Big Data analysis of *ab initio* molecular integrals in the neglect of diatomic differential overlap approximation

Xin Wu   Pavlo O. Dral   Axel Koslowski   Walter Thiel [*]

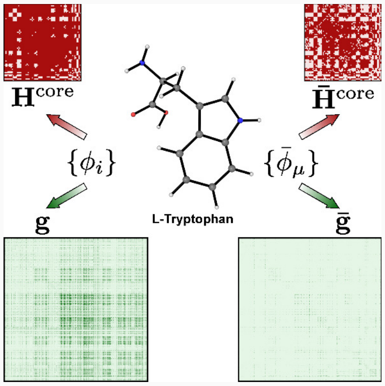
February 25, 2019

Abstract

Most modern semiempirical quantum-chemical (SQC) methods are based on the neglect of diatomic differential overlap (NDDO) approximation to *ab initio* molecular integrals. Here we check the validity of this approximation by computing all relevant integrals for 32 typical organic molecules using Gaussian-type orbitals and various basis sets (from valence-only minimal to all-electron triple-ζ basis sets) covering in total more than 15.6 million one-electron (1-e) and 10.3 billion two-electron (2-e) integrals. The integrals are calculated in the nonorthogonal atomic basis and then transformed by symmetric orthogonalization to the Löwdin basis. In the case of the 1-e integrals, we find strong orthogonalization effects that need to be included in SQC models, for example by strategies such as those adopted in the available OMx methods. For the valence-only minimal basis, we confirm that the 2-e Coulomb integrals in the Löwdin basis are quantitatively close to their counterparts in the atomic basis and that the 2-e exchange integrals can be safely neglected in line with the NDDO approximation. For larger all-electron basis sets, there are strong multi-shell orthogonalization effects that lead to more irregular patterns in the transformed 2-e integrals and thus cast doubt on the validity of the NDDO approximation for extended basis sets. Focusing on the valence-only minimal basis, we find that some of the NDDO-neglected integrals are reduced but remain sizable after the transformation to the Löwdin basis; this is true for the two-center 2-e hybrid integrals, the three-center 1-e nuclear attraction integrals, and the corresponding three-center 2-e hybrid integrals. We consider a scheme with a valence-only minimal basis that includes such terms as a possible strategy to go beyond the NDDO integral approximation in attempts to improve SQC methods.

Keywords: semiempirical quantum-chemical methods, neglect of diatomic differential overlap, neglect of atomic exchange, *ab initio* molecular integrals, symmetric orthogonalization. ■

[*] E-mail: thiel@mpi-muelheim.mpg.de
Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany



The evaluation of 2-*e* integrals in the nonorthogonal atomic basis ($\{\phi_i\}$) dominates the computational cost of *ab initio* Hartree-Fock calculations. We demonstrate that both the 1-*e* and 2-*e* integrals in the molecular Hamiltonian are reduced significantly in the Löwdin basis ($\{\bar{\phi}_\mu\}$), albeit to different extent for different types of integrals. Based on our analysis we consider the novel NAX scheme (Neglect of Atomic eXchange) for molecular integrals that goes beyond the semiempirical NDDO approximation.

Introduction

Quantum-chemical methods for molecules normally expand the wavefunction in terms of atom-centered basis functions and thus involve the calculation of molecular integrals over these basis functions.[1] Gaussian-type orbitals (GTOs)[2] are the most popular choice of basis functions since all necessary integrals can be computed analytically.[3] Repeated evaluation of the two-electron (2-e) repulsion integrals in the self-consistent-field (SCF) procedure[4] is the primary computational bottleneck of the *ab initio* Hartree-Fock (HF) method.[5–8] For large molecules, the formal quartic scaling of 2-e integral evaluation with respect to system size, e.g. the number of GTO basis functions (N_g), can be reduced to quadratic or even near-linear scaling by a variety of algorithms and techniques developed over the past decades. These include integral prescreening, the resolution-of-identity (RI) approach in combination with suitable auxiliary basis functions, the pseudospectral method, the fast multipole method, and several other algorithms for speeding up the calculation of the Coulomb matrix and the HF exchange matrix. Excellent reviews of these approaches are available.[9–11] However, even with such low-order scaling methods, the *ab initio* evaluation of the 2-e repulsion integrals remains costly.

Semiempirical quantum-chemical (SQC) methods follow a different strategy to solve the 2-e integral problem. They neglect most of the multi-center molecular integrals and attempt to correct for the associated errors by introducing suitable empirical representations of the remaining terms with adjustable parameters, which are then optimized through extensive parameterizations to reproduce theoretical[12] or experimental[13] reference data. Nowadays, current SQC methods are successfully applied in a number of areas, for example to calculate the electronic structure of huge molecules with up to several tens of thousands of atoms,[14] to simulate the long-time dynamics of molecules in the ground state[15] and in excited states,[16] and to perform high-throughput virtual screening in drug discovery.[17] Furthermore, taking advantage of modern cutting-edge high-performance computer architectures[18–22] the capacity of the SQC methods is expanding towards ever larger systems in chemistry, pharmacology, and materials science. Several recent reviews survey the development of SQC methods and the scope of possible applications.[23–25]

The neglect of diatomic differential overlap (NDDO) approximation for molecular valence-electron integrals,[26] an extension of the zero differential overlap (ZDO) approximation for π-electrons,[27] is the foundation of most modern SQC models. In the NDDO framework, the differential overlap of two basis functions ϕ^{A} and ϕ^{B} belonging to different atoms A and B, respectively, equals zero

$$\phi^{\mathrm{A}}\phi^{\mathrm{B}}\,\mathrm{d}\mathbf{r} \equiv 0 \qquad \forall\,\mathrm{A} \neq \mathrm{B}, \tag{1}$$

where $\mathbf{r}$ denotes the coordinates of an electron (omitted as argument of the basis functions to simplify notation). As a consequence, NDDO-based SQC models neglect most of the molecular integrals, e.g. all three-center (3-c) and four-center (4-c) integrals and also some types of two-center (2-c) 2-e repulsion integrals. There were several early endeavors to justify the NDDO approximation for the symmetrically orthogonalized basis (called Löwdin basis in the following)[28] through polynomial expansions of the $\mathbf{S}^{-\frac{1}{2}}$ matrix, where $\mathbf{S}$ is the overlap matrix in the original nonorthogonal atomic basis (denoted as atomic basis in this article). These early efforts were either based on purely mathematical analysis[29–37] or on numerical validation for a small number of integrals in small molecules, e.g. dinitrogen,[32] water and ethane,[38] and benzene.[39]

Nowadays, a much more extensive analysis of the NDDO approximation is technically feasible. This is the subject of the present "Big Data" study of molecular integrals. We compute 15.6 million 1-e integrals and 10.3 billion 2-e integrals for 32 closed-shell organic molecules and classify them into 14 different types. We first establish quantitative correlations between the *ab initio* 2-e integrals in the atomic basis and in the Löwdin basis. The emerging patterns are rationalized by examining the Löwdin basis, the symmetric orthogonalization matrix, and the pertinent 2-e integrals in the linear H_4 model system. Thereafter, we turn to the 1-e integrals in the core Hamiltonian and demonstrate the strong orthogonalization effects on these integrals. We address the connection between the multi-center 1-e nuclear attraction integrals in core Hamiltonian and the contracted 2-e part in the Fock matrix, and quantify the coupling effects in terms of 3-c coupled potential integrals and 2-c penetration integrals. Finally, we perform a systematic series of HF-SCF calculations with GTOs, for basis sets ranging from a valence-only minimal to an all-electron triple-ζ basis, using seven distinct schemes of integral approximation in the Löwdin basis that take us from an NDDO to a full *ab initio* HF-SCF calculation. This allows us to quantify the effect of the NDDO-neglected integrals on the electronic energy. On the basis of this extensive analysis, we consider the NAX scheme (Neglect of Atomic eXchange) for SQC calculations with a minimal valence-electron GTO basis, which goes beyond NDDO by including the most important NDDO-neglected integrals.

After submission of this paper, another comprehensive analysis of the NDDO approximation was published online, which compares the values of 2-e integrals in the atomic basis and the Löwdin basis, examines the effect of the NDDO approximation on molecular energies, and proposes system-specific error corrections in the two-electron matrices that enter the Fock operator.[40] This analysis is complementary to our present study but targets different objectives.

Methodology

A total of 32 typical organic molecules (see Figure 1) were chosen and divided into six groups representing simple aliphatic compounds, homocyclic and heterocyclic aromatic molecules, saturated carbocycles and heterocycles, and common amino acids. Equilibrium geometries of these molecules were obtained at the B3LYP/6-31G(d) level of theory by using the ORCA program.[41]

Molecular integrals over GTOs were computed at these geometries using an in-house modified version of the LIBCINT library.[42] These integrals were stored on disk for statistical analysis. Single-point HF-SCF calculations were performed for different levels of integral approximation with our in-house LÖWDIN program. The convergence threshold was set to 1.0×10^{-10} for the maximum variation of the density matrix elements in successive SCF iterations.

The integral evaluations and the HF-SCF calculations were done for six basis sets, namely the valence-only minimal basis set CEP-4G[43] used in SQC methods and five all-electron basis sets used in *ab initio* work: single-ζ STO-3G and STO-6G,[44] double-ζ 3-21G[45] and 6-31G,[46] and triple-ζ 6-311G,[47] in order to complete a systematic survey. Due to technical constraints it was impossible to store all 6-311G integrals for tryptophan, the largest molecule in the test set. Altogether, more than 15.6 million 1-e integrals and 10.3 billion 2-e integrals were eventually collected for the Big Data analysis. The exact number of computed integrals for each type of integral and each basis set is reported in the Supporting Information (see Tables S1 and S2). For data visualization we focus on the integrals in the valence-only CEP-4G basis, since these are most relevant for NDDO-based SQC models.

