated with RAS or MRCI, known as inter-space excitations, are separated and utilized for the creation of configuration categories.

Configuration Category

Applying inter-space excitations, also referred to as categorical excitations, on the reference electron configurations leads to the formation of different electron configuration categories (see Figure 1 for an example). These categories will from now on be referred to as either configuration categories or simply as categories.

A configuration category $\mathcal{A} = \{\mathbb{X}_\mu^{\mathcal{A}}, \mathbb{X}_\nu^{\mathcal{A}}, ...\}$ is constructed from a set of DASs with unique electron occupation numbers. Note that the DAS orbital partitioning is unchanged. Therefore, it is important to associate each DAS with its parent category through the notation, $\mathbb{X}_\mu^{\mathcal{A}}$. Categorical excitation operators are defined as:

$$\begin{aligned}
\hat{\mathcal{E}}_{pq}^{\mathbb{X}_\mu \mathbb{X}_\nu} &= \left\{ \hat{E}_{pq} : p \in \mathbb{X}_\mu, q \in \mathbb{X}_\nu \right\} \\
\hat{\mathcal{E}}_{pqrs}^{\mathbb{X}_\mu \mathbb{X}_\nu \mathbb{X}_\lambda \mathbb{X}_\kappa} &= \{ \hat{e}_{pqrs} : p \in \mathbb{X}_\mu, q \in \mathbb{X}_\nu, r \in \mathbb{X}_\lambda, s \in \mathbb{X}_\kappa \} \\
&\cdots
\end{aligned}$$

Figure 1 shows three examples of different categories generated from categorical excitations.

A full-space determinant in category $\mathcal{A}$ can be constructed using sub-determinants from each DAS space

$$\left|K^{\mathcal{A}}\right\rangle = \left|\mathbb{K}_\mu^{\mathcal{A}}\right\rangle \oplus \left|\mathbb{K}_\nu^{\mathcal{A}}\right\rangle \oplus \cdots \tag{13}$$

The total number of determinants in category $\mathcal{A}$ is:

$$N^{\mathcal{A}} = N_\mu^{\mathcal{A}} \cdot N_\nu^{\mathcal{A}} \cdots \tag{14}$$

Considering the definition of a configuration category and its constituent DASs, it becomes apparent how they relate to the overall number of determinants (N), the total electron count (n_e), and the complete set of orbitals ($\{\phi\}$) in the entire correlation space:

$$N = \sum_{\mathcal{A}} N^{\mathcal{A}} \tag{15}$$

$$n_e = \sum_{\mu} n_\mu^{\mathcal{A}} \tag{16}$$

$$\{\phi\} = \{\phi\}_\mu \cup \{\phi\}_\nu \cup ... \tag{17}$$

It's important to note that the notation for categories is not applied in the DAS orbital representation, ϕ_μ, since every category employs the same scheme for orbital partitioning.

3.4 Small-Tensor Addressing in String-Based DAS

The objective is to develop a STP addressing algorithm that exclusively utilizes and reuses local address strings, thereby avoiding the need for explicitly constructing and storing the complete global list of excitations. In the STP-DAS framework, a three-level addressing protocol is adapted from the Kozlowski and Pulay's RAS addressing scheme.[57]

For a determinant $|K\rangle$ that belongs to category $\mathcal{B}$, its global address can be defined as

$$A(K) = A(\mathcal{B}) + A(K^{\mathcal{B}}) \tag{18}$$

Here, we define A as an address function. $A(\mathcal{B})$ returns the global offset for all determinants in category $\mathcal{B}$. It can be easily calculated as

$$A(\mathcal{B}) = \sum_{\mathcal{A}<\mathcal{B}} N^{\mathcal{A}} \tag{19}$$

which naturally arises from Eq. (15). Figure 2 illustrates the address mapping strategy utilized in the STP-DAS framework. In order to compute the global offset for each category, it is necessary to systematically organize all categories for efficient record-keeping. $A(K^{\mathcal{B}})$ returns the address of the determinant K in category $\mathcal{B}$.

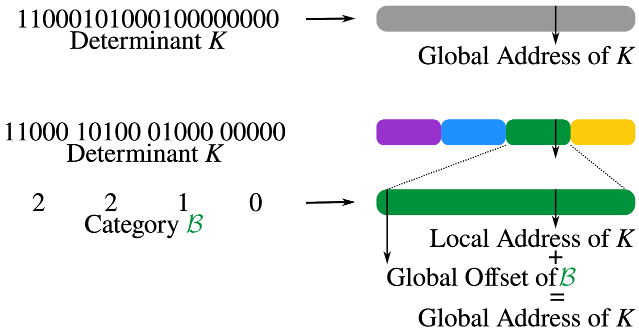

Figure 2. A comparison between a traditional indexing scheme and the indexing scheme used in STP-DAS. In DAS, the global address of a determinant is found by combining a global offset of the category of a given determinant with the local index of that determinant within the category. Boxes symbolize determinant categories.

Eq. (18) serves as the outer loop in constructing the STP-DAS $\boldsymbol{\sigma}$ build, where it is also possible to achieve effective load balancing across categories in a parallel computing environment. Although Eq. (18) successfully divides global addresses into more manageable categorical address sections, it does not reduce the overall size of the excitation list. This is due to the uniqueness of each category, requiring that each $A(K^{\mathcal{B}})$ be stored individually.

From Eq. (13), we can effectively write the categorical address $A(K^{\mathcal{B}})$ via separations of active and inactive DASs, based on its usage in the $\boldsymbol{\sigma}$ build. Since only the one-electron excitation list is saved, we only need to consider local addresses associated with

$$\left\langle L^{\mathcal{A}} \right| \hat{E}_{pq} \left| K^{\mathcal{B}} \right\rangle, \quad p \in \mathbb{X}_\mu^{\mathcal{A}}, \quad q \in \mathbb{X}_\nu^{\mathcal{B}} \tag{20}$$

Only two scenarios are possible:

- $\mu = \nu$, *i.e.*, braket excitations between an identical DAS in different categories.

$$\begin{aligned} A(L^{\mathcal{A}}) &= \cdots + A(\mathbb{L}_\mu^{\mathcal{A}}) \cdot \prod_{\kappa<\mu} N_\kappa^{\mathcal{A}} + \dots \\ A(K^{\mathcal{B}}) &= \cdots + A(\mathbb{K}_\mu^{\mathcal{B}}) \cdot \prod_{\kappa<\mu} N_\kappa^{\mathcal{B}} + \dots \end{aligned} \tag{21}$$

where "…" refers to concurrent looping over addresses in DASs that are not associated with the excitations, *i.e.* inactive DASs, for both $A(L^{\mathcal{A}})$ and $A(K^{\mathcal{B}})$.

- $\mu \neq \nu$, *i.e.*, braket excitations between two different DASs in different categories.

$$\begin{aligned} A(L^{\mathcal{A}}) &= \cdots + A(\mathbb{L}_\mu^{\mathcal{A}}) \cdot \prod_{\kappa<\mu} N_\kappa^{\mathcal{A}} + \dots \\ &+ A(\mathbb{L}_\nu^{\mathcal{A}}) \cdot \prod_{\kappa<\nu} N_\kappa^{\mathcal{A}} + \dots \\ A(K^{\mathcal{B}}) &= \cdots + A(\mathbb{K}_\mu^{\mathcal{B}}) \cdot \prod_{\kappa<\mu} N_\kappa^{\mathcal{B}} + \dots \\ &+ A(\mathbb{K}_\nu^{\mathcal{B}}) \cdot \prod_{\kappa<\nu} N_\kappa^{\mathcal{B}} + \dots \end{aligned} \tag{22}$$

Eq. (21) and Eq. (22) efficiently separate out the inactive DASs linked to the excitation list in $\left\langle L^{\mathcal{A}} \right| \hat{E}_{pq} \left| K^{\mathcal{B}} \right\rangle$. For single-electron excitations, the local addresses within the inactive DASs must be identical for both bra and ket. This requirement facilitates an efficient tensor looping design, where a single outer loop iterates over the addresses of determinants within the inactive DASs, enabling concurrent tensor looping for both bra and ket. The address structure of Eq. (21) and Eq. (22) allow the inner tensor loop to use only the local addresses of subdeterminants $A(\mathbb{L}_\mu^{\mathcal{A}})$ and $A(\mathbb{K}_\mu^{\mathcal{B}})$, leading to the small-tensor-product algorithm.

