The orbital-specific-virtual local coupled cluster singles and doubles method

Jun Yang, Garnet Kin-Lic Chan, Frederick R. Manby, Martin Schütz, and Hans-Joachim Werner

Citation: The Journal of Chemical Physics **136**, 144105 (2012); doi: 10.1063/1.3696963
View online: http://dx.doi.org/10.1063/1.3696963
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/136/14?ver=pdfcov
Published by the AIP Publishing

The orbital-specific-virtual local coupled cluster singles and doubles method

Jun Yang,[1,a)] Garnet Kin-Lic Chan,[1,b)] Frederick R. Manby,[2,c)] Martin Schütz,[3,d)] and Hans-Joachim Werner[4,e)]
[1]*Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA*
[2]*Center for Computational Chemistry, School of Chemistry, University of Bristol, Bristol BS8 1TS, United Kingdom*
[3]*Institut für Physikalische und Theoretische Chemie, Universität Regensburg, Regensburg, D-93040, Germany*
[4]*Institut für Theoretische Chemie, Universität Stuttgart, Stuttgart D-70569, Germany*

(Received 7 January 2012; accepted 6 March 2012; published online 10 April 2012)

We extend the orbital-specific-virtual tensor factorization, introduced for local Møller-Plesset perturbation theory in Ref. [J. Yang, Y. Kurashige, F. R. Manby and G. K. L. Chan, J. Chem. Phys. **134**, 044123 (2011)], to local coupled cluster singles and doubles theory (OSV-LCCSD). The method is implemented by modifying an efficient projected-atomic-orbital local coupled cluster program (PAO-LCCSD) described recently, [H.-J. Werner and M. Schütz, J. Chem. Phys. **135**, 144116 (2011)]. By comparison of both methods we find that the compact representation of the amplitudes in the OSV approach affords various advantages, including smaller computational time requirements (for comparable accuracy), as well as a more systematic control of the error through a single energy threshold. Overall, the OSV-LCCSD approach together with an MP2 correction yields small domain errors in practical calculations. The applicability of the OSV-LCCSD is demonstrated for molecules with up to 73 atoms and realistic basis sets (up to 2334 basis functions). © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.3696963]

I. INTRODUCTION

Ab initio quantum chemistry defines hierarchies of correlation theories, such as perturbation theory (PT), coupled cluster (CC), and configuration interaction. Despite much progress, conventional correlation treatments are still too expensive to apply to large systems, due to high scaling of computational effort with system size. For example, conventional second-order Møller-Plesset perturbation theory (MP2) requires N^5 cost for the energy (where N measures system size); coupled cluster singles and doubles (CCSD) theory requires N^6 cost; and the "gold standard," CC with perturbative triples CCSD(T), requires N^7 cost.

The steep computational scalings stem from the high tensor rank of the mathematical objects in the theories, and the delocalized nature of the underlying orbital basis. By tensor, we mean an array of numbers, written $\mathcal{T}_{n_1 n_2 \ldots}$, where n_i are the tensor indices, and the number of indices is the tensor rank. These objects include wavefunction amplitudes as well as integral intermediates.

To reduce the computational complexity, we can impose special structures on the tensors. Such a structure can be interpreted as defining a tensor factorization, where a high rank tensor is written as products of tensors,[1,2] with possible contractions over auxiliary indices. Matrix factorizations, such as the Cholesky decomposition, and density fitting (DF) [sometimes also called resolution of the identity] (Refs. 3–15) are obvious examples, but methods which define new occupied and virtual orbital sets, such as the projected atomic orbital (PAO),[16–20] frozen natural orbital,[21–25] and pair natural orbital (PNO) (Refs. 26–33) methods, can also be understood in this mathematical language.

PAO methods define a single global set of occupied and virtual orbitals for correlation, while PNO methods define an adapted set of virtual orbitals for every pair of correlated occupied orbitals. We recently described a tensor factorization that lies between the PAO and PNO methods, the orbital-specific virtual (OSV) method.[1,34] The OSV factorization associates a set of virtual orbitals with each occupied orbital ("orbital specific") rather than to each orbital pair ij as in the PNO method. This leads to simplifications relative to the PNO method, in particular in the integral transformation.

The adaption of the virtual orbitals to the occupied space provides a more compact description of correlation space with OSVs than with PAOs, which in most cases leads to computational savings if one aims at the same accuracy. More importantly, the OSV scheme allows to improve the accuracy systematically based on a single parameter, and at least in principle the canonical result can be approached smoothly. Other theoretical improvements offered by the OSV method include the recovery of smooth potential energy surfaces (without the need for unphysical smoothing schemes[35–37]) even for modest numbers of OSVs, particularly when fully optimized OSVs are used.[34] It should be noted, however, that there are also disadvantages relative to PAOs, such as higher memory requirements and less straightforward generalization to open-shell cases.

a)Electronic mail: jy459@cornell.edu.
b)Electronic mail: gc238@cornell.edu.
c)Electronic mail: fred.manby@bris.ac.uk.
d)Electronic mail: martin.schuetz@chemie.uni-regensburg.de.
e)Electronic mail: werner@theochem.uni-stuttgart.de.

The orbital-specific-virtual local coupled cluster singles and doubles (OSV-LCCSD) ansatz improves the treatment of the virtual space only, and without further approximations still exhibits a relatively steep computational scaling [mostly $\mathcal{O}(m^3)$, some terms scale even as $\mathcal{O}(m^4)$] with respect to the number m of occupied orbitals. This problem is exactly as in other local correlation methods.[16,17,38,39] One way to avoid this is to use local pair approximations, as first introduced by Pulay[16–20] and used in many later PAO methods.[12,15,40–46] Local pair approximations provide a simple way to reach linear scaling and to extend OSV-LCCSD to large systems. However, as pointed out in previous work,[15,46,47] the application of local approximations requires a careful balancing of errors (e.g., local pair error versus domain error) and in the current work we reconsider these issues in the context of the OSV method.

The work presented in this paper can be separated into three parts. First, we will outline the theory. We will then investigate the performance of the OSV method without pair approximations. Here, we primarily contrast the compactness and cost of OSV-LCCSD with that of PAO-LCCSD, and study the accuracy and computational cost requirements as a function of the number of OSVs. In the third part, pair approximations are introduced and their impact on the accuracy and efficiency is demonstrated. Finally, we present some applications that demonstrate the applicability of the method to real problems.

II. THEORY

In this section, we first briefly introduce the coupled cluster equations in order to define the relevant quantities and notation. Subsequently, we will discuss various choices of virtual orbitals and introduce the OSV-LCCSD method. In the following indices i, j, k, l will denote localized occupied molecular orbitals (LMOs), and a, b, c, d canonical virtual orbitals (VMOs). It will be assumed that the occupied orbitals are orthonormal and that the occupied and virtual orbital spaces are mutually orthogonal, i.e., $\langle i|a\rangle = 0\ \forall\, a,\ i$. Other choices of virtual orbitals will be denoted by indices r, s, t, u.

A. Definition of the CCSD wavefunction

The CCSD wavefunction in an orthonormal orbital basis is defined as

$$\Psi = e^{\hat{T}}\Phi_0, \tag{1}$$

where Φ_0 is the closed-shell Hartree-Fock Slater determinant, and $\hat{T} = \hat{T}_1 + \hat{T}_2$ is the singles and doubles cluster operator

$$\hat{T}_1 = \sum_i \sum_a t_a^i \hat{E}_i^a, \tag{2}$$

$$\hat{T}_2 = \frac{1}{2}\sum_{i,j}\sum_{ab} T_{ab}^{ij}\hat{E}_i^a\hat{E}_j^b. \tag{3}$$

$\hat{E}_i^a$ are spin-summed one-electron excitation operators, and t_a^i, T_{ab}^{ij} are the singles and doubles amplitudes, respectively. These quantities can be considered as second- and fourth-order tensors, respectively. However, since in local treatments with pair approximations the list of pairs ij is very sparse, we prefer to denote them as vectors and matrices, respectively, where the superscripts denote different matrices (upper case quantities) or vectors (lower case quantities), and the subscripts their elements. Such vectors and matrices will be written in bold face if reference to the individual elements is not needed, e.g., $T_{ab}^{ij} = [\mathbf{T}^{ij}]_{ab}$. The elements of such matrices always correspond to virtual orbitals. Since $\hat{E}_i^a$ and $\hat{E}_j^b$ commute, $T_{ab}^{ij} = T_{ba}^{ji}$. The amplitudes are determined by solving the CC amplitude equations

$$r_a^i = \langle \tilde{\Phi}_i^a | e^{-\hat{T}}\hat{H}e^{\hat{T}}|\Phi_0\rangle = 0 \qquad \forall\, a, i, \tag{4}$$

$$R_{ab}^{ij} = \langle \tilde{\Phi}_{ij}^{ab} | e^{-\hat{T}}\hat{H}e^{\hat{T}}|\Phi_0\rangle = 0 \qquad \forall\, i \geq j, a, b. \tag{5}$$

The quantities r_a^i and R_{ab}^{ij} are called *residual* vectors and matrices, respectively. They vanish for the optimized amplitudes. The (contravariant) configurations $\tilde{\Phi}_i^a$ and $\tilde{\Phi}_{ij}^{ab}$ are defined as

$$|\tilde{\Phi}_i^a\rangle = \frac{1}{2}\hat{E}_{ai}|\Phi_0\rangle, \tag{6}$$

$$|\tilde{\Phi}_{ij}^{ab}\rangle = \frac{1}{6}\left(2\hat{E}_i^a\hat{E}_j^b + \hat{E}_j^a\hat{E}_i^b\right)|\Phi_0\rangle. \tag{7}$$

They have the property that

$$t_a^i = \langle \tilde{\Phi}_i^a | \Psi\rangle, \tag{8}$$

$$C_{ab}^{ij} = \langle \tilde{\Phi}_{ij}^{ab} | \Psi\rangle, \tag{9}$$

where

$$C_{ab}^{ij} = T_{ab}^{ij} + t_a^i t_b^j. \tag{10}$$

The choice of projection functions in Eqs. (6) and (7) leads to the most compact form of the CCSD equations.[48–50] For convenience in later expressions we also define the corresponding contravariant doubles amplitudes

$$\tilde{T}_{ab}^{ij} = 2T_{ab}^{ij} - T_{ab}^{ji}. \tag{11}$$

The two-electron integrals can be represented by matrices and vectors as well, for example

$$J_{ab}^{ij} = (ab|ij), \tag{12}$$

$$K_{ab}^{ij} = (ai|bj), \tag{13}$$

$$L_{ab}^{ij} = 2K_{ab}^{ij} - K_{ba}^{ij}. \tag{14}$$

In terms of these quantities, the CCSD correlation energy is

$$E_{\mathrm{corr}} = \sum_{ij}\sum_{ab} C_{ab}^{ij}L_{ab}^{ij} = \sum_{i\geq j}(2-\delta_{ij})\mathrm{tr}[\mathbf{C}^{ij}\mathbf{L}^{ji}]. \tag{15}$$

B. General transformations for the virtual orbitals

We now consider transformations of the CCSD equations to different virtual orbital representations. The new orbitals, which may in general be non-orthonormal, will be labeled by indices r, s. In general, different orbital sets can be defined for each pair ij, and this will then be indicated by superscripts,

$$|r^{ij}\rangle = \sum_a |a\rangle Q_{ar}^{ij}. \tag{16}$$

The amplitudes transform as

$$t_a^i = \sum_r Q_{ar}^{ii} t_r^i, \tag{17}$$

$$T_{ab}^{ij} = \sum_{rs} Q_{ar}^{ij} T_{rs}^{ij} Q_{bs}^{ij}. \tag{18}$$

Single excitations from an LMO i are made into the same virtual orbitals as for the "diagonal" double excitation ($i = j$) from the same LMO. Inserting these expressions into the CCSD residual equations and transforming the residuals to the new basis, i.e.,

$$r_r^i = \sum_a r_a^i Q_{ar}^{ii}, \tag{19}$$

$$R_{rs}^{ij} = \sum_{ab} Q_{ar}^{ij} R_{ab}^{ij} Q_{bs}^{ij}, \tag{20}$$

yields equations in which all integrals and amplitudes are in the new virtual basis. They differ formally from the equations in an orthonormal virtual basis only by the multiplications with the overlap matrix,

$$\langle r^{ij}|s^{kl}\rangle = [\mathbf{S}^{ij,kl}]_{rs} = [\mathbf{Q}^{ij\dagger}\mathbf{Q}^{kl}]_{rs}, \tag{21}$$

in all places where an amplitude index is not matched by an integral label. The explicit form of the resulting equations can be found in Appendix B of Ref. 15.

