This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

The Journal of Chemical Physics

GronOR: Massively Parallel and GPU-Accelerated Non-Orthogonal Configuration Interaction for Large Molecular Systems

T. P. Straatsma,[1, a)] R. Broer,[2] S. Faraji,[2] R. W. A. Havenith,[2, 3] L. E. Aguilar Suarez,[2] R. K. Kathir,[2] M. Wibowo,[2] and C. de Graaf[2, 4, 5]

[1)]*National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6373, U. S. A.*

[2)]*Theoretical Chemistry Group, Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*

[3)]*Stratingh Institute for Chemistry, University of Groningen, Groningen, The Netherlands*

[4)]*Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, C. Marcel·lí Domingo 1, 43007 Tarragona, Spain*

[5)]*ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain*

(Dated: 27 January 2020)

GronOR is a program package for non-orthogonal configuration interaction calculations for an electronic wave function built in terms of anti-symmetrized products of multi-configuration molecular fragment wave functions. The two-electron integrals that have to be processed may be expressed in terms of atomic orbitals or in terms of an orbital basis determined from the molecular orbitals of the fragments. The code has been specifically designed for execution on distributed memory massively parallel and GPU-accelerated computer architectures, using an MPI+OpenACC/OpenMP programming approach. The task-based execution model used in the implementation allows for linear scaling with the number of nodes on the largest pre-exascale architectures available, provides hardware fault resiliency, and enables effective execution on systems with distinct CPU-only and GPU-accelerated partitions. The code interfaces with existing multi-configuration electronic structure codes that provide optimized molecular fragment orbitals, configuration interaction coefficients and the required integrals. Algorithm and implementation details, parallel and accelerated performance benchmarks, and an analysis of the sensitivity of the accuracy of results and computational performance to thresholds used in the calculations are presented.

[a)]Electronic mail: str@ornl.gov

I. INTRODUCTION

The quantum mechanical description of the electronic structure of molecular systems can be broadly categorized by two complementary approaches. In the valence bond (VB) method, molecular systems are described in terms of atomic orbitals (AO) with double or single occupation, and chemical bonds are thought of as arising from partially occupied overlapping atomic orbitals. This is an attractive approach from the perspective of an intuitive presentation of chemical bonding, as evidenced by the ubiquitous use of presenting molecules as Lewis structures, but leads to the need to use non-orthogonal orbital methods. Molecules with many partially occupied orbitals with minimal overlap, such as those containing lanthanides, actinides or transition metals, are among those effectively treated using such methods. However, the use of non-orthogonal methods increases the computational complexity of evaluating Hamiltonian matrix element contributions.

The second approach is to describe chemical bonding in terms of delocalized molecular orbitals (MO) expressed as expansions in terms of atom centered basis functions. The orthogonality of the MOs reduces the computational complexity but the wave function expansion in terms of a linear combination of Slater determinants is usually much longer than in a VB approach, and hence, the interpretation of chemical bonding in this approach is less intuitive. Also, the calculation of the Hamiltonian matrix elements is greatly simplified by the orthogonality condition, since one only has to consider determinant pairs that differ by at most two occupied orbitals. As a result, methods based on delocalized orthogonal MOs became more prevalent in computational chemistry research.

Recent methodological developments and the availability of unprecedented computational capabilities have stimulated renewed interest in applying methods based on non-orthogonal orbitals.[1,2] Orbital optimization and non-orthogonal configuration interaction (CI) methods, and ways to include dynamic correlation contributions using variational or perturbatation approaches are being designed that overcome some of the main challenges for non-orthogonal methodologies. A first challenge is that CI expansions of non-orthogonal determinants quickly become computationally expensive. Second, non-orthogonal determinants included in the CI expansion are typically hand-selected based on the attributes considered important for a specific problem. This may lead to shorter CI expansions, but it relies on the chemical intuition of the user. Third, a non-orthogonal approach relies on the description of each electronic configuration in its own optimal set of molecular orbitals. The convergence of such self-consistent single state wave func-

tions could become difficult for excited states. In an orthogonal CI approach, this problem is circumvented by using a common set of orbitals optimized for an average of the states under consideration.

As an alternative method to reliably finding the lowest electronic states in systems with many low-lying SCF solutions, Thom and Head-Gordon developed meta-dynamics method for locating multiple solutions to the SCF equations based on the use of biasing potentials to avoid convergence to already determined solutions.[3] These Hartree-Fock (HF) solutions resemble the diabatic electronic states of the systems and form a natural basis for CI calculations to produce adiabatic states, but the lack of orthogonality between the orbitals of different SCF solutions requires a non-orthogonal configuration interaction (NOCI) treatment.[4] This work has subsequently been extended to the calculation of multi-electron excitations[5] and core-excited states.[6,7]

A method for describing strongly correlated systems based on defining a linear combination of determinants generated from all possible spin-flip excitations of a high spin restricted open-shell Hartree-Fock (ROHF) wave function and for which, independently, all non-active-space orbitals were allowed to relax was proposed by Mayhall *et al.*,[8] and avoids potential difficulties with converging excited states.

A novel approach to orbital optimization for non-orthogonal wave functions based on the evaluation of the Hamiltonian matrix multiplied by a vector was developed by Olsen.[9] This method allows for the CI vector to be expressed in a bi-orthonormal basis resulting in a reduction of the computational complexity approaching that found in standard orthogonal approaches. Kähler and Olsen describe improved perturbational and variational approaches to include dynamic correlation in non-orthogonal reference states,[10,11] as implemented in LUCIA.[12]

A multi-reference strategy to include both static and dynamic correlation was proposed in the work by van Voorhis *et al.* in which a set of self-consistent HF determinants, modified by a first-order Møller Plesset (MP) perturbation treatment, are used to construct a Hamiltonian for a NOCI calculation.[13] HF optimized and perturbation-corrected ground and excited states of a molecule can thus be treated with a small number of non-orthogonal reference wave functions.

Here the algorithm as well as implementation and performance details of GronOR, a NOCI code developed by and named for a collaboration between the University of Groningen and Oak Ridge National Laboratory (ORNL), are presented. GronOR combines orthogonal and non-orthogonal approaches for describing the electronic structure of molecular assemblies in terms of individual molecular wave functions or molecular fragment wave functions. The molecular

component wave functions are generated for a range of excited or ionized states using any multi-configuration MO based approach such as complete active space SCF (CASSCF). Spin-adapted anti-symmetrized combinations of products of these molecular fragment wave functions provide the many-electron basis functions (MEBF) for subsequent NOCI calculations. This approach enables a convenient description of inter-molecular electron excitation, electron transfer, and energy transfer processes in terms of molecular states. Prediction of electron mobilities between molecules is possible through the calculation of electronic couplings between states of the assembly.

II. METHODOLOGY

In state-averaged CASSCF calculations, one single set of MOs is used to compute several the states of a given spatial and spin symmetry, with each state described by specific, optimized CI coefficients. For example, in a molecular system consisting of molecules A and B, ground (G) and first excited (X) states are described by

$$\begin{aligned}\Psi_G &= C_{G1}|...a_A^2 b_A^0 ... a_B^2 b_B^0 ...| + ... \\ \Psi_X &= C_{XA1}|...a_A^1 b_A^1 ... a_B^2 b_B^0 ...| + ... \\ &\quad + C_{XB1}|...a_A^2 b_A^0 ... a_B^1 b_B^1 ...| + ...\end{aligned} \tag{1}$$

in which the orthonormal MO sets $\{a_A, b_A\}$ for molecule A and $\{a_B, b_B\}$ for molecule B are the same for each state. The method described here is illustrated for only two molecules A and B, but the method and the implementation in GronOR is valid for any number of molecules.

In a non-orthogonal multi-configuration approach molecular states are separately optimized using any multi-reference wave function method such as CASSCF. For example, for molecules A and B individual wave functions for ground and excited states are determined

$$\begin{aligned}\Psi_{GA} &= C_{GA1}|...a_A^2 b_A^0 ...| + ... \\ \Psi_{XA} &= C_{XA1}|...{a'_A}^1 {b'_A}^1 ...| + ... \\ \Psi_{GB} &= C_{GB1}|...a_B^2 b_B^0 ...| + ... \\ \Psi_{XB} &= C_{XB1}|...{a'_B}^1 {b'_B}^1 ...| + ...\end{aligned} \tag{2}$$

which are then combined to

$$\begin{aligned}
\Psi_G &= \hat{A}|\Psi_{GA}\Psi_{GB}| \\
&= C_{G1}|...a_A^2 b_A^0 ... a_B^2 b_B^0 ...| + ... \\
\Psi_{XG} &= \hat{A}|\Psi_{XA}\Psi_{GB}| \\
&= C_{XG1}|...{a'_A}^1 {b'_A}^1 ... a_B^2 b_B^0 ...| + ... \\
\Psi_{GX} &= \hat{A}|\Psi_{GA}\Psi_{XB}| \\
&= C_{GX1}|...a_A^2 b_A^0 ... {a'_B}^1 {b'_B}^1 ...| + ...
\end{aligned} \tag{3}$$

leading to orbitals $\{a_A, b_A\}$ for ground states and $\{a'_A, b'_A\}$ for excited states which are mutually non-orthogonal. Moreover, the orbitals for molecule A are also not orthogonal to those of molecule B.

This approach allows for full orbital optimization for individual molecular states, and MOs as well as molecular CI expansion coefficients can be obtained from any multi-configuration method, like CASSCF. Proper inclusion of orbital relaxation and local correlation effects lead to a proper and more intuitive description of physical processes such as photo-excitation, induced charge separation, and the "singlet fission", process of formation of two molecular triplet states from a single high-energy photo-excitation.

The first step in GronOR is the reading of multi-reference molecular wave functions and construction of the spin-adapted anti-symmetrized product wave functions for the full molecular assembly, leading to wave functions as a, potentially extremely large, expansion in terms of determinants. Consider a molecular monomer CASSCF wave function with 500 determinants for each of two molecules in a molecular system. Combining these for the ensemble of two molecules leads to MEBFs with potentially 250,000 terms in the expansion. Calculation of a 4x4 Hamiltonian matrix over 4 such MEBFs would potentially involve 625 billion determinant pairs. Fortunately, this number can be significantly reduced by removing those determinants pairs for which the CI coefficient product falls below a certain threshold.

For the evaluation of individual matrix elements, first transformation matrices are determined that perform a corresponding orbitals transformation of the two orbital sets to two new biorthogo-

nal of so-called corresponding orbitals, i.e. from

$$\begin{aligned} \Delta_1 &= |\phi_1 \phi_2 ... \phi_N| \\ \Delta_2 &= |\psi_1 \psi_2 ... \psi_N| \\ S_{ij} &= < \phi_i | \psi_j > \end{aligned} \tag{4}$$

to

$$\begin{aligned} \Delta_1 &= |\phi'_1 \phi'_2 ... \phi'_N| \\ \Delta_2 &= |\psi'_1 \psi'_2 ... \psi'_N| \\ S'_{ij} &= < \phi'_i | \psi'_j > = \delta_{ij} \lambda_i \end{aligned} \tag{5}$$

The evaluation of individual matrix elements between non-orthogonal determinants is based on the method as implemented in the General Non-Orthogonal Matrix Element (GNOME) code.[14] Matrix elements of Hermitian one- and two-electron operators $O_1 = \sum_i O_1(i)$ and $O_{12} = \sum_{i<j} O_{12}(i,j)$ between non-orthogonal determinants Δ_1 and Δ_2, expressed in corresponding orbitals $\{\phi'\}$ and $\{\psi'\}$ respectively, can be written in terms of elements of the first- and second-order cofactor matrices $S(i,j)$ and $S(ik,jl)$ of the overlap matrix as follows

$$\begin{aligned} I_1 &= \sum_i^N \sum_j^N < \phi'_i | O_1 | \psi'_j > S(i,j) \\ I_2 &= \sum_{k>i}^N \sum_{l>j}^N < \phi'_i \phi'_k | (1 - p_{12}) O_{12} | \psi'_j \psi'_l > S(ik,jl) \end{aligned} \tag{6}$$

Factorization of the cofactor matrix and expression of the AOs ϕ' and ψ' in terms of atomic basis functions $\{\chi\}$ and $\{\chi'\}$ respectively,

$$\begin{aligned} \phi'_i &= \sum_p^m \chi_p c_{pi} \\ \psi'_i &= \sum_q^n \chi'_q c_{qi} \end{aligned} \tag{7}$$

leads to

$$\begin{aligned} I_1 &= \sum_p^m \sum_q^n < \chi_p | O_1 | \chi'_q > B(p,q) \\ I_2 &= \sum_{r>p}^m \sum_{s>q}^n < \chi_p \chi_r | (1 - p_{12}) O_{12} | \chi'_q \chi'_s > B(pr,qs) \end{aligned} \tag{8}$$

The fourth-order scaling in number of basis functions of the second order cofactor matrix $B(pr,qs)$ can be reduced by expression in factorized form,[15]

$$B(pr,qs) = (1-p_{pr})(1-p_{qs})F_{pq}(\omega)G_{rs}(\omega) \tag{9}$$

in which the functional form of $F(\omega)$ and $G(\omega)$ depends on the number of singularities ω in S' as follows. Note, that this formulation implies that a transformation towards integrals in terms of corresponding orbitals does not need to be carried out. Moreover, the one- and two-electron integral sets can be expressed in any suitable AO based or MO based basis set.