It is evident to see that when two DASs share the same configuration, characterized by an identical number of orbitals and electrons, they can utilize the same set of small-tensor address strings. This is true even if the orbitals and electrons between the two DASs differ in physical character. This principle is fundamental to the shift from a single, large global excitation list to numerous, smaller, shared local excitation lists.

In adopting this approach to streamline determinant addressing, a new challenge emerges. The CI excitation list features a symbolic matrix $\langle L^{\mathcal{A}} | \hat{E}_{pq} | K^{\mathcal{B}} \rangle$, where each element is determined by the global positions of the orbitals involved in the excitation operator. This global dependency poses a conflict with the local addressing framework. To address this, the subsequent section will present an algorithm designed to differentiate between global and local phase factors during the construction of the symbolic matrix $\langle L^{\mathcal{A}} | \hat{E}_{pq} | K^{\mathcal{B}} \rangle$, enabling the storage of solely local excitation lists in small tensor products.

3.5 Local One-Electron Excitation List

To account for the sparsity of the Hamiltonian, one-electron excitation list that contains the information of non-zero elements of $\langle L | \hat{E}_{tu} | K \rangle$, is pre-computed and stored in memory. Only elements that are non-zero in the excitation list are summed in Eq. (8). Based on the addressing scheme shown in Eqs. (20) to (22), the local excitation list can be defined and the symbolic matrix element $\langle \mathbb{L}^{\mathcal{A}}_{\mu} | \hat{E}_{pq} | \mathbb{K}^{\mathcal{B}}_{\nu} \rangle$ can be computed as

- $\mu = \nu$:

  $$\left\{ \begin{array}{c} A(\mathbb{L}^{\mathcal{A}}_{\mu}) \\ p \in \mathbb{X}^{\mathcal{A}}_{\mu} \\ q \in \mathbb{X}^{\mathcal{B}}_{\nu} \\ A(\mathbb{K}^{\mathcal{B}}_{\nu}) \\ \langle \mathbb{L}^{\mathcal{A}}_{\mu} | \hat{E}_{pq} | \mathbb{K}^{\mathcal{B}}_{\nu} \rangle \end{array} \right\} \tag{23}$$

  In this intra-space one-electron excitation scenario, for each $|\mathbb{K}^{\mathcal{B}}_{\nu}\rangle$, there are $n^{-}_{\mu}(n^{+}_{\mu}+1)$ number of non-zero elements, where n^{-}_{μ} and n^{+}_{μ} are the number of electrons (occupied orbitals) and holes (unoccupied orbitals), respectively, in DAS $\mathbb{X}_{\mu} = \mathbb{X}_{\nu}$. The non-zero element is computed as

  $$\begin{aligned} &\forall p > q : \langle \mathbb{L}^{\mathcal{A}}_{\mu} | \hat{E}_{pq} | \mathbb{K}^{\mathcal{B}}_{\nu} \rangle \\ &= \begin{cases} +1, \text{ if } \left(\sum_{i=q+1}^{p-1} b_i\right) \text{ is even} \\ -1, \text{ if } \left(\sum_{i=q+1}^{p-1} b_i\right) \text{ is odd} \end{cases} \end{aligned} \tag{24}$$

  where $b_i = 0$ or 1 is the electron occupancy.

- $\mu \neq \nu$: This inter-space one-electron excitation list involves four sub-determinants in different DASs,

  $$\left\{ \begin{array}{c} A(\mathbb{L}^{\mathcal{A}}_{\mu}), A(\mathbb{L}^{\mathcal{A}}_{\nu}) \\ p \in \mathbb{X}^{\mathcal{A}}_{\mu} \\ q \in \mathbb{X}^{\mathcal{B}}_{\nu} \\ A(\mathbb{K}^{\mathcal{B}}_{\mu}), A(\mathbb{K}^{\mathcal{B}}_{\nu}) \\ \langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} | \hat{E}_{pq} | \mathbb{K}^{\mathcal{B}}_{\mu} \oplus \mathbb{K}^{\mathcal{B}}_{\nu} \rangle \end{array} \right\} \tag{25}$$

  The non-zero one-electron symbolic matrix elements can be computed as:

  $$\begin{aligned} &\forall \mu < \nu : \langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} | \hat{E}_{pq} | \mathbb{K}^{\mathcal{B}}_{\mu} \oplus \mathbb{K}^{\mathcal{B}}_{\nu} \rangle \\ &= \begin{cases} +1, \text{ if } \left(\sum_{i=p+1}^{M_{\mu}} b_i + \sum_{j=0}^{q-1} b_j\right) \text{ is even} \\ -1, \text{ if } \left(\sum_{i=p+1}^{M_{\mu}} b_i + \sum_{j=0}^{q-1} b_j\right) \text{ is odd} \end{cases} \end{aligned} \tag{26}$$

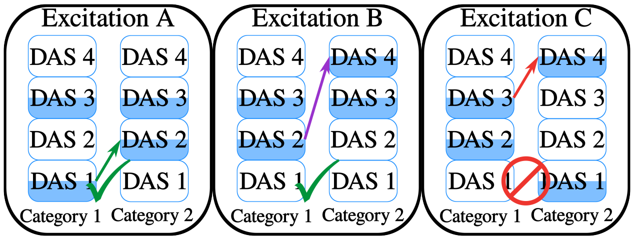

Figure 3. Examples of excitations between categories. Boxes symbolize active regions. Within these boxes, shaded portions indicate electron occupations, while white areas represent empty orbitals. All partially filled DASs have an identical number of electrons and holes.

Eq. (23) and Eq. (25) define the one-electron excitation lists in the context of STP-DAS using local addresses. These lists are free of global addresses, enabling their reuse across excitations with identical DAS configurations. As illustrated in Figure 3, the green and purple excitations (excitations A and B), although associated with different orbitals and electrons, have identical DAS configurations, allowing them to share the same local excitation lists used in the small tensor product. This commonality highlights that dividing the correlation space evenly among numerous DASs can significantly reduce the number of unique local excitation lists. Achieving this efficiency is a principal objective in the conceptualization of STP-DAS.

3.6 Global Phase Factor in the Small Tensor Product algorithm

Equipped with determinantal addresses and one-electron excitation lists completely defined in STP-DAS, the CI $\boldsymbol{\sigma}$ build can be ideally performed with small tensor products using local excitation lists. Nevertheless, two critical aspects need to be considered for enabling the $\boldsymbol{\sigma}$ matrix assembly with small tensor products using these local lists.

One issue is distinguishing between different types of excitations, such as excitations A and B, as shown Figure 3, during the $\boldsymbol{\sigma}$ matrix construction. This process requires global symbolic matrix elements, which are derived from the global positioning of the orbitals. Another challenge is identifying invalid excitations, such as excitation C in Figure 3. The excitation lists mainly provide local DAS specifics, leading to a loss of global context. However, STP-DAS's architecture facilitates the straightforward retrieval of global address-dependent information during the CI $\boldsymbol{\sigma}$ matrix construction, resolving both issues.

The distinction between excitations A and B in Figure 3 stems from a difference in the phase factor between the global $\left\langle L^{\mathcal{A}}\right| \hat{E}_{pq} \left|K^{\mathcal{B}}\right\rangle$ and local $\left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu}\right| \hat{E}_{pq} \left|\mathbb{K}^{\mathcal{B}}_{\mu} \oplus \mathbb{K}^{\mathcal{B}}_{\nu}\right\rangle$ symbolic matrix element. To address this, we define a global phase factor $P_{\mu\nu}$ as:

$$
\begin{gathered}
p \in \mathbb{X}^{\mathcal{A}}_{\mu},\ q \in \mathbb{X}^{\mathcal{B}}_{\nu},\ \forall \mu \leq \nu : \\
P_{\mu\nu} = \begin{cases} +1, \text{ if } \left(\sum_{\mathbb{X}_{\kappa}=\mathbb{X}_{\nu+1}}^{\mathbb{X}_{\nu-1}} n^{-}_{\kappa}\right) \text{ is even} \\ -1, \text{ if } \left(\sum_{\mathbb{X}_{\kappa}=\mathbb{X}_{\nu+1}}^{\mathbb{X}_{\nu-1}} n^{-}_{\kappa}\right) \text{ is odd} \end{cases}
\end{gathered}
\tag{27}
$$

where n^{-}_{κ} is the number of electrons in DAS $\mathbb{X}_{\kappa}$. $P_{\mu\nu}$ takes into account the contributions of DASs that are not directly involved in the excitation to the symbolic matrix element.