So far, there is no advantage of these transformations. However, if the virtual orbitals are suitably chosen, the number of amplitudes and residual equations can be strongly reduced by introducing *domain approximations,*

$$t_a^i \approx \sum_{r\in[i]} Q_{ar}^{ii} t_r^i, \tag{22}$$

$$T_{ab}^{ij} \approx \sum_{rs\in[ij]} Q_{ar}^{ij} T_{rs}^{ij} Q_{bs}^{ij}. \tag{23}$$

The subset of orbitals $|r^{ij}\rangle$ used to approximate the amplitude matrix T_{ab}^{ij} is denoted as *pair domain* $[ij]$. The domains $[i]$ for single excitations correspond to the pair domains of the diagonal pairs, i.e., $[i] = [ii]$. The residual equations have then to be solved only for the same domains

$$r_r^i = 0 \quad \forall\, r \in [i], \tag{24}$$

$$R_{rs}^{ij} = 0 \quad \forall\, r,s \in [ij], \tag{25}$$

and the correlation energy is given by

$$E_{\text{corr}} = \sum_{i\geq j}(2-\delta_{ij}) \sum_{rs\in[ij]} C_{rs}^{ij} L_{rs}^{ij}. \tag{26}$$

For large molecules, the domain approximation leads to a strong reduction of the computational effort and its scaling with molecular size. Note that the domain approximation involves only the virtual labels r, s in the tensor quantities. *Pair approximations* allow us to use different levels of theory based on the occupied labels i, j. For example, only *strong pairs*, which contribute most to the correlation energy, are included in the LCCSD; the remaining *weak* or *distant* pairs are either approximated by LMP2 or neglected. It is then possible to achieve linear scaling of the computational effort with molecular size.[15,40,43,51] We discuss pair approximations further in Sec. II H. The important point is that the convergence of the correlation energy and other molecular properties as a function of domain sizes depends crucially on the choice of the transformation matrices $\mathbf{Q}^{ij}$. In Subsections II C–II E, we will discuss three different choices and their implications on the computational efficiency.

C. Projected atomic orbitals (PAOs)

Pulay[16] suggested spanning the virtual orbital space by projected atomic orbitals,

$$|r\rangle = \sum_a |a\rangle Q_{ar}, \tag{27}$$

$$Q_{ar} = \langle a|\chi_r^{\text{AO}}\rangle, \tag{28}$$

where $|\chi_r^{\text{AO}}\rangle$ are atomic orbitals (AOs). Usually, the AOs are taken to be the contracted Gaussian type orbitals (CGTOs), and then each PAO is associated to a CGTO. For example, the local correlation methods of Pulay and Saebø[17–20] and of Werner, Schütz and co-workers[12,15,40–46,51–55] are based on PAOs. The PAOs are local by construction, pair-independent and nonorthogonal

$$\langle r|s\rangle = [\mathbf{S}]_{rs} = [\mathbf{Q}^{\dagger}\mathbf{Q}]_{rs}. \tag{29}$$

The standard way to select domains in the PAO-LCCSD method is either to use the method of Boughton and Pulay[56] (BP) or natural population analysis (NPA).[53,57] Both methods can be used with any localization scheme, e.g., Pipek-Mezey[58] (PM) or natural localized orbitals (NLMOs).[53,59] Unless otherwise noted, we will use the NLMO/NPA method[53] in the current paper.

The domain selection with the BP or NPA methods depends on thresholds l_{bp} and l_{npa}, respectively. l_{bp} is a completeness criterion, and with $l_{bp} = 1$ domains that span the full virtual space are obtained. l_{npa} refers to the natural charge of a center in a given orbital, and all centers are included which have charges larger then l_{npa}. In the current work we use $l_{npa} = 0.07$. Smaller values yield larger domains, and in this case $l_{npa} = 0$ gives full domains. However, when these thresholds are close to 1 or 0, respectively, the domains may become unphysical, and therefore a variation of these thresholds is not very suitable to approach the canonical limit systematically.

One way to overcome this problem is to use the BP or NPA methods just to determine "standard" (or "primary") domains, which include the most important atoms for each orbital and usually correspond to chemical intuition. The accuracy can be improved by extending the domains by adding all PAOs at shells of neighboring centers.[47] The fraction of correlation energy then converges quickly towards 100%. The disadvantage of this method is, however, that it is quite coarse grained and the domains grow rapidly.

In order to achieve a more fine-grained variation of the domain sizes we have adopted an approach that is based on contributions of individual centers to the correlation energies of the diagonal pairs *ii* (similar to the OSV case, see later). Initially, an LMP2 calculation is carried out, in which the domains of the diagonal pairs are extended by several shells of neighboring atoms. The energy contributions of individual centers A to the pair energy ϵ_{ii} are then evaluated as

$$\epsilon_{ii}^{A} = \sum_{r\in[A]} \sum_{s} T_{rs}^{ii} K_{rs}^{ii}, \tag{30}$$

where the sum over r is restricted to PAOs at center A. Equivalent to this would be to partition the pair energy to contributions ϵ_{ii}^{AB}, and to assign half of these to the centers A and B. The orbital domains $[i]$ include all centers that yield energy contributions larger than an energy threshold l_{pao}. Unfortunately, due to the non-orthogonality of the PAOs, this selection procedure slightly depends on the domains used in the initial LMP2. In the current work, we have used complete domains for the diagonal pairs, and standard domains for the remaining pairs.

Pair domains $[ij]$ are then taken to be the union of the orbital domains $[i]$ and $[j]$. As will be shown in Sec. III, on the average typically 120–150 (250–300) PAOs per pair are needed to recover around 99% (99.8%) of the canonical correlation energy for an augmented triple-ζ basis set. These domain sizes grow linearly with the size of the basis set per atom. However, the domain sizes are (asymptotically) independent of the molecular size.

D. Pair natural orbitals (PNOs)

Much better convergence of the correlation energy as a function of the domain sizes can be achieved with pair-specific virtual orbitals. An excellent choice is to use MP2 pair natural orbitals, where $\mathbf{Q}^{ij}$ is defined by diagonalizing the MP2-like density matrix

$$\mathbf{D}^{ij} = \frac{1}{1+\delta_{ij}} (\tilde{\mathbf{T}}^{ij\dagger} \mathbf{T}^{ij} + \tilde{\mathbf{T}}^{ij} \mathbf{T}^{ij\dagger}), \tag{31}$$

$$\left[\mathbf{Q}^{ij\dagger} \mathbf{D}^{ij} \mathbf{Q}^{ij}\right]_{rs} = n_r^{ij} \delta_{rs}, \tag{32}$$

for pair ij. The amplitudes in Eq. (31) are computed as

$$T_{ab}^{ij} = -\frac{K_{ab}^{ij}}{\epsilon_a + \epsilon_b - f_{ii} - f_{jj}}. \tag{33}$$

Here the virtual orbitals are assumed to be canonical, i.e., $f_{ab} = \epsilon_a \delta_{ab}$, and f_{ii} are the diagonal elements of the Fock matrix in the LMO basis. The domain $[ij]$ can then be determined by neglecting orbitals that have natural occupation numbers n_r^{ij} below a certain threshold.[60] This is the approach used by Neese and co-workers.[31–33] The PNOs for a given pair ij are orthonormal, but PNOs of different pairs are non-orthogonal.

Using this ansatz one typically needs only 30–40 PNOs per pair in order to recover 99.8% of the canonical CCSD correlation energy (again for a triple-ζ basis set and independent of the molecular size). However, a severe disadvantage of the PNO method is that the total number of virtual orbitals may become very large; for example, if 1000 pairs ij are correlated, one needs about 40 000 virtual orbitals. This leads to difficulties in the integral transformations and storage of the integral matrices unless drastic approximations are used, as described in Sec. II F.

E. Orbital specific virtuals (OSVs)

Recently, Yang *et al.*[1] have proposed orbital-specific virtual orbitals as a compromise between the pair-specific PNOs and the pair-independent PAOs. In this case, a set of virtual orbitals is associated with each LMO. An excellent choice is to generate the OSVs by singular value decomposition (SVD) of the *diagonal* MP2 pair amplitudes,

$$[\mathbf{Q}^{i\dagger} \mathbf{T}^{ii} \mathbf{Q}^{i}]_{rs} = t_r^{ii} \delta_{rs}, \tag{34}$$

$$|r^i\rangle = \sum_a |a\rangle Q_{ar}^i. \tag{35}$$

The amplitudes T_{ab}^{ii} are approximated according to Eq. (33). Since the diagonal amplitude matrices are symmetrical, the left and right singular vectors are identical, and SVD is equivalent to diagonalization of $\mathbf{T}^{ii}$. The OSVs $|r^i\rangle$ are also identical to the PNOs $|r^{ii}\rangle$, and $n_r^{ii} = (t_r^i)^2$. Based on the magnitude of the eigenvalues t_r^{ii} or of the occupation numbers n_r^{ii}, a domain $[i]$ of OSVs can be selected for each LMO i. Alternatively, here we will use an energy criterion. The diagonal MP2 pair correlation energies are written as

$$\epsilon_{ii} = \sum_{ab} T_{ab}^{ii} K_{ab}^{ii} = \sum_r t_r^{ii} k_r^{ii}, \tag{36}$$

$$k_r^{ii} = \sum_{ab} Q_{ar}^i K_{ab}^{ii} Q_{br}^i, \tag{37}$$

and as many orbitals $|r^i\rangle$ are included in the domain $[i]$ as needed to make the error of $\sum_{r\in[i]} t_r^{ii} k_r^{ii}$ relative to the exact pair energy smaller than a threshold l_{osv} (the orbitals are ordered according to decreasing t_r^{ii}). Note that l_{pao} and l_{osv} are not directly comparable, as the former is a threshold on the contribution of a center and its set of PAOs to the diagonal pair energy, while the latter is a threshold on the contribution of a single orbital to the diagonal pair energy. Consequently, for the same error in ϵ_{ii}, l_{pao} will typically be larger than l_{osv}. As in PAO methods, pair domains $[ij]$ are then formed as the union of the orbital domains $[i]$ and $[j]$. Thus, the transformation matrix Q_{ar}^{ij} that generates the pair domain from the canonical orbitals can be written in block form as

$$\left(\mathbf{Q}^{ij}\right) = \left(\mathbf{Q}^{i}\mathbf{Q}^{j}\right), \tag{38}$$

which indicates that the columns of $\mathbf{Q}^i$ are collated with those of $\mathbf{Q}^j$. The canonical amplitudes can then be approximated as in Eqs. (22) and (23).

The OSVs for a given LMO are orthonormal, but OSVs for different LMOs are non-orthogonal. Thus, the overlap matrix $\mathbf{S}^{ij,\,ij}$ for a pair domain is block diagonal. However, this sparsity is not exploited in our current implementation. It should be noted that the orbitals $|r^{ij}\rangle$ in a pair domain $[ij]$ may become (nearly) linear dependent. Such linear dependencies are removed by diagonalizing $\mathbf{S}^{ij,\,ij}$ and removing eigenvectors that correspond to very small eigenvalues. This is exactly as in the PAO case and technical details can be found in Ref. 15.

As will be demonstrated in Sec. III, typically 100 OSVs per pair are needed to recover 99.8% of the canonical correlation energy. This is one third to one half of the number of PAOs required for the same accuracy, but about twice as many PNOs. The advantage of using OSVs rather than PNOs is that the total number of virtual orbitals is very much smaller.

Finally, it should be noted that the generation of the OSVs scales as $\mathcal{O}(N^4)$, where N is a measure of the molecular size (e.g., the number of correlated electrons). This scaling is steeper than of all other terms in an OSV-LCCSD calculation, but since efficient density fitting methods are used to generate the necessary integrals K^{ii}_{ab} this step did not present a bottleneck in any of the calculations presented in this paper.