The symmetric orthogonalization of *ab initio* molecular integrals is a linear transformation between two finite function spaces, i.e. $L : \{\phi_i\} \mapsto \{\bar{\phi}_\mu\}$, where $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ represent the nonorthogonal atomic basis and the orthonormal Löwdin basis, respectively. The transformation matrix $\mathbf{X}$ is the inverse square root of the overlap matrix in the $\{\phi_i\}$ basis

$$\bar{\phi}_\mu = \sum_{i=1}^{N_g} \phi_i X_{i\mu}, \quad \mathbf{X} = \mathbf{S}^{-\frac{1}{2}}. \tag{2}$$

$\mathbf{X}$ is a real symmetric matrix.

The 1-e kinetic integrals, 1-e nuclear attraction integrals (NAIs), and 2-e repulsion integrals in the $\{\bar{\phi}_\mu\}$ basis, i.e. $\bar{T}_{\mu\nu}$, $\bar{V}_{\mu\nu,\mathrm{A}}$, and $\bar{g}_{\mu\nu,\kappa\tau}$, respectively, are linked to the corresponding integrals in the $\{\phi_i\}$ basis by

$$\bar{T}_{\mu\nu} = \sum_{ij} X_{i\mu} T_{ij} X_{j\nu} \tag{3}$$

$$\bar{V}_{\mu\nu,\mathrm{A}} = \sum_{ij} X_{i\mu} V_{ij,\mathrm{A}} X_{j\nu} \tag{4}$$

$$\bar{g}_{\mu\nu,\kappa\tau} = \sum_{ij,kl} X_{i\mu} X_{j\nu} g_{ij,kl} X_{k\kappa} X_{l\tau}. \tag{5}$$

An integral in the Löwdin basis is denoted throughout this article by a bar above the corresponding symbol, for the sake of convenience.

In the single-point HF-SCF calculations (see above) the molecular integrals were transformed from the atomic basis to the Löwdin basis by symmetric orthogonalization, see Eqs. (3), (4), and (5). The two different sets of molecular integrals were used in separate HF-SCF calculations, and identical results were obtained in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations for each molecule and each basis set. This verified the correctness of our Löwdin program by confirming that the molecular electronic energy is invariant to symmetric orthogonalization.

Results and Discussion

This Big Data study of *ab initio* molecular integrals is primarily aimed at providing insights into the approximations underlying NDDO-based SQC models. The computed molecular integrals are therefore classified into seven and five different types of 2-e and 1-e integrals, respectively, according to the NDDO integral convention that is widely adopted in many popular SQC methods, e.g. MNDO,[48] AM1,[49] PMx,[50–53] and the OMx series.[54–57] In addition, we define two extra types of molecular integrals associated with the interplay between attractive 1-e and repulsive 2-e terms in the molecular Hamiltonian. The classification of all these integrals is summarized in Table 1. The 2-e integrals are explicitly labeled in terms of the involved atomic centers. The 1-e integrals and the combined 1-e and 2-e integrals are likewise labeled according to the NDDO convention, which specifies the affiliation of a basis function to an atom as $\{\phi_\mu^{\mathrm{A}}, \phi_\nu^{\mathrm{A}}\} \in \mathrm{A}$, $\{\phi_\lambda^{\mathrm{B}}\} \in \mathrm{B}$, and $\{\phi_\rho^{\mathrm{C}}, \phi_\sigma^{\mathrm{C}}\} \in \mathrm{C}$, where ϕ_μ^{A} and ϕ_ν^{A} are different basis functions on atom A, ϕ_ρ^{C} and ϕ_σ^{C} may refer to the same basis function on atom C, and A, B, and C symbolize three distinct atoms.

2-e integrals

The statistical measures for each type of *ab initio* 2-e molecular integrals in the Löwdin basis are listed in Table 2. These 2-e integrals can be categorized into three distinct groups in accordance with the convention used by Roothaan and Rüdenberg,[58–60] namely Coulomb, exchange, and hybrid integrals.

Coulomb integrals

The Coulomb integrals are always retained in the NDDO approximation. The correlation diagrams for $g_{1\mathrm{cc}}$ and $g_{2\mathrm{cc}}$ in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations are shown in Figure 2 for the valence-only minimal basis. In both cases, good linear correlations are evident, with correlation coefficients r^2 of 0.996 and 0.994 for $\bar{g}_{1\mathrm{cc}}$ and $\bar{g}_{2\mathrm{cc}}$, respectively (see Table 2). Closer inspection of Figure 2 reveals that $\bar{g}_{1\mathrm{cc}}$ is generally shifted up considerably with respect to $g_{1\mathrm{cc}}$: the associated standard deviation σ of 0.75 eV is significantly larger than the corresponding value of 0.16 eV for $\bar{g}_{2\mathrm{cc}}$ vs. $g_{2\mathrm{cc}}$. In current NDDO-based SQC methods such as OMx, the 1-c 2-e integrals are derived from experimental atomic data while the 2-c 2-e integrals are represented by $g_{2\mathrm{cc}}$ values that are scaled to account for dynamic correlation effects in an average manner.

Corresponding correlation diagrams are shown in the Supporting Information for the Coulomb integrals obtained with the larger all-electron basis sets (see Figures S1 to S5). In contrast to the valence-only minimal basis set, no regular patterns are found for these larger all-electron basis sets. In an attempt to find correlations in the case of the triple-ζ 6-311G set, we divided the Coulomb integrals into several sub-types according to the involved $1S$, $2SP$, $2S'P'$, and $2S''P''$ shells (see Figures S6 to S11 in the Supporting Information) but there were still no satisfactory linear correlations for any of these sub-types. Hence, when using multi-shell all-electron basis functions, the symmetric orthogonalization among different shells can give rise to fairly strong and non-uniform changes in the Coulomb integrals, which will make it difficult to devise useful approximate expressions for these integrals in the Löwdin basis. This may be an obstacle for attempts to improve SQC models by introducing extended multi-shell basis function.

Exchange integrals

The 2-e exchange integrals include $\bar{g}_{2\mathrm{cx}}$, $\bar{g}_{3\mathrm{cx}}$, and $\bar{g}_{4\mathrm{cx}}$. All these are completely neglected in NDDO-based SQC models. As shown in Figure 3 these integrals are non-negligible in the atomic basis: their absolute magnitudes range up to 2 eV, 4 eV, and more than 5 eV for $g_{4\mathrm{cx}}$, $g_{3\mathrm{cx}}$, and $g_{2\mathrm{cx}}$, respectively. By contrast, all these exchange integrals are very small in the Löwdin basis (see Figure 3). The statistical measures given in Table 2 quantify this observation: the standard deviations σ from zero are 0.033, 0.004, and 0.001 eV for $\bar{g}_{2\mathrm{cx}}$, $\bar{g}_{3\mathrm{cx}}$, and $\bar{g}_{4\mathrm{cx}}$, respectively; the absolute magnitude of these integrals is less than 10^{-5} eV for 29%, 30%, and 48% of $\bar{g}_{2\mathrm{cx}}$, $\bar{g}_{3\mathrm{cx}}$, and $\bar{g}_{4\mathrm{cx}}$, respectively (see Table S3 in the Supporting Information). Therefore the neglect of the 2-e exchange integrals in NDDO-based SQC models is justified for the valence-only minimal basis.

For the larger all-electron GTO basis sets, there are billions of 2-e exchange integrals to be computed and analyzed for our 32 test molecules. We refrain from a detailed analysis but just report the most negative and the most positive values of these integrals for the five all-electron basis sets (see Table S4 in the Supporting Information). Evidently, the exchange integrals in the $\{\phi_i\}$ basis are much too large to be neglected; for example, in the 6-311G case, the maximum absolute values exceed 4, 5, and 7 eV for $g_{4\mathrm{cx}}$, $g_{3\mathrm{cx}}$, and $g_{2\mathrm{cx}}$, respectively. In the Löwdin basis, these integrals are much reduced through the symmetric orthogonalization but they do not become generally negligible: for example, in the double-ζ (3-21G, 6-31G) and triple-ζ (6-311G) case, the maximum values of $\bar{g}_{2\mathrm{cx}}$ reach 1.1 and 1.3 eV, respectively. Hence, when using multi-shell all-electron basis functions, the neglect of the 2-e exchange integrals in NDDO-based SQC models is not well supported by considering their values in the Löwdin basis.

Hybrid integrals

The 2-e hybrid integrals $\bar{g}_{2\mathrm{ch}}$ and $\bar{g}_{3\mathrm{ch}}$ describe the interaction between a Coulomb density on one atom (A) and the exchange density between two atomic centers (A-B or B-C). The correlation diagrams of the 2-c and 3-c hybrid integrals in the atomic basis and the Löwdin basis are shown in Figure 4. As in the case of the exchange integrals, the absolute magnitude of the hybrid integrals is significantly decreased by symmetric orthogonalization. The standard deviations from zero are fairly small in the Löwdin basis, 0.120 eV for $\bar{g}_{2\mathrm{ch}}$ and 0.033 eV for $\bar{g}_{3\mathrm{ch}}$ (see Table 2). However, individual hybrid integrals may remain fairly large in the Löwdin basis, as indicated by maximum values greater than 2 eV for $\bar{g}_{2\mathrm{ch}}$ and 1 eV for $\bar{g}_{3\mathrm{ch}}$. Hence, the NDDO approximation is less well justified for hybrid integrals than for exchange integrals.

This conclusion is reinforced when going from the valence-only minimal basis (considered above) to larger all-electron GTO basis sets, where the maximum values of the hybrid integrals in the Löwdin basis are found to range from 3.0 to 3.5 eV for $\bar{g}_{2\mathrm{ch}}$ and from 1.0 to 2.4 eV for $\bar{g}_{3\mathrm{ch}}$ (see Table S4 in the Supporting Information).