In Figure 3, excitation C is an invalid excitation because it cannot be defined with one-electron excitation lists. To determine the validity of an excitation, $\left\langle L^{\mathcal{A}}\right| \hat{E}_{pq} \left|K^{\mathcal{B}}\right\rangle$, we need to check if the non-excitation DASs between categories are identical. Here, we introduce a Kronecker δ function $\delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu}\bar{\mathbb{X}}^{\mathcal{B}}_{\nu}}$ where $\bar{\mathbb{X}}^{\mathcal{A}}_{\mu}$ refers to all but $\mathbb{X}_{\mu}$ DASs in category $\mathcal{A}$. The δ function is non-zero only when non-excitation DASs between categories are identical.

3.7 $\boldsymbol{\sigma}$ Build using Small Tensor Products

Taking everything discussed above into consideration, we now have the final working expressions for the CI $\boldsymbol{\sigma}$ build using the STP algorithm with only local excitation lists:

$$\sigma_{L^{\mathcal{A}}} = {}^{1\text{e}}\sigma_{L^{\mathcal{A}}} + {}^{2\text{e}}\sigma_{L^{\mathcal{A}}} \tag{28}$$

$$
\begin{aligned}
{}^{1\text{e}}\sigma_{L^{\mathcal{A}}} = &\sum_{\mathcal{B}} \sum_{\mathbb{K}^{\mathcal{B}}_{\mu}\oplus\mathbb{K}^{\mathcal{B}}_{\nu}} \sum_{pq} P_{\mu\nu}\delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu\nu}\bar{\mathbb{X}}^{\mathcal{B}}_{\mu\nu}} \\
&h'_{pq} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu}\right| \hat{E}_{pq} \left|\mathbb{K}^{\mathcal{B}}_{\mu} \oplus \mathbb{K}^{\mathcal{B}}_{\nu}\right\rangle C_{L^{\mathcal{A}}K^{\mathcal{B}}}
\end{aligned}
\tag{29}
$$

$$
\begin{aligned}
{}^{2\text{e}}\sigma_{L^{\mathcal{A}}} = &\frac{1}{2}\sum_{\mathcal{B}\mathcal{C}} \sum_{\mathbb{J}^{\mathcal{B}}_{\mu}\oplus\mathbb{J}^{\mathcal{B}}_{\nu}} \sum_{\mathbb{J}^{\mathcal{B}}_{\kappa}\oplus\mathbb{J}^{\mathcal{B}}_{\lambda}} \sum_{\mathbb{K}^{\mathcal{C}}_{\kappa}\oplus\mathbb{K}^{\mathcal{C}}_{\lambda}} \sum_{pqrs} P_{\mu\nu}P_{\kappa\lambda} \\
&\delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu\nu}\bar{\mathbb{X}}^{\mathcal{B}}_{\mu\nu}} \delta_{\bar{\mathbb{X}}^{\mathcal{B}}_{\kappa\lambda}\bar{\mathbb{X}}^{\mathcal{C}}_{\kappa\lambda}} g_{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu}\right| \hat{E}_{pq} \left|\mathbb{J}^{\mathcal{B}}_{\mu} \oplus \mathbb{J}^{\mathcal{B}}_{\nu}\right\rangle \\
&\left\langle \mathbb{J}^{\mathcal{B}}_{\kappa} \oplus \mathbb{J}^{\mathcal{B}}_{\lambda}\right| \hat{E}_{rs} \left|\mathbb{K}^{\mathcal{C}}_{\kappa} \oplus \mathbb{K}^{\mathcal{C}}_{\lambda}\right\rangle C_{L^{\mathcal{A}}K^{\mathcal{B}}}
\end{aligned}
\tag{30}
$$

where $p \in \mathbb{X}^{\mathcal{A}}_{\mu},\ q \in \mathbb{X}^{\mathcal{B}}_{\nu}, r \in \mathbb{X}^{\mathcal{B}}_{\kappa},\ s \in \mathbb{X}^{\mathcal{C}}_{\lambda}$.

Eqs. (28) to (30) are completely formulated in terms of small tensor products using local DAS one-electron excitation lists. These expressions can be adapted to any $\boldsymbol{\sigma}$ build algorithm in any CI-based method. In this work, we used the Knowles-Handy algorithm[47] for the categorical $\boldsymbol{\sigma}$ build. The outer loop is over $\mathbb{J}^{\mathcal{B}}_{\mu\nu\kappa\lambda}$, and maximum sparsity is utilized to avoid large intermediates.

4 Computational Details

All calculations in this work are performed with a development version of the Chronus Quantum software package[58] with the STP-DAS framework. The speed of light utilized in this study is 137.035999074 a.u. All calculations utilized the standard Gaussian nuclear model.[59] Relativistic calculations are done in the Kramers unrestricted exact-two-component (X2C) framework where all spinor orbitals are singly occupied. The one-electron X2C transformation is a one-electron-Hamiltonian-based one-step procedure that "folds" small component wave function into a pseudo-large component so that the four-component Dirac equation becomes an effective two-component eigenfunction problem.[22,43,60–77] The one-electron X2C approach makes use of the effective one-electron spin–orbit Hamiltonian and avoids the four-component self-consistent-field procedure. In this work, we use the new Dirac–Coulomb–Breit-parameterized effective one-electron spin–orbit Hamiltonian in the X2C approach.[78]

STP-DAS stands as a scalable, high-performance CI framework supporting various electronic structure theories – including CAS, RAS, GAS, ORMAS, MRCI, and MRPT2 – in both relativistic and non-relativistic domains. The fundamental characteristics and dimensionalities of these electronic structure methods remain unchanged within this framework. In the current work, we benchmark the performance and analyze the memory requirement of the DAS framework applied to X2C-CASCI calculations.

To benchmark the STP-DAS framework, we use molecular thallium monohydride (TlH), the heaviest stable monohydride species observed experimentally.[79,80] An accurate electronic structure characterization of TlH requires the use of a relativistic many-body approach.[81,82] In this study, we concentrate on evaluating the algorithm's performance as the correlation space expands. Furthermore, we investigate the load balance among nodes within a high-performance computing environment, as well as the reduction in memory demands when computing resources are constrained, such as on a laptop. In the following benchmark calculations, the Tl atom used the x2c-TZVPall-2c basis set[83] and the H atom used the aug-cc-pVTZ basis set.[84,85]

When forming Slater determinants in the Kramers' unrestricted two-component or four-component no-virtual-pair CASCI method, the number of possible determinants is given by

$$N_{\text{CAS}} = \binom{M_{\text{spinor}}}{n_e} \tag{31}$$

where M_{spinor} is the number of active spinor orbitals. Since the spin-symmetry is no longer enforced, for a same number of active electrons and molecular orbitals, the CI dimension in the Kramers' unrestricted relativistic CAS is much bigger than that in the non-relativistic calculation. In relativistic computations, the floating point operations (FLOP) count for constructing the $\boldsymbol{\sigma}$ matrix experiences a sixfold increase, arising from complex-valued arithmetic, compared to a non-relativistic (NR) calculation.

The calculation of the memory requirement in all test cases takes into account both the sparsity of the excitation list, *i.e.*, non-zero elements only, and bit-wise compression of the determinant address, representing the minimum requirement in a CAS calculation. For X2C-CASCI calculations, the size of this list (in Bytes) can be calculated as $n_{\text{save}} \times n_e \times (n_h + 1) \times N_{\text{CAS}}$, where n_e is the number of electrons and n_h is the number of holes in the complete active space. Here, n_{save} represents the amount of information saved in the excitation list, including the address of small tensors, p, p, and the phase factor. Depending on the level of bit-wise compression, n_{save} varies from 4 to 11 Bytes for most applications.

5 Results and Discussion

5.1 A Large-Scale X2C-CASCI Calculation on a Laptop

A significant benefit of the STP-DAS framework lies in its capacity to facilitate large-scale CI computations using constrained computational resources, like a laptop. This was illustrated through the performance of X2C-CAS calculations on the TlH system, utilizing the STP-DAS framework on an Apple M3 Max laptop equipped with 14 compute cores and 128 GB of RAM.