F. The OSV-LCCSD residuals

In this section, we will discuss the solution of the coupled-cluster equations and the required integral transformations when using OSVs. Our implementation of the OSV-LCCSD method is based on the DF-LCCSD method that has recently been described by two of us,[15] and that is part of the MOLPRO program package.[61,62] Formally, the OSV-LCCSD equations are exactly the same as given in Appendix B of Ref. 15 for PAO-LCCSD. However, larger integral and overlap matrices are needed. In order to illustrate this, we consider a typical contribution in the doubles residual,

$$\mathbf{R}^{ij}_{[ij,ij]} = \ldots \sum_k \mathbf{S}_{[ij,ik]} \tilde{\mathbf{T}}^{ik}_{[ik,ik]} \mathbf{Y}^{kj}_{[ik,ij]}, \tag{39}$$

$$\mathbf{Y}^{kj}_{[\bar{k},\bar{j}]} = \mathbf{K}^{kj}_{[\bar{k},\bar{j}]} + \frac{1}{4}\left[\sum_l \mathbf{L}^{kl}_{[\bar{k},lj]} \tilde{\mathbf{T}}^{lj}_{[lj,lj]}\right] \mathbf{S}_{[\bar{j},\bar{j}]} + \ldots. \tag{40}$$

The first and second labels in square brackets indicate the domains of the rows and columns of the matrices, respectively, and obviously these must match in the matrix multiplications. This makes it necessary to use gather operations to extract the appropriate blocks from the overlap and integral matrices. For example, in order to evaluate Eq. (39), the block $[ij, ik]$ is extracted from the full overlap matrix $\mathbf{S}$, and the block $[ik, ij]$ is extracted from the intermediate matrix $\mathbf{Y}^{kj}$. Since $\mathbf{Y}^{kj}$ can be used to compute all residuals $\mathbf{R}^{ij}$ for a fixed j, it is computed in the *united domains* $[\bar{k}, \bar{j}]$. The united domain $[\bar{k}]$ is the union of all pair domains $[ik]$ that share the same k, and the united domain $[\bar{j}]$ is the union of all $[ij]$ for fixed j. Thus, the blocks $\mathbf{Y}^{kj}_{[ik,ij]}$ can be extracted from the larger matrix $\mathbf{Y}^{kj}_{[\bar{k},\bar{j}]}$ for all i. Note that for large molecules the united domains are independent of the molecular size (if pair approximations are applied), and therefore linear scaling is automatically achieved for these terms.

In Eq. (40) each term in the summation over l involves a different domain $[lj]$. Therefore, the block $\mathbf{L}^{kl}_{[\bar{k},lj]}$ must be extracted from the integral matrix $\mathbf{L}^{kl}$, and the matrix product must be added to the appropriate blocks in $\mathbf{Y}^{kj}_{[\bar{k},\bar{j}]}$. The result in square brackets has then dimension $[\bar{k}, \bar{j}]$ and is finally multiplied with $\mathbf{S}_{[\bar{j},\bar{j}]}$.

The matrix multiplications in the residual equations can be carried out in various possible orders. For example, one could also evaluate the second term of $\mathbf{Y}^{kj}$ as

$$+\frac{1}{4}\left[\sum_l \mathbf{L}^{kl}_{[\bar{k},lj]} \tilde{\mathbf{T}}^{lj}_{[lj,lj]} \mathbf{S}_{[lj,\bar{j}]}\right], \tag{41}$$

i.e., the multiplication with $\mathbf{S}$ is now done within the loop over l. However, the number of operations is in this case larger than Eq. (40) since the union of all $[lj]$ (for fixed j) equals the union of all $[l]$, while the sum of all dimensions of $[lj]$ is larger (since it contains $[j]$ repeatedly). Therefore Eq. (40) is used. Similar considerations apply for other contributions to the residuals.

In our program, the whole overlap matrix $\mathbf{S}$ of all OSVs is kept in memory. The required blocks are obtained when needed by gather operations as described above. The total dimension of the overlap matrix is $\sum_i n_i$, where n_i is the number of OSVs for LMO i. On the average, $n_i \approx 50$ in order to recover 99.8 % of the correlation energy. Thus, in a calculation with 100 correlated LMOs the dimension of $\mathbf{S}$ is about 5000. Note that this dimension can be larger than the number of virtual orbitals and depends on the OSV selection threshold l_{osv}. For example, in the polyglycine $(Gly)_8$ calculation that will be presented in Sec. III there are 92 correlated LMOs, 1757 VMOs, and in total 2819 ($l_{osv} = 1.0 \times 10^{-4}$) or 3855 ($l_{osv} = 3.2 \times 10^{-5}$) OSVs. In contrast, in a PAO calculation the dimension of the overlap matrix can never be larger than the number of basis functions (in our example this is 1882). Similar considerations hold for the integral matrices $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$. If all pairs are included in the LCCSD, one needs all $m(m + 1)/2$ matrices of each type in the full OSV basis, where m is the number of correlated LMOs.

The number of matrices, as well as their dimensions, are reduced if pair approximations are introduced, i.e., if weak pairs are approximated by MP2 (cf. Sec. II H). Then i, j, k, l must all be within a finite distance, since i is close to j through pair ij; i close to k though pair ik; and j close to l through pair jl. Using such considerations one can form operator lists and operator domains, as discussed previously for PAO-LCCSD.[43] Despite the fact that more PAOs than OSVs are needed to reach a certain accuracy, one usually needs more integrals $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$ for OSV-LCCSD than for PAO-LCCSD (in particular for small values of the threshold l_{osv}). This means that in most cases the reduced CPU-time of the OSV-LCCSD iterations comes at the expense of greater memory requirements.

This situation is even much more pronounced in the PNO-LCCSD method. Even though one needs only about 40 PNOs per pair to recover 99.8% of the canonical correlation energy, the overlap and integral matrices in the basis of all PNOs would be huge. If no pairs were neglected the total dimension of the overlap matrix would for the above example be $(40 \times 92 \times 93)/2 = 171\,120$. In practice, one can approximate weak pairs by LMP2, and then only about 1000 pairs need to be included in the LCCSD. But the total dimension of $\mathbf{S}$ would still be ≈40 000 (about 6 GB if stored in triangular form). Again, similar considerations hold for the integral matrices $\mathbf{J}^{kl}$, $\mathbf{K}^{kl}$, and it would obviously be quite impossible to store them. In order to overcome this problem, Neese *et al.*[32] have introduced some rather drastic approximations: in some terms the operators $\mathbf{K}^{kl}_{[ik,lj]}$ are projected onto the domain [*kl*, *kl*], i.e.,

$$\mathbf{K}^{kl}_{[ik,lj]} \approx \mathbf{S}_{[ik,kl]}\mathbf{K}^{kl}_{[kl,kl]}\mathbf{S}_{[kl,lj]}. \tag{42}$$

These approximations introduce errors which eliminate some of the advantages of the systematic convergence achievable using PNOs. Alternatively, the integrals are stored in the canonical molecular orbital (MO) basis and transformed on the fly into the PNO basis when needed. As the canonical MO basis is involved, this loses the local scaling.[63]

G. OSV-LCCSD integrals

The computation of two-electron integrals in the OSV representations forms a large part of the cost of the OSV-LCCSD method. All required integrals are computed by DF approximations as described in Ref. 15. However, somewhat different restrictions to the virtual orbital labels apply. In the PAO-LCCSD method the necessary integrals are defined by quadruplets of centers (for PAOs) and/or LMOs.[14,15,45] Since the OSVs are not related to centers but only to LMOs, center labels are now replaced by LMO labels. Local density fitting approximations as described previously for PAO methods[12,15,64] should be possible for OSVs as well, but have not yet been implemented in our program.

We first consider the contributions of integrals over four OSVs (in the following denoted 4-ext integrals):

$$R^{ij}_{rs} \mathrel{+}= \sum_{t,u\in[ij]} (rt|su)C^{ij}_{tu} \qquad \forall\, r,s \in [ij]. \tag{43}$$

All four labels *r*, *s*, *t*, *u* are associated to the same pair domain [*ij*]. Consequently, it is sufficient to generate the 4-ext integrals where *r*, *s*, *t*, *u* are OSVs for LMOs *i* or *j*. Thus, there are only four integral classes, namely $(r^it^i|s^iu^i)$, $(r^it^j|s^iu^j)$, $(r^it^i|s^ju^j)$, and $(r^it^i|s^iu^j)$. The total number of unique 4-external integrals is approximately $(7N_p/4 - 13m/8)L^4$, where *L* is the average number of OSVs per LMO, *m* is the number of correlated LMOs, and N_p is the number of pairs included in the LCCSD. Since both *m* and N_p are proportional to the molecular size and *L* is independent of the molecular size the total number of integrals scales linearly with molecular size.

Similar considerations apply to the 3-external integrals (*rs*|*tk*). In this case *r*, *s*, *t* must be OSVs for the LMOs *i*, *j*, or *k*. In addition, *ij* and *ik* or *jk* must be strong pairs (see Appendix B of Ref. 15 for more details). It follows that the number of 3-external integrals scales linearly as well (provided distant pairs are neglected).

Lastly, 0-ext, 1-ext, and 2-ext integrals appear in a number of different contractions with the singles and doubles amplitudes. In most terms the LMO labels are related by strong pair conditions, as, e.g., discussed in Sec. II F. There are a few terms, however, where *ij* and *kl* are not directly related. For example, this is the case for the contribution

$$R^{ij}_{rs} \mathrel{+}= \sum_{kl} \alpha_{ij,kl}[\mathbf{S}\mathbf{C}^{kl}\mathbf{S}]_{rs}, \tag{44}$$

with

$$\alpha_{ij,kl} = K^{ij}_{kl} + \sum_{r,s\in[ij]} C^{ij}_{rs}K^{kl}_{rs} + \sum_{r\in[i]} t^i_r(rk|lj) + \sum_{r\in[j]} t^j_r(rl|ki). \tag{45}$$

However, the sum over *kl* can be restricted since the integrals (*ik*|*jl*), (*rk*|*lj*), and (*rl*|*ki*) decay exponentially with the distance of *i* and *k* or *j* and *l*. Furthermore, the integrals (*rk*|*ls*) become small if the *r*, *s* ∈ [*ij*] are far from *k*, *l*. Finally, the overlap integrals in Eq. (44) become also small if *ij* is far from *kl*. Similar considerations apply to some other terms. In the current work, we use exactly the same restrictions as described in detail in Ref. 43.

Overall, the generation and storage of the 3-ext and 4-ext integrals is the major bottleneck of OSV-LCCSD. As will be shown in Sec. III, the total number of these integrals depends crucially on the threshold l_{osv}. Depending on this threshold, the number of integrals and the computational effort may be smaller or larger than for PAO-LCCSD.

H. Local pair approximations

If all pairs are included in the LCCSD, the CPU time as well as the disk space scale formally as $\mathcal{O}(N^4)$. However, both can be reduced to linear scaling by introducing pair approximations. In principle, it would be sufficient to neglect very distant pairs which have negligible contributions to the correlation energy. However, the cross-over point to low-order scaling then occurs only for quite large molecular sizes and will not be reached in most applications involving medium-size molecules (50–100 atoms). In the past, additional approximations were therefore introduced for weak pairs,[15–17,40,43] which have small but non-negligible contributions to the correlation energy. The current work follows the earlier developments. We classify the orbital pairs according to the distance of the atoms that contribute to the primary domains. This can be done either by distance or connectivity criteria. Here we use the latter, in which the pair classes depend on the minimum number of bonds between any atom in domain [*i*] and any atom in domain [*j*]. We distinguish *strong*, *close*, *weak*, and *very distant* pairs. The latter are entirely neglected. The amplitudes of strong pairs are fully optimized by LCCSD, while the remaining amplitudes are determined by LMP2.