Without singularities, $\lambda_\alpha \neq 0$ for $\alpha = 1,..,N$, and

$$\begin{aligned} F(0)_{pq} &= \frac{1}{2}\sum_i c_{ip}d_{qi}\lambda_i^{-1} \\ G(0)_{pq} &= 2F(0)_{pq}\prod_\alpha^N \lambda_\alpha = 2|S|F(0)_{pq} \end{aligned} \tag{10}$$

With one singularity, $\lambda_\mu = 0$ and $\lambda_\alpha \neq 0$,

$$\begin{aligned} B(pq,rs) &= (1-p_{pr})(1-p_{qs})c_{p\mu}d_{\mu q}\sum_{i\neq\mu}^N c_{ir}d_{si}\sum_{\alpha\neq\mu,i}^N \lambda_\alpha \\ F(1)_{pq} &= \sum_i^N c_{ip}d_{qi}\lambda_i^{-1} \\ G(1)_{pq} &= c_{\mu p}d_{q\mu}\prod_{\alpha\neq\mu}^N \lambda_\alpha \end{aligned} \tag{11}$$

With two singularities, $\lambda_\mu = \lambda_\nu = 0$ and $\lambda_\alpha \neq 0$,

$$\begin{aligned} B(pq,rs) &= (1-p_{pr})(1-p_{qs})F(2)_{pq}G(2)_{rs} \\ F(2)_{pq} &= c_{\nu p}d_{q\nu} \\ G(2)_{pq} &= c_{\mu p}d_{q\mu}\prod_{\alpha\neq\mu,\neq\nu}^N \lambda_\alpha \end{aligned} \tag{12}$$

In case of multi-configuration molecular wave functions, the vast majority of Slater determinant combinations lead to two or more singularities and hence to $B(pq,rs) = 0$. This reduces enormously the number of determinant pairs contributing to I_2. In the original code, the basis sets $\{\chi\}$ and $\{\chi'\}$ are chosen to be identical and consisting of the atomic basis functions in terms of which the original orbital sets $\{\phi\}$ and $\{\psi\}$ are expanded.

Advantages of using NOCI includes the efficient evaluation of effective electronic couplings γ between adiabatic states as one of the important parameters in determining excitation energy and electron transfer rates, which are approximated by[16]

$$\gamma_{AB} \approx \frac{<\Psi_A|H|\Psi_B> - H^{av} <\Psi_A|\Psi_B>}{1-(<\Psi_A|\Psi_B>)^2} \tag{13}$$

with

$$H^{av} = \frac{<\Psi_A|H|\Psi_A> + <\Psi_B|H|\Psi_B>}{2} \tag{14}$$

The computational complexity can be significantly reduced by transformation of the one- and two-electron integrals to a MO basis. In the case of non-orthogonal wave functions, this is not as straightforward as in the case of orthogonal wave functions, but can be accomplished by transformation to a common MO basis as follows.

The superposition of the occupied (inactive + active) MOs trivially provides a complete basis to describe the different, non-orthogonal wave functions under consideration in the NOCI. However, this basis is obviously not the optimal choice, it contains linear dependencies and its dimension can become larger than the AO basis when many non-orthogonal states are considered. Therefore, a more compact basis is constructed by, for each molecule or fragment, diagonalizing the overlap matrix of all occupied MOs. Next, all eigenvectors with eigenvalues smaller than a certain threshold (see below) are discarded. The remaining eigenvectors are expressed in the AO basis and then used to re-express the non-orthogonal molecular states in the new common basis.[17] Simultaneously, the integrals can be transformed from the AO to the common MO basis. Note that although the new one-electron basis is made of orthogonal functions, the electronic states under consideration do not lose their non-orthogonality, they are only expressed in a new basis.

Construction of the molecular wave function from CASSCF or other multi-configuration fragment wave functions provides the means for treatment of static correlation effects. This approach does not, however, appropriately include dynamic correlation effects. Dynamic correlation basically affects the NOCI in two different ways. In the first place, the relative energies of the different MEBFs considered in the calculation can be rather strongly influenced by the inclusion of dynamic correlation. The largest effects are expected in the final NOCI wave function which may lead to an indirect effect on the interaction between the electronic states. Those MEBFs with a relative energy that is lowered by the dynamic correlation will gain importance. A second effect can be foreseen to arise from the change in the wave function by including dynamic correlation in the generation of the MEBFs, which does have a direct effect on the electronic coupling between the different electronic states.

At present, we are studying both effects in a detailed manner. Ideally one would use MEBFs constructed from the fragment wave functions with dynamic correlation, such as those obtained in a multi-configurational CI (MRCI) treatment, but this is problematic for two reasons. MRCI is typically only applicable for relatively small systems, and secondly, the correlated wave function is typically a linear combination of thousands (or millions) of determinants, much larger than the 500 determinants in the example described above. Therefore, we first consider the effect of the relative energies of MEBFs in the NOCI by simply replacing the diagonal matrix elements by the energies obtained with dynamic correlation (CASPT2, NEVPT2, or any other appropriate method), keeping the wave functions at the CASSCF level. This shifting of the diagonal elements is done in the orthogonal basis of the NOCI matrix, followed by either a full diagonalization to obtain the NOCI wave functions or a back transformation to the non-orthogonal basis of diabatic states to see how the dynamic correlation affects the coupling between the states.

To study the effect of dynamic correlation in the wave function we rely on effective Hamiltonian theory. The effect of the determinants that account for dynamic correlation is effectively mapped on the CAS in such a way that the diagonalization of the 'dressed' CAS gives eigenvalues that are the same as those of the full calculation. As a consequence, the corresponding eigenvectors (projections of the complete eigenvectors on the CAS) are no longer those of the 'undressed' CAS but have incorporated the effect of dynamic correlation. Among the different approaches to perform the dressing of the CAS matrix elements, the dynamic correlation dressed CAS(2) (DCD-CAS(2)) by Pathak, Lang and Neese[18] is currently under investigation to include the effects of dynamic correlation in the CAS wave function and subsequently in the coupling of the non-orthogonal MEBFs.

Although it is undeniable that dynamic correlation affects relative energies of ground and excited states, it should be noted that in some cases the change in the energy difference is in fact at least partly caused caused by the inability to describe a collection of electronic states with one set of MOs. A nice example is given by the study of the inter-valence charge transfer state in a bi-nuclear Fe(II)/Fe(III) complex by Domingo *et al.*[19] The CASPT2 correction to the state average CASSCF wave function is more than 80% and the excitation energy is completely unreliable (large variation with the level shift in an attempt to eliminate intruder states), while a state specific CASSCF approach (optimized orbitals for both electronic states) drastically reduces the effect of the CASPT2 correction to approximately 15% in the wave function and less than 0.1 eV in the relative energy without need of applying level shifts. In the NOCI approach this orbital relaxation

is already fully accounted for in the monomer wave function used to construct the MEBFs, and hence, counts with important advantages compared to standard orthogonal methods.

III. IMPLEMENTATION

The implementation of the NOCI method in GronOR was designed to be interfaced with electronic structure codes that can provide fragment wave functions in terms of Slater determinants and the one and two-electron integrals for the full molecular system. To provide the capability to treat large molecular systems, the implementation was designed from the outset for use on massively parallel and GPU-accelerated architectures.[20]

The first part of a NOCI calculation is to generate anti-symmetrized product determinants and their associated coefficients from the molecular fragment wave functions provided by previously carried out multi-configuration SCF calculations such as CASSCF or one of its variants. Whereas the calculation of the fragment ground state wave functions does in general not imply any special difficulty, the optimization of the orbitals and the CI expansion for excited states may become more troublesome when this state is not the lowest excited state. In these cases, root flipping may avoid the straightforward convergence to the single-state solution and special attention needs to be paid to the optimization procedure. One can either follow the strategy described in Refs. 21 and 22, or rely on root selection in the orbital optimization step based on maximum overlap with the previous iteration.

From molecular fragment wave functions with different spin states, combinations can be constructed using the appropriate spin-coupling coefficients that result in a targeted overall spin state of the MEBFs describing the molecular system. The generation of these MEBFs is computationally inexpensive.

The second and computationally most demanding step is the calculation of the Hamiltonian and overlap matrix elements. These calculations involve the processing of large numbers of two-electron integrals. Since the Hamiltonian matrix elements contain a large number of determinants pair contributions that can be evaluated independently, this part of the calculation is relatively straightforwardly parallelizable. The task-based master-worker model in GronOR asynchronously processes these determinant pair contributions in batches by the groups of worker processes. The number of determinant pair contributions per batch is user-specified and should be chosen sufficiently large so that the number of communication operations to the master process is not leading

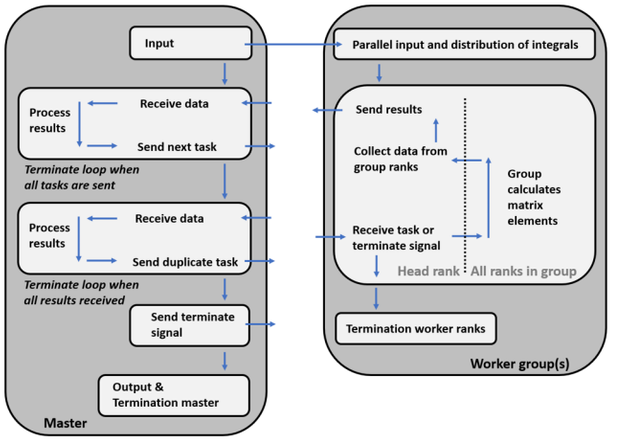


FIG. 1. Schematic of the task-based workflow implemented in GronOR illustrating the sending of tasks from the master to the worker groups, followed by the sending of duplicates of still outstanding tasks that guarantees completion of the task list upon hardware failure.