Linear H_4 model

To understand the patterns identified in the preceding Big Data analysis we scrutinize a simple model, linear H_4 with an internuclear distance of 1.40 Bohr described by the minimal

STO-3G basis. Figure 5 contains plots of the Löwdin basis ($\{\bar{\phi}_\mu\}$, in solid lines) and of the nonorthogonal atomic basis ($\{\phi_i\}$, in dotted lines) as well as the coefficient matrix of the symmetric orthogonalization. The $\{\bar{\phi}_\mu\}$ basis is designed to resemble the parental $\{\phi_i\}$ basis as closely as possible by minimizing the sum of the squared deviations between the two sets of basis functions.[28] Since each $\{\bar{\phi}_\mu\}$ is comprised of contributions from all $\{\phi_i\}$ of a molecule, see Eq. (2), it is usually deemed to be more delocalized.

However, inspection of the plots in Figure 5 shows that the $\{\bar{\phi}_\mu\}$ basis seems somewhat "slimmer" than the parental $\{\phi_i\}$ basis in the H_4 model: the solid lines representing the $\{\bar{\phi}_\mu\}$ basis are more squeezed toward the atomic centers and decay even faster than the dotted lines (the $\{\phi_i\}$ basis) in the covalent region between two neighboring hydrogen atoms. The Coulomb densities are thus quite localized in the $\{\bar{\phi}_\mu\}$ representation, even more so than in the parental $\{\phi_i\}$ representation, and both will be dominated by the contributions from the region of the corresponding atoms. This explains the good correlation between the Coulomb integrals in the two representations, and also why the Coulomb integrals tend to be somewhat larger in the $\{\bar{\phi}_\mu\}$ basis.

The absolute magnitude of the 2-e exchange and hybrid integrals is generally reduced in the $\{\bar{\phi}_\mu\}$ basis (see Figures 3 and 4). This can be traced back to typical patterns in the coefficient matrix of the symmetric orthogonalization, i.e. $\mathbf{X}$ in Eq. (2). In the case of the linear H_4 model (Figure 5) we can identify the following patterns: i) The diagonal elements $X_{\mu\mu}$ are close to but slightly larger than unity. ii) The off-diagonal elements $X_{\mu,\mu\pm1}$ for neighboring atoms are negative; their absolute values are smaller than unity but still sizable. iii) When going from the closest neighboring atom to distant atoms, the off-diagonal elements $X_{\mu\nu}$ alternate in sign, and their absolute values decay steeply. A term-by-term analysis of the symmetric orthogonalization in the linear H_4 model shows that the negative $X_{\mu,\mu\pm1}$ coefficients (ii) are largely responsible for cancellations in $\bar{g}_{\mu\nu,\kappa\tau}$, see Eq. (5), and thus for the decrease in the absolute magnitudes of the exchange and hybrid integrals.

1-e core Hamiltonian

The 1-e core Hamiltonian $\mathbf{H}^{\text{core}}$ represents the electron kinetic energy and the nuclear-electron attraction energy. Instead of decomposing $\mathbf{H}^{\text{core}}$ into individual $T_{\mu\nu}$ and $V_{\mu\nu,\text{A}}$

contributions for each atom, the 1-e integrals are classified as diagonal one-center integrals $U_{\mu\mu}$, 2-c NAIs $V_{\mu\nu,\mathrm{B}}$, 2-c resonance integrals $\beta_{\mu\lambda}$, off-diagonal one-center integrals $U_{\mu\nu}$, and 3-c NAIs $V_{\mu\lambda,\mathrm{C}}$, in accord with the integral conventions adopted in most NDDO-based SQC methods.[48–57] A complete list of these integrals and their mathematical definitions is given in Table 1. Having demonstrated that the NDDO approximation is less justified for the 2-e integrals when using multi-shell all-electron GTO basis sets, we will focus in the following discussion on the 1-e integrals obtained with the valence-only CEP-4G basis set.

NDDO-retained integrals

In NDDO-based SQC models, the integrals $\bar{U}_{\mu\mu}$, $\bar{V}_{\mu\nu,\mathrm{B}}$, and $\bar{\beta}_{\mu\lambda}$ are retained. The correlation diagrams of these integrals in the atomic and Löwdin bases and the associated statistical measures are given in Figure 6 and Table 2, respectively.

We find a roughly linear correlation for $\bar{U}_{\mu\mu}$ ($r^2 = 0.983$) but there are large deviations between the values in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations, with a standard deviation σ of 7.0 eV. Evidently, $\bar{U}_{\mu\mu}$ is severely affected by orthogonalization effects, which should be taken into account explicitly in SQC models (as e.g. in the OMx methods[54–57]).

There is a satisfactory linear correlation for $\bar{V}_{\mu\nu,\mathrm{B}}$ ($r^2 = 0.994$). The deviations between the values in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations are smaller than those for $\bar{U}_{\mu\mu}$ but not negligible ($\sigma = 0.77$ eV). Hence, it is reasonable to add suitable orthogonalization corrections to the 2-c NAIs in NDDO-based SQC methods (as e.g. in the OMx methods[54–57]).

The resonance integrals $\bar{\beta}_{\mu\lambda}$ are retained in all SQC models because they describe the strength of bonding between the orbitals ϕ_μ^{A} and $\phi_\lambda^{\mathrm{B}}$ on atoms A and B. They are most severely affected by the symmetric orthogonalization, being often reduced by about one order of magnitude (Figure 6). However, they are clearly not negligible in the $\{\bar{\phi}_\mu\}$ representation (note that the NDDO approximation does not apply to the resonance integrals because of the presence of the kinetic energy operator, see Table 1). Given the lack of a clear correlation between $\beta_{\mu\lambda}$ and $\bar{\beta}_{\mu\lambda}$ ($r^2 = 0.740$) and the order-of-magnitude difference between their values, it is understandable that NDDO-based SQC methods normally attempt to model the resonance integrals directly through suitable empirical functions (rather than relating them to their analytical counterparts). Nevertheless, it has been found advantageous in the OM2 and OM3 methods to include small 3-c orthogonalization corrections in the resonance integrals to properly account for subtle environmental effects.[55–57]

NDDO-neglected integrals

In NDDO-based SQC methods, $U_{\mu\nu}$ and $V_{\mu\lambda,\mathrm{C}}$ are neglected on different grounds: $U_{\mu\nu}$ is zero by symmetry in the atomic basis, while $V_{\mu\lambda,\mathrm{C}}$ is zero according to the NDDO approximation since it involves the charge distribution of orbitals ϕ_μ^{A} and $\phi_\lambda^{\mathrm{B}}$ on different atoms A and B.

The correlation diagrams of $U_{\mu\nu}$ and $V_{\mu\lambda,\mathrm{C}}$ in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations are shown in Figure 7. It is obvious that the off-diagonal one-center energies are no longer zero in the $\{\bar{\phi}_\mu\}$ basis but adopt rather large values upon symmetric orthogonalization. The standard deviation from zero is 0.9 eV for $\bar{U}_{\mu\nu}$ (see σ in Table 2), and individual values can be as large as ±6.5 eV (see Figure 7). The neglect of $\bar{U}_{\mu\nu}$ in conventional NDDO-based SQC methods[48–53] may result in intrinsic errors, which may be avoided in methods that include corresponding orthogonalization corrections.[54–57]

Concerning the 3-c nuclear attractions integrals, the correlation diagram of $V_{\mu\lambda,\mathrm{C}}$ in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations shows no simple pattern (see Figure 7). The orthogonalization again leads to a significant decrease in the values of the integrals: $V_{\mu\lambda,\mathrm{C}}$ and $\bar{V}_{\mu\lambda,\mathrm{C}}$ range up to ±28 eV and ±4.7 eV, respectively. The standard deviation from zero is 0.2 eV for $\bar{V}_{\mu\lambda,\mathrm{C}}$ (see σ in Table 2). Hence, $\bar{V}_{\mu\lambda,\mathrm{C}}$ does not vanish as assumed by the NDDO approximation, which may be an intrinsic source of error in current NDDO-based SQC models.

Combined 1-e and 2-e integrals

The Fock matrix of a molecular system is composed of the 1-e core Hamiltonian and a 2-e part containing 2-e integrals contracted with electron density matrix elements. Overall, the former is attractive and the latter is repulsive. In SQC models it is essential to achieve a proper balance between attractive and repulsive terms, and hence it is appropriate to explore combinations of such terms when analyzing SQC integral approximations.

We first consider the total 3-c contribution to the Fock matrix arising from the 3-c 1-e and 2-e integrals that are normally neglected in NDDO-based SQC models. To quantify this contribution we define the coupled potential (CP) integral as

$$V^{\mathrm{CP}}_{\mu\lambda,\mathrm{C}} = V_{\mu\lambda,\mathrm{C}} + \sum_{\rho}^{\mathrm{C}} \sum_{\sigma}^{\mathrm{C}} D_{\rho\sigma} \left(g_{\mu\lambda,\rho\sigma} - \frac{1}{2} g_{\mu\sigma,\rho\lambda} \right),$$

where $V_{\mu\lambda,\mathrm{C}}$ is the 3-c 1-e NAI, $g_{\mu\lambda,\rho\sigma}$ and $g_{\mu\sigma,\rho\lambda}$ denote the 3-c 2-e hybrid and exchange integrals, respectively, and $D_{\rho\sigma}$ is a converged density matrix element at atom C.