The weak pair approximation usually leads to an overestimation of the correlation energy.[15,47] In most cases this is not caused by an overshooting of the LMP2 weak pair

correlation energies, but due to the neglect of the weak pair amplitudes in the LCCSD equations for the strong pairs. The overshooting due to the pair approximation partly compensates the error caused by the domain approximation (*domain error*), since the latter reduces the correlation energies. This error compensation is favorable if one uses standard PAO domains and medium size basis sets. However, if the accuracy of the domain approximation is improved by extending the domains, using OSVs, or by including explicitly correlated terms in the wavefunction,[46] the error compensation is lost and the error of the pair approximation dominates. It is then necessary to include more pairs in the LCCSD. Furthermore, the error can be significantly reduced with small additional cost by including the LMP2 close pair amplitudes in the LCCSD residual equations for the strong pairs.[15,46]

If connectivity criteria are used, the pair approximation can be specified by three integers w, c, and k. w and c specify the minimum number of bonds between pairs of orbitals that form weak and close pairs, respectively. $k = 1$ means that close pairs are included in the LCCSD residuals for the strong pairs. For $k = 0$ this is not done, and in the absence of triple excitations there is then no difference between close and weak pairs. The default for standard LCCSD calculations with triple-ζ basis sets is $wck = 210$. This means that in strong pairs the two orbital domains must share at least one atom, close pairs are separated by 1 bond, and weak pairs are separated by at least 2 bonds. Very distant pairs are neglected if the distance between the two orbitals exceeds 15 a_0. If MP2 corrections are applied or explicitly correlated wavefunctions are used, the lists of strong and close pairs must be increased, and $wck = 321$ has been recommended as a good compromise between accuracy and cost.[15,46]

One may very well argue that the use of distance criteria is incompatible with the goal of avoiding physically motivated *ad hoc* approximations in the OSV-LCCSD method. The main reason for still employing the same distance criteria as in the previous work was to be able to directly compare the OSV and PAO values to previous results, using exactly the same pair approximations. One could equally well determine the pair classes solely on the basis of LMP2 pair energies, so that no definitions of distances or bonds between LMOs would be necessary any more. It would then be possible to control the whole calculation by a single energy threshold.

III. BENCHMARK CALCULATIONS

In this section, we will investigate the dependence of the LCCSD correlation energy and the computational cost as a function of the domain sizes, using PAOs or OSVs. In order to isolate the effect of the domains, we begin with calculations in which all pairs are treated at the LCCSD level (Sec. III A). Additional pair approximations will be considered in Sec. III B. A benchmark for reaction energies will be presented in Sec. III C, and applications to larger molecular systems in Sec. IV.

In the following, we will first investigate the convergence of the correlation energies as a function of the thresholds l_{pao} and l_{osv} towards the canonical CCSD limit obtained with the same basis set. We will denote basis sets consisting of cc-pVxZ for hydrogen atoms and aug-cc-pVxZ for other atoms as aVxZ. It will be shown that this convergence can be much improved by adding an MP2 domain correction[15,47]

$$E_{\mathrm{CCSD}} \approx E_{\mathrm{LCCSD}} + E_{\mathrm{MP2}} - E_{\mathrm{LMP2}}, \tag{46}$$

where all energies are computed with the same basis set. In the following, we denote this as ΔMP2 correction. One might argue that the calculation of the canonical MP2 energy E_{MP2} leads to $\mathcal{O}(N^5)$ scaling and might therefore dominate the computational cost in large molecules. However, the DF-MP2 method is very efficient and well parallelized. For example, the DF-MP2 calculation for $(\mathrm{Gly})_{10}$ (2334 basis functions, 228 correlated electrons) took 429 minutes elapsed time on a single core. Using 12 cores and 2 Nvidia C2070 graphics processor units on the same machine, this can be reduced to just 11.5 min elapsed time (without HF).

With increasing domain sizes, the above procedure should converge to the canonical CCSD limit for a given basis set. However, the basis set error is often larger than the domain error, and the goal should be to approach the CCSD complete basis set (CBS) limit. This can be achieved by replacing, say a triple-zeta MP2 energy by the CBS limit, i.e.,

$$E_{\mathrm{CCSD/CBS}} \approx E_{\mathrm{LCCSD/aVTZ}} + E_{\mathrm{MP2/CBS}} - E_{\mathrm{LMP2/aVTZ}}\,. \tag{47}$$

This corrects both for domain and basis set incompleteness errors. The performance of this approximation for reaction energies will be investigated in Sec. III C. We note that the expensive MP2 extrapolations could be avoided by using the explicitly correlated LMP2-F12 method,[64] which scales much lower with molecular size and should be at least as accurate as aVQZ/aV5Z extrapolated MP2 values.

A. Dependence of the correlation energy and computational cost on the domain sizes

As test examples we have chosen a subset of the molecules used by Neese *et al.*[32] in their benchmarks of the PNO-LQCISD and PNO-LCCSD methods, namely pyrazole, 2-hydroxypyridine, cyclooctatetraene, neopentane, vinyl acetate, and vinylcyclopropane. All the geometries were obtained based on MP2/VTZ optimizations.[65] In all cases the aVTZ basis set was used. The corresponding aVTZ/MP2FIT basis sets of Weigend *et al.*[66] were used in the density fitting for all integrals except for the 4-external ones. As shown in Ref. 15, the cardinal number should be increased by one for the latter integral class in order to keep the fitting error on the absolute correlation energies small. For example, in the case of pyrazine the correlation energies are overestimated by about 0.05% when the aVTZ/MP2FIT fitting sets are used for the 4-external integrals. Thus, we have used the aVQZ/MP2FIT sets for these integrals, and then the fitting errors are negligible. The reference CCSD calculations did not involve any density fitting approximations, but in the MP2, LMP2, and LCCSD calculations all integrals were obtained by density fitting.

TABLE I. Average pair domain sizes AVD (including redundant functions), correlation energies (in E_h), computation times (in min), and file sizes (in GB) for various molecules and domain selection thresholds. The percentage of correlation energy relative to the canonical CCSD value is given in parenthesis. All pairs are included in the LCCSD. Basis set: hydrogen atoms cc-pVTZ, other atoms aug-cc-pVTZ. Calculations were carried out on a single core Xeon X5690 @ 3.47 GHz. Timings for the complete LCCSD calculations (including integral evaluation and transformations, 10 iterations).

		Correlation energies			Timings		File sizes	
Molecule	AVD	LMP2	LCCSD	LCCSD+ΔMP2	CPU	WALL	3-ext.	4-ext.
			OSV, $l_{osv} = 1.0 \times 10^{-4}$					
Vinylcyclopropane	58	−0.828460	−0.872329 (98.90)	−0.884794 (100.31)	6.2	6.7	2.0	2.9
Pyrazole	62	−0.885997	−0.895608 (98.88)	−0.908675 (100.32)	5.4	6.0	2.0	3.7
Neopentane	56	−0.893418	−0.955388 (98.73)	−0.971244 (100.37)	8.7	9.4	2.5	2.8
Vinylacetate	57	−1.117724	−1.147004 (98.95)	−1.161542 (100.20)	10.3	11.1	3.3	3.8
2-hydroxypyridine	64	−1.215459	−1.231890 (98.75)	−1.251569 (100.32)	18.9	20.3	5.6	7.9
Cyclooctatetraene	63	−1.255072	−1.302675 (98.72)	−1.324573 (100.38)	28.4	30.0	7.0	7.8
			OSV, $l_{osv} = 3.2 \times 10^{-5}$					
Vinylcyclopropane	79	−0.835851	−0.877888 (99.53)	−0.882962 (100.11)	14.8	16.2	4.9	9.9
Pyrazole	85	−0.893944	−0.901541 (99.53)	−0.906660 (100.10)	14.3	16.0	5.1	13.4
Neopentane	76	−0.902641	−0.962382 (99.45)	−0.969016 (100.14)	20.8	22.4	6.4	10.0
Vinylacetate	76	−1.126264	−1.153745 (99.53)	−1.159744 (100.05)	23.1	24.8	7.7	11.9
2-hydroxypyridine	88	−1.227604	−1.241121 (99.49)	−1.248655 (100.09)	50.7	57.5	15.0	30.8
Cyclooctatetraene	85	−1.267911	−1.312207 (99.44)	−1.321267 (100.13)	66.6	71.2	17.7	27.2
			OSV, $l_{osv} = 1.0 \times 10^{-5}$					
Vinylcyclopropane	102	−0.838929	−0.880193 (99.79)	−0.882190 (100.02)	32.2	35.8	10.7	27.9
Pyrazole	110	−0.897182	−0.903996 (99.80)	−0.905878 (100.01)	33.5	37.8	11.2	39.2
Neopentane	100	−0.906558	−0.965336 (99.76)	−0.968053 (100.04)	46.2	50.1	14.5	29.8
Vinylacetate	97	−1.129896	−1.156693 (99.78)	−1.159059 (99.99)	48.8	59.1	15.9	31.0
2-hydroxypyridine	114	−1.232073	−1.244548 (99.76)	−1.247614 (100.01)	123.5	182.1	32.7	90.3
Cyclooctatetraene	110	−1.273342	−1.316225 (99.74)	−1.319854 (100.02)	148.0	200.5	38.7	78.2
			PAO, $l_{pao} = 1.0 \times 10^{-3}$					
Vinylcyclopropane	126	−0.829325	−0.871671 (98.83)	−0.883271 (100.14)	14.8	15.7	2.2	8.6
Pyrazole	143	−0.889414	−0.897179 (99.05)	−0.906829 (100.11)	11.9	12.5	1.2	6.5
Neopentane	116	−0.894701	−0.954651 (98.66)	−0.969224 (100.16)	20.0	21.1	3.9	9.9
Vinylacetate	141	−1.122412	−1.150099 (99.21)	−1.159949 (100.06)	24.4	25.4	3.1	10.8
2-hydroxypyridine	153	−1.220476	−1.234267 (98.94)	−1.248930 (100.11)	37.3	38.9	4.2	16.7
Cyclooctatetraene	138	−1.256212	−1.301524 (98.63)	−1.322283 (100.20)	51.3	54.6	8.5	20.6
			PAO, $l_{pao} = 3.2 \times 10^{-4}$					
Vinylcyclopropane	167	−0.832466	−0.874455 (99.14)	−0.882914 (100.10)	22.7	23.8	2.2	11.7
Pyrazole	219	−0.896650	−0.903571 (99.75)	−0.905985 (100.02)	24.8	25.5	1.2	7.8
Neopentane	185	−0.900953	−0.960242 (99.23)	−0.968563 (100.09)	41.9	43.5	3.9	16.8
Vinylacetate	209	−1.127343	−1.154661 (99.61)	−1.159580 (100.03)	52.2	53.6	3.1	16.0
2-hydroxypyridine	244	−1.229159	−1.242015 (99.56)	−1.247995 (100.04)	89.7	91.8	4.2	24.0
Cyclooctatetraene	215	−1.263339	−1.307198 (99.06)	−1.320831 (100.09)	128.0	133.7	8.5	44.8
			PAO, $l_{pao} = 1.0 \times 10^{-4}$					
Vinylcyclopropane	236	−0.837956	−0.879321 (99.69)	−0.882290 (100.03)	42.6	43.9	2.2	14.2
Pyrazole	239	−0.897597	−0.904454 (99.85)	−0.905921 (100.01)	29.3	30.0	1.2	7.8
Neopentane	204	−0.903514	−0.962538 (99.47)	−0.968298 (100.07)	56.7	58.7	3.9	24.3
Vinylacetate	262	−1.129876	−1.156985 (99.81)	−1.159371 (100.01)	79.6	81.3	3.1	17.8
2-hydroxypyridine	311	−1.233045	−1.245567 (99.84)	−1.247660 (100.01)	147.8	150.2	4.2	25.3
Cyclooctatetraene	338	−1.272886	−1.315892 (99.72)	−1.319976 (100.03)	293.2	299.6	8.5	54.5

The results for the six test molecules are presented in Table I, which shows the convergence of the correlation energy and of the computational resources (CPU and elapsed times, disk space) as a function of the domain selection thresholds l_{osv} and l_{pao} for OSV-LCCSD and PAO-LCCSD, respectively. As an example, the convergence of the OSV-LCCSD and OSV-LCCSD+ΔMP2 correlation energies as a function of the domain sizes are shown for pyrazole in Fig. 1. For all molecules one can observe that much smaller domains are sufficient with OSVs than with PAOs to achieve a certain accuracy of the correlation energy. Typically, with OSVs 99.5% of the correlation energy is recovered with average pair domain sizes of 80–90 ($l_{osv} = 3.2 \times 10^{-5}$). For the same accuracy, the PAO domains need to be 2–3 times larger (l_{pao}). If the MP2 domain correction is added, the total correlation energies overestimate the canonical limit, indicating that the domain error is somewhat larger for LMP2 than for LCCSD. This overshooting is more pronounced with OSVs