Table 1 illustrates the STP-DAS framework's capability to reduce storage demands, thereby

Table 1. The storage needs for the one-electron excitation list in a $\binom{44}{36}$ X2C-CASCI calculation (Apple M3 Max laptop with 14 compute cores and 128 GB of RAM). The total number of X2C determinants is 177×10^6, which is comparable to the computational cost of 1.1×10^9 determinants in non-relativistic calculations when measured by the number of FLOPs.

# of DASs (Orbital Partitions)	Excitation List Storage[a]	$\boldsymbol{\sigma}$ Build Time
1 (44)	402. GB	–
2 (22,22)	210. GB	–
6 (7,7,7,7,7,9)	7.91×10^{-3} GB	2981 s
9 (5,5,5,5,5,5,5,5,4)	9.06×10^{-5} GB	5788 s

[a] Only non-zero elements are considered in the one-electron excitation list. Bit-wise compression of the determinant address is used.

enabling large-scale CI calculations on a laptop. For a $\binom{44}{36}$ X2C-CASCI calculation with STP-DAS, the resulting 177×10^6 X2C determinants (1.1×10^9 non-relativistic determinants equivalent in terms of FLOP count) would require 402 GB of RAM to hold the excitation list with non-zero elements. This requirement makes the task impractical for a personal laptop and many smaller workstations.

As shown in Table 1, partitioning the correlation space into multiple DASs significantly reduces the storage needs. Once the number of DASs hits 6, the storage demands for one-electron excitations become minimal, making it feasible to conduct the X2C-CASCI calculations on a laptop with ease. Each $\boldsymbol{\sigma}$ build only takes 0.8 hours on the latest Apple laptop for a 177×10^6 determinant X2C-CASCI calculation.

With the introduction of additional STP-DAS for space partitioning using small tensor products, the storage requirement continues to decrease. However, this leads to an increase in the time required to build the $\boldsymbol{\sigma}$ vector. The primary reason for this is the increased overhead associated with extensive tensor looping necessary to locate each local address within the global framework, as indicated in Eq. (21) and Eq. (22). This analysis reveals that while STP-DAS is effective in reducing memory requirements, excessively fine partitioning might result in added costs associated with small tensor mapping from local to global addresses.

5.2 A Large-Scale X2C-CASCI Calculation on a Supercomputer

In high-performance computing, especially for large CI calculations on a supercomputer, it's crucial to ensure that the workload is evenly distributed across all compute nodes. This is where the STP-DAS framework is particularly effective. The STP-DAS framework is capable of decreasing memory demands while concurrently leveraging distributed memory on a massively parallel high-performance computing system.

As illustrated in Table 2 with a $\binom{44}{29}$ X2C-CASCI calculation example, the STP-DAS approach achieves memory reduction by employing localized excitation lists within the distributed active space. Expanding the X2C-CASCI of $\binom{44}{29}$ results in over 230 billion determinants (1.4×10^{12} non-relativistic determinants equivalent in terms of FLOP count), requiring 1,173 terabytes (TB) of memory to maintain the one-electron excitation list within the conventional CAS setup. Assuming that each compute node allocates 0.5 TB of memory for this purpose, it would require over 2,346 compute nodes with distributed memory systems to perform a CI calculation. A configuration with 4 DASs can lower the memory needs to a manageable 1.7 GB for a small distributed computing system.

Table 2. The storage needs for the one-electron excitation list in a $\binom{44}{29}$ X2C-CASCI calculation. The total number of X2C determinants is 230×10^9, which is comparable to the computational cost of 1.4×10^{12} determinants in non-relativistic calculations when measured by the number of FLOPs.

# of DASs (Orbital Partitions)	Excitation List Storage[a]	# of Categories
1 (44)	1,173 TB	1
2 (22,22)	563 TB	16
4 (11,11,11,11)	2 GB	736
6 (7,7,7,7,7,9)	1×10^{-2} GB	11,292
8 (6,6,6,6,6,6,6,2)	4×10^{-4} GB	80,823
9 (5,5,5,5,5,5,5,5,4)	9×10^{-5} GB	260,656

[a] Only non-zero elements are considered in the one-electron excitation list. Bit-wise compression of the determinant address is used.

In the STP-DAS framework, we allocate

workloads among the compute nodes according to different configuration categories. As we introduce more categories, we can achieve a more balanced distribution of the total workload for constructing the $\boldsymbol{\sigma}$ matrix. The effectiveness of the STP-DAS framework is illustrated in Table 2, which demonstrates a rapid increase in the number of configuration categories as additional DASs are introduced into the system for space and tensor partitioning. This expansion significantly enhances the efficiency of the $\boldsymbol{\sigma}$ matrix construction. A figure of merit for computational work distribution is the percent difference of the median number of determinants per node relative to the theoretical ideal mean distribution of determinants. The observed distribution of determinants to each node in a 600-node calculation is illustrated in Figure 4. The 8 DAS distribution is markedly more extended with a much longer tail. This results in a few nodes with many more determinants than all of the other nodes shifting the median number of determinants further from the ideal distribution. When increasing the number of DASs from 8 to 9, the percent difference in the work distribution drops from 4% to 0.15% resulting in a nearly ideal division of determinants among the nodes. With the resulting balanced workload, each $\boldsymbol{\sigma}$ build for the $\binom{44}{29}$ X2C-CASCI calculation with 9 DASs containing 230×10^9 determinants (1.4×10^{12} NR equivalent determinants) only took 7 hours. This calculation was run on the Department of Energy's Perlmutter high-performance super computer with a total of 16000 compute cores (AMD EPYC 7763 Milan, 200 GB/s NIC, 1 MPI per node and 16 SMP threads per MPI process). This analysis shows the scalability and efficiency improvements in computational performance that can be realized through optimized workload distribution in the STP-DAS framework.

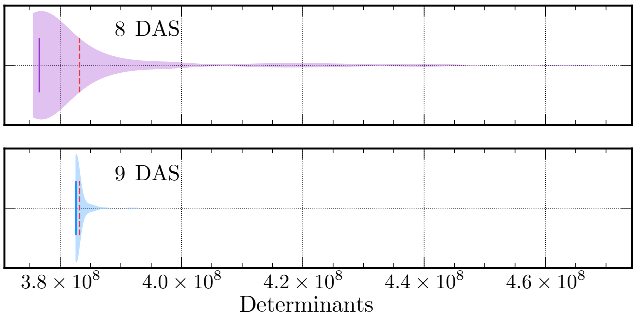

Figure 4. Distribution of determinants per node for a $\binom{44}{29}$ X2C-DASCI calculation using 600 nodes with (top) 8 DAS spaces and (bottom) 9 DAS spaces. The theoretical ideal distribution of determinants is denoted by the red line in both plots, and the solid line denotes the median of the observed distributions.

5.3 Massively Parallel Performance

In this section, we examine how the STP-DAS framework functions within extensive supercomputing environments. In order to demonstrate the massively parallel performance of the STP-DAS algorithm, we performed X2C-CASCI calculations on the TlH molecule with increasing size active spaces of 35 to 40 spin orbitals using 24 correlated electrons.

Table 3. Number of determinants in an X2C-CASCI calculation using the STP-DAS framework of TlH with 24 correlated electrons and a varying number of active spinor orbitals.

# of Orbitals	# of X2C Determinants (NR Equivalent[a])	Excitation List Storage[b]
35	417,225,900 ($2.5{\times}10^9$)	841 GB
36	1,251,677,700 ($7.5{\times}10^9$)	2,733 GB
37	3,562,467,300 ($2.1{\times}10^{10}$)	8,378 GB
38	9,669,554,100 ($5.8{\times}10^{10}$)	38 TB
39	25,140,840,660 ($1.5{\times}10^{11}$)	106 TB
40	62,852,101,650 ($3.8{\times}10^{11}$)	282 TB

[a] Non-relativistic (NR) equivalent number of determinants in non-relativistic calculations when measured by the number of FLOPs.
[b] Only non-zero elements are considered in the one-electron excitation list. Bit-wise compression of the determinant address is used.

Table 3 lists the total number of determinants in a X2C-CASCI calculation and the equivalent number of non-relativistic determinants measured by the number of FLOPs. The minimal memory storage requirement for each case is also computed should all active orbitals and electrons be included in a single active space, *e.g.*, in a conventional CASCI calculation. Table 3 shows that as the active space expands, the memory needed to store non-zero elements of the excitation list increases rapidly, becoming exceedingly demanding.

Table 4 lists the STP-DAS partition schemes

Table 4. STP-DAS space partition schemes and the resulting numbers of configuration categories used in the benchmark for the massively parallel performance of the STP-DAS framework.