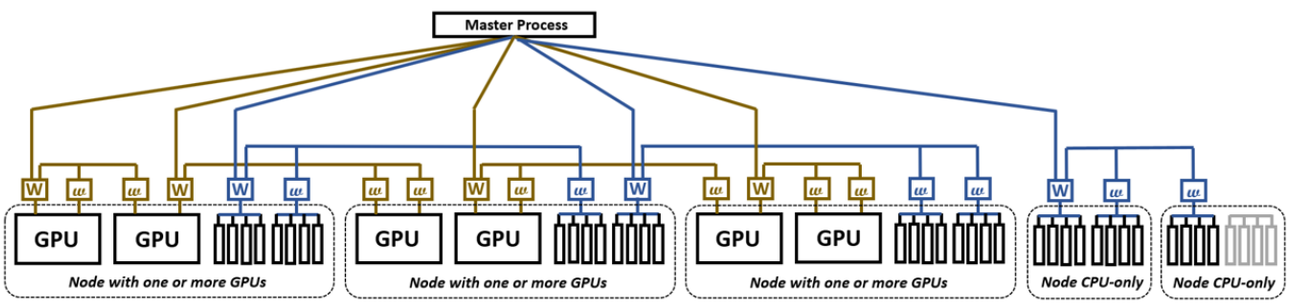


FIG. 2. Groups of MPI processes consisting of a single head process (W) and multiple worker processes (*w*) can span across nodes. Within one group all processes have the same compute resources. GPU-enabled groups are shown in gold, and CPU-only groups in blue. In the current implementation all groups have the same number of MPI processes.

to network contention, but sufficiently small to benefit from the load-balancing enabled from asynchronous processing.

GronOR uses a two-tiered task-based programming model, illustrated in Figure 1, with one master process determining the calculations performed on and collecting results from groups of worker processes. This approach enables effective load balancing between the groups of worker processes, and provides opportunities for hard-fault resiliency. Inter-process communication is implemented using the MPI message passing interface, while computations within a process are or-

chestrated using OpenMP[23] for CPU threading and OpenACC[24] for GPU-accelerator off-loading. The high-level process layout is illustrated in Figure 2. Each group of worker processes is controlled by a single head-process that handles all communication with the master process. Each worker process within a group has the same compute resources available, to avoid load imbalance within a group. Groups can span across nodes as illustrated. In the current implementation each group has the same user-defined number of worker processes. The list of determinant pairs that contribute to the Hamiltonian matrix elements is available to all worker processes. Each time a worker process group is available for the evaluation of a batch of contributions, its head-process requests from the master process an index into this list and the number of contributions to be calculated, and sends this information to the other processes within the group. After processing the batch of contributions, the head process collects all results from the other processes in the group and returns to the master process the combined contribution to the Hamiltonian and overlap matrix elements. Organized in this way, for each batch of calculated contributions only two small messages are required to be sent, namely two integer values at the start and two real values at the end.

Typically the number of processes within a group is determined by the amount of available memory per process to hold the two-electron integrals. The current implementation requires the number of processes per group to be equal for each group. This allows each process in one particular group to read the integrals in parallel from the two-electron integral file(s) and broadcast the integrals to the corresponding process in all other groups.

The largest data structure is the list of two-electron integrals. These integrals are read from file once at the start of a job and stored in memory for the duration of the calculation. Storing integrals in the AO basis can lead to significant memory requirements. The code can be compiled to store integrals in single rather than double precision, but since labels and integrals are stored this reduces the memory requirements to only 75%. Performing a transformation to a MO basis significantly reduces the number of integrals, making the NOCI methodology applicable to much larger molecular systems than heretofore possible. For GPU-accelerated systems the memory use is typically determined by the available memory on the accelerator. In the current implementation MPI processes that have access to a GPU-accelerator will use local DDR memory to store the integrals in addition to the copy that is kept in High Bandwidth Memory (HBM) memory on the GPU.

For most molecular systems of interest the number of determinant pair contributions to the

Hamiltonian and overlap matrices can be so large that each worker process group has a large number of batches to evaluate. Since each new batch is assigned by the master process as soon as a worker group has completed the previous batch, each worker group remains busy and the calculation is overall well load balanced. GPU-accelerated computer systems typically have separate memory domains for the CPU and the GPU. By storing the two-electron integral data in both memory domains, for processes that have access to a GPU-accelerator the calculation of individual determinant pair contributions can be balanced between both the CPU and GPU. This provides an additional level of dynamic load balancing in GronOR.

The assignment of tasks by the master process to the worker groups is completely asynchronous. This is not only an efficient load balancing scheme, but it also facilitates fault resilient execution. On the master process a list is kept of all outstanding determinant pair batches. When towards the end of the run all batches have been assigned, the master process is assigning duplicates of still outstanding tasks whenever a request for a new task comes from one of the worker groups. The master process continues assigning duplicates until all expected matrix element contributions have been completed and returned. If one of the worker process groups fails as a result of some hardware fault, the last batch it was assigned will be reassigned to another group. Only after all contributions have been received by the master process, it signals all worker processes asynchronously to exit. Using this scheme, the code has been demonstrated to be hardware fault resilient.

The calculation of the determinant pair contributions has been ported to GPU-accelerators using the OpenACC directives programming model. Work is currently ongoing to develop a version of GronOR in which OpenMP target off-loading directives will be available for computer systems with non-NVIDIA accelerators.

The latest GPU-accelerators have access to increasing amounts of local memory. For example, the latest NVIDIA V-100 GPU has 32 GB of HBM memory. We have extensively tested the use of NVIDIA's multi-process server of MPS capability, which allows multiple MPI processes to share the GPU, giving the opportunity for additional computational performance.

The implemented algorithm has two solvers for which GronOR can use external libraries. A singular value decomposition and a matrix diagonalization can be carried out on the GPU using the CUSOLVER[25] library.

GronOR requires an interface to an electronic structure code for the multi-configuration SCF vectors and state-specific orbitals for molecular fragments, as well as the one and two-electron

integrals for the full molecular system. The current version of the code provides for interfaces to GAMESS-UK[26] and MolCAS[27] for the coefficients, orbitals and integrals, with SYMOL[28] as an alternative for the integrals. GAMESS-UK and MOLCAS can also be used for the integral transformation to a common MO basis.

The largest data set to be read is the set of labeled two-electron integrals. These integrals can be read from multiple files in a parallel fashion as described above. One electron integrals are read from a separate file, as are geometry and basis set information.

IV. VALIDATION

In order to validate results calculated with GronOR, a set of single point energy verification test runs against an existing commercially available package was performed. For that purpose, the energies of the ground (S_0) and lowest singlet excited (S_1 states for a set of test molecules (pyridine, furan, butadiene and benzene shown in Figure 3) were calculated with GronOR and MolCAS 8.0.[27] Table I reports the state energies calculated with GronOR and ΔE which represents the difference between the state energies calculated with GronOR and Molcas. Results show that GronOR is able to reproduce accurately state energies for a diversity of molecules when compared with MolCAS.

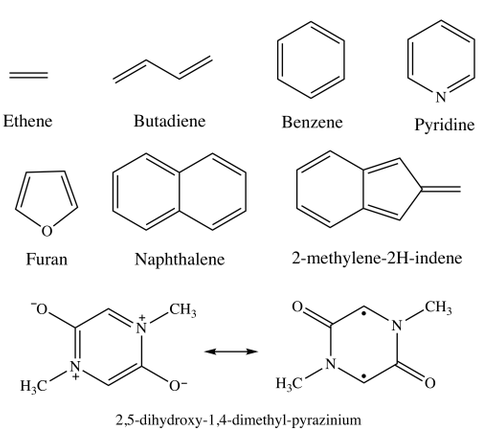


FIG. 3. Structures of compounds discussed in this contribution.

The spin coupling schemes implemented in GronOR were tested using an ethene dimer (monomer structure is shown in Figure 3). For each monomer named as A and B, the molecular wave func-

TABLE I. Comparison of the ground (S_0) and singlet excited (S_1) state energies (in a.u.) calculated with GronOR and MolCAS for a set of test molecules. ΔE = E(GronOR) - E(MolCAS)

Molecule	State	GronOR	ΔE
Pyridine	S_0	-246.59215	-3.8×10^{-8}
	S_1	-246.07947	3.2×10^{-7}
Furan	S_0	-228.54493	5.6×10^{-8}
	S_1	-228.06416	5.3×10^{-7}
Butadiene	S_0	-154.92293	-4.1×10^{-8}
	S_1	-154.59647	3.9×10^{-8}
Benzene	S_0	-230.66621	2.0×10^{-8}
	S_1	-230.35349	6.7×10^{-8}

tions for the S_0 and S_1 states were obtained at the CASSCF(2,2)/6-31G level of theory, whereas a HF wave function was used to describe the T_1 state. These molecular wave functions were subsequently used to construct the following MEBFs: $|\Psi_A^{S_1}\Psi_B^{S_0}>$, $|\Psi_A^{S_0}\Psi_B^{S_1}>$, and $|\Psi_A^{T_1}\Psi_B^{T_1}>$. The Hamiltonian and overlap matrix elements between the MEBFs were calculated with GronOR and TURTLE,[29] a program of the GAMESS-UK package. The GronOR results are shown in Table II. TURTLE gives exactly the same values as those shown in the two tables, which verifies the correct implementation of the spin coupling schemes in GronOR.

As any quantum chemical code, GronOR counts with several thresholds to control the precision of the results. The most significant ones in GronOR are the threshold for considering an integral to be zero, the threshold on singularities in the co-factors, the linear dependency threshold of the common MO basis and the threshold for considering a pair of determinants in the calculation of the matrix elements based on the product of the CI coefficients of the two determinants. Because the evaluation of the matrix element contributions is implemented in a fully asynchronous manner, and with minimal communication required for batches of such contributions, variations in the first two thresholds, while reducing the time to solution, do not affect the computational parallel

TABLE II. Hamiltonian H (in a.u.) and overlap S matrix elements for the ethene dimer.

H	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	$\|\Psi_A^{T_1}\Psi_B^{T_1}>$
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	-155.46532		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	0.03199	-155.51962	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	-0.37510	-0.36960	-155.74927
S	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	$\|\Psi_A^{T_1}\Psi_B^{T_1}>$
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	1.000000		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	-0.000205	1.000000	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	0.002400	0.001180	1.000000

efficiency. Moreover, the number of two-electron integrals with an absolute value below 10^{-9} is usually rather small, less than 5% in all the cases that we have treated so far. Therefore, using this threshold for considering an integral as zero will not reduce the computational effort. Comparable thresholds in other programs are at least two orders of magnitude smaller (10^{-14} in MolCAS and $2.5 \cdot 10^{-11}$ in Orca), which makes it not recommendable to use larger threshold to reduce the computational cost. Two other thresholds, τ_{MO} and τ_{det}, have a stronger impact on the accuracy of the calculation and choosing these thresholds wisely is essential to balance precision and time to solution. Their influence on the results has been carefully calibrated in Ref. 17 and here we will shortly summarize the most important findings for one of the six test systems, namely the naphthalene dimer with CAS(4,4) monomer wave functions for the τ_{MO} study and CAS(6,6) wave functions to establish the dependence on τ_{det}.

The common orbital threshold τ_{MO} controls the size of the common molecular orbital basis. Larger thresholds leads to a smaller basis to express the different non-orthogonal states, but also reduces the number of integrals and speeds up the calculation. Figure 4 illustrates how the electronic coupling V_{ij} of the $|\Psi_A^{S_1}\Psi_B^{S_0}\rangle$ and $|\Psi_A^{T_1}\Psi_B^{T_1}\rangle$ states of the naphthalene dimer evolves as function of τ_{MO} together with the evolution of the computer time required to calculate the coupling. Up to values of 10^{-4}, the result is practically the same as the one obtained with the full AO basis marked

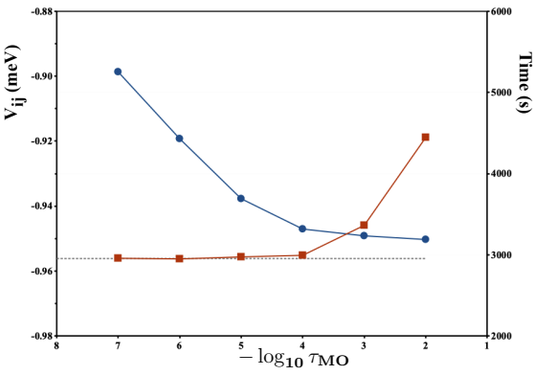


FIG. 4. Electronic coupling V_{ij} (in meV, red squares) of the $|\Psi_A^{S_1}\Psi_B^{S_0}\rangle$ and $|\Psi_A^{T_1}\Psi_B^{T_1}\rangle$ states of the naphthalene dimer and wall-clock time (in s, blue circles) as function of τ_{MO}. The dashed grey line is the coupling obtained with the AO basis.

in the graph by the horizontal dashed line. The coupling starts to deviate for larger thresholds, but even for $\tau_{MO} = 10^{-2}$ the deviation from the reference value is still rather small; 0.92 versus 0.96 meV. The computer time steadily decreases with increasing thresholds and levels off around 10^{-4}, the calculation in the AO basis takes 9120 sec. Much larger savings were observed for bigger systems, for which as a matter of fact the calculation in the AO basis could not be performed within a reasonable execution time.