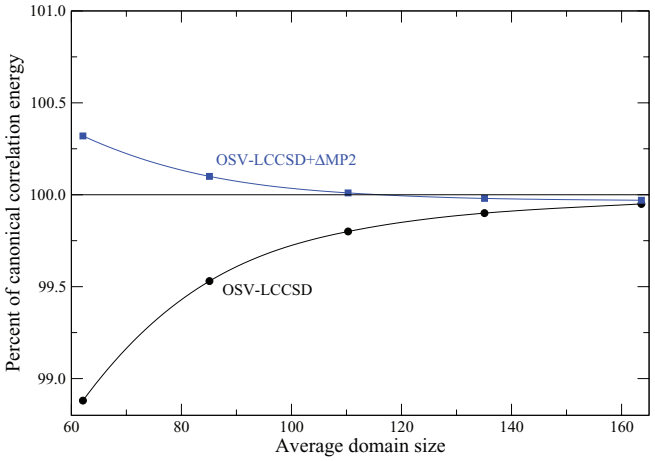

FIG. 1. Fraction of the OSV-LCCSD and OSV-LCCSD + ΔMP2 correlation energies relative to the CCSD value as a function of the average domain size for pyrazole (basis aVTZ, see text). No pair approximations are applied.

than with PAOs for small domains. Already with the smallest PAO domains considered here ($l_{pao} = 1.0 \times 10^{-3}$), the MP2-corrected error is below 0.2% of the canonical correlation energy (typical domain sizes 140). For a comparable accuracy an OSV threshold of $l_{osv} = 3.2 \times 10^{-5}$ is required (average domain size 85). It is very satisfying that the average OSV domain sizes as well as the percentage of correlation energy for a given threshold are very similar for all molecules. In the PAO case the fraction of correlation energy varies considerably more, since the domain selection is less fine grained.

We now consider the computation times and the required disk space. Because of the smaller pair domain sizes, the OSV calculations can be faster than the PAO calculations for a similar accuracy. For example, for an accuracy of 99.5% in the correlation energy ($l_{pao} = 3.0 \times 10^{-4}$ and $l_{osv} = 3.2 \times 10^{-5}$), the OSV calculations are up to about a factor of two faster. However, it is obvious that there must be a cross-over point between PAOs and OSVs at some accuracy. This is due to the fact as we approach higher accuracies the total space of OSVs (for all LMOs) eventually becomes larger than the number of virtual orbitals, and therefore the number of required transformed integrals strongly increases with decreasing OSV threshold. With PAOs the number of integrals becomes independent of the threshold much earlier (in fact, the number of 3-external integrals is independent of the threshold if all pairs are included). This is because in the PAO case different pairs share the PAOs at certain centers, and the minimum number of 4-external integrals is determined by center quadruplets $(AB|CD)$, where the PAOs at all four centers A, B, C, D must belong to the same pair domain.[15,43] In the OSV case this is replaced by the condition that the four OSVs for a pair ij belong either to orbital i or to orbital j, i.e., the smallest number of OSVs that can be shared by different pairs is determined by the number of OSVs per orbital, cf. Sec. II G.

The above also means that with decreasing energy thresholds the I/O times grow more rapidly for OSV-LCCSD than for PAO-LCCSD. This can be observed, e.g., for cyclooctatetraene or 2-hydroxypyridine, where for the lowest thresholds the difference between the elapsed and CPU times is much larger in the OSV calculation. It should be noted that the calculations were carried out on a machine with rather large memory (96 GB) and fast file systems (two RAID0 file systems with a bandwidth of about 300 MB/s each). If a machine with less memory and slower I/O would be used, the elapsed times would be more dominated by the I/O overhead, and this would change the relative timings for small thresholds in favor of the PAO method.

B. Dependence of the correlation energy and computational cost on pair approximations

The above calculations included all pairs. A very significant speedup can be achieved by pair approximations. Table II shows the results for the same molecules and domain selection thresholds as used in Table I, but now including pair approximations ($wck = 321$). As expected, the savings by the pair approximations increase with molecular size. In the case of cyclooctatetraene both the timings as well as the file sizes are reduced by about a factor of two. Clearly, the savings would be more pronounced for larger molecules. As for the calculation with all pairs, the OSV results for a given threshold yield more consistent fractions of the canonical correlation energies than the PAO ones. It can also be seen that the pair approximations lead to a slight overestimation of the correlation energies, as discussed in Sec. II H. Note that the LMP2 correlation energies are smaller than the LCCSD ones for all six test molecules, and therefore the overestimation is not due to an overestimation of the LMP2 weak pair energies.

Table III shows the percentage of correlation energy recovered by OSV-LCCSD and PAO-LCCSD relative to canonical CCSD calculations for glycine polypeptides $(Gly)_n$, $n = 1 - 4$ (with our computational resources, larger calculations are not possible with canonical CCSD). The basis set and other computational details are the same as in Sec. III A. The results for different choices of the weak pair parameters wck demonstrate again the overestimation of the correlation energies by the pair approximations. Comparison of the calculations for $wck = 210$ and $wck = 211$ shows that this is significantly reduced by setting $k = 1$ (cf. Section II H). It is further reduced if the close and weak pair parameters are increased. For the larger peptides, the percentage of correlation energy that is recovered for each choice of wck is nearly independent of the molecular size. This means that the absolute error is proportional to the molecular size. However, in most situations this error will cancel to a large extent when energy differences are considered. This is similar to the basis set error on the correlation energies, which also increases with molecular size.

The dependence of the computational resources on the molecular size is demonstrated in Table IV. As before, the aVTZ basis set was used and the weak pair parameters were varied. The scaling of the OSV-LCCSD and PAO-LCCSD calculation as a function of the molecular size is very similar, and therefore the larger calculations have only been done with OSV-LCCSD. Since wall-clock (elapsed) times depend strongly on the machine configuration (memory, disk), we only report pure user-CPU times. The disk space given is the maximum used at any stage of the calculation. The

TABLE II. Average pair domain sizes AVD (including redundant functions), correlation energies (in E_h), computation times (in min), and file sizes (in GB) for various molecules. The percentage of correlation energy relative to the canonical CCSD value is given in parenthesis. Pairs are classified using `wck=321` (see text). Basis set: hydrogen atoms cc-pVTZ, other atoms aug-cc-pVTZ. Calculations were carried out on a single core Xeon X5690 @ 3.47 GHz. Timings for the complete LCCSD calculations (including integral evaluation and transformations, 10 iterations).

		Correlation energies			Timings		File sizes	
Molecule	AVD	LMP2	LCCSD	LCCSD+ΔMP2	CPU	WALL	3-ext.	4-ext.
			OSV, $l_{osv} = 1.0 \times 10^{-4}$					
Vinylcyclopropane	58	−0.828460	−0.872686 (98.94)	−0.885151 (100.35)	4.3	4.7	1.3	2.0
Pyrazole	62	−0.885997	−0.895835 (98.90)	−0.908902 (100.34)	5.0	5.5	1.8	3.4
Neopentane	56	−0.893418	−0.955810 (98.78)	−0.971666 (100.41)	5.9	6.3	1.6	1.7
Vinylacetate	57	−1.117724	−1.147728 (99.01)	−1.162265 (100.26)	5.9	6.4	1.8	2.3
2-hydroxypyridine	64	−1.215458	−1.232607 (99.80)	−1.252287 (100.38)	13.3	14.3	4.0	6.2
Cyclooctatetraene	63	−1.255072	−1.305784 (98.95)	−1.327682 (100.61)	14.7	15.6	3.5	4.4
			OSV, $l_{osv} = 3.2 \times 10^{-5}$					
Vinylcyclopropane	79	−0.8358514	−0.8782953 (99.58)	−0.8833694 (100.15)	10.3	11.2	3.4	7.0
Pyrazole	85	−0.8939442	−0.9017939 (99.56)	−0.9069134 (100.12)	13.1	14.6	4.7	12.6
Neopentane	76	−0.9026407	−0.9628540 (99.50)	−0.9694876 (100.19)	13.5	14.5	4.3	6.3
Vinylacetate	76	−1.1262639	−1.1545054 (99.59)	−1.1605035 (100.11)	13.1	14.2	4.3	7.2
2-hydroxypyridine	88	−1.2276038	−1.2419553 (99.55)	−1.2494894 (100.16)	35.8	39.2	11.0	25.0
Cyclooctatetraene	85	−1.2679105	−1.3155698 (99.69)	−1.3246298 (100.38)	32.3	34.6	8.9	15.3
			PAO, $l_{pao} = 1.0 \times 10^{-3}$					
Vinylcyclopropane	126	−0.829325	−0.872129 (98.88)	−0.883729 (100.19)	11.0	11.7	2.0	6.3
Pyrazole	143	−0.889414	−0.897436 (99.08)	−0.907086 (100.14)	11.2	11.8	1.2	6.1
Neopentane	116	−0.894701	−0.955179 (98.71)	−0.969752 (100.22)	16.7	17.7	3.5	8.3
Vinylacetate	141	−1.122412	−1.150883 (99.28)	−1.160733 (100.13)	17.6	18.5	2.7	8.5
2-hydroxypyridine	153	−1.220476	−1.235166 (99.01)	−1.249829 (100.18)	28.4	29.7	4.0	13.0
Cyclooctatetraene	138	−1.256212	−1.304864 (98.88)	−1.325623 (100.46)	30.7	32.2	6.4	11.8
			PAO, $l_{pao} = 3.2 \times 10^{-4}$					
Vinylcyclopropane	167	−0.832466	−0.874876 (99.19)	−0.883335 (100.15)	17.2	18.1	2.1	9.4
Pyrazole	219	−0.896650	−0.903816 (99.78)	−0.906229 (100.05)	23.3	24.1	1.2	7.8
Neopentane	185	−0.900953	−0.960789 (99.29)	−0.969111 (100.15)	34.1	35.6	3.8	13.7
Vinylacetate	209	−1.127343	−1.155413 (99.67)	−1.160332 (100.10)	33.9	35.2	2.9	13.0
2-hydroxypyridine	244	−1.229159	−1.242842 (99.62)	−1.248821 (100.10)	69.3	71.4	4.2	21.7
Cyclooctatetraene	215	−1.263338	−1.310778 (99.33)	−1.324411 (100.36)	87.5	91.3	8.3	38.3

number of correlated electrons (N_{el}) is taken as a measure for the molecular size, but virtually the same scalings would be obtained by using the number of basis functions, since in the current case $N_{el} \approx 10\,N_{\mathrm{AO}}$. We find that for all choices of wck the scaling behaviour is nearly the same. The overall OSV-LCCSD CPU-times [evaluated from $(\mathrm{Gly})_6$ and $(\mathrm{Gly})_8$] scale cubically [$\mathcal{O}(N_{el}^{3.1})$ to $\mathcal{O}(N_{el}^{3.2})$], while the disk space scales quadratically [$\mathcal{O}(N_{el}^{2.0})$ to $\mathcal{O}(N_{el}^{2.1})$]. These are the expected values, since no local density fitting approximations are applied.

TABLE III. Comparison of OSV and PAO correlation energies relative to the canonical CCSD correlation energy for linear polyglycine chains (in percent). The pair selection thresholds wck (see text) are varied. The aVTZ basis set has been used. Average domain size (AVD) includes redundant functions.

		LCCSD				LCCSD+ΔMP2			
Molecule	AVD	$wck = 210$	$wck = 211$	$wck = 311$	$wck = 321$	$wck = 210$	$wck = 211$	$wck = 311$	$wck = 321$
				OSV ($l_{osv} = 1.0 \times 10^{-4}$)					
$(\mathrm{Gly})_1$	56	99.62	99.21	99.13	99.03	100.80	100.39	100.31	100.21
$(\mathrm{Gly})_2$	58	99.64	99.13	99.01	98.89	101.04	100.53	100.42	100.29
$(\mathrm{Gly})_3$	59	99.65	99.11	98.98	98.84	101.13	100.58	100.46	100.32
$(\mathrm{Gly})_4$	60	99.66	99.10	98.96	98.82	101.18	100.61	100.48	100.33
				PAO ($l_{pao} = 1.0 \times 10^{-3}$)					
$(\mathrm{Gly})_1$	121	99.85	99.42	99.32	99.22	100.73	100.29	100.20	100.10
$(\mathrm{Gly})_2$	143	99.96	99.43	99.29	99.15	100.96	100.42	100.28	100.15
$(\mathrm{Gly})_3$	145	99.95	99.37	99.22	99.07	101.05	100.47	100.32	100.17
$(\mathrm{Gly})_4$	147	99.96	99.36	99.19	99.04	101.10	100.50	100.34	100.18

TABLE IV. Comparison of OSV and PAO computational resources for linear polyglycine chains. CPU times are user times, disk space is the total disk space used by the calculations. The pair selection thresholds wck (see text) are varied. The aVTZ basis set has been used. N_{AO} is the number of basis functions (CGTOs), N_{el} is the number of correlated electrons.