The correlation diagram of the CP integrals in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations is plotted in Figure 8. Compared to the 3-c NAIs $V_{\mu\lambda,\mathrm{C}}$ and $\bar{V}_{\mu\lambda,\mathrm{C}}$ in Figure 7, the corresponding CP integrals are much smaller due to the cancellation of 1-e and 2-e contributions of nearly equal magnitude and opposite sign. As a result, $\bar{V}^{\mathrm{CP}}_{\mu\lambda,\mathrm{C}}$ deviates much less from zero than $\bar{V}_{\mu\lambda,\mathrm{C}}$, as indicated by a standard deviation σ of 0.05 eV. Given this approximate mutual balance between the 3-c 1-e and 2-e contributions to the Fock matrix, it may be qualitatively reasonable to neglect all these 3-c terms in NDDO-based SQC methods (as commonly done). However, closer inspection of Figure 8 shows that there exist quite a few $\bar{V}^{\mathrm{CP}}_{\mu\lambda,\mathrm{C}}$ integrals with values around ±1 eV. For further improving current SQC methods, one might thus consider to include parametric terms representing the total 3-c contribution to the Fock matrix.

The 2-c penetration integral (PI) was introduced first for planar π-electron systems[61,62] and later for valence-electron SQC treatments.[63] It is defined as the sum of the attractive 2-c electron-nucleus attraction $V_{\mu\nu,\mathrm{B}}$ and a repulsive 2-c term $Z_{\mathrm{B}} \cdot g_{\mu^{\mathrm{A}}\nu^{\mathrm{A}},s^{\mathrm{B}}s^{\mathrm{B}}}$, i.e. the valence nuclear charge Z_{B} multiplied by the 2-e interaction between the $\phi^{\mathrm{A}}_{\mu}\phi^{\mathrm{A}}_{\nu}$ and $s^{B}s^{B}$ charge distributions:

$$V^{\mathrm{PI}}_{\mu\nu,\mathrm{B}} = V_{\mu\nu,\mathrm{B}} + Z_{\mathrm{B}} \cdot g_{\mu^{\mathrm{A}}\nu^{\mathrm{A}},s^{\mathrm{B}}s^{\mathrm{B}}}$$

The penetration integrals are often neglected in NDDO-based SQC methods,[48–53] in the tradition of CNDO/2.[63] The assumption is that $V_{\mu\nu,\mathrm{B}}$ and $Z_{\mathrm{B}} \cdot g_{\mu^{\mathrm{A}}\nu^{\mathrm{A}},s^{\mathrm{B}}s^{\mathrm{B}}}$ are exactly equal but of opposite signs so that $V_{\mu\nu,\mathrm{B}}$ is represented by $-Z_{\mathrm{B}} \cdot g_{\mu^{\mathrm{A}}\nu^{\mathrm{A}},s^{\mathrm{B}}s^{\mathrm{B}}}$ in the 1-e core Hamiltonian.

The penetration integrals in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations are visualized for all test molecules in Figure 9. They are uniformly negative in the atomic basis, and mostly negative in the Löwdin basis (with very few exceptions that are almost invisible in Figure 9). They are fairly large in both representations, with negative values reaching −8.3 eV and −6.0 eV in the $\{\phi_i\}$ and $\{\bar{\phi}_\mu\}$ representations, respectively. The penetration effect thus does not vanish upon symmetric orthogonalization, and it is therefore advisable to include penetration integrals explicitly in NDDO-based SQC models, as e.g. in the OMx methods.[54–57]

Induced error in molecular electronic energy

Next we address the question of how much the NDDO-neglected molecular integrals would contribute to the molecular electronic energy if they were included. Since the NDDO approximation is best justified in the Löwdin basis we investigate this issue through calculations in this basis. We utilize our insights into the relative magnitude and importance of the various molecular integrals in the Löwdin basis to consider eight distinct schemes that take us from the classic NDDO model[26] to the full *ab initio* HF-SCF level. These schemes are characterized as follows:

$$\mathcal{E}^{\mathrm{NDDO}} = \mathrm{neg}(E^{\mathrm{HF}}_{\mathrm{SCF}}, \bar{U}_{\mu\nu}, \bar{g}_{\mathrm{2ch}}, \bar{g}_{\mathrm{2cx}}, \bar{V}_{\mu\lambda,\mathrm{C}}, \bar{g}_{\mathrm{3ch}}, \bar{g}_{\mathrm{3cx}}, \bar{g}_{\mathrm{4cx}}) \tag{6}$$

$$\mathcal{E}^{\bar{U}_{\mu\nu}} = \mathrm{aug}(\mathcal{E}^{\mathrm{NDDO}}, \bar{U}_{\mu\nu}) \tag{7}$$

$$\mathcal{E}^{\bar{g}_{\mathrm{2ch}}} = \mathrm{aug}(\mathcal{E}^{\bar{U}_{\mu\nu}}, \bar{g}_{\mathrm{2ch}}) \tag{8}$$

$$\mathcal{E}^{\bar{g}_{\mathrm{2cx}}} = \mathrm{aug}(\mathcal{E}^{\bar{g}_{\mathrm{2ch}}}, \bar{g}_{\mathrm{2cx}}) \tag{9}$$

$$\mathcal{E}^{\bar{V}_{\mu\lambda,\mathrm{C}}} = \mathrm{aug}(\mathcal{E}^{\bar{g}_{\mathrm{2cx}}}, \bar{V}_{\mu\lambda,\mathrm{C}}) \tag{10}$$

$$\mathcal{E}^{\bar{g}_{\mathrm{3ch}}} = \mathrm{aug}(\mathcal{E}^{\bar{V}_{\mu\lambda,\mathrm{C}}}, \bar{g}_{\mathrm{3ch}}) \tag{11}$$

$$\mathcal{E}^{\bar{g}_{\mathrm{3cx}}} = \mathrm{aug}(\mathcal{E}^{\bar{g}_{\mathrm{3ch}}}, \bar{g}_{\mathrm{3cx}}) \tag{12}$$

$$\mathcal{E}^{\bar{g}_{\mathrm{4cx}}} = \mathrm{aug}(\mathcal{E}^{\bar{g}_{\mathrm{3cx}}}, \bar{g}_{\mathrm{4cx}}) \equiv E^{\mathrm{HF}}_{\mathrm{SCF}}, \tag{13}$$

where $E^{\mathrm{HF}}_{\mathrm{SCF}}$ denotes the standard HF-SCF molecular electronic energy. In all cases, the HF-SCF converged electron density in the Löwdin basis is used for the molecular electronic energy calculations (Eqs. from (6) to (13)) so that every quantity considered herein is fully consistent with the Löwdin basis.

The function neg() in Eq. (6) returns $\mathcal{E}^{\mathrm{NDDO}}$, the electronic energy evaluated in an *ab initio* manner while invoking the standard NDDO integral approximation. The various types of the NDDO-neglected integrals (see Tables 1 and 2) are then incrementally introduced in a stepwise manner, in accordance with their relative importance; this is symbolized by the aug() function in Eqs. (7) to (13). Since all the NDDO-neglected molecular integrals are eventually recovered in Eq. (13), $\mathcal{E}^{\bar{g}_{4\mathrm{cx}}}$ must be identical to the standard $E_{\mathrm{SCF}}^{\mathrm{HF}}$. Finally, in order to establish quantitative estimates of the contributions of the NDDO-neglected integrals to the HF-SCF molecular electronic energy, the induced error ΔE^i relative to $E_{\mathrm{SCF}}^{\mathrm{HF}}$ is defined by

$$\Delta E^i = \mathcal{E}^i - E_{\mathrm{SCF}}^{\mathrm{HF}}$$

where the superscript i stands for the seven distinct neglect schemes introduced here (see Eqs. from (6) to (12), respectively).

In our analysis, we ignore electron correlation effects and thus do not strive for chemical accuracy, which is impossible to achieve at the HF-SCF level.[1] Instead, we consider the ΔE^i values as a suitable systematic measure to assess the energetic effect of neglecting certain types of molecular integrals when applying the NDDO approximation. The results obtained with the valence-only minimal CEP-4G basis set are plotted for all test molecules in Figure 10. The mean induced errors ($\bar{\Delta} E^i$) are listed in Table 3 for all the GTO basis sets employed in this study. Detailed numerical results are collected in Tables S5 to S36 in the Supporting Information. In the following discussion, we will go through the seven schemes defined above and discuss the relevant energy terms (i.e. $\mathcal{E}^i$, ΔE^i, and $\bar{\Delta} E^i$) first for the valence-only basis set and then for the all-electron basis sets.

The energies $\mathcal{E}^{\mathrm{NDDO}}$ calculated from a parameter-free NDDO model show huge deviations from the reference energies $E_{\mathrm{SCF}}^{\mathrm{HF}}$, with errors for the CEP-4G basis up to 181 eV for tryptophan (see Figure 10). The errors ΔE^{NDDO} tend to increase with molecular size. The behavior of the parameter-free NDDO model is even more erratic when using the all-electron GTO basis sets, especially for 6-311G (see $\bar{\Delta} E^{\mathrm{NDDO}}$ in the first column of Table 3).

Compared to the parameter-free NDDO model, the energies $\mathcal{E}^{\bar{U}_{\mu\nu}}$ calculated with the valence-only CEP-4G basis offer almost no improvement over $\mathcal{E}^{\mathrm{NDDO}}$ (see the nearly overlapping red and black lines in Figure 10); the mean induced errors are very similar (+79.1 eV vs. +78.2 eV, Table 3). When using multi-shell all-electron GTO basis sets, the mean deviations from the corresponding $E_{\mathrm{SCF}}^{\mathrm{HF}}$ values are even larger (see Table 3).

The 2-c hybrid integrals $\bar{g}_{2\mathrm{ch}}$ are neglected in current NDDO-based SQC methods. It is obvious from Figure 10 that inclusion of these integrals ($\bar{g}_{2\mathrm{ch}}$) leads to a dramatic improvement and a drastic drop of the mean induced error ($\bar{\Delta} E^{\bar{g}_{2\mathrm{ch}}}$).