The second important threshold to control the balance between accuracy and computational efficiency is the τ_{det} parameter, which filters the pairs of determinants in the *bra* and *ket* of the NOCI matrix elements based on the product of the CI coefficients in the MEBFs. Only if this product is larger than the threshold, the determinant pair will be included in the calculation of the matrix element between the MEBFs under consideration. The red squares in Figure 5 depict the number of determinant pairs in the calculation of the $|\Psi_A^{S_0}\Psi_B^{S_0}\rangle$ diagonal matrix element of the naphthalene dimer as function of the τ_{det} parameter, using $\tau_{MO} = 10^{-4}$. Although the total energy of the $\Psi_A^{S_0}\Psi_B^{S_0}$ MEBF steadily increases with increasing threshold ($+1.7 \cdot 10^3 E_\mathrm{h}$ for $\tau_{det} = 10^{-5}$), the relative energies stay within 10 meV of the the reference calculation up to thresholds of 10^{-4}. The off-diagonal matrix elements are even more stable with variations smaller than 1 meV for $\tau_{det} = 10^{-4}$, indicating that indeed a large part of the determinant pairs can be safely ignored leading to a substantial saving in the computational time as indicated by the blue circles in Figure 5.

Finally, the extent to which memory requirements can be lowered was tested by considering

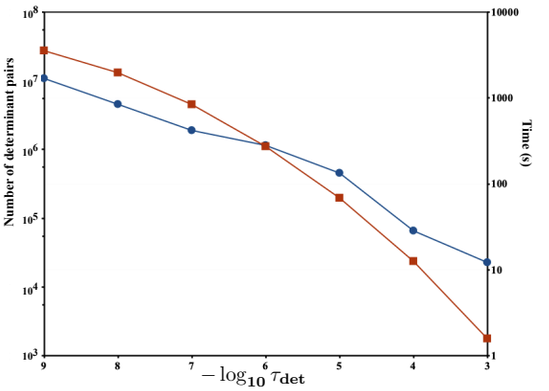


FIG. 5. Number of determinant pairs (red squares) and wall-clock time (in s, blue circles) in the calculation of the $\langle \Psi_A^{S_0}\Psi_B^{S_0}|\hat{H}|\Psi_A^{S_0}\Psi_B^{S_0}\rangle$ matrix element of the naphthalene dimer as function of the τ_{det} parameter.

the integrals as reals with single precision. Comparison of the matrix elements did not show significant changes compared to the standard double precision algorithm, however, the time gain was very little and unless performing calculations on a machine with very little memory, the single precision integrals do not provide any additional performance improvement.

V. SCALABILITY AND PERFORMANCE

Summit is a 200 PFlop IBM/NVIDIA/Mellanox supercomputer in the Oak Ridge Leadership Computing Facility (OLCF) at ORNL in Oak Ridge, Tennessee. Summit ranked number one on the Top500 list of supercomputers in November 2018 and again in June 2019.[30] It consists of 4608 nodes with dual socket IBM Power9 processors and six NVIDIA V-100 GPU accelerators. Each of the two Power9 CPUs is linked through an on-node NVLINK interconnect to three of the GPUs. Each GPU has 16 GB of HBM memory, and the node memory consists of 512 GB DDR4 and 1,600 GB non-volatile NVRAM.

Benchmarking on Summit was carried out for a molecular system consisting of two naphthalene molecules at 7 a.u. separation, as shown in Figure 6. The naphthalene geometry was optimized at the density functional theory (DFT) level using the B3LYP functional.[31,32] CASSCF calculations using a 6-311G basis set were carried out with eight electrons in eight orbitals (CAS(8,8)) for one of the molecules (A) and four electrons in four orbitals (CAS(4,4)) for the other (B), labeled in the tables and figures as CAS(8,8;4,4). For the individual molecules, CASSCF calculations were performed for the S_0, $S_1(^1B_{1u})$ and T_1 states from which four MEBFs $|\Psi_A^{S_0}\Psi_B^{S_0}>$, $|\Psi_A^{S_1}\Psi_B^{S_0}>$,

$|\Psi_A^{S_0}\Psi_B^{S_1}>$, and $|\Psi_A^{T_1}\Psi_B^{T_1}>$ were constructed. Using a DZ basis, the number basis functions is 308, leading to a total of 1.1 billion two-electron integrals of which 149 million are non-zero. With labels, this results in 5.6 GB of integral data. Calculation of the single matrix element $<\Psi_A^{S_0}\Psi_B^{S_0}|H|\Psi_A^{S_0}\Psi_B^{S_0}>$ scales linear with the number of Summit nodes used, as illustrated in Figure 7 by improved benchmark results from our earlier reported timings[20], and the performance improvement from using the GPU-accelerators is a factor of 6.8 when using 1024 nodes.

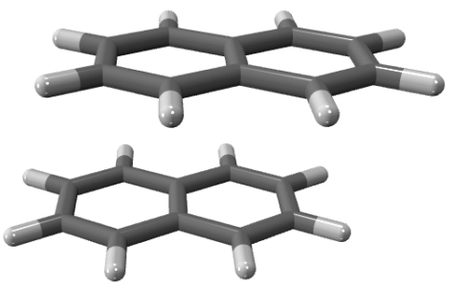


FIG. 6. Relative orientation of the naphthalene dimer, with the molecular planes at 7 a.u. separation.

The total number of determinant pair contributions to be evaluated for the 4x4 Hamiltonian is 2.1 billion, and the individual contributions for each of the Hamiltonian and overlap matrix elements are given in Table III. Resulting electronic couplings are given in Table IV. The time to solution as a function of the number of Summit nodes is given in Table V and Figure 8 and illustrates the near-linear scaling that results from the fully asynchronous evaluation of determinant pair contributions. On Summit nodes, roughly 90% of the floating point operations are provided by the GPU accelerators, such that from a floating point perspective a GPU-accelerated run could achieve a 10-fold speedup compared to CPU-only execution. The factor of 6.8 found for GronOR comparing 28 MPI ranks on the CPU with 6 ranks using the GPU per node, which would be a 32-fold speedup comparing execution on a single rank with and without GPU acceleration, compares favorably with the applications ported as part of the Summit Center for Accelerated Application Readiness.[33] For example, reported speedups for computational chemistry codes LS-Dalton, NWChem, and NAMD are 2.1, 5.0, and 5.7, respectively.

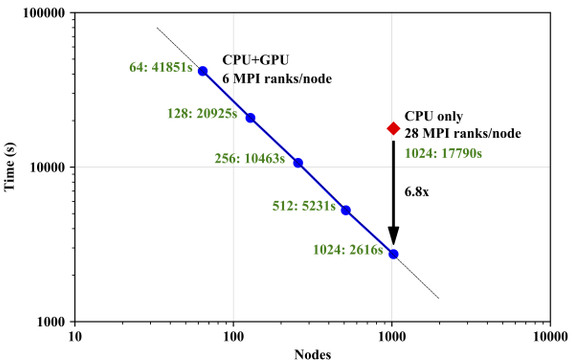


FIG. 7. Scaling of a non-orthogonal CI calculation of the single Hamiltonian matrix element $< \Psi_A^{S_0}\Psi_B^{S_0}|H|\Psi_A^{S_0}\Psi_B^{S_0} >$ for a CAS(8,8;4,4) naphthalene dimer obtained as time to solution in seconds as a function of the number of nodes on Summit, illustrating near-linear scalability and GPU-acceleration with a factor of 6.8 on 1024 nodes.

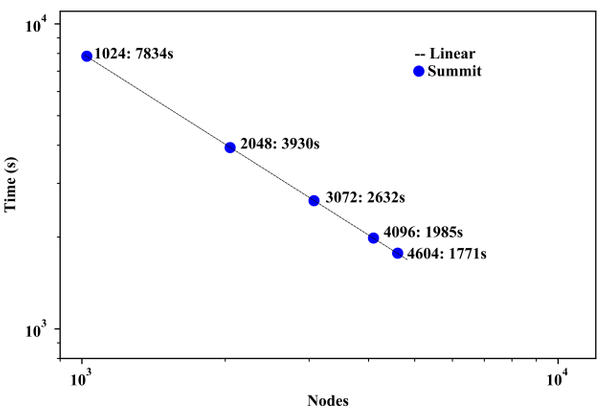


FIG. 8. Scaling of a NOCI calculation of the 4x4 Hamiltonian matrix for a CAS(8,8;4,4) naphthalene dimer obtained as time to solution in seconds as a function of the number of nodes on Summit.

VI. APPLICATIONS

Among the first applications of NOCI based on the GNOME algorithm were the studies of Broer and Nieuwpoort of oxygen core and valence hole states in CrO_4^-.[34,35] For O-1s core hole states, symmetry adapted HF wave functions do not give the most adequate description of the core-hole state because lifting the symmetry restrictions and allowing for the hole to localize on one of the oxygen atoms results in an important energy lowering. To restore the symmetry four equivalent wave functions were generated with the hole localized on one of the four atoms. These non-orthogonal wave functions were then combined through NOCI to obtain a wave function with

TABLE III. Number N of determinant pair contributions per Hamiltonian element, Hamiltonian H (in a.u.) and overlap S matrix elements for the naphthalene dimer at CAS(8,8;4,4).

N	$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	$\|\Psi_A^{T_1}\Psi_B^{T_1}>$
$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	112,867,800			
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	219,230,208	106,470,528		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	150,480,384	146,153,472	50,165,136	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	386,537,472	375,422,976	257,691,648	330,977,856
H	$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	$\|\Psi_A^{T_1}\Psi_B^{T_1}>$
$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	-766.68211			
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	0.66647	-766.42812		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	0.82800	-0.05931	-766.44162	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	-0.38469	0.26580	0.23744	-766.39464
S	$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	$\|\Psi_A^{T_1}\Psi_B^{T_1}>$
$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	1.00000			
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	-0.00087	1.00000		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	-0.00108	0.00008	1.00000	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	0.00050	-0.00035	-0.00031	1.00000

the correct spatial symmetry, and with full orbital relaxation due to the presence of the core-hole. Whereas the interaction between the four wave functions that describe these localized core-holes is not very large, the situation is quite different in the case of valence holes. In this case, orbital relaxation due to the localization of the hole on one of the atoms and the subsequent symmetry restoration by NOCI are of equal importance and treating both in a rigorous way allowed the

TABLE IV. Electronic couplings (in a.u.) for the naphthalene dimer at CAS(8,8;4,4).

	$\|\Psi_A^{S_0}\Psi_B^{S_0}>$	$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	$\|\Psi_A^{S_0}\Psi_B^{S_1}>$
$\|\Psi_A^{S_1}\Psi_B^{S_0}>$	0.00131		
$\|\Psi_A^{S_0}\Psi_B^{S_1}>$	0.00155	-0.00159	
$\|\Psi_A^{T_1}\Psi_B^{T_1}>$	-0.00022	0.00018	0.00017

TABLE V. Time to solution (in sec) for the naphthalene dimer at CAS(8,8;4,4) obtained as a function of nodes with 6 MPI processes per node on Summit.