			CPU-times (min)				Disk space (GB)			
Molecule	N_{AO}	N_{el}	$wck = 210$	$wck = 211$	$wck = 311$	$wck = 321$	$wck = 210$	$wck = 211$	$wck = 311$	$wck = 321$
					OSV ($l_{osv} = 1.0 \times 10^{-4}$)					
$(Gly)_1$	300	30	2.0	2.2	2.5	3.9	3.4	3.6	3.9	6.2
$(Gly)_2$	526	52	10.6	11.2	13.2	18.3	9.9	10.6	13.0	19.2
$(Gly)_3$	752	74	31.4	33.3	38.9	51.1	20.0	21.6	27.5	40.0
$(Gly)_4$	978	96	75.5	81.2	92.2	117.9	34.8	38.1	48.9	70.0
$(Gly)_6$	1430	140	271.9	284.4	303.1	374.6	75.9	80.0	100.9	144.9
$(Gly)_8$	1882	184	655.4	682.5	744.1	873.8	134.1	140.5	175.4	250.4
$(Gly)_{10}$	2334	228	1308.9	1342.7	1448.6	1688.9	211.5	220.5	272.9	386.6
					PAO ($l_{pao} = 1.0 \times 10^{-3}$)					
$(Gly)_1$	300	30	4.2	4.5	5.2	8.9	4.9	5.0	5.4	8.5
$(Gly)_2$	526	52	19.0	20.9	25.2	37.3	14.5	14.9	17.5	29.2
$(Gly)_3$	752	74	55.0	58.3	70.8	103.2	23.1	24.1	29.9	50.3
$(Gly)_4$	978	96	123.2	134.8	159.2	216.2	36.0	37.6	47.0	77.5

In more detail, the scaling exponents x for the 2-ext/3-ext/4-ext integral transformations ($wck = 321$) are $x = 3.1/3.2/3.3$. The total disk space is dominated by the 3-index integrals needed in the density fitting integral transformations. As expected, the scaling of the final 3-ext and 4-ext file sizes (not shown) is close to linear ($x = 1.1$). This affects the I/O time in the iterations, and indeed the elapsed times for the iterations ($x = 2.1$) scale better than the CPU-times ($x = 2.5$). Overall, the iterations take only about 15% of the total CPU-time for $(Gly)_8$. About half of the iteration time is required to compute the matrix

$$[\mathbf{G}(\mathbf{E})]_{rs} = \sum_i \sum_u [2(rs|iu) - (ri|su)]t_u^i. \qquad (48)$$

This matrix is needed for *all* r, s, and is therefore computed directly using density fitting, in analogy to a Fock matrix (for details see Ref. 15). Without local fitting, this part scales with $x = 2.6$. In Refs. 15 and 43 an approximation has been proposed that neglects the contributions of $\mathbf{G}(\mathbf{E})$ (along with some other terms) entirely, but overall the saving by this approximation is rather small. The scaling of most other terms in the iterations is close to linear.

As a final example we present in Table V some calculations for penicillin. The structure is the same as in the PNO-CCSD calculations of Neese *et al.*,[32] and also the basis set is (almost) the same (we used def2-TZVPP, 1009 CGTOs, Neese *et al.* used TZV(2df, 2pd), 999 CGTOs). Their calculations took 1004–1170 min total elapsed time (depending on the approximation used, using Intel Xeon 3.0 GHZ CPUs). Taking into account that our machine is probably 10% –20% faster (Intel Xeon 3.47 GHZ CPUs), these times are comparable with our most accurate calculations. In practice, the calculations with the largest thresholds in Table V should be sufficient if the MP2 correction is applied. These calculations are more than four times faster. Unfortunately, the correlation energies are not given in Neese's paper. It is also interesting to compare the iteration times: In the PNO-CCSD calculations, 55% –65% of the time was spent in the iterations. In our OSV calculations, the iterations took 40% ($l_{osv} = 1.0 \times 10^{-4}$) to 45% ($l_{osv} = 1.0 \times 10^{-5}$) of the total time. Note that these

TABLE V. Average pair domain sizes AVD,[a] correlation energies (in E_h), computation times (in min), and file sizes (in GB) for various domain selection thresholds for penicillin, using the TZVPP basis set and $wck = 321$. The relative contribution of the MP2 correction (in percent) is given in parenthesis. Calculations were carried out on a single core Xeon X5690 @ 3.47 GHz. Timings for the complete LCCSD calculations (including integral evaluation and transformations, 11 iterations)

		Correlation energies			Timings		File sizes	
THR	AVD	LMP2	LCCSD	LCCSD+ΔMP2	CPU	WALL	3-ext.	4-ext.
				OSV				
1.0×10^{-4}	55	−4.234663	−4.331603	−4.412904 (1.88)	164.6	172.8	12.6	8.8
3.2×10^{-5}	73	−4.279226	−4.366733	−4.403473 (0.84)	323.8	374.1	31.2	30.5
1.0×10^{-5}	97	−4.301803	−4.384751	−4.398913 (0.32)	634.2	806.8	76.0	100.5
				PAO				
1.0×10^{-3}	110	−4.244243	−4.333514	−4.405236 (1.66)	177.2	183.8	21.4	15.8
3.2×10^{-4}	165	−4.270262	−4.356595	−4.402298 (1.05)	296.1	312.5	35.1	33.4
1.0×10^{-4}	259	−4.295844	−4.379334	−4.399456 (0.46)	704.1	816.5	72.6	92.7

[a]The domain sizes include redundant functions. $E_{\text{DF-HF}} = -1497.525214\ E_h$, $E_{\text{DF-MP2}} = -1501.841179\ E_h$.

fractions are larger than for the linear glycine chains. This is mainly due to the larger effort to compute the intermediates $\mathbf{Y}^{kj}$ and $\mathbf{Z}^{kj}$ (cf. Ref. 15) since the united domains (cf. Sec. II F) are larger in molecules with a compact structure than in the linear glycine chains. About half of the time to compute these intermediates is needed for the singles contributions involving 1-ext integrals. Overall, the timings for penicillin are very similar for the PAO-LCCSD and OSV-LCCSD calculations, but for the chosen thresholds the OSV-LCCSD correlation energies are slightly more accurate. This is also seen by the fact that the MP2 corrections (given in parenthesis in %) are smaller in the OSV case.

C. Reaction energies

In order to test the accuracy of relative energies as a function of the domain sizes, we carried out calculations for the benchmark of 52 reactions presented in Ref. 46. In order to make the calculations comparable, exactly the same basis sets, orbitals, and pair lists as in Ref. 46 were used. The basis set is VTZ-F12,[67] the localization Pipek-Mezey (with the contributions of the most diffuse functions of each angular momentum removed in the localization criterion, `CPLDEL=1`), and the pair approximation is $wck = 321$. This yields results that are very close to those without any pair approximations. The aug-cc-pVTZ/MP2FIT sets was used for all integrals, even for the 4-external ones. It has been verified that this has only a negligible effect (0.1 kJ mol^{-1}) on the statistical values of the relative energies. This is consistent with the findings of Ref. 15.

Table VI summarizes the results relative to the CCSD reference for the same basis set. This monitors just the domain errors as a function of the domain selection thresholds for the given basis set. It is found that acceptable maximum (MAX) and root mean square (RMS) deviations from the CCSD values require tight domain selection thresholds. Only with $l_{osv} = 10^{-5}$ is a satisfactory accuracy (MAX deviation ca 1 kcal mol^{-1}). With PAOs, comparable accuracy is achieved with $l_{pao} = 10^{-4}$. However, when the MP2 domain correction is applied the convergence with domain size is strongly accelerated, and the same overall accuracy is already achieved with $l_{osv} = 10^{-4}$ and $l_{pao} = 10^{-3}$, respectively.

Table VII shows similar results, but this time relative to the CCSD/CBS values. The latter were taken from Ref. 46.

TABLE VI. Maximum (MAX) and root mean square (RMS) errors of PAO-LCCSD and OSV-LCCSD calculations (in kJ mol^{-1}) relative to canonical CCSD values for the 52 reactions of Ref. 46 for different values of the domain selection threshold THR. The VTZ-F12 basis set has been used. ΔMP2=MP2−LMP2 is a correction for domain errors. The LMP2 values are computed with the same domains as the corresponding LCCSD values.

	PAO		OSV		PAO+ΔMP2		OSV+ΔMP2	
THR	MAX	RMS	MAX	RMS	MAX	RMS	MAX	RMS
1.0×10^{-3}	16.3	4.7	41.5	14.6	3.8	1.0	11.4	3.3
3.2×10^{-4}	15.0	3.8	17.5	7.8	4.1	0.9	6.5	1.8
1.0×10^{-4}	4.3	1.5	12.6	4.3	2.3	0.6	5.3	1.4
3.2×10^{-5}	3.3	0.8	4.3	1.8	2.2	0.5	3.4	0.9
1.0×10^{-5}	2.2	0.6	3.0	0.9	1.8	0.5	2.4	0.7

TABLE VII. Maximum (MAX) and root mean square (RMS) errors of PAO-LCCSD and OSV-LCCSD calculations relative to canonical CCSD/CBS values[a] for the 52 reactions of Ref. 46 for different values of the domain selection threshold THR. The VTZ-F12 basis set has been used. ΔMP2=MP2/CBS−LMP2 is a correction for both basis set and domain errors. The LMP2 values are computed with the same domains as the corresponding LCCSD values.

	PAO		OSV		PAO+ΔMP2		OSV+ΔMP2	
THR	MAX	RMS	MAX	RMS	MAX	RMS	MAX	RMS
1.0×10^{-3}	21.7	7.2	45.1	17.3	4.2	1.6	10.9	3.4
3.2×10^{-4}	20.4	6.6	21.1	10.0	4.5	1.6	6.1	1.8
1.0×10^{-4}	12.4	4.6	14.2	6.5	4.2	1.5	4.7	1.4
3.2×10^{-5}	12.7	4.4	13.9	5.1	4.4	1.5	3.8	1.3
1.0×10^{-5}	12.9	4.3	13.5	4.5	4.0	1.5	3.8	1.2
Canonical	12.2	4.4	12.2	4.4	3.6	1.3	3.6	1.3

[a]See text for details of the CBS estimates.

They were obtained by extrapolating CCSD-F12b correlation energies as described in Ref. 68 and should be as accurate as CCSD aVQZ/aV5Z extrapolations. The MP2 corrections have been computed using Eq. (47). The MP2 CBS values were obtained by extrapolating the DF-MP2 aVQZ and aV5Z correlation energies using $E_n = E_{\mathrm{CBS}} + An^{-3}$, where n are the cardinal numbers (4,5 in the current case).[69] The Hartree-Fock reference energies were extrapolated using the method of Karton and Martin.[70]

The last line shows the results obtained with canonical CCSD, and this reflects the pure basis set error. It amounts on the average to about 1 kcal mol^{-1} (4.5 kJ mol^{-1}) with maximum errors of 3 kcal mol^{-1}. Again, the MP2 correction strongly improves the results, and reduces the MAX errors to below 1 kcal mol^{-1}. Satisfactory accuracy is again achieved with $l_{osv} = 10^{-4}$ and $l_{pao} = 10^{-3}$. This is consistent with the results in Table VI and the findings in Sec. III A. As already mentioned, for these thresholds the OSV calculations are about a factor of two faster than the PAO ones.

IV. ILLUSTRATIVE APPLICATIONS

In this section, we will present a few further calculations that illustrate that the OSV-LCCSD method can be applied to larger systems of chemical interest.