Further addition of the 2-c exchange integrals $\bar{g}_{2\mathrm{cx}}$ only causes minor changes in the individual deviations (CEP-4G, see Figure 10 where the blue line ($\Delta E^{\bar{g}_{2\mathrm{ch}}}$) largely traces the orange line ($\Delta E^{\bar{g}_{2\mathrm{cx}}}$)) and in the mean induced errors (all basis sets, see Table 3).

Next we consider incorporating the 3-c NAIs $\bar{V}_{\mu\lambda,\mathrm{C}}$ which has been regarded as a promising approach to improve NDDO-based models.[64,65] However, this turns out to be disappointing since the induced errors $\Delta E^{\bar{V}_{\mu\lambda,\mathrm{C}}}$ (see the green line in Figure 10) increase again compared with the more approximate schemes without 3-c NAIs, i.e. $\Delta E^{\bar{g}_{2\mathrm{ch}}}$ and $\Delta E^{\bar{g}_{2\mathrm{cx}}}$. The reason is obvious from our preceding Big Data analysis on the 3-c NAIs and the CP integrals: including only the mostly attractive 3-c $\bar{V}_{\mu\lambda,\mathrm{C}}$ terms while still neglecting the repulsive 3-c 2-e terms will lead to an imbalance in the Fock matrix, which will ultimately deteriorate the computed total energies. The cure for this problem is then to restore the balance by also including the 3-c 2-e integrals ($\bar{g}_{3\mathrm{ch}}$ and $\bar{g}_{3\mathrm{cx}}$) in the Fock matrix. This will substantially increase the formal computational scaling of integral evaluation from $\mathcal{O}({N_g}^2)$ to $\mathcal{O}({N_g}^3)$.

Following this route we first incorporate the 3-c hybrid integrals $\bar{g}_{3\mathrm{ch}}$ because our preceding analysis has shown them to be much more important than the 3-c exchange integrals. This reduces the errors $\Delta E^{\bar{g}_{3\mathrm{ch}}}$ in the computed energies significantly (CEP-4G, see the cyan line in Figure 10) and leads to consistently good agreement with the reference energies $E^{\mathrm{HF}}_{\mathrm{SCF}}$. Moreover, at this stage, the mean induced errors $\bar{\Delta} E^{\bar{g}_{3\mathrm{ch}}}$ become reasonably small and uniform for all basis sets considered (typically in the range from -2 to -4 eV, see Table 3). This suggests that the 3-c interactions are now treated in a balanced manner.

Finally, the 3-c terms in the Fock matrix can be fully restored by including the 3-c 2-e exchange integrals. As expected, this further diminishes the deviations of the computed energies $\mathcal{E}^{\bar{g}_{3\mathrm{cx}}}$ from the reference energies $E^{\mathrm{HF}}_{\mathrm{SCF}}$: the mean induced errors $\bar{\Delta} E^{\bar{g}_{3\mathrm{cx}}}$ drop to -0.92 eV for the valence-only minimal CEP-4G basis set and to values between -0.36 and -1.34 eV for the all-electron GTO basis sets. The remaining errors are due to the neglect of many small 4-c 2-e terms. Unfortunately, the complete inclusion of all 3-c integrals will hardly lead to efficient SQC models in practice since retaining a large portion of all 2-e integrals will severely limit the possible speedup compared with a full HF treatment.

The NAX scheme

Based on the considerations in the preceding section we will now discuss an improved SQC scheme for the treatment of molecular integrals called NAX (Neglect of Atomic eXchange). As the name implies, the NAX scheme neglects atomic exchange in the Löwdin basis completely, i.e. $\bar{g}_{\text{2cx}}$, $\bar{g}_{\text{3cx}}$, and $\bar{g}_{\text{4cx}}$. In a straightforward implementation of this concept, typically $\sim 15\%$ of the 2-e molecular integrals are still kept in our test molecules, most of them 3-c hybrid integrals $\bar{g}_{\text{3ch}}$.

For computational efficiency further simplifications are desirable. In particular, we should attempt to avoid the costly computation of $\bar{g}_{\text{3ch}}$ which would reduce the number of retained 2-e molecular integrals to typically $\sim 3\%$ in our test molecules; at the same time, however, we should still include their effect in the Hamiltonian. This might be achieved by introducing effective 3-c interaction terms in the 1-e core Hamiltonian consisting of the 3-c NAIs $\bar{V}_{\mu\lambda,\text{C}}$ and a term that accounts for the counterbalancing contributions from $\bar{g}_{\text{3ch}}$ (see the preceding section). To test the viability of this idea, we contract $\bar{g}_{\text{3ch}}$ with the converged density matrix, absorb this term into the 3-c NAIs, and use the resulting $\bar{V}^{\text{eff}}_{\mu\lambda,\text{C}}$ integrals in the 1-e core Hamiltonian:

$$\bar{V}^{\text{eff}}_{\mu\lambda,\text{C}} = \bar{V}_{\mu\lambda,\text{C}} + \sum_{\rho}^{\text{C}} \sum_{\sigma}^{\text{C}} \bar{D}_{\rho\sigma} \bar{g}_{\mu\lambda,\rho\sigma}$$

Compared with the OMx methods, the resulting NAX scheme entails two additional types of molecular integrals, namely 2-c hybrid $\bar{g}_{\text{2ch}}$ and 3-c combined $\bar{V}^{\text{eff}}_{\mu\lambda,\text{C}}$ integrals. In the notation of the preceding section, the NAX energy is given by:

$$\mathcal{E}^{\text{NAX}} = \text{aug}(\mathcal{E}^{\bar{g}_{\text{2ch}}}, \bar{V}^{\text{eff}}_{\mu\lambda,\text{C}}).$$

We have computed all test molecules using this scheme. The corresponding errors (ΔE^{NAX}) relative to the HF-SCF reference energies are plotted in Figure 11 for all basis sets used (see Table S37 for numerical data). A comparison of ΔE^{NAX} against ΔE^{i} shows that the NAX scheme is generally less accurate than schemes including the 3-c 2-e integrals explicitly ($\Delta E^{\bar{g}_{\text{3ch}}}$ and $\Delta E^{\bar{g}_{\text{3cx}}}$) but far more accurate than all others. Closer inspection of Figure 11 reveals that the errors ΔE^{NAX} increase notably when going from minimal (or single-ζ) via double-ζ to triple-ζ basis sets, and they also tend to increase with molecular size, especially for the largest basis set. These trends may be related to the number of neglected 2-e integrals which increases with the number of basis functions. Hence, the NAX scheme is expected to work best for a minimal basis set.

Before closing this section, we pinpoint the major hindrance that may prevent the NAX scheme from being readily implemented in practice. In our tests, the computation of the effective 3-c interaction terms $\bar{V}^{\mathrm{eff}}_{\mu\lambda,\mathrm{C}}$ makes use of the 3-c hybrid integrals $\bar{g}_{3\mathrm{ch}}$ and the density matrix, while any efficient practical implementation will need a realistic estimate of $\bar{V}^{\mathrm{eff}}_{\mu\lambda,\mathrm{C}}$ to be computed before the SCF procedure (without knowledge of $\bar{g}_{3\mathrm{ch}}$ and the density matrix). A second less severe problem is the efficient semiempirical computation of the 2-c hybrid integrals $\bar{g}_{2\mathrm{ch}}$ (neglected in NDDO but included here) which may be achieved by suitable scaling strategies. Given the documented importance of these extra terms, it would seem worthwhile to explore possible implementations of the NAX scheme with the aim to enhance the intrinsic accuracy of SQC models.

Conclusions

In this Big Data analysis of *ab initio* molecular integrals we computed 15.6 million 1-e integrals and 10.3 billion 2-e integrals for 32 typical organic molecules and classified them into 14 different types, e.g. seven for 2-e integrals, five for 1-e integrals, and two for combined 1-e and 2-e integrals.

When using the valence-only minimal CEP-4G basis for integral evaluation, the 2-e Coulomb integrals in the Löwdin basis are quantitatively close to their counterparts in the atomic basis and the 2-e exchange integrals can be safely neglected in line with the NDDO approximation. These patterns can be rationalized by scrutinizing the $\{\bar{\phi}_\mu\}$ basis, the symmetric orthogonalization matrix, and the pertinent 2-e integrals in the linear H_4 model system. The symmetric orthogonalization makes the 2-e hybrid integrals smaller but not entirely negligible (as assumed at the NDDO level), which may limit the accuracy of NDDO-based SQC methods.

When using large all-electron GTO basis sets for integral evaluation, there are strong multi-shell orthogonalization effects that lead to more irregular patterns in the transformed

2-e integrals; for example, the exchange integrals become smaller in the Löwdin basis but remain sizable so that it would seem inappropriate to neglect them. Thus the NDDO approximation for the 2-e integrals is less well justified for large all-electron basis sets, and it appears doubtful whether the *intrinsic* accuracy of NDDO-based SQC models can be improved by using more extended basis sets.

Our Big Data analysis confirms that the 1-e integrals in the core Hamiltonian are strongly affected by orthogonalization effects, especially the resonance integrals that are critical for chemical bonding. The analysis provides further support for the strategies to deal with these effects that have been introduced in the OMx series of NDDO-based SQC models.

In NDDO approximation, 3-c integrals are generally neglected. Our analysis shows that the symmetric orthogonalization indeed makes most of these integrals negligibly small in the Löwdin basis; however, the 1-e nuclear attraction integrals $\bar{V}_{\mu\lambda,\mathrm{C}}$ and the 2-e hybrid integrals $\bar{g}_{\mathrm{3ch}}$ are found to differ from zero appreciably also in the Löwdin basis. The contributions from these two types of 3-c integrals tend to compensate each other to some extent when building the Fock matrix. This effect can be quantified by considering the 3-c coupled potential integral that is defined as the sum of corresponding 3-c 1-e NAIs and 2-e terms involving 3-c hybrid integrals contracted with the density matrix. The analysis demonstrates that it may be qualitatively satisfactory to simultaneously neglect the 3-c integrals in NDDO-based SQC methods, but it also indicates that the accuracy of such methods might be enhanced by taking these terms into account.