Summit nodes	MPI processes	Wall Clock Time (sec)
1,024	6,144	7,834
2,048	12,288	3,930
3,072	18,432	2,632
4,096	24,576	1,985
4,604	27,624	1,771

authors to give a satisfactory description of the X-ray photo-electron spectroscopy (XPS) spectrum of Na_2CrO_4.

Similar inspections were made to rationalize the Ni-3s and Mn-3s XPS spectra in NiO and MnO, respectively.[36,37] There it was shown that the screening of the core-hole by the oxygen ligands can effectively be treated by NOCI. The wave function of the ionized system was built as a linear combination of TM-$3s^1 3d^n$ and TM-$3s^1 3d^{n+1}L^{-1}$ determinants, where TM is Ni or Mn and L^{-1} denotes an electronic configuration with a hole in the O-2p orbitals. By separately optimizing the charge transfer and non charge transfer determinants, a full account of the orbital relaxation could be obtained. In the 7x7 NOCI (the TM-$3s^1 3d^n$ determinant and six charge transfer determinants, one for each of the six oxygen atoms around the TM^{2+} ion) the symmetry was restored and the interaction between the determinants accounted for. The relative energies of the

main peak and the satellites, and their relative intensities estimated by the sudden approximation were in quite good agreement with experiment. Moreover, quantitative estimates could be given of the importance of the screening effects in the different final states observed in the XPS spectra.

Next in the early applications, are the NOCI estimates of the magnetic coupling parameter in the La_2CuO_4 and related compounds.[38–40] The isotropic magnetic coupling between two localized, spatially separated spin moments is adequately described by the Heisenberg Hamiltonian in most cases. This model Hamiltonian reads $\hat{H} = -J\hat{S}_1\hat{S}_2$ and *ab initio* calculations are widely used to estimate J, the coupling strength. Oxygen to copper charge transfer configurations are known to play an important role in magnetically connecting neighboring copper ions in these compounds, but this effect is not easily incorporated in a wave function based on orthogonal orbitals. Only when these charge-transfer configurations are combined with excitations from inactive to virtual orbitals, they gain significant weight in the wave function.[41,42] To avoid such lengthy wave functions expansions, the NOCI approach takes a different route. Ground state and charge transfer configurations are both expressed in their own optimal orbitals and mixing the ground state configuration with the relaxed charge transfers leads to magnetic coupling parameters that are in remarkably good agreement with experiment for such a small wave function expansion: less than 10 determinants in NOCI versus several millions in the approaches based on orthogonal orbitals.

The third early application focused on the calculation of the hopping probability of electrons (or holes) in strongly correlated materials and the subsequent construction of a many-electron band structure.[43–45] NOCI was applied to calculate the electron coupling between the A^+—B and A—B^+ states, often referred to as the hopping parameter t. Neutral and ionized fragment wave functions were generated with a standard CASSCF approach. These A, A^+, B and B^+ wave functions were then combined as spin-adapted anti-symmetrized linear combinations to form wave functions for the whole cluster. Typically the fragments overlap in space and a corresponding orbital transformation was performed to remove the orbitals that appear in both fragments. The calculation of the interaction matrix element $\langle AB^+|\hat{H}|A^+B\rangle$ and the overlap integral between the two non-orthogonal MEBF leads to an estimate of t_{AB}. After calculating the overlap and hopping parameter for different combinations of sites (along the different crystallographic directions, nearest and next-nearest neighbor hopping), one can determine the energy dependence of the N-electron states as function of the momentum vectors k in a tight-binding approach. The resulting band structure differs from the usual band structures by the fact that here many-electron bands are calculated taking explicitly into account orbital relaxation and electron correlation. Normally

these effects are hidden in the effective one-electron model adopted in band structure calculations.

More recent application of NOCI focused on the calculation of effective electronic couplings between diabatic states applied to singlet fission process.[46] The main advantage of employing NOCI approach to study singlet fission is a clear chemical interpretation of the diabatic state in terms of molecular states. Further, it allows one to investigate the effect of charge transfer states on the computed coupling. NOCI was applied to calculate the effective electronic couplings in a biradicaloid molecule, namely the bis(inner salt) of 2,5-dihydroxy-1,4-dimethyl-pyrazinium (Figure 3), which on the basis of quantum chemical calculations, satisfies the energetic criteria for a singlet fission chromophore. The computed couplings on the pair of molecules with π-stack arrangement are sufficiently large for singlet fission to occur.

GronOR was also employed to study the singlet fission process in 2-methylene-2H-indene (Figure 3), a new recognized singlet fission molecule, using a NOCI approach.[47] Four different pair arrangements were identified within a theoretically predicted crystal structure of the molecule. Calculated effective electronic couplings on the four pairs of molecules suggest the efficient formation of the so-called ^1TT state in the crystal structure, which is promising for applications in singlet fission-enhanced solar cells. Additionally, in that contribution, a comparison of the NOCI results with two other theoretical approaches, i.e. restricted active space with two spin flips and the *ab initio* Frenkel–Davydov exciton model as implemented in Q-Chem electronic structure package,[48] reveals that the NOCI approach is able to differentiate between pairs of molecules with low and high singlet fission probabilities. Very recently, the method has also been used in the study to rationalize the factors that maximize the S_0S_1 to ^1TT conversion in molecules with extended π systems[49].

VII. DISCUSSION

The implementation of NOCI in GronOR using a batched and task-based algorithm with directive-based offloading to GPUs has been demonstrated to achieve near-perfect scalability and good accelerated performance on Summit, the largest supercomputer available for open science in the world. Methodological improvements, such as the transformation to a molecular orbital basis, have further reduced the time to solution for such calculations. The benchmarking results presented above illustrate that NOCI calculations with very large numbers of determinants are now feasible for molecular assemblies of interest for, among others, energy materials applications.

The current version of the GronOR code forms an excellent basis for several implementation and methodological developments that will further improve its accelerated performance and portability to other architectures, as well as the availability of new technical features and properties.

The first step in a further reduction of the computational cost is the implementation of a frozen core. The orbitals of the core electrons can be considered to a large extent to be identical in all the electronic states on the same fragment and virtually orthogonal to the core orbitals on other fragments. Hence, the contribution to the NOCI matrix elements of the core electrons can be determined with standard orthogonal approaches without having to use the heavy machinery that is needed when considering non-orthogonal orbitals. To further increase the efficiency of GronOR, the use of Cholesky decomposed integrals[50,51] will be implemented in the NOCI approach. This will not only improve the performance of the NOCI itself, but also drastically reduce the computational cost of the transformation of the integrals to the common MO basis. Doing this transformation in the conventional way can become prohibitive when considering systems with a large number of electrons. Alternatively, the possibility to use other schemes based on the resolution of the identity (RI-methods) to lower the computational cost of the approach will also be explored.

When studying inter-molecular electron transport and exciton delocalization in molecular crystals such as those that show singlet fission described in the previous section, the electronic states that are used to construct the MEBFs are localized on discrete molecules and as such clearly defined identities. This becomes slightly more complicated when one decides to study intramolecular charge transfer and energy transport processes. Taking the charge transfer as example, a molecule can be thought of a donor (D) and an acceptor (A) part, connected through a linker (L). By dividing the molecule A–L–B in two parts A–L and L–B, different electron states can be calculated on the acceptor (neutral and anionic, A and A^-, for example) and the donor part of the molecule (neutral and cationic, B and B^+) taking into account the full orbital relaxation. Next, these fragment wave functions are combined into the relevant MEBFs such as the neutral ground state and the charge transfer state, where the orbitals of the overlapping fragment L are identified and removed by a corresponding orbital transformation of the two fragments. This overlapping fragment approach has been used before in transition metal oxides[43], but needs to be generalized and tested for non-ionic compounds, where covalent bonds need to be broken for dividing up the systems in fragments. These bonds will be saturated with hydrogen atoms.

A high priority development objective is to ready the GronOR application for the upcoming exascale systems that will be deployed in the next two to three years. Frontier, the next OLCF

high performance computing system announced for delivery in 2021 by Cray and AMD, will be based on high-performance AMD EPYC CPU and AMD Radeon Instinct GPU technology and Cray's new Shasta architecture and Slingshot interconnect, with an expected performance greater than 1.5 exa-flops. In the same time frame, Aurora, the next ALCF system will be delivered by Intel and Cray, and will include a future generation of Intel Xeon Scalable processor, Intel's Xe compute architecture, and a future generation of Intel Optane DC persistent memory, also with Cray's Shasta architecture Slingshot high-performance scalable interconnect, with an expected performance of 1 exa-flops. Both of these exascale machines will be based on non-NVIDIA GPU accelerators. Work is underway to develop an OpenMP port of GronOR for off-loading to these novel accelerators. Another option that will be investigated is to design parts of the code to use CUDA and CUDA/HIP for offloading to NVIDIA and AMD GPUs respectively.

While the implementation of GronOR is hardware fault resilient, the inclusion of a checkpoint restart capability is planned to be able to break up single calculations of very large chemical systems into multiple jobs.

GronOR is available to the scientific community as an open source code under the Apache 2.0 license.[52]

ACKNOWLEDGMENTS

This work used resources of the Oak Ridge Leadership Computing Facility (OLCF) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This work was supported by the (Shell NWO) research program of the Foundation for Fundamental Research on Matter (FOM), which is part of the Netherlands Organization for Scientific Research (NWO); innovational research incentives scheme Vidi 2017 with project number 016.Vidi.189.044, which is (partly) financed by the NWO; the European Joint Doctorate (EJD) in Theoretical Chemistry and Computational Modeling (TCCM), which has been financed under the framework of the Innovative Training Networks (ITN) of the MARIE Skłodowska-CURIE Actions (ITN-EJD642294-TCCM).

Financial support has also been provided by the Spanish Administration (Project CTQ2017-83566-P) and the Generalitat de Catalunya (Project 2017-SGR629).

The authors thank Jeff Larkin of NVIDIA for his assistance with using PGI-compiler specific

OpenACC directives and ongoing work to use the CUSOLVER library for additional GPU off-loading, and Eric Luo of IBM for investigating using OpenMP target directives for accelerator off-loading.

REFERENCES

1. P. C. Hiberty and S. Shaik, J. Comp. Chem. **28**, 137 (2007).
2. W. Wu, P. Su, S. Shaik, and P. C. Hiberty, Chem. Rev. **111**, 7557–7593 (2011).
3. A. J. W. Thom and M. Head-Gordon, Phys. Rev. Lett. **101**, 193001 (2008).
4. A. J. W. Thom and M. Head-Gordon, J. of Chem. Phys. **131**, 124113 (2009).
5. E. J. Sundstrom and M. Head-Gordon, J. of Chem. Phys. **140**, 114103 (2014).
6. K. J. Oosterbaan, A. F. White, and M. Head-Gordon, J. Chem. Phys. **149**, 044116 (2018).
7. K. J. Oosterbaan, A. F. White, and M. Head-Gordon, J. Chem. Theory Comput. **15**, 2966 (2019).
8. N. J. Mayhall, P. R. Horn, E. J. Sundstrom, and M. Head-Gordon, Phys. Chem. Chem. Phys. **16**, 22694 (2014).
9. J. Olsen, J. Chem. Phys. **143**, 144104 (2015).
10. S. Kähler and J. Olsen, J. Chem. Phys. **147**, 174106 (2017).
11. S. Kähler and J. Olsen, J. Chem. Phys. **149**, 144104 (2018).
12. J. Olsen, LUCIA, a Correlation Program (2019).
13. S. R. Yost, T. Kowalczyk, and T. van Voorhis, J. Chem. Phys. **139**, 174104 (2013).
14. R. Broer, *Localized orbitals and broken symmetry in molecules: theory and applications to the chromate ion and para-benzoquinone*, Ph.D. thesis, University of Groningen (1981).
15. J. T. van Montfort, *Photo-electron spectroscopy. General theoretical aspects and the calculation of peak positions and intensities in some simple systems*, Ph.D. thesis, University of Groningen (1980).
16. C.-P. Hsu, Accounts of Chemical Research **42**, 509 (2009).
17. R. K. Kathir, C. de Graaf, R. Broer, and R. W. A. Havenith, submitted (2019).
18. S. Pathak, L. Lang, and F. Neese, J. Chem. Phys. **147**, 234109 (2017).
19. A. Domingo, C. Angeli, C. de Graaf, and V. Robert, J. Comput. Chem. **36**, 861 (2015).
20. T. P. Straatsma, R. Broer, S. S. Faraji, and R. W. A. Havenith, Annual Reports in Computational Chemistry **14**, 77 (2018).