A. Barrier height of PHBH

Our first example concerns the barrier height in *p*-hydroxybenzoate hydroxylase (PHBH), which has been extensively studied using local correlation methods previously.[71,72] We use the same mixed quantum-mechanical/molecular mechanics (QM/MM) hybrid scheme as previously, were the QM system comprises 49 atoms (cf. Fig. 2) and the environment is described by 19 233 point charges. Only the structures of snapshot 3 of Ref. 71 are considered. The basis set is again aVTZ (1678 CGTOs). For comparison we have carried out the calculations with Pipek-Mezey (PM) (Ref. 73) localized orbitals as well as with NLMOs.[53] In both cases the atom domains (only needed to classify the pairs) were determined using the NPA method.[53] We have only carried out OSV-LCCSD calculations since the

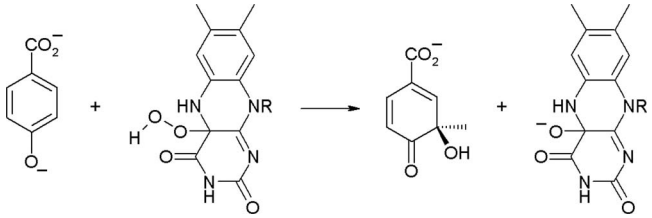

FIG. 2. The PHBH reaction.

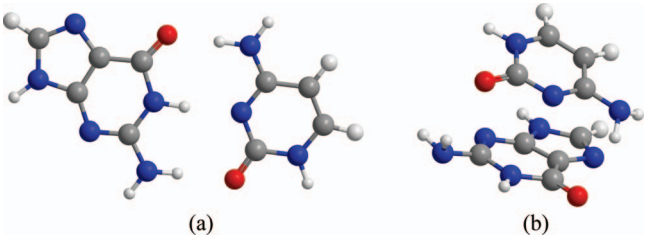

FIG. 3. The G-C dimer structures: Watson-Crick (left) and Stacked (right).

performance of PAO-LCCSD(T) was extensively studied in the earlier papers.

First of all it should be noted that the total correlation effect on the barrier height is very large: it amounts to -15 kcal mol^{-1} and -25 kcal mol^{-1} for LCCSD and LMP2, respectively. As has been shown in the earlier papers, triples excitations have a large effect as well (-7 kcal mol^{-1}).[71] Therefore, the LCCSD barrier heights without triples are much too high, but the only purpose of the current calculations is to demonstrate the convergence as a function of the domain parameter l_{osv}. In all calculations distance criteria were used since connectivity criteria do not work for the extended bond lengths at the transition state geometry. We use in all cases $R_w = 5\,a_0$ and $R_c = 3\,a_0$ and `keepcls=1`, as recommended in Ref. 71. No pairs were neglected in the LMP2.

The results are presented in Table VIII. For both choices of LMOs increasing of the domains (i.e., reducing l_{osv}) lowers the barrier heights. For a given threshold the domains are larger with NLMOs than with PM LMOs, and, consistent with this, the NLMO barrier heights are slightly lower. This effect is more pronounced at the LMP2 than at the LCCSD level; apparently the weak pair approximations compensate some of the domain effects. In view of the difficulties to get balanced results with the PAO-LCCSD method, where domains at the two structures had to be fixed or merged,[52,71] it is very satisfying to see that the OSV-LCCSD values converge smoothly without any special treatment; also the average domain sizes at the two structures are very similar, despite the strongly differing electronic structures. In contrast to the OSV-LCCSD values, the MP2 corrected ones converge from below to the limits. Most likely this is due to the overshooting of the MP2 correction, as discussed in Sec. III. However, the domain effect on the MP2 corrected values is very small. Interestingly, the basis set effect is opposite to the domain effect. This might be due to basis set superposition effects (BSSE), that should lower the barrier height. It is to be expected (and will be shown in Sec. IV B), that the BSSE increases with decreasing l_{osv}, and thus part of the effect of the threshold on the barrier height may be due to BSSE. The final OSV-LCCSD+ΔMP2 results are in good agreement with the PAO-LCCSD+ΔMP2 value of $23.2 + 0.9 = 24.1$ kcal mol^{-1}, taken from Table IV of Ref. 71. Note that the latter value was obtained with a somewhat smaller basis set without diffuse functions on the carbon atoms.

B. Intermolecular interactions

In order to compare the performance of OSV-LCCSD vs. PAO-LCCSD in the context of intermolecular interactions we performed calculations on the guanine-cytosine Watson-Crick (G-C/WC) and stacked (G-C/S) dimers. The geometries were taken from the JSCH-2005 benchmark set presented in Ref. 74. G-C/WC and G-C/S represent hydrogen-bonded and π-stacked complexes (cf. Fig. 3), respectively. The aVTZ AO basis set together with the related aVTZ/MP2FIT fitting basis as specified before was used for these calculations. The local approximation (domains, pair lists, number of redundancies in the pair specific virtual spaces) were determined at large intermolecular separation and kept fixed, as recommended in Refs. 75 and 76 for the treatment of intermolecular interactions. Therefore, in contrast to most other calculations, Pipek-Mezey LMOs were used here, since in our program only these allow for a proper restart at the dimer geometry after the

TABLE VIII. OSV-LMP2 and OSV-LCCSD barrier heights for PHBH, using the aVTZ basis set and two different localization methods. The energy difference contains the MM contribution of -2.8 kcal mol^{-1} (see text). The Hartree-Fock value is 38.74 kcal mol^{-1}. The structures correspond to snapshot 3 in Ref. 71.

THR	AVD(RS)	AVD(TS)	LMOs	LMP2	LCCSD	LCCSD+ΔMP2
1.0×10^{-4}	63	64	NLMO	14.9	24.5	23.1
3.2×10^{-5}	87	87	NLMO	14.3	23.9	23.2
1.0×10^{-5}	112	113	NLMO	14.0	23.8	23.3
3.2×10^{-6}	138	139	NLMO	13.7	...	...
1.0×10^{-4}	65	65	PM	15.4	24.4	22.5
3.2×10^{-5}	90	90	PM	14.5	23.8	22.7
1.0×10^{-5}	116	116	PM	14.2	23.5	22.8
MP2	1566	1566		13.4		
CBS				14.1[a]	24.2[b]	23.5[b]

[a]Extrapolated using the (aug)-cc-pVQZ and (aug)-cc-pV5Z correlation energies (diffuse functions only on N,O).
[b]Estimated using MP2 basis set correction.

TABLE IX. Counterpoise-corrected LCCSD/aVTZ interaction energies in (kcal mol^{-1}) for G-C base pairs. The BSSE estimate according to the CP correction is given in parenthesis. The appended letters (W), (C), and (S) specify the treatment of the intermolecular pairs (see text). Note that the CP correction is identical for (W), (C), and (S).

Method	LMP2	LCCSD(W)	LCCSD(C)	LCCSD(S)	LCCSD(W) +ΔMP2	LCCSD(C) +ΔMP2	LCCSD(S) +ΔMP2
				Watson-Crick			
OSV, 1.0×10^{-4}	−27.42 (0.39)	−28.16	−26.56	−26.77 (0.22)	−30.44	−28.83	−29.04 (1.69)
OSV, 3.2×10^{-5}	−28.26 (0.86)	−29.02	−27.23	−27.48 (0.68)	−30.46	−28.71	−28.92 (1.68)
OSV, 1.0×10^{-5}	−28.71 (1.27)	−29.48	−27.65	−27.87 (1.03)	−30.47	−28.64	−28.86 (1.63)
PAO, IEXT=0	−28.52 (0.01)	−29.30	−27.46	−27.67 (0.04)	−30.47	−28.64	−28.85 (1.82)
PAO, IEXT=1	−29.34 (0.77)	−30.11	−28.14	−28.37 (0.35)	−30.47	−28.50	−28.73 (1.44)
Canonical	−29.70 (1.86)						
				Stacked			
OSV, 1.0×10^{-4}	−16.06 (1.46)	−16.87	−13.50	−12.27 (1.35)	−20.97	−17.61	−16.39 (2.97)
OSV, 3.2×10^{-5}	−17.58 (1.98)	−18.43	−14.77	−13.43 (1.79)	−21.01	−17.35	−16.01 (2.90)
OSV, 1.0×10^{-5}	−18.48 (2.42)	−19.34	−15.53	−14.14 (2.17)	−21.02	−17.02	−15.82 (2.84)
PAO, IEXT=0	−18.88 (0.36)	−19.68	−15.82	−14.47 (0.35)	−20.96	−17.10	−15.75 (3.08)
PAO, IEXT=1	−19.89 (0.72)	−20.75	−16.67	−15.23 (0.67)	−21.02	−16.94	−15.50 (3.04)
Canonical	−20.16 (3.09)						

initial specification of the domains and pair lists at a large intermolecular distance. In order to make sure that the domain information is consistent at both geometries, the localized orbitals at the dimer geometry were rotated to yield maximum overlap with those at the initial large distance.

The results obtained with OSV-LCCSD and the PAO-LCCSD are compiled in Tables IX. Three different OSV selection thresholds have been tried, i.e., $l_{osv} = 1.0 \times 10^{-4}$, 3.2×10^{-5}, and 1.0×10^{-5}. The PAO-LCCSD calculations were performed with standard BP domains (IEXT=0), and with extended domains (IEXT=1, BP domains augmented by all centers one chemical bond away from the closest center of the initial BP domain). To correct for the domain error, the MP2 correction [cf. Eq. (46)] has been applied to both the OSV-LCCSD and PAO-LCCSD interaction energies. As is evident from Table IX, this correction works remarkably well also here, rendering the LCCSD+ΔMP2 interaction energies much less dependent on the domain approximation than the bare LCCSD ones.

Full LCCSD calculations of intermolecular interaction energies are expensive, since the many intermolecular pairs should normally be included in the LCCSD. In order to reduce the computational cost, we tested two approximations in which the intermolecular pairs are treated as weak or close pairs. In Table IX the following three cases are compared:
LCCSD(W): all intermolecular pairs are treated as weak pairs (amplitudes optimized at the LMP2 level, not included in the LCCSD amplitude equations),
LCCSD(C): all intermolecular pairs are treated as close pairs (amplitudes optimized at the LMP2 level, included in the strong LCCSD pair amplitude equations).
LCCSD(S): all intermolecular pairs are treated as strong pairs (fully treated in the LCCSD amplitude equations). In this case the only pair approximation is the intramolecular wck=321 restriction described above.

The H-bonded G-C/WC dimer is rather insensitive to the intermolecular pair approximation, since MP2 is already very accurate for H-bonded complexes.[74,77–79] Interestingly, the pure LMP2 result is even slightly better than LCCSD(W). The π-stacked G-C/S dimer, on the other hand, exhibits a notable dependence on the pair approximation. LCCSD(W) overestimates the interaction energy relative to LCCSD(S) by 5.5 kcal mol^{-1}. LCCSD(C) is much better, but still deviates from LCCSD(S) by 1.5 kcal mol^{-1}. The weak or close pair approximations lead to a strong reduction of the CPU-time and disk space. For example, the 3- and 4-external integrals in OSV-LCCSD (W), (C), and (S) calculations with $l_{osv} = 3.2 \times 10^{-5}$ required 92, 148, and 329 GB of disk space. The corresponding PAO, IEXT=0 values are 106, 194, and 442 GB. It might be possible to get better interaction energies without the expensive full LCCSD treatment of intermolecular pairs by treating the latter at the level of the local random phase approximation, rather than LMP2. This is presently being explored by one of us.

As expected, the BSSE increases with increasing domain sizes (note that it is independent of the treatment of the intermolecular pairs). Comparing OSV-LCCSD and PAO-LCCSD results without the ΔMP2 correction one can see that the PAO-LCCSD contains less BSSE and converges quicker with respect to domain extensions. For example, the OSV-LCCSD(S) result for G-C/S with $l_{osv} = 1.0 \times 10^{-5}$ yields −14.14 kcal mol^{-1}, which compares to −14.47 kcal mol^{-1} for the much cheaper PAO-LCCSD calculation with standard BP domains. The BSSE of the OSV calculation amounts to 2.17 kcal mol^{-1}, not much less than for a canonical calculation. The BSSE of the PAO calculation, on the other hand, is only 0.35 kcal mol^{-1}. Although the PAO domain extensions rapidly converge the interaction energies, the smaller BSSE formally reflects domain incompleteness. For example, the ΔMP2 correction, which removes most of the domain error, restores the BSSE of both PAO and OSV calculations to close to the canonical value. The slow convergence of the counterpoise-corrected OSV intermolecular energy reflects the fact that the OSV orbitals are chosen to reproduce

global MP2 amplitudes and thus the total correlation energy, of which BSSE is a significant part.