# of Orbitals (Orbital Partitions)	# of Categories
35 (5,5,5,5,5,5,5)	9,142
36 (5,5,5,5,5,5,5,1)	21,259
37 (5,5,5,5,5,5,5,2)	36,526
38 (5,5,5,5,5,5,5,3)	54,853
39 (5,5,5,5,5,5,5,4)	75,846
40 (5,5,5,5,5,5,5,5)	98,813

employed in the following HPC benchmark study. The excitation list storage requirement is not presented because it is reduced to less than 100 KB for all test cases. Two high-performance computing (HPC) systems were used in this benchmark study. The first is a medium size HPC system, Hyak, managed by the University of Washington (UW). Each Hyak node has two Intel Xeon 6230 Gold CPUs with a single 100 GB/s network interface card. A maximum of 250 nodes on this medium sized HPC system were available to the authors. The second system is the Department of Energy's Perlmutter high-performance super computer with up to 512 nodes (AMD EPYC 7763 (Milan), 200 GB/s NIC).

5.3.1 Strong scaling

We first study the strong scaling of the X2C-CASCI calculation using the STP-DAS framework on a medium size HPC system, the UW Hyak. Figure 5 shows the strong scaling performance of the STP-DAS $\boldsymbol{\sigma}$ build on varying active space sizes for the TlH test case. For the log-log presentation of runtime versus number of nodes, an ideally scaling algorithm would be represented by straight, decreasing line. As such, one can observe excellent strong scaling of the STP-DAS implementation for a variety of problem sizes.

For real world applications, an important feature that can be extracted from a strong scaling plot is a stagnation point where one can observe that speedups have ceased despite an increase in computational power. This is usually a sign that the amount of work is insufficient for the number of nodes. Focusing on the $\binom{35}{24}$ problem, one can observe this stagnation point at around 75 nodes. As one increases the problem size, this stagnation point shifts to a greater node count as more work is available to divide amongst the nodes. This can be observed for the $\binom{39}{24}$ X2C-CASCI problem, which still benefits from extra computational resources and scales up to 250 nodes.

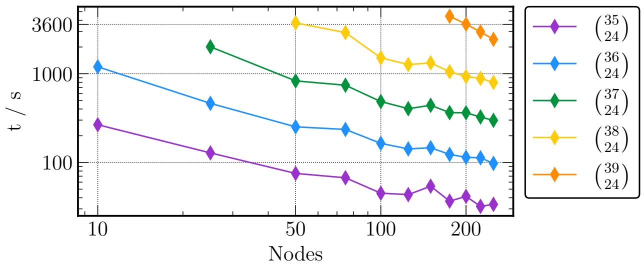

Figure 5. Execution time of a single $\boldsymbol{\sigma}$ build using the STP-DAS framework for the TlH test case with increasing CAS space sizes with respect to the number of nodes (1 MPI per node, 20 SMP threads per MPI).

5.3.2 Relative speed-up

A strong scaling analysis is helpful at understand the overall behavior of an algorithm, but this does not yield an understanding of the origins of the stagnation point. For a detailed understanding of the STP-DAS framework, we turn to a detailed analysis of the speed up of a certain problem size with respect to the number of nodes. Ideally the concept of speedup would require a definition of a serial run time, however since the size of the CI calculations described in Table 3 would not fit on a single node, we instead present a strong scaling as a relative speed-up with respect to the performance on 16 nodes. We study this speedup in detail for two problem sizes $\binom{37}{24}$ and $\binom{38}{24}$ to extract the salient features of the STP-DAS algorithm.

Figure 6 (top) shows the relative speed up of an X2C-CASCI calculation using the STP-DAS framework of TlH with an active space of (37o,24e). We additionally plot the relative speedup of the computation of the $\boldsymbol{\sigma}$ build and the MPI communication idle time as the total runtime is a combination of these two times. It is immediately apparent that for this problem size the strong scaling stagnation onset occurs

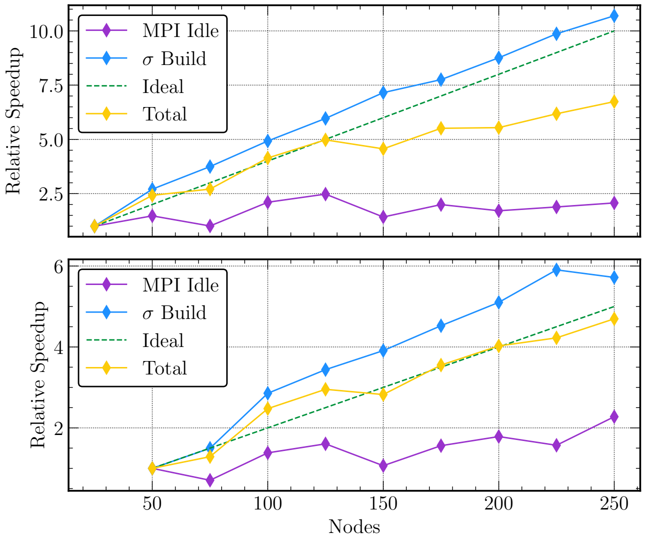

Figure 6. Relative speedup of a single $\boldsymbol{\sigma}$ build for TlH (top) $\binom{37}{24}$ and (bottom) $\binom{38}{24}$ with respect to the number of nodes (1 MPI per node, 20 SMP threads per MPI). Relative speed-up is defined with respect to the performance on the smallest node count capable of solving the problem.

around 125 nodes. Additionally, one can observe that the MPI communication idle time does not scale with the number of nodes and that the actual computation time scales extremely well with the number of nodes. It is important to recognize that a superlinear scaling of the computation, the $\boldsymbol{\sigma}$ build time, is not an indication of a superlinear scaling of the algorithm as this does not include the MPI communication idle time. Additionally, it is important to stress that the onset of the strong scaling stagnation is not an indication that the calculation takes the same amount of time irrespective of the number of nodes beyond that point. Beyond the stagnation point, one observes deviation from ideal scaling, but there is still a reduction in total runtime as the computation scales with the number of nodes while the MPI communication idle time does not.

By increasing the problem size, one shifts the location of the strong scaling stagnation point. Figure 6 (bottom) shows the relative speedup of an X2C-CASCI calculation using the STP-DAS framework of TlH with an active space of (38o, 24e). The behavior now differs from the smaller (37o, 24e) case. For the range of calculations presented, ideal scaling is observed for the full range of the number of nodes. One may often see scaling plots such as this one rather than the previous case, which does not show the stagnation point. However, it is important to note that this does not imply that the stagnation point does not exist only that it has shifted to larger node counts.

5.3.3 Computation time, MPI communication idle time, and load balancing

To further understand the communication versus computation time of the STP-DAS algorithm, we plot the raw execution time as a function of problem size. These calculations were run on the large Department of Energy's Perlmutter HPC system. From Figure 7, it is evident that the total execution time exhibits a linear relationship with the number of determinants. This linear scaling is due to the fact that, with a fixed number of electrons, the number of non-zero elements in the one-electron excitation list increases linearly as the number of virtual orbitals grows. Additionally, it is observed that the MPI communication idle time grows very slowly with problem size. In the largest test case $\binom{40}{24}$, the MPI communication idle time is only 18% of the total $\boldsymbol{\sigma}$ build time.

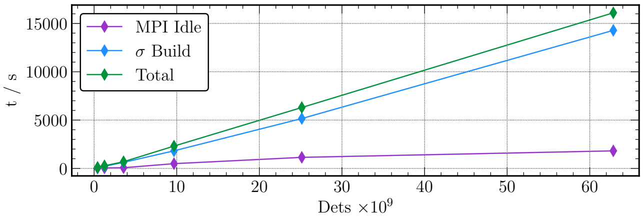

Figure 7. Execution time of a single $\boldsymbol{\sigma}$ build using the STP-DAS framework for the TlH test case with increasing CAS space sizes (Perlmutter 256 Nodes, 1 MPI per node, 64 SMP threads per MPI).