21. A. Domingo, M. A. Carvajal, C. de Graaf, K. Sivalingam, F. Neese, and C. Angeli, Theor. Chem. Acc. **131**, 1264 (2012).
22. B. Meyer, A. Domingo, T. Krah, and V. Robert, Dalton Trans. **43**, 11209 (2014).
23. OpenMP Architecture Review Board, "OpenMP Application Programming Interface Version 5.0," (2018).
24. OpenACC-Standard.org, "The OpenACC® Application Programming Interface Version 2.7," (2018).
25. "https://docs.nvidia.com/cuda/cusolver/index.html," (2019).
26. M. F. Guest, I. J. Bush, H. J. J. van Dam, P. Sherwood, J. M. H. Thomas, J. H. van Lenthe, R. W. A. Havenith, and J. Kendrick, Mol. Phys. **103**, 719 (2005).
27. F. Aquilante, J. Autschbach, R. K. Carlson, L. F. Chibotaru, M. G. Delcey, L. D. Vico, I. F. Galván, N. Ferré, L. M. Frutos, L. Gagliardi, M. Garavelli, A. Giussani, C. E. Hoyer, G. L. Manni, H. Lischka, D. Ma, P. Å. Malmqvist, T. Müller, A. Nenov, M. Olivucci, T. B. Pedersen, D. Peng, F. Plasser, B. Pritchard, M. Reiher, I. Rivalta, I. Schapiro, J. Segarra-Martí, M. Stenrup, D. G. Truhlar, L. Ungur, A. Valentini, S. Vancoillie, V. Veryazov, V. P. Vysotskiy, O. Weingart, F. Zapata, and R. Lindh, J. Comp. Chem. **37**, 506 (2016).
28. G. A. van der Velde, *Electron correlation in molecules. Theoretical and numerical analysis of cluster expansions of electronic wavefunctions*, Ph.D. thesis, University of Groningen (1974).
29. J. Verbeek, J. H. Langenberg, C. P. Byrman, F. Dijkstra, and J. H. van Lenthe, "TURTLE an *ab initio* VB/VBSCF program (1988-2000),".
30. "https://www.top500.org/," (2019).
31. A. Becke, J. Chem. Phys. **98**, 5648 (1993).
32. C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
33. L. Luo, T. P. Straatsma, L. E. A. Suarez, R. Broer, D. Bykov, E. F. D'Azevedo, S. S. Faraji, K. C. Gottiparthi, C. de Graaf, J. A. Harris, R. W. A. Havenith, H. J. A. Jensen, W. Joubert, R. K. Kathir, J. Larkin, Y. Li, D. I. Lyakh, O. E. B. Messer, M. R. Norman, J. C. Oefelein, R. Sankaran, A. F. Tillack, A. Barnes, L. Visscher, J. C. Wells, and M. Wibowo, IBM Journal of Research and Development, *accepted for publication,* (2020).
34. R. Broer and W. C. Nieuwpoort, Chem. Phys. **54**, 291 (1981).
35. R. Broer and W. C. Nieuwpoort, Theor. Chim. Acta **73**, 405 (1988).
36. C. de Graaf, R. Broer, W. C. Nieuwpoort, and P. S. Bagus, Chem. Phys. Lett. **272**, 341 (1997).
37. A. H. de Vries, L. Hozoi, R. Broer, and P. S. Bagus, Phys. Rev. B **66**, 035108 (2002).

38.A. B. van Oosten, R. Broer, and W. C. Nieuwpoort, Chem. Phys. Lett. **257**, 207 (1996).

39.A. B. van Oosten and F. Mila, Chem. Phys. Lett. **295**, 359 (1998).

40.R. Broer, L. Hozoi, and W. C. Nieuwpoort, Mol. Phys. **101**, 233 (2003).

41.C. J. Calzado, C. Angeli, D. Taratiel, R. Caballol, and J.-P. Malrieu, J. Chem. Phys. **131**, 044327 (2009).

42.J.-P. Malrieu, R. Caballol, C. J. Calzado, C. de Graaf, and N. Guihéry, Chem. Rev. **114**, 429 (2014).

43.A. Stoyanova, C. Sousa, C. de Graaf, and R. Broer, Int. J. Quantum Chem. **106**, 2444 (2006).

44.A. Stoyanova, *Delocalized and Correlated Wave Functions for Excited States in Extended Systems*, Ph.D. thesis, University of Groningen (2006).

45.A. Stoyanova, C. de Graaf, and R. Broer, in *Computation in Modern Science and Engineering*, Vol. 2, edited by G. Maroulis and T. E. Simos (Springer, Berlin, 2007) pp. 163–166.

46.M. Wibowo, R. Broer, and R. W. A. Havenith, Comp. Theor. Chem. **1116**, 190 (2017).

47.L. E. Aguilar Suarez, R. K. Kathir, E. Siagri, R. W. A. Havenith, and S. Faraji, Adv. Quantum Chem. **79**, 263 (2019).

48.Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kús, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. W. III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. D. Jr., H. Dop, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, P. A. Pieniazek, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, N. Sergueev, S. M. Sharada, S. Sharmaa, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, V. Vanovschi, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhou,

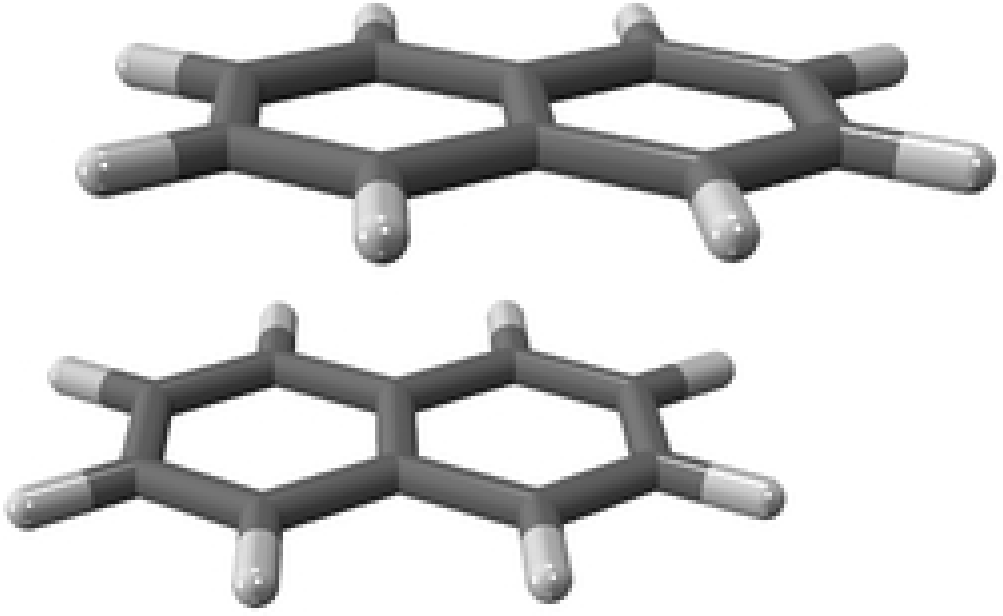
B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. G. III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. S. III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xua, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill, and M. Head-Gordon, Mol. Phys. **113**, 184–215 (2015).

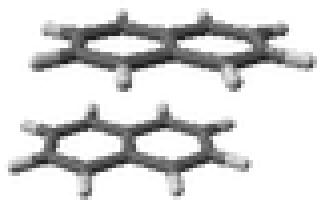
49. A. Zaykov, P. Felkel, E. A. Buchanan, M. Jovamovic, R. W. A. Havenith, R. K. Kathir, R. Broer, Z. Havlas, and J. Michl, J. Am. Chem. Soc. **141**, 17729 (2019).

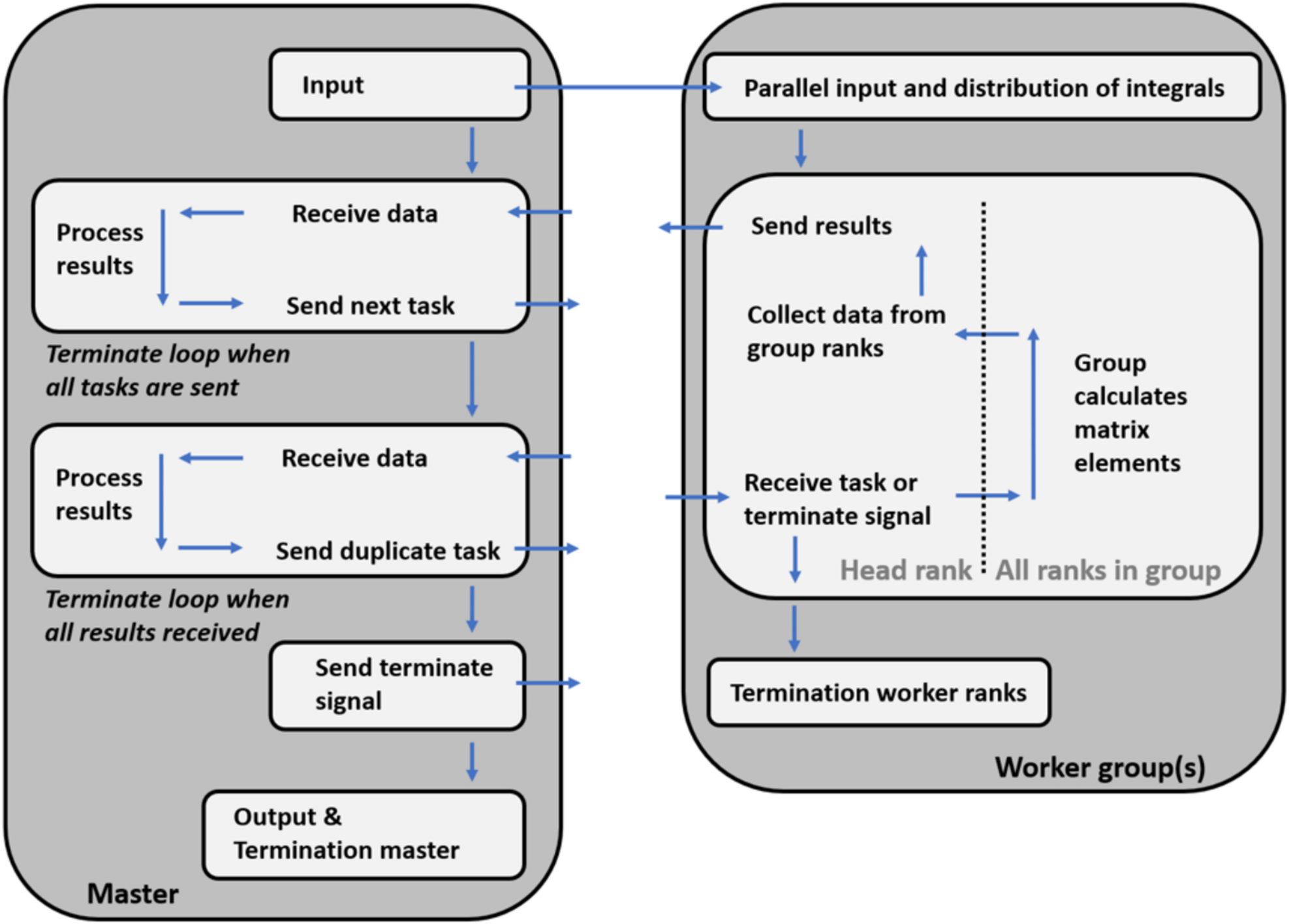
50. H. Koch, A. Sánchez de Merás, and T. B. Pedersen, J. Chem. Phys. **118**, 9481 (2003).