V. CONCLUSIONS

We have presented and extensively tested the OSV-LCCSD method as an alternative to the well established PAO-LCCSD method. The OSV treatment allows to fine-tune the accuracy of the local domain approximation by a single energy parameter l_{osv}, and the selection of domains is free of any *ad hoc* assumptions. The convergence of correlation energies and energy differences (reaction energies, barrier heights, intermolecular interaction energies) on the threshold l_{osv} has been extensively tested. While the pure OSV-LCCSD results converge rather slowly with increasing domains, very rapid convergence is obtained if an MP2 domain correction is added. Then in most cases a threshold $l_{osv} = 10^{-4}$ is sufficient. For this value OSV-LCCSD calculations are up to twice as fast as PAO-LCCSD calculations of comparable accuracy. In addition, the required disk space is also smaller. However, for very small thresholds l_{osv} (large domains) OSV-LCCSD is less efficient than PAO-LCCSD since the total number of virtual orbitals may become much larger than the number of PAOs. Our benchmarks demonstrate that the OSV-LCCSD method works very well for reaction energies or barrier heights, while it seems to be less well suited than PAO-LCCSD for intermolecular interaction energies. This is partly due to much larger BSSE effects.

The computational effort of our current OSV-LCCSD implementation scales cubically with molecular size. This could be reduced to linear if local density fitting approximations were applied. Future work is necessary to implement and test such approximations. Furthermore, it will be of utmost importance also to add a perturbative (T) energy correction for triple excitations. It should be quite straightforward to implement this on the basis of the existing PAO-LCCSD(T) program.[15,42] Future work will also focus on adding explicitly correlated terms. It has recently be demonstrated for PAO-LCCSD(T)-F12 that these not only strongly reduce the basis set incompleteness errors, but also the domain errors.[46,65,80,81] This might then replace the MP2 domain correction and the related $\mathcal{O}(N^5)$ overhead. Indeed, instead of canonical MP2 it is already now possible to use a low-order scaling LMP2-F12 method[64] to compute the domain/basis set correction. The accuracy and efficiency of these approaches will be demonstrated in future work.

ACKNOWLEDGMENTS

G.K.C. acknowledges support from the Department of Energy (DOE), Office of Science Award DE-FG02-07ER46432. M.G.S. acknowledges support from the Deutsche Forschungsgemeinschaft (DFG). H.J.W. acknowledges support from the DFG within the SimTech Cluster of Excellence at the University of Stuttgart.

[1] J. Yang, Y. Kurashige, F. R. Manby, and G. K. L. Chan, J. Chem. Phys. **134**, 044123 (2011).
[2] U. Benedikt, A. A. Auer, M. Espig, and W. Hackbusch, J. Chem. Phys. **134**, 054118 (2011).
[3] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).
[4] N. H. F. Beebe and J. Linderberg, Int. J. Quantum Chem. **7**, 683 (1977).
[5] D. W. O'Neal and J. Simons, Int. J. Quantum Chem. **36**, 673 (1989).
[6] H. Koch, A. Sánchez de Merás, and T. B. Pedersen, J. Chem. Phys. **118**, 9481 (2003).
[7] T. Kinoshita, O. Hino, and R. J. Bartlett, J. Chem. Phys. **119**, 7756 (2003).
[8] F. Aquilante, T. B. Pedersen, and R. Lindh, J. Chem. Phys. **126**, 194106 (2007).
[9] F. Aquilante and T. B. Pedersen, Chem. Phys. Lett. **449**, 354 (2007).
[10] F. Weigend, M. Kattannek, and R. Ahlrichs, J. Chem. Phys. **130**, 164106 (2009).
[11] T. S. Chwee and E. A. Carter, J. Chem. Phys. **132**, 074104 (2010).
[12] H.-J. Werner, F. R. Manby, and P. J. Knowles, J. Chem. Phys. **118**, 8149 (2003).
[13] F. R. Manby, J. Chem. Phys. **119**, 4607 (2003).
[14] M. Schütz and F. R. Manby, Phys. Chem. Chem. Phys. **5**, 3349 (2003).
[15] H.-J. Werner and M. Schütz, J. Chem. Phys. **135**, 144116 (2011).
[16] P. Pulay, Chem. Phys. Lett. **100**, 151 (1983).
[17] S. Saebø and P. Pulay, Chem. Phys. Lett. **113**, 13 (1985).
[18] P. Pulay and S. Saebø, Theor. Chim. Acta **69**, 357 (1986).
[19] S. Saebø and P. Pulay, J. Chem. Phys. **86**, 914 (1987).
[20] S. Saebø and P. Pulay, J. Chem. Phys. **88**, 1884 (1988).
[21] T. L. Barr and E. R. Davidson, Phys. Rev. A **1**, 644 (1970).
[22] A. G. Taube and R. J. Bartlett, Collect. Czech. Chem. Commun. **70**, 837 (2005).
[23] A. G. Taube and R. J. Bartlett, J. Chem. Phys. **128**, 164101 (2008).
[24] A. Landau, K. Khistyaev, S. Dolgikh, and A. I. Krylov, J. Chem. Phys. **132**, 014109 (2010).
[25] Z. Rollik and M. Kállay, J. Chem. Phys. **135**, 104111 (2011).
[26] C. Edmiston and M. Krauss, J. Chem. Phys. **42**, 1119 (1965).
[27] W. Meyer, Int. J. Quantum Chem. **S5**, 341 (1971).
[28] W. Meyer, J. Chem. Phys. **58**, 1017 (1973).
[29] R. Ahlrichs, F. Driessler, H. Lischka, V. Staemmler, and W. Kutzelnigg, J. Chem. Phys. **62**, 1235 (1975).
[30] V. Staemmler and R. Jaquet, Theor. Chim. Acta **59**, 487 (1981).
[31] F. Neese, F. Wennmohs, and A. Hansen, J. Chem. Phys. **130**, 114108 (2009).
[32] F. Neese, A. Hansen, and D. G. Liakos, J. Chem. Phys. **131**, 064103 (2009).
[33] A. Hansen, D. G. Liakos, and F. Neese, J. Chem. Phys. **135**, 214102 (2011).
[34] Y. Kurashige, J. Yang, G. K. L. Chan, and F. R. Manby, "Optimization of orbital-specific virtuals in local Møller-Plesset perturbation theory," J. Chem. Phys. (submitted).
[35] J. E. Subotnik and M. Head-Gordon, J. Chem. Phys. **123**, 064108 (2005).
[36] J. E. Subotnik, A. Sodth, and M. Head-Gordon, J. Chem. Phys. **125**, 074116 (2006).
[37] J. E. Subotnik, A. Sodth, and M. Head-Gordon, J. Chem. Phys. **128**, 034103 (2008).
[38] G. E. Scuseria and P. Y. Ayala, J. Chem. Phys. **111**, 8330 (1999).
[39] A. Auer and M. Nooijen, J. Chem. Phys. **125**, 024104 (2006).
[40] C. Hampel and H.-J. Werner, J. Chem. Phys. **104**, 6286 (1996).
[41] M. Schütz and H.-J. Werner, Chem. Phys. Lett. **318**, 370 (2000).
[42] M. Schütz, J. Chem. Phys. **113**, 9986 (2000).
[43] M. Schütz and H.-J. Werner, J. Chem. Phys. **114**, 661 (2001).
[44] M. Schütz, J. Chem. Phys. **116**, 8772 (2002).
[45] M. Schütz, Phys. Chem. Chem. Phys. **4**, 3941 (2002).
[46] T. B. Adler and H.-J. Werner, J. Chem. Phys. **135**, 144117 (2011).
[47] H.-J. Werner and K. Pflüger, Annu. Rep. Comp. Chem. **2**, 53 (2006).
[48] P. Pulay, S. Saebø, and W. Meyer, J. Chem. Phys. **81**, 1901 (1984).
[49] G. E. Scuseria, C. L. Janssen, and H. F. Schaefer III, J. Chem. Phys. **89**, 7382 (1988).
[50] C. Hampel, K. A. Peterson, and H.-J. Werner, Chem. Phys. Lett. **190**, 1 (1992).
[51] M. Schütz, G. Hetzer, and H.-J. Werner, J. Chem. Phys. **111**, 5691 (1999).
[52] R. Mata and H.-J. Werner, J. Chem. Phys. **125**, 184110 (2006).
[53] R. Mata and H.-J. Werner, Mol. Phys. **105**, 2753 (2007).
[54] D. Kats and M. Schütz, J. Chem. Phys. **131**, 124117 (2009).
[55] K. Freundorfer, D. Kats, T. Korona, and M. Schütz, J. Chem. Phys. **133**, 244110 (2010).
[56] J. W. Boughton and P. Pulay, J. Comput. Chem. **14**, 736 (1993).
[57] A. E. Reed, R. B. Weinstock, and F. Weinhold, J. Chem. Phys. **83**, 735 (1985).
[58] J. Pipek and J. Ladik, Chem. Phys. **102**, 445 (1986).

[59]A. E. Reed and F. Weinhold, J. Chem. Phys. **83**, 1736 (1985).
[60]Note that Neese *et al.* used a different normalization in Eqs. (18), (19) of Ref. 31 for which we do not have a theoretical explanation. The normalization in Eq. (31) yields faster convergence of the correlation energy as a function of the average domain sizes than the normalization of Neese.
[61]H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, WIRES Comput. Mol. Sci. **2**, 242 (2012).
[62]H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, *et al.* MOLPRO, development version 2010.2, a package of *ab initio* programs (2011), see http://www.molpro.net.
[63]Note that in Eqs. (31) of Ref. 31 and (22) of Ref. 32 transformation matrices $\mathbf{d}^{ij}$ (which correspond to our $\mathbf{Q}^{ij}$) are missing between amplitudes and integrals.
[64]T. B. Adler, H.-J. Werner, and F. R. Manby, J. Chem. Phys. **130**, 054106 (2009).
[65]See supplementary material at http://dx.doi.org/10.1063/1.3696963 for relevant cartesian coordinates of studied molecules.
[66]F. Weigend, A. Köhn, and C. Hättig, J. Chem. Phys. **116**, 3175 (2001).
[67]K. A. Peterson, T. B. Adler, and H.-J. Werner, J. Chem. Phys. **128**, 084102 (2008).
[68]J. G. Hill, K. A. Peterson, G. Knizia, and H.-J. Werner, J. Chem. Phys. **131**, 194105 (2009).
[69]T. Helgaker, W. Klopper, H. Koch, and J. Noga, J. Chem. Phys. **106**, 9639 (1997).
[70]A. Karton and J. Martin, Theor. Chem. Acc. **115**, 330 (2006).
[71]R. A. Mata, H.-J. Werner, S. Thiel, and W. Thiel, J. Chem. Phys. **128**, 025104 (2008).
[72]F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, and H.-J. Werner, Angew. Chem. **118**, 7010 (2006).
[73]J. Pipek and P. G. Mezey, J. Chem. Phys. **90**, 4916 (1989).
[74]P. Jurečka, J. Šponer, J. Černý, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).
[75]M. Schütz, G. Rauhut, and H.-J. Werner, J. Phys. Chem. A **102**, 5997 (1998).
[76]J. G. Hill, J. A. Platts, and H.-J. Werner, Phys. Chem. Chem. Phys. **8**, 4072 (2006).
[77]M. Schütz, W. Klopper, and H.-P. Lüthi, J. Chem. Phys. **103**, 6114 (1995).
[78]O. Marchetti and H.-J. Werner, Phys. Chem. Chem. Phys. **10**, 3400 (2008).
[79]O. Marchetti and H.-J. Werner, J. Phys. Chem. A **113**, 11580 (2009).
[80]H.-J. Werner, J. Chem. Phys. **129**, 101103 (2008).
[81]T. B. Adler and H.-J. Werner, J. Chem. Phys. **130**, 241101 (2009).