In the STP-DAS algorithm the $\boldsymbol{\sigma}$ build computation time and the MPI communication idle time vary slightly per MPI process as each MPI process receives a different set of categories to process. To study the load balancing performance of the STP-DAS algorithm, Figure 8 (top) shows the build time for each of the X2C-CASCI calculations. The solid line represents the mean of the $\boldsymbol{\sigma}$ build times from each process, and the shaded region represents the histogram of the $\boldsymbol{\sigma}$ build times from each MPI

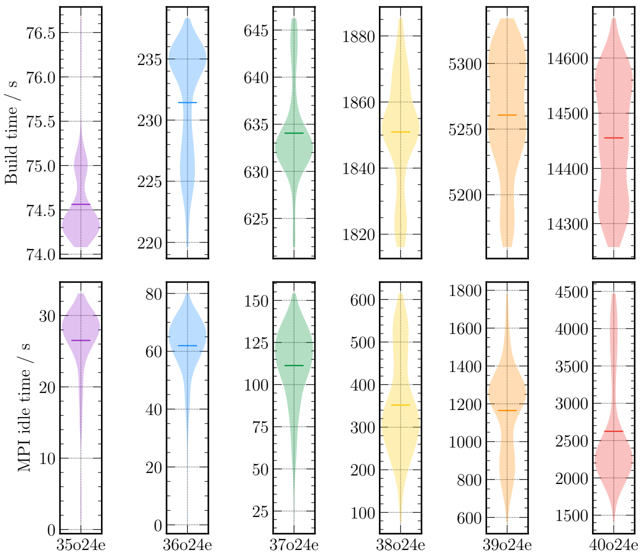

Figure 8. (top) $\boldsymbol{\sigma}$ build time (seconds) and (bottom) the distribution of the MPI communication idle time of each process during the $\boldsymbol{\sigma}$ build process of the TlH test case with a increasing numbers of correlated orbitals (Perlmutter 256 Nodes, 1 MPI per node, 64 SMP threads per MPI).

process. The small range of times over which the histogram is spread, irrespective of the calculation size, is an indicator of the strong load balancing in the current iteration of the STP-DAS algorithm. This shows that breaking the problem into smaller DAS spaces allows for a balanced distribution of categories of determinants.

Once each MPI process has completed the assigned work for the $\boldsymbol{\sigma}$ build, it must idle and communicate its result with the other processes. The distribution of the MPI communication idle times is represented in Figure 8 (bottom). The MPI communication idle time tends to be closely grouped around the average for most problem sizes. However, for the largest calculation, $\binom{40}{24}$, the distribution exhibits a minor tail. It is anticipated that the incorporation of Remote Memory Access into MPI will enhance both the efficiency of MPI communication idle time and the load balancing, especially in cases of such substantial computational magnitude.

5.3.4 Effect of the active space partitioning

To test the effect of active space partitioning on the performance of the STP-DAS algorithm, we used four different DAS partitionings for a small test case $\binom{28}{18}$ (13 million X2C determinants) to run the calculations on a single node. The runtimes for one $\boldsymbol{\sigma}$-build are collected in Table 5.

We observed a decrease in the $\boldsymbol{\sigma}$-build time from one to four DAS partitionings. This reduction is due to shifting work from internal loops (over excitations within DAS spaces) to the outer loop over categories. However, when the space is partitioned into seven DASs, the runtime increases significantly due to the extra work required to locate each local address, as seen previously in Section 5.1. This observation suggests that there is an optimal condition for STP-DAS, but it is strongly dependent on the system size and the nature of the computing architecture.

Table 5. The storage needs for the one-electron excitation list and runtimes for one σ build in a $\binom{28}{18}$ X2C-CASCI calculation.

# of DASs (Orbital Partitions)	Excitation List Storage[a]	$\boldsymbol{\sigma}$ Build Time
1 (28)	18.12 GB	876 s
2 (14,14)	0.86 GB	404 s
4 (7,7,7,7)	1.98×10^{-3} GB	86.7 s
7 (4,4,4,4,4,4,4)	1.06×10^{-5} GB	365.4 s

[a] Only non-zero elements are considered in the one-electron excitation list. Bit-wise compression of the determinant address is used. 2 Intel Xeon Gold 6148 processors with 20 physical cores each and 250 GB of memory.

6 Conclusion

In this work, we introduced a small tensor product distributed active space (STP-DAS) framework, characterized by adjustable space partitioning and many small tensor products with an efficient tensor loop used in the $\boldsymbol{\sigma}$ build. This framework is designed to support a variety of configuration interaction (CI) methodologies. It is also compatible with both relativistic (2-component and 4-component) and non-relativistic electronic structure methods.

The CI engine within the STP-DAS framework leverages this adjustable partitioning to significantly reduce the memory requirements for large-scale multiconfigurational calculations, offering scalability from single workstations to massively parallel computing environments.

Our benchmark tests, conducted on two different supercomputers using realistic X2C-CASCI calculations using the STP-DAS framework with determinant numbers ranging from 10^9 (billion) to 10^{12} (trillion), consistently demonstrated robust parallel scalability and excellent load balancing.

A standout feature of the STP-DAS framework is its capacity to facilitate extensive CI calculations with limited computational resources. Illustrating this, we performed a relativistic CI calculation involving 177 million X2C determinants, a task computationally equivalent to 1.1 billion non-relativistic determinants based on FLOP count, on a laptop. This capability showcases the STP-DAS framework's potential to broaden the applicability of CI methods in computational science research.

Although determinant-based algorithms vary in their contraction order and the size of the intermediates they form, they all benefit from the STP-DAS framework through significantly reduced excitation list sizes and the ability to express contractions solely in terms of local subdeterminants. Benchmarking these different algorithms requires fine-tuning each one within the STP-DAS framework, an ongoing research endeavor that will be presented in a future publication.

The STP-DAS algorithm provides an opportunity to optimize CI-based methods for specific hardware architectures. Although the optimal STP-DAS scheme cannot be determined a priori, benchmarking various space partitioning schemes enables users to tailor the algorithm to a particular large-scale high-performance computing facility. This approach helps to leverage hardware configurations effectively and minimize the impact of communication latency on overall computational efficiency.

Acknowledgement X.L. acknowledge the support to develop reduced scaling computational methods from the Scientific Discovery through Advanced Computing (SciDAC) program sponsored by the Offices of Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) of the U.S. Department of Energy (Grant No. DE-SC0022263). The development of variational relativistic multireference methods is supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, in the Heavy-Element Chemistry program (Grant No. DE-SC0021100). The development of the Chronus Quantum computational software is supported by the Office of Advanced Cyberinfrastructure, National Science Foundation (Grant No. OAC-2103717).

References

(1) Shavitt, I. In *Methods of Electronic Structure Theory*; Schafer, H. F. I., Ed.; Plentum: New York, 1977; pp 189–275.

(2) Shavitt, I. The History and Evolution of Configuration Interaction. *Mol. Phys.* **1998**, *94*, 3–17.

(3) Sherrill, C. D.; Schaefer, H. F. The Configuration Interaction Method: Advances in Highly Correlated Approaches. *Adv. Quantum Chem.* **1999**, *34*, 143–269.

(4) Čársky, P. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons, Ltd, 2002.

(5) Karwowski, J. A.; Shavitt, I. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; John Wiley & Sons, Ltd, 2003.

(6) Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. Multiconfiguration Self-Consistent Field and Multireference Configuration Interaction Methods and Applications. *Chem. Rev.* **2012**, *112*, 108–181.

(7) Roos, B. O. The Complete Active Space Self-Consistent Field Method and its Applications in Electronic Structure Calculations. **1987**, *69*, 399–445.

(8) Olsen, J.; Roos, B. O.; Jørgensen, P.; Jensen, H. J. A. Determinant Based Configuration Interaction Algorithms for Complete and Restricted Configuration Interaction Spaces. *J. Chem. Phys.* **1988**, *89*, 2185–2192.

(9) Liu, B. Ab Initio Potential Energy Surface for Linear H_3. *J. Chem. Phys.* **1973**, *58*, 1925–1937.

(10) Lischka, H.; Shepard, R.; Brown, F. B.; Shavitt, I. New Implementation of the Graphical Unitary Group Approach for Multireference Direct Configuration Interaction Calculations. *Int. J. Quant. Chem.* **1981**, *20*, 91–100.

(11) Werner, H.-J.; Knowles, P. J. An Efficient Internally Contracted Multiconfiguration-Reference Configuration Interaction Method. *J. Chem. Phys.* **1988**, *89*, 5803–5814.

(12) Hirao, K., Ed. *Recent Advances in Multireferences Methods*; World Sientific, 1999.

(13) Lischka, H.; Müller, T.; Szalay, P. G.; Shavitt, I.; Pitzer, R. M.; Shepard, R. Columbus - a Program System for Advanced Multireference Theory Calculations. *WIREs Comput. Mol. Sci.* **2011**, *1*, 191–199.