Based on the insights gained from the Big Data analysis of integrals, we consider the novel NAX scheme with a valence-only minimal basis set as a possible strategy to go beyond the NDDO integral approximation in attempts to improve SQC methods while retaining their computational efficiency.

Acknowledgments

We are grateful to the European Research Council for financial support through an ERC Advanced Grant (OMSQC). We thank Markus Reiher for sending us a preprint of their analysis.[40]

References

[1] T. Helgaker, P. Jørgensen, J. Olsen, *Molecular Electronic-Structure Theory*, Wiley, **2000**.

[2] S. F. Boys, *Proc. R. Soc. Lond. A* **1950**, *200*, 542 – 554.

[3] H. Taketa, S. Huzinaga, K. O-ohata, *J. Phys. Soc. Jpn.* **1966**, *21*, 2313 – 2324.

[4] J. Almlöf, K. Faegri, K. Korsell, *J. Comput. Chem.* **1982**, *3*, 385 – 399.

[5] D. R. Hartree, *Math. Proc. Camb. Philos. Soc.* **1928**, *24*, 89 – 110.

[6] V. Fock, *Z. Phys.* **1930**, *61*, 126 – 148.

[7] C. C. J. Roothaan, *Rev. Mod. Phys.* **1951**, *23*, 69 – 89.

[8] G. G. Hall, *Proc. R. Soc. Lond. A* **1951**, *205*, 541 – 552.

[9] G. E. Scuseria, *J. Phys. Chem. A* **1999**, *103*, 4782 – 4790.

[10] C. Ochsenfeld, J. Kussmann, D. S. Lambrecht in *Linear-scaling methods in quantum chemistry*, Wiley, **2007**, Chapter 1, pp. 1 – 82.

[11] J. Kussmann, M. Beer, C. Ochsenfeld, *WIREs Comput. Mol. Sci.* **2013**, *3*, 614 – 636.

[12] J. A. Pople, D. L. Beveridge, *Approximate Molecular Orbital Theory*, McGraw-Hill, **1970**.

[13] M. J. S. Dewar, *The Molecular Orbital Theory of Organic Chemistry*, McGraw-Hill, **1969**.

[14] J. J. P. Stewart, *J. Mol. Model.* **2009**, *15*, 765 – 805.

[15] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, W. Yang, *Proteins* **2001**, *44*, 484 – 489.

[16] G. Cui, W. Thiel, *Angew. Chem. Int. Ed.* **2013**, *52*, 433 – 436.

[17] M. Hennemann, A. Friedl, M. Lobell, J. Keldenich, A. Hillisch, T. Clark, A. H. Göller, *ChemMedChem* **2009**, *4*, 657 – 669.

[18] X. Wu, A. Koslowski, W. Thiel, *J. Chem. Theory Comput.* **2012**, *8*, 2272 – 2281.

[19] X. Wu, A. Koslowski, W. Thiel in *Electronic Structure Calculations on Graphics Processing Units* (Eds.: R. C. Walker, A. W. Götz), Wiley, **2016**, Chapter 11, pp. 239 – 258.

[20] J. D. C. Maia, G. A. Urquiza Carvalho, C. P. Mangueira, S. R. Santana, L. A. F. Cabral, G. B. Rocha, *J. Chem. Theory Comput.* **2012**, *8*, 3072 – 3081.

[21] M. Hennemann, T. Clark, *J. Mol. Model.* **2014**, *20*, 2331.

[22] J. T. Margraf, M. Hennemann, B. Meyer, T. Clark, *J. Mol. Model.* **2015**, *21*, 144.

[23] W. Thiel, *WIREs Comput. Mol. Sci.* **2014**, 145 – 157.

[24] A. S. Christensen, T. Kubař, Q. Cui, M. Elstner, *Chem. Rev.* **2016**, *116*, 5301 – 5337.

[25] T. Bredow, K. Jug in *Handbook of Solid State Chemistry, Vol. 5* (Eds.: R. Dronskowski, S. Kikkawa, A. Stein), Wiley, **2017**, Chapter 6, pp. 159 – 202.

[26] J. A. Pople, D. P. Santry, G. A. Segal, *J. Chem. Phys.* **1965**, *43*, S129 – S135.

[27] R. G. Parr, *J. Chem. Phys.* **1952**, *20*, 1499 – 1499.

[28] P.-O. Löwdin, *J. Chem. Phys.* **1950**, *18*, 365 – 375.

[29] R. D. Brown, K. R. Roby, *Theor. Chim. Acta* **1970**, *16*, 175 – 193.

[30] K. R. Roby, *Chem. Phys. Lett.* **1971**, *11*, 6 – 10.

[31] K. R. Roby, *Chem. Phys. Lett.* **1972**, *12*, 579 – 582.

[32] G. S. Chandler, F. E. Grader, *Theor. Chim. Acta* **1980**, *54*, 131 – 144.

[33] K. Neymeyr, F. F. Seelig, *Int. J. Quantum Chem.* **1995**, *53*, 515 – 518.

[34] K. Neymeyr, F. F. Seelig, *International Journal of Quantum Chemistry* **1995**, *53*, 519 – 535.

[35] K. Neymeyr, K. Engel, *Int. J. Quantum Chem.* **1995**, *53*, 537 – 540.

[36] K. Neymeyr, *Int. J. Quantum Chem.* **1995**, *53*, 541 – 552.

[37] K. Neymeyr, *Int. J. Quantum Chem.* **1995**, *53*, 553 – 568.

[38] D. B. Cook, P. C. Hollis, R. McWeeny, *Mol. Phys.* **1967**, *13*, 553 – 571.

[39] I. Fischer-Hjalmars, *J. Chem. Phys.* **1965**, *42*, 1962 – 1972.

[40] T. Husch, M. Reiher, *J. Chem. Theory Comput.* **2018**.

[41] N. Frank, *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

[42] Q. Sun, *J. Comput. Chem.* **2015**, *36*, 1664 – 1671.

[43] W. J. Stevens, H. Basch, M. Krauss, *J. Chem. Phys.* **1984**, *81*, 6026 – 6033.

[44] W. J. Hehre, R. F. Stewart, J. A. Pople, *J. Chem. Phys.* **1969**, *51*, 2657 – 2664.

[45] J. S. Binkley, J. A. Pople, W. J. Hehre, *J. Am. Chem. Soc.* **1980**, *102*, 939 – 947.

[46] W. J. Hehre, R. Ditchfield, J. A. Pople, *J. Chem. Phys.* **1972**, *56*, 2257 – 2261.

[47] R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.* **1980**, *72*, 650 – 654.

[48] M. J. S. Dewar, W. Thiel, *J. Am. Chem. Soc.* **1977**, *99*, 4899 – 4907.

[49] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902 – 3909.

[50] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 209 – 220.

[51] J. J. P. Stewart, *J. Mol. Model.* **2004**, *10*, 6 – 12.

[52] J. J. P. Stewart, *J. Mol. Model.* **2007**, *13*, 1173 – 1213.

[53] J. J. P. Stewart, *J. Mol. Model.* **2013**, *19*, 1 – 32.

[54] M. Kolb, W. Thiel, *J. Comput. Chem.* **1993**, *14*, 775 – 789.

[55] W. Weber, W. Thiel, *Theor. Chem. Acc.* **2000**, *103*, 495 – 506.

[56] M. Scholten, PhD thesis, Universität Düsseldorf, Düsseldorf, **2003**.

[57] P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, W. Weber, R. Steiger, M. Scholten, W. Thiel, *J. Chem. Theory Comput.* **2016**, *12*, 1082 – 1096.

[58] C. C. J. Roothaan, *J. Chem. Phys.* **1951**, *19*, 1445 – 1458.

[59] K. Rüdenberg, *J. Chem. Phys.* **1951**, *19*, 1459 – 1477.

[60] K. Ruedenberg, C. C. J. Roothaan, W. Jaunzemis, *J. Chem. Phys.* **1956**, *24*, 201 – 220.

[61] M. Goeppert-Mayer, A. L. Sklar, *J. Chem. Phys.* **1938**, *6*, 645 – 652.

[62] A. L. Sklar, R. H. Lyddane, *J. Chem. Phys.* **1939**, *7*, 374 – 379.

[63] J. A. Pople, G. A. Segal, *J. Chem. Phys.* **1965**, *43*, S136 – S151.

[64] D. N. Laikov, *J. Chem. Phys.* **2011**, *135*, 134120.

[65] K. R. Briling, *J. Chem. Phys.* **2017**, *147*, 157101.