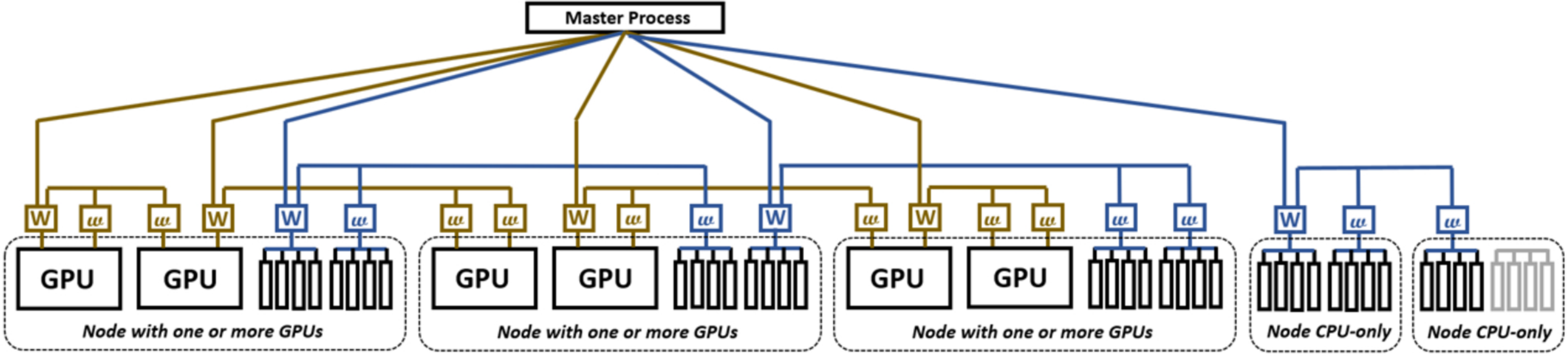
51. S. D. Folkestad, E. F. Kjønstad, and H. Koch, J. Chem. Phys. **150**, 194112 (2019).

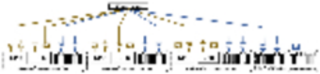
52. R. Broer, S. S. Faraji, C. de Graaf, R. W. A. Havenith, T. P. Straatsma, L. E. Aguilar Suarez, M. A. Izquierdo Morelos, R. K. Kathir, and M. K. Wibowo, "GronOR non-orthogonal configuration interaction, http://gronor.org," (2018).





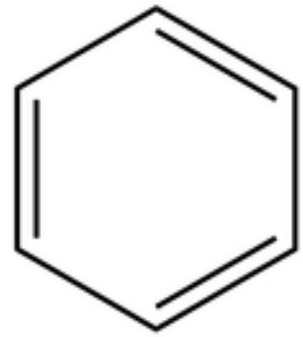

Input
Parallel input and distribution of integrals
Process results
Receive data
Send next task
Terminate loop when all tasks are sent
Send results
Collect data from group ranks
Group calculates matrix elements
Receive task or terminate signal
Head rank
All ranks in group
Process results
Receive data
Send duplicate task
Terminate loop when all results received
Send terminate signal
Termination worker ranks
Output & Termination master
Master
Worker group(s)


Master Process
W
GPU
GPU
Node with one or more GPUs
GPU
GPU
Node with one or more GPUs
GPU
GPU
Node with one or more GPUs
Node CPU-only
Node CPU-only