(14) Hu, H.; Jenkins, A. J.; Liu, H.; Kasper, J. M.; Frisch, M. J.; Li, X. Relativistic Two-Component Multireference Configuration Interaction Method with Tunable Correlation Space. *J. Chem. Theory Comput.* **2020**, *16*, 2975–2984.

(15) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-order Perturbation Theory with a CASSCF Reference Function. *J. Chem. Phys.* **1990**, *94*, 5483–5488.

(16) Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order Perturbation Theory with a Complete Active Space Self-Consistent Field Reference Function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(17) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-â.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. Multiconfigurational Perturbation Theory: Applications in Electronic Spectroscopy. **1996**, *93*, 219–331.

(18) Celani, P.; Werner, H.-J. Multireference Perturbation Theory for Large Restricted and Selected Active Space Reference Wave Functions. *J. Chem. Phys.* **2000**, *112*, 5546–5557.

(19) Shiozaki, T.; Mizukami, W. Relativistic Internally Contracted Multireference Electron Correlation Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4733–4739.

(20) Ma, D.; Li Manni, G.; Olsen, J.; Gagliardi, L. Second-Order Perturbation Theory for Generalized Active Space Self-Consistent-Field Wave Functions. *J. Chem. Theory Comput.* **2016**, *12*, 3208–3213.

(21) Vlaisavljevich, B.; Shiozaki, T. Nuclear Energy Gradients for Internally Contracted Complex Active Space Second-Order Perturbation Theory: Multistate Extensions. *J. Chem. Theory Comput.* **2016**, *12*, 3781–3787.

(22) Lu, L.; Hu, H.; Jenkins, A. J.; Li, X. Exact-Two-Component Relativistic Multireference Second-Order Perturbation Theory. *J. Chem. Theory Comput.* **2022**, *18*, 2983–2992.

(23) Malmqvist, P. Å.; Roos, B. O.; Schimmelpfennig, B. The Restricted Active Space (RAS) State Interaction Approach with Spin-orbit Coupling. *Chem. Phys. Lett.* **2002**, *357*, 230–240.

(24) Klene, M.; Robb, M. A.; Blancafort, L.; Frisch, M. J. A New Efficient Approach to the Direct Restricted Active Space Self-Consistent Field Method. *J. Chem. Phys.* **2003**, *119*, 713–728.

(25) Ivanic, J. Direct configuration interaction and multiconfigurational self-consistent-field method for multiple active spaces with variable occupations. I. Method. *J. Chem. Phys.* **2003**, *119*, 9364–9376.

(26) Ivanic, J. Direct configuration interaction and multiconfigurational self-consistent-field method for multiple active spaces with variable occupations. II. Application to oxoMn(salen) and N2O4. *J. Chem. Phys.* **2003**, *119*, 9377–9385.

(27) Fleig, T.; Olsen, J.; Marian, C. M. The Generalized Active Space Concept for The Relativistic Treatment of Electron Correlation. I. Kramers-Restricted Two-Component Configuration Interaction. *J. Chem. Phys.* **2001**, *114*, 4775–4790.

(28) Fleig, T.; Olsen, J.; Visscher, L. The Generalized Active Space Concept for The Relativistic Treatment of Electron Correlation. II. Large-Scale Configuration Interaction Implementation Based on Relativistic 2- and 4-Spinors And Its Application. *J. Chem. Phys.* **2003**, *119*, 2963–2971.

(29) Fleig, T.; Jensen, H. J. A.; Olsen, J.; Visscher, L. The Generalized Active Space Concept for The Relativistic Treatment of Electron Correlation. III. Large-Scale Configuration Interaction And Multiconfiguration Self-Consistent-Field Four-Component Methods with Application to UO2. *J. Chem. Phys.* **2006**, *124*, 104106.

(30) Knecht, S.; Jensen, H. J. A.; Fleig, T. Large-Scale Parallel Configuration Interaction. I. Nonrelativistic And Scalar-Relativistic General Active Space Implementation with Application to (Rb–Ba)$^+$. *J. Chem. Phys.* **2008**, *128*, 014108.

(31) Knecht, S.; Jensen, H. J. A.; Fleig, T. Large-Scale Parallel Configuration Interaction. II. Two- And Four-Component Double-Group General Active Space Implementation with Application to BiH. *J. Chem. Phys.* **2010**, *132*, 014108.

(32) Ma, D.; Li Manni, G.; Gagliardi, L. The Generalized Active Space Concept in Multiconfigurational Self-Consistent Field Methods. *J. Chem. Phys.* **2011**, *135*, 044128.

(33) Vogiatzis, K. D.; Li Manni, G.; Stoneburner, S. J.; Ma, D.; Gagliardi, L. Systematic Expansion of Active Spaces beyond the CASSCF Limit: A GASSCF/SplitGAS Benchmark Study. *J. Chem. Theory Comput.* **2015**, *11*, 3010–3021.

(34) Jensen, H. J. A.; Dyall, K. G.; Saue, T.; Fægri, K. Relativistic Four-component Multiconfigurational Self-Consistent-Field Theory for Molecules: Formalism. *J. Chem. Phys.* **1996**, *104*, 4083–4097.

(35) Thyssen, J.; Fleig, T.; Jensen, H. J. A. A Direct Relativistic Four-component Multiconfiguration Self-Consistent-Field Method for Molecules. *J. Chem. Phys.* **2008**, *129*, 034109.

(36) Knecht, S.; Legeza, O.; Reiher, M. Communication: Four-component Density Matrix Renormalization Group. *J. Chem. Phys.* **2014**, *140*, 041101.

(37) Bates, J. E.; Shiozaki, T. Fully Relativistic Complete Active Space Self-Consistent Field for Large Molecules: Quasi-second-order Minimax Optimization. *J. Chem. Phys.* **2015**, *142*, 044112.

(38) Almoukhalalati, A.; Knecht, S.; Jensen, H. J. A.; Dyall, K. G.; Saue, T. Electron Correlation within the Relativistic No-pair Approximation. *J. Chem. Phys.* **2016**, *145*, 074104.

(39) Reynolds, R. D.; Yanai, T.; Shiozaki, T. Large-scale Relativistic Complete Active Space Self-Consistent Field with Robust Convergence. *J. Chem. Phys.* **2018**, *149*, 014106.

(40) Battaglia, S.; Keller, S.; Knecht, S. Efficient Relativistic Density-Matrix Renormalization Group Implementation in a

Matrix-Product Formulation. *J. Chem. Theory Comput.* **2018**, *14*, 2353–2369.

(41) Knecht, S.; Jensen, H. J. A.; Saue, T. Relativistic Quantum Chemical Calculations Show that the Uranium Molecule U_2 has a Quadruple Bond. *Nat. Chem.* **2019**, *11*, 40–44.

(42) Jenkins, A. J.; Hu, H.; Lu, L.; Frisch, M. J.; Li, X. Two-Component Multireference Restricted Active Space Configuration Interaction for the Computation of L-Edge X-ray Absorption Spectra. *J. Chem. Theory Comput.* **2022**, *18*, 141–150.

(43) Sharma, P.; Jenkins, A. J.; Scalmani, G.; Frisch, M. J.; Truhlar, D. G.; Gagliardi, L.; Li, X. Exact-Two-Component Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2022**, *18*, 2947–2954.

(44) Hoyer, C. E.; Lu, L.; Hu, H.; Shumilov, K. D.; Sun, S.; Knecht, S.; Li, X. Correlated Dirac–Coulomb–Breit Multiconfigurational Self-Consistent-Field Methods. *J. Chem. Phys.* **2023**, *158*, 044101.

(45) Dyall, K. G.; Fægri, Jr., K. *Introduction to Relativistic Quantum Chemistry*; Oxford University Press, 2007.

(46) Reiher, M.; Wolf, A. *Relativistic Quantum Chemistry*, 2nd ed.; Wiley-VCH, 2015.

(47) Knowles, P.; Handy, N. A New Determinant-based Full Configuration Interaction Method. *Chem. Phys. Lett.* **1984**, *111*, 315 – 321.

(48) Olsen, J.; Roos, B. O.; Jørgensen, P. Determinant Based Configuration Interaction Algorithms for Complete and Restricted Configuration Interaction Spaces. *J. Chem. Phys.* **1988**, *89*, 2185–2192.