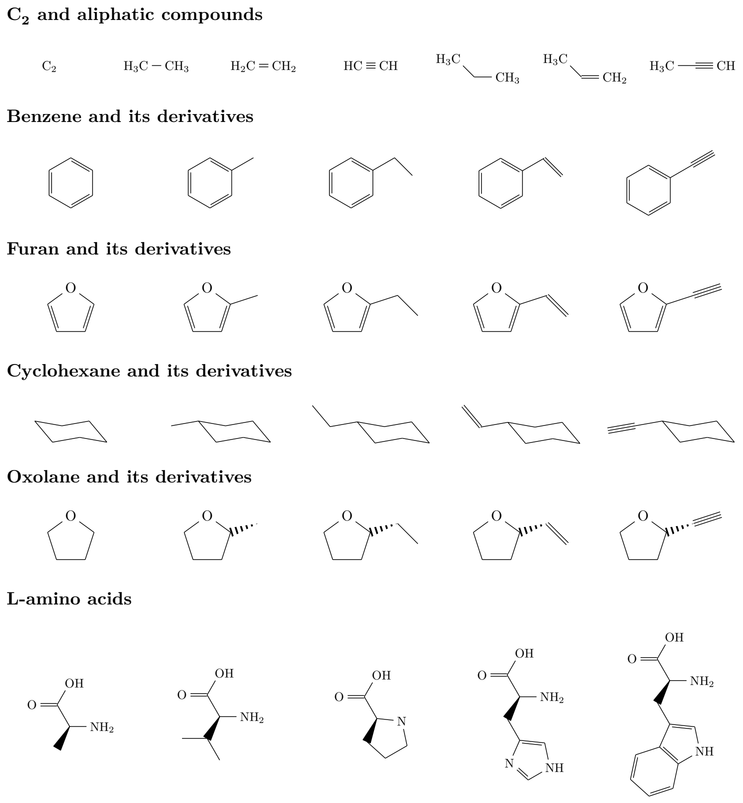
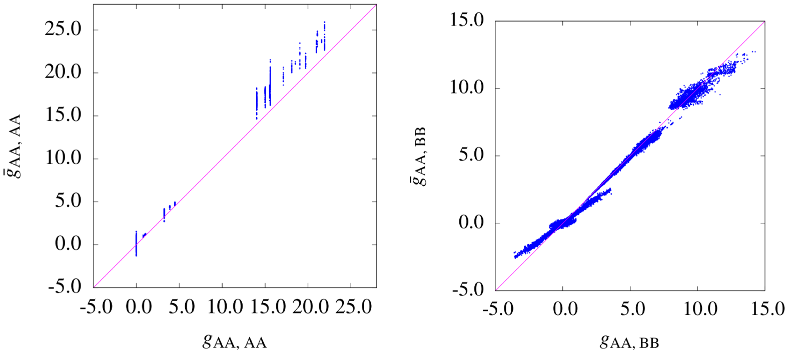

C_2 and aliphatic compounds
C_2
H_3C-CH_3
$H_2C=CH_2$
$HC\equiv CH$
H_3C
CH_3
H_3C
CH_2
$H_3C-\!\!\equiv CH$
Benzene and its derivatives
Furan and its derivatives
O
Cyclohexane and its derivatives
Oxolane and its derivatives
O
L-amino acids
OH
O
NH_2
N
N
NH
NH

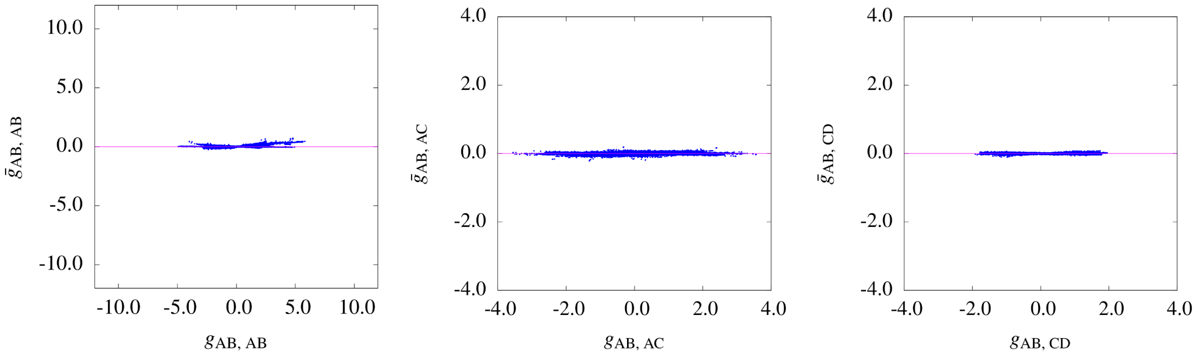
Figure 1: Molecules for Big Data analysis of *ab initio* molecular integrals.



(a) 1-c integral

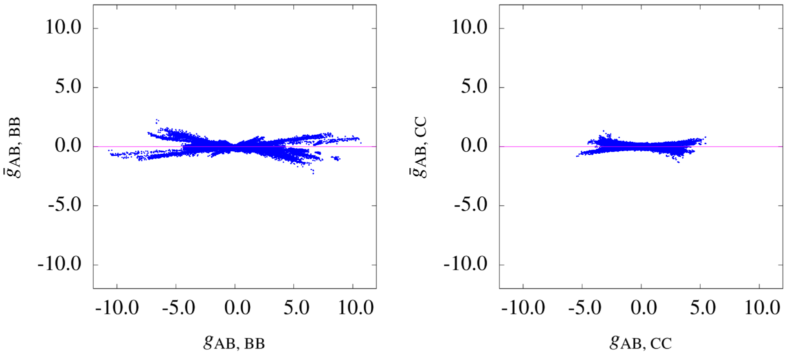
(b) 2-c Coulomb integral

Figure 2: Correlation diagrams of the NDDO-retained *ab initio* 2-*e* Coulomb integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation. The diagonal lines represent ideal correlation lines with slope 1 going through the origin.



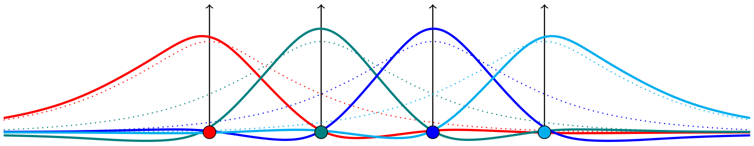
(a) 2-c exchange integral (b) 3-c exchange integral (c) 4-c integral

Figure 3: Correlation diagrams of the NDDO-neglected *ab initio* 2-*e* exchange integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation.



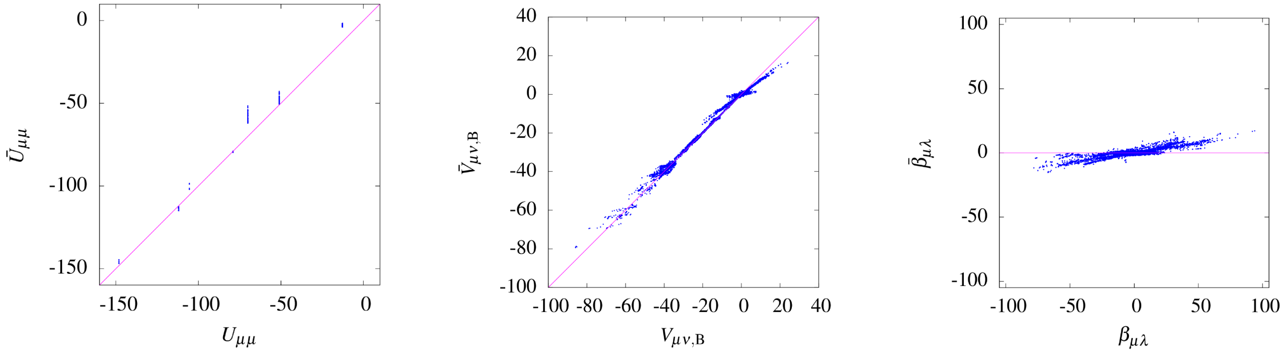
(a) 2-c hybrid integral (b) 3-c hybrid integral

Figure 4: Correlation diagrams of the NDDO-neglected *ab initio* 2-*e* hybrid integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation.



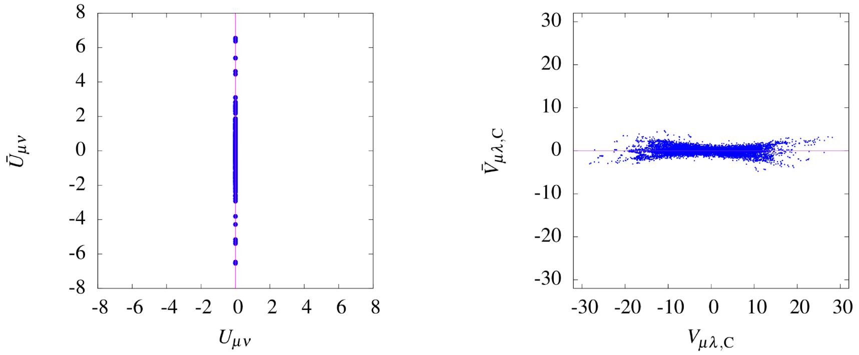
	ϕ_1	ϕ_2	ϕ_3	ϕ_4
$\bar{\phi}_1 =$	$+1.28$	-0.57	$+0.13$	-0.02
$\bar{\phi}_2 =$	-0.57	$+1.65$	-0.67	$+0.13$
$\bar{\phi}_3 =$	$+0.13$	-0.67	$+1.65$	-0.57
$\bar{\phi}_4 =$	-0.02	$+0.13$	-0.57	$+1.28$

Figure 5: The atomic basis ($\{\phi_i\}$, dotted lines) and the Löwdin basis ($\{\bar{\phi}_\mu\}$, solid lines) for the linear H_4 model system with internuclear distances of 1.40 Bohr. The standard STO-3G basis set is used. The coefficient matrix for the symmetric orthogonalization is listed for each $\{\bar{\phi}_\mu\}$ in different colors.



(a) diagonal atomic integral (b) 2-c nuclear attraction integral (c) resonance integral

Figure 6: Correlation diagrams of the NDDO-retained *ab initio* 1-*e* integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation. The diagonal lines represent ideal correlation with slope 1 going through the origin.



(a) off-diagonal atomic integral (b) 3-c nuclear attraction integral

Figure 7: Correlation diagrams of the NDDO-neglected *ab initio* 1-*e* integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation.