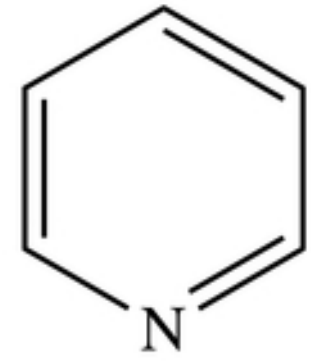


Ethene

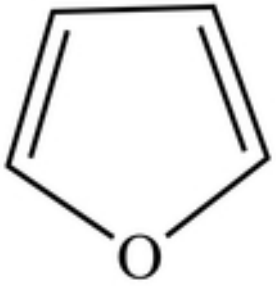
Butadiene



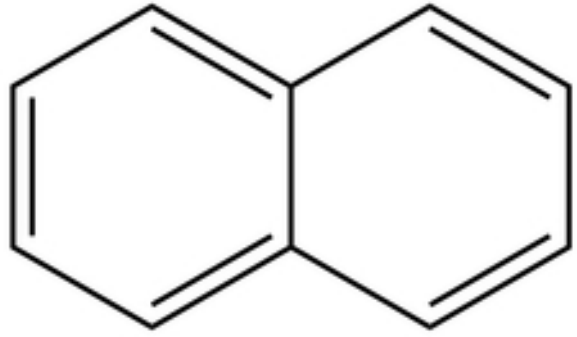
Benzene



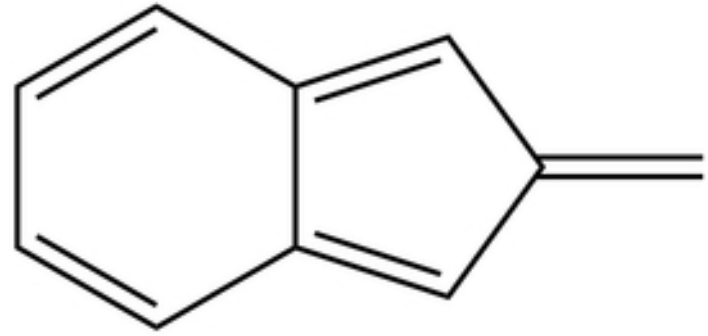
Pyridine



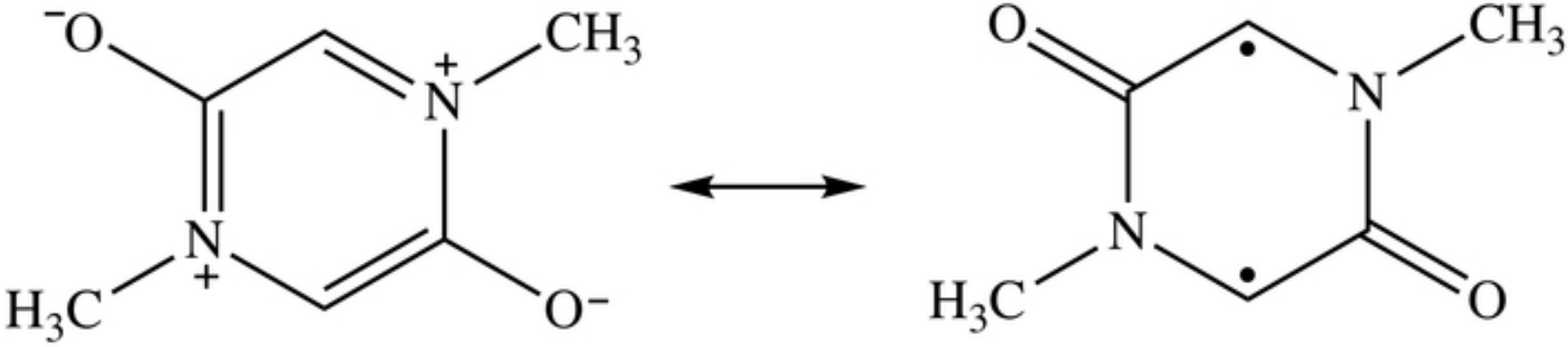
Furan



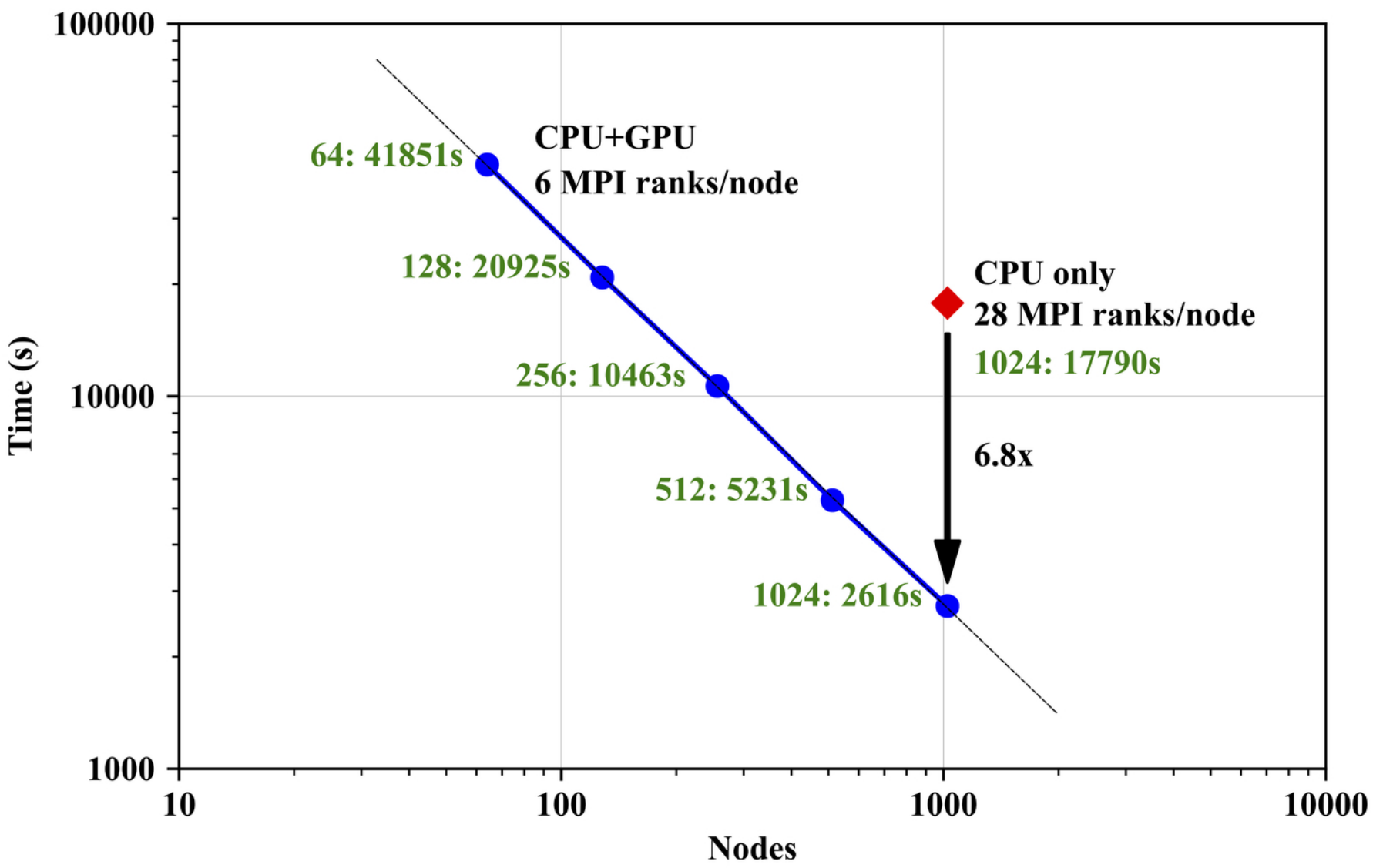
Naphthalene

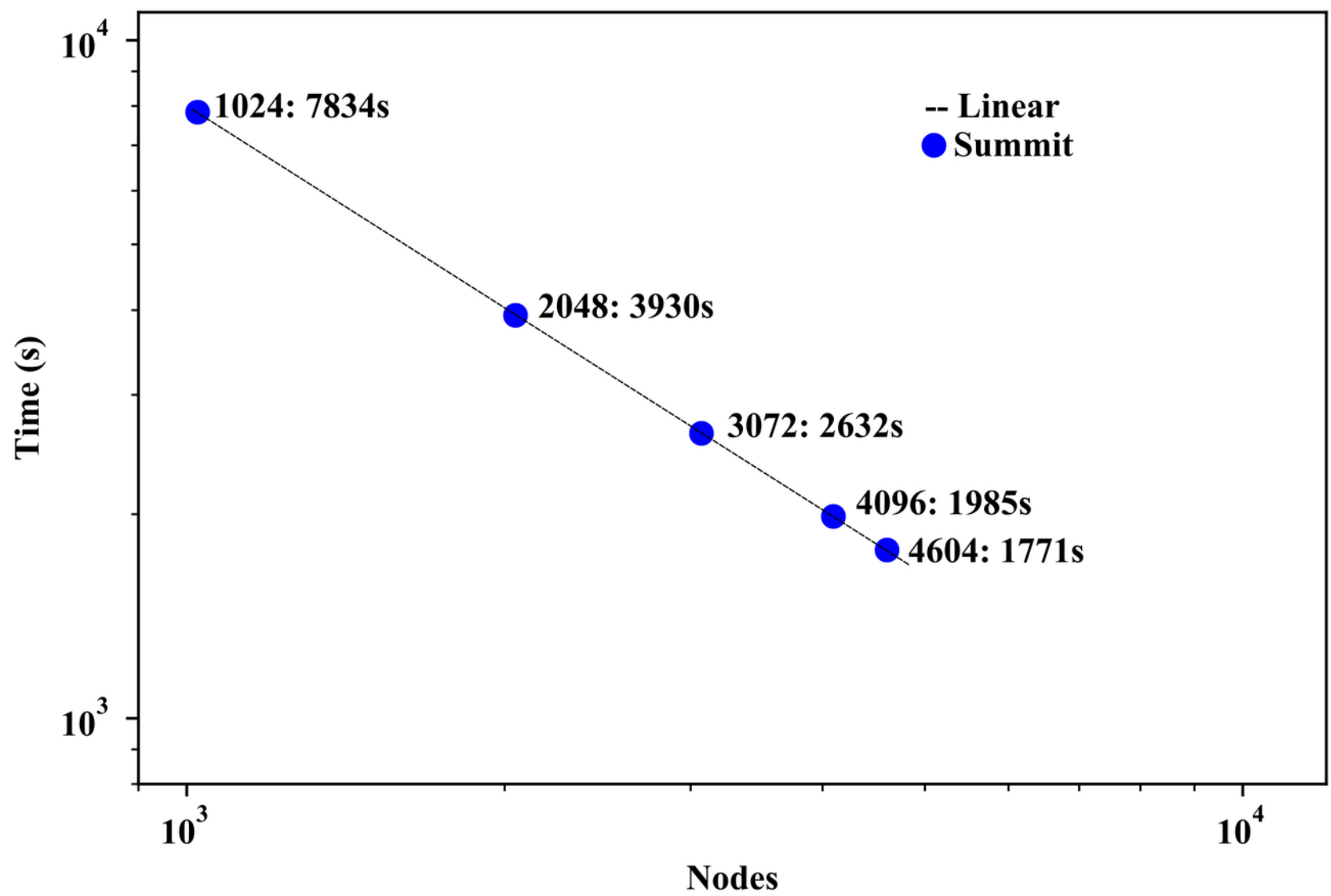


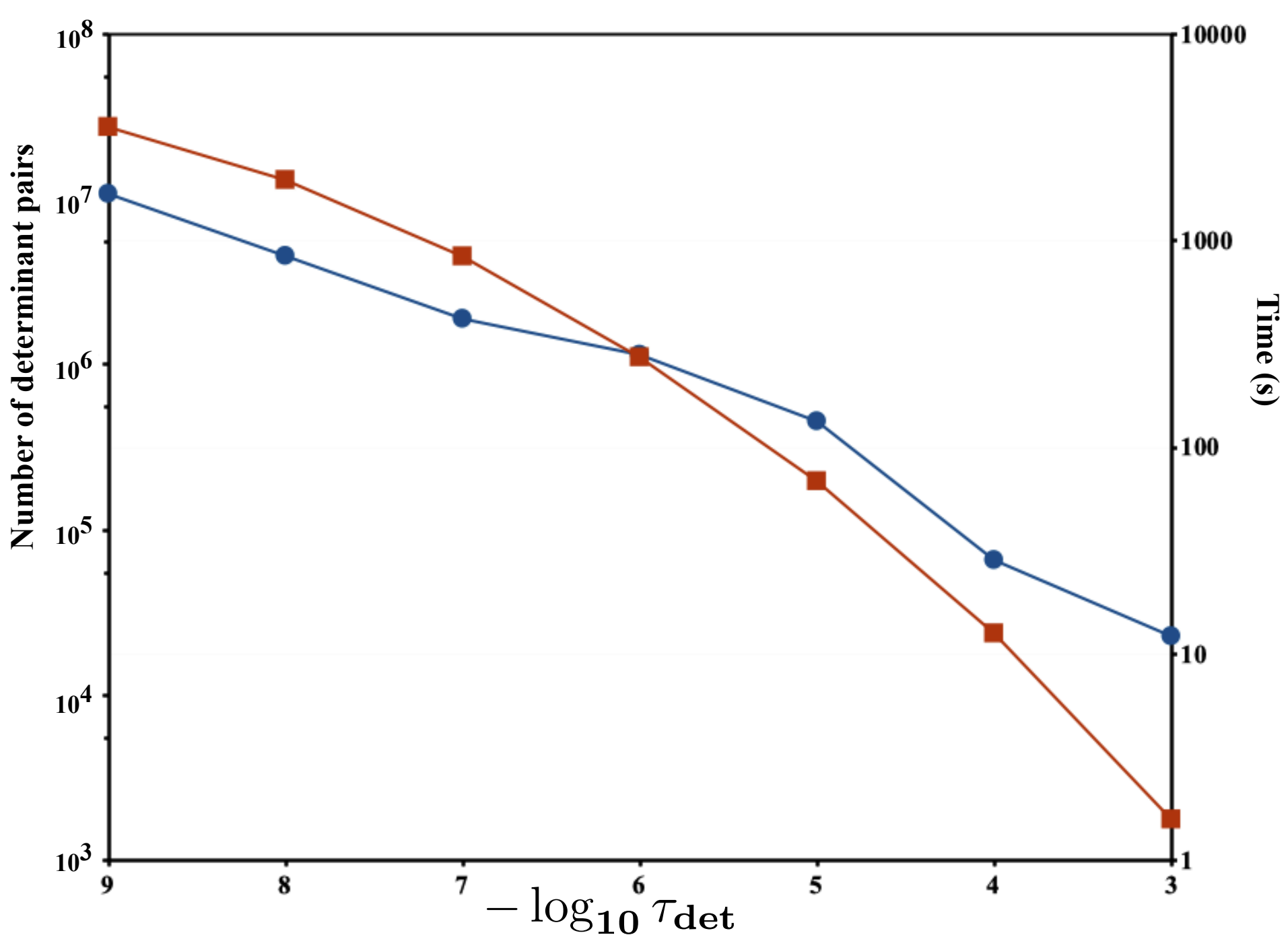
2-methylene-2H-indene

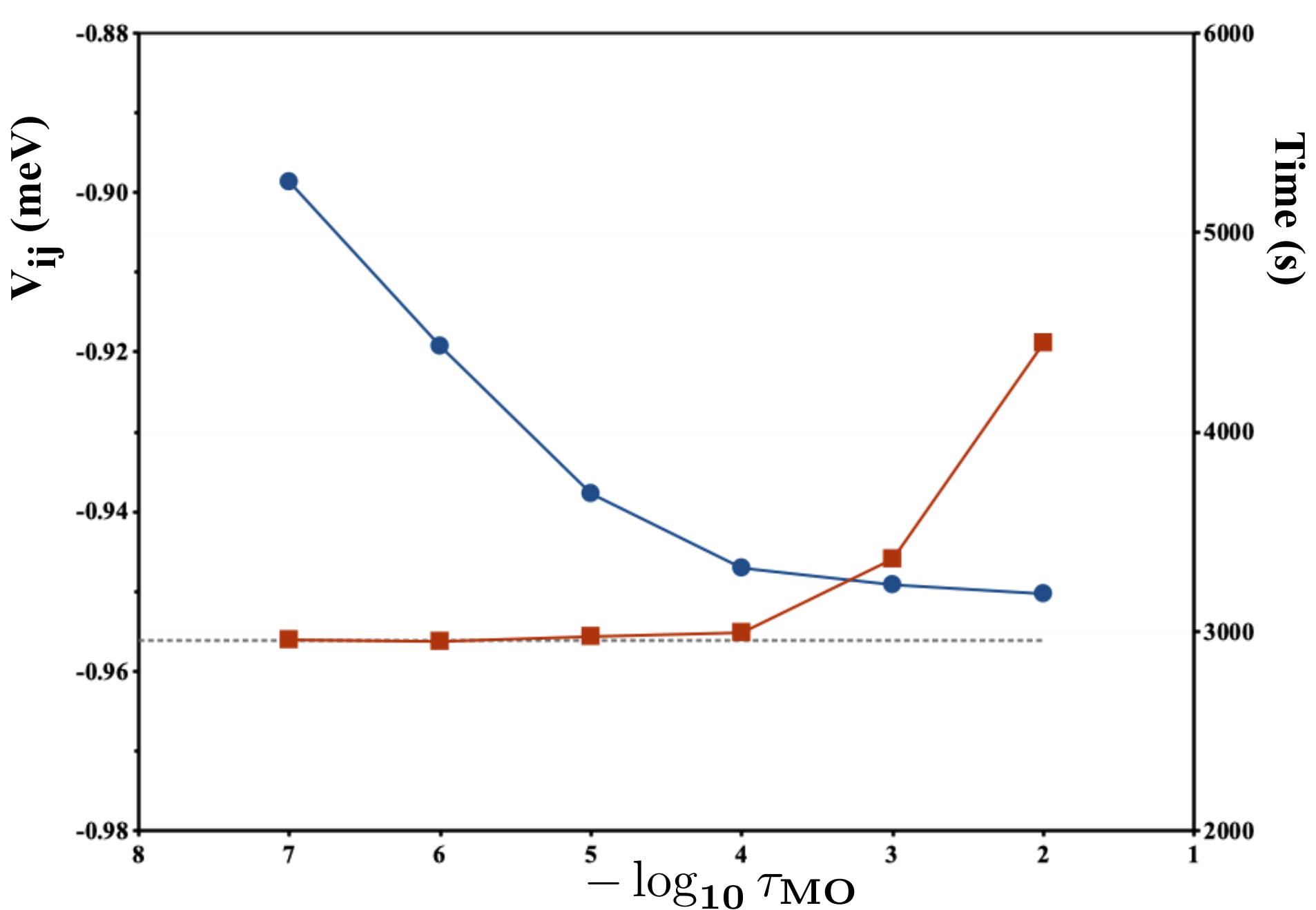


2,5-dihydroxy-1,4-dimethyl-pyrazinium


100000
10000
1000
Time (s)
10
100
1000
10000
Nodes
64: 41851s
CPU+GPU
6 MPI ranks/node
128: 20925s
256: 10463s
512: 5231s
1024: 2616s
CPU only
28 MPI ranks/node
1024: 17790s
6.8x


10^4
10^3
Time (s)
1024: 7834s
2048: 3930s
3072: 2632s
4096: 1985s
4604: 1771s
-- Linear
Summit
10^3
10^4
Nodes


Number of determinant pairs
Time (s)
10^8
10^7
10^6
10^5
10^4
10^3
10000
1000
100
10
1
9
8
7
6
5
4
3
$-\log_{10}\tau_{\mathbf{det}}$


-0.88
-0.90
-0.92
-0.94
-0.96
-0.98
V_{ij} (meV)
6000
5000
4000
3000
2000
Time (s)
8
7
6
5
4
3
2
1
$-\log_{10}\tau_{MO}$