(49) Siegbahn, P. E. A new direct CI method for large CI expansions in a small orbital space. *Chem. Phys. Lett.* **1984**, *109*, 417–423.

(50) Zarrabian, S.; Sarma, C.; Paldus, J. Vectorizable approach to molecular CI problems using determinantal basis. *Chem. Phys. Lett.* **1989**, *155*, 183–188.

(51) Fales, B. S.; Levine, B. G. Nanoscale Multireference Quantum Chemistry: Full Configuration Interaction on Graphical Processing Units. *J. Chem. Theory Comput.* **2015**, *11*, 4708–4716.

(52) Fales, B. S.; Martínez, T. J. Efficient Treatment of Large Active Spaces through Multi-GPU Parallel Implementation of Direct Configuration Interaction. *J. Chem. Theory Comput.* **2020**, *16*, 1586–1596.

(53) Dobbyn, A. J.; Knowles, P. J.; Harrison, R. J. Parallel Internally Contracted Multireference Configuration Interaction. *J. Comput. Chem.* **1998**, *19*, 1215–1228.

(54) Vogiatzis, K. D.; Ma, D.; Olsen, J.; Gagliardi, L.; de Jong, W. A. Pushing Configuration-Interaction to The Limit: Towards Massively Parallel MCSCF Calculations. *J. Chem. Phys.* **2017**, *147*, 184111.

(55) Gao, H.; Imamura, S.; Kasagi, A.; Yoshida, E. Distributed Implementation of Full Configuration Interaction for One Trillion Determinants. *J. Chem. Theory Comput.* **2024**, *20*, 1185–1192.

(56) Davidson, E. R. The Iterative Calculation of a Few of the Lowest Eigenvalues and Corresponding Eigenvectors of Large Real-symmetric Matrices. *J. Comp. Phys.* **1975**, *17*, 87–94.

(57) Kozlowski, P. M.; Pulay, P. The Unrestricted Natural Orbital-Restricted Active Space Method: Methodology and Implementation. *Theor. Chem. Acc.* **1998**, *100*, 12–20.

(58) Williams-Young, D. B.; Petrone, A.; Sun, S.; Stetina, T. F.; Lestrange, P.;

Hoyer, C. E.; Nascimento, D. R.; Koulias, L.; Wildman, A.; Kasper, J.; Goings, J. J.; Ding, F.; DePrince III, A. E.; Valeev, E. F.; Li, X. The Chronus Quantum (ChronusQ) Software Package. *WIREs Comput. Mol. Sci.* **2020**, *10*, e1436.

(59) Visscher, L.; Dyall, K. Dirac-Fock Atomic Electronic Structure Calculations using Different Nuclear Charge Distributions. *Atomic Data and Nuclear Data Tables* **1997**, *67*, 207 – 224.

(60) Dyall, K. G. Interfacing Relativistic and Nonrelativistic Methods. I. Normalized Elimination of the Small Component in the Modified Dirac Equation. *J. Chem. Phys.* **1997**, *106*, 9618–9626.

(61) Dyall, K. G. Interfacing Relativistic and Nonrelativistic Methods. II. Investigation of a Low-Order Approximation. *J. Chem. Phys.* **1998**, *109*, 4201–4208.

(62) Dyall, K. G.; Enevoldsen, T. Interfacing Relativistic and Nonrelativistic Methods. III. Atomic 4-Spinor Expansions and Integral Approximations. *J. Chem. Phys.* **1999**, *111*, 10000–10007.

(63) Dyall, K. G. Interfacing Relativistic and Nonrelativistic Methods. II. One- and Two-Electron Scalar Approximations. *J. Chem. Phys.* **2001**, *115*, 9136–9143.

(64) Kutzlenigg, W.; Liu, W. Quasirelativistic Theory Equivalent to Fully Relativistic Theory. *J. Chem. Phys.* **2005**, *123*, 241102.

(65) Liu, W.; Peng, D. Infinite-Order Quasirelativistic Density Functional Method Based on the Exact Matrix Quasirelativistic Theory. *J. Chem. Phys.* **2006**, *125*, 044102.

(66) Peng, D.; Liu, W.; Xiao, Y.; Cheng, L. Making Four- and Two-Component Relativistic Density Functional Methods Fully Equivalent Based on the Idea of From Atoms to Molecule. *J. Chem. Phys.* **2007**, *127*, 104106.

(67) Ilias, M.; Saue, T. An Infinite-Order Relativistic Hamiltonian by a Simple One-Step Transformation. *J. Chem. Phys.* **2007**, *126*, 064102.

(68) Liu, W.; Peng, D. Exact Two-component Hamiltonians Revisited. *J. Chem. Phys.* **2009**, *131*, 031104.

(69) Liu, W. Ideas of Relativistic Quantum Chemistry. *Mol. Phys.* **2010**, *108*, 1679–1706.

(70) Li, Z.; Xiao, Y.; Liu, W. On the Spin Separation of Algebraic Two-Component Relativistic Hamiltonians. *J. Chem. Phys.* **2012**, *137*, 154114.

(71) Peng, D.; Middendorf, N.; Weigend, F.; Reiher, M. An Efficient Implementation of Two-Component Relativistic Exact-Decoupling Methods for Large Molecules. *J. Chem. Phys.* **2013**, *138*, 184105.

(72) Egidi, F.; Goings, J. J.; Frisch, M. J.; Li, X. Direct Atomic-Orbital-Based Relativistic Two-Component Linear Response Method for Calculating Excited-State Fine Structures. *J. Chem. Theory Comput.* **2016**, *12*, 3711–3718.

(73) Goings, J. J.; Kasper, J. M.; Egidi, F.; Sun, S.; Li, X. Real Time Propagation of the Exact Two Component Time-Dependent Density Functional Theory. *J. Chem. Phys.* **2016**, *145*, 104107.

(74) Konecny, L.; Kadek, M.; Komorovsky, S.; Malkina, O. L.; Ruud, K.; Repisky, M. Acceleration of Relativistic Electron Dynamics by Means of X2C Transformation: Application to the Calculation of Nonlinear Optical Properties. *J. Chem. Theory Comput.* **2016**, *12*, 5823–5833.

(75) Egidi, F.; Sun, S.; Goings, J. J.; Scalmani, G.; Frisch, M. J.; Li, X. Two-Component Non-Collinear Time-Dependent Spin Density Functional

Theory for Excited State Calculations. *J. Chem. Theory Comput.* **2017**, *13*, 2591–2603.

(76) Liu, J.; Cheng, L. Relativistic Coupled-Cluster and Equation-of-Motion Coupled-Cluster Methods. *WIREs Comput. Mol. Sci.* **2021**, *11*, 1536.

(77) Hoyer, C. E.; Hu, H.; Lu, L.; Knecht, S.; Li, X. Relativistic Kramers-Unrestricted Exact-Two-Component Density Matrix Renormalization Group. *J. Phys. Chem. A* **2022**, *126*, 5011–5020.

(78) Ehrman, J.; Martinez-Baez, E.; Jenkins, A. J.; Li, X. Improving One-Electron Exact-Two-Component Relativistic Methods with the Dirac–Coulomb–Breit-Parameterized Effective Spin–Orbit Coupling. *J. Chem. Theory Comput.* **2023**, *19*, 5785–5790.

(79) Urban, R.-D.; Bahnmaier, A. H.; Magg, U.; Jones, H. The Diode Laser Spectrum of Thallium Hydride (^{205}TlH and ^{203}TlH) in its Ground Electronic State. *Chem. Phys. Lett.* **1989**, *158*, 443–446.

(80) Wang, X.; Souter, P. F.; Andrews, L. Infrared Spectra of Antimony and Bismuth Hydrides in Solid Matrixes. *J. Phys. Chem. A* **2003**, *107*, 4244–4249.

(81) Schwerdtfeger, P. Metal-metal Bonds in Thallium(I)–Thallium(I) Compounds: Fact or Fiction? *Inorg. Chem.* **1991**, *30*, 1660–1663.

(82) Fægri Jr, K.; Visscher, L. Relativistic Calculations on Thallium Hydride. *Theor. Chem. Acc.* **2001**, *105*, 265–267.

(83) Pollak, P.; Weigend, F. Segmented Contracted Error-Consistent Basis Sets of Double- and Triple-ζ Valence Quality for One- and Two-Component Relativistic All-Electron Calculations. *J. Chem. Theory Comput.* **2017**, *13*, 3696–3705.

(84) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(85) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.