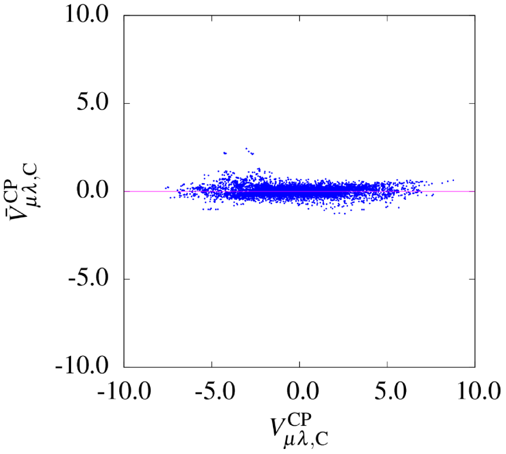
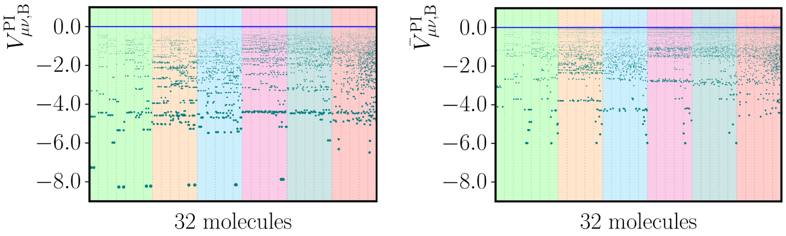


Figure 8: Correlation diagram of *ab initio* coupled potential integrals (all in eV) in the atomic basis (abscissa) and the Löwdin basis (ordinate). The valence-only minimal CEP-4G basis function is employed for integral evaluation. The correlation coefficient r^2 and the standard deviation σ with respect to the zero line are 0.969 and 0.052 eV, respectively.



(a) $V^{\mathrm{PI}}_{\mu\nu,\mathrm{B}}$ in $\{\phi_i\}$ (b) $\bar{V}^{\mathrm{PI}}_{\mu\nu,\mathrm{B}}$ in $\{\bar{\phi}_\mu\}$

Figure 9: Comparison of *ab initio* penetration integrals (all in eV) for 32 organic molecules (a) in the atomic basis ($\{\phi_i\}$) and (b) in the Löwdin basis ($\{\bar{\phi}_\mu\}$). The valence-only minimal CEP-4G basis function is employed for integral evaluation. Repulsive (almost invisible) and attractive interactions are depicted as red and green dots, respectively.

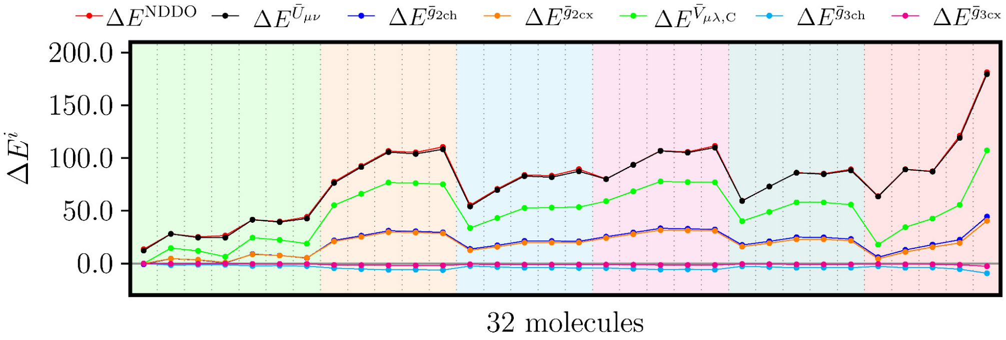


Figure 10: Induced error (ΔE^i, in eV) relative to the HF-SCF molecular electronic energy. See the text for the seven distinct schemes of neglecting molecular integrals. The valence-only minimal CEP-4G basis function is employed for integral evaluation.

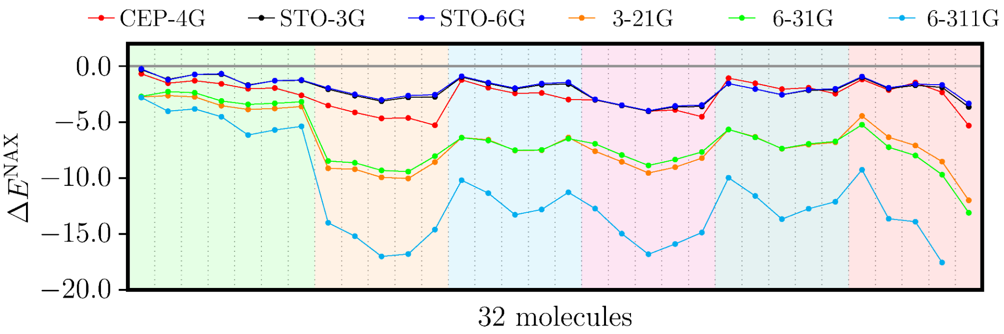


Figure 11: Error (ΔE^{NAX}, in eV) relative to the HF-SCF molecular electronic energy calculated with the CEP-4G, STO-3G, STO-6G, 3-21G, 6-31G, and 6-311G basis functions. The NAX scheme is employed in the calculations. The NAX result for tryptophan in 6-311G is not available due to technical constraints.

2-e integrals

notation	mathematical definition	N_A	NDDO	description
g_{1cc}	$(\phi^\mathrm{A}\phi^\mathrm{A}, \phi^\mathrm{A}\phi^\mathrm{A})$	1	Y	1-c Coulomb integral
g_{2cc}	$(\phi^\mathrm{A}\phi^\mathrm{A}, \phi^\mathrm{B}\phi^\mathrm{B})$	2	Y	2-c Coulomb integral
g_{2cx}	$(\phi^\mathrm{A}\phi^\mathrm{B}, \phi^\mathrm{A}\phi^\mathrm{B})$	2	N	2-c exchange integral
g_{3cx}	$(\phi^\mathrm{A}\phi^\mathrm{B}, \phi^\mathrm{A}\phi^\mathrm{C})$	3	N	3-c exchange integral
g_{4cx}	$(\phi^\mathrm{A}\phi^\mathrm{B}, \phi^\mathrm{C}\phi^\mathrm{D})$	4	N	4-c exchange integral
g_{2ch}	$(\phi^\mathrm{A}\phi^\mathrm{B}, \phi^\mathrm{B}\phi^\mathrm{B})$	2	N	2-c hybrid integral
g_{3ch}	$(\phi^\mathrm{A}\phi^\mathrm{B}, \phi^\mathrm{C}\phi^\mathrm{C})$	3	N	3-c hybrid integral

1-e integrals

notation	mathematical definition	N_A	NDDO	description
$U_{\mu\mu}$	$(\phi^\mathrm{A}_\mu \mid \mathcal{T} + \mathcal{V}_\mathrm{A} \mid \phi^\mathrm{A}_\mu)$	1	Y	diagonal atomic integral
$V_{\mu\nu,\mathrm{B}}$	$(\phi^\mathrm{A}_\mu \mid \mathcal{V}_\mathrm{B} \mid \phi^\mathrm{A}_\nu)$	2	Y	2-c nuclear attraction integral
$\beta_{\mu\lambda}$	$(\phi^\mathrm{A}_\mu \mid \mathcal{T} + \mathcal{V}_\mathrm{A} + \mathcal{V}_\mathrm{B} \mid \phi^\mathrm{B}_\lambda)$	2	Y	resonance integral
$U_{\mu\nu}$	$(\phi^\mathrm{A}_\mu \mid \mathcal{T} + \mathcal{V}_\mathrm{A} \mid \phi^\mathrm{A}_\nu)$	1	N	off-diagonal atomic integral
$V_{\mu\lambda,\mathrm{C}}$	$(\phi^\mathrm{A}_\mu \mid \mathcal{V}_\mathrm{C} \mid \phi^\mathrm{B}_\lambda)$	3	N	3-c nuclear attraction integral

combined 1-e and 2-e integrals

notation	mathematical definition	N_A	description
$V^{\mathrm{CP}}_{\mu\lambda,\mathrm{C}}$	$V_{\mu\lambda,\mathrm{C}} + \sum_\rho^\mathrm{C} \sum_\sigma^\mathrm{C} G_{\mu\lambda}[D_{\rho\sigma}]$	3	coupled potential integral
$V^{\mathrm{PI}}_{\mu\nu,\mathrm{B}}$	$V_{\mu\nu,\mathrm{B}} + Z_\mathrm{B} \cdot (\phi^\mathrm{A}_\mu\phi^\mathrm{A}_\nu, \phi^\mathrm{B}_s\phi^\mathrm{B}_s)$	2	penetration integral

Table 1: Classification of *ab initio* molecular integrals. N_A denotes the number of the atomic centers involved. “Y” and “N” are abbreviations for “Yes” and “No” specifying whether the integral is kept or not in NDDO approximation. The subscript in the notation for the 2-e integrals indicates the number of involved atomic centers and the integral type; the latter is symbolized by the last character “c”, “x”, and “h” representing Coulomb, exchange, and hybrid integrals, respectively. NDDO conventions (see the text for details) are used for denoting the 1-e integrals and the combined 1-e and 2-e integrals. The 2-e integral part of $V^{\mathrm{CP}}_{\mu\lambda,\mathrm{C}}$ is abbreviated using $\mathbf{G}$ matrix elements and $D_{\rho\sigma}$ stands for a converged density matrix element belonging to atom C. Z_B is the valence nuclear charge on atom B.

2-*e* integrals							
	NDDO-retained		NDDO-neglected				
	Coulomb integral		exchange integral			hybrid integral	
notation	$\bar{g}_{1\mathrm{cc}}$	$\bar{g}_{2\mathrm{cc}}$	$\bar{g}_{2\mathrm{cx}}$	$\bar{g}_{3\mathrm{cx}}$	$\bar{g}_{4\mathrm{cx}}$	$\bar{g}_{2\mathrm{ch}}$	$\bar{g}_{3\mathrm{ch}}$
r^2	0.996	0.994	0.989	0.997	0.999	0.938	0.975
σ	0.749	0.157	0.033	0.004	0.001	0.120	0.033
$\max_-$	-1.301	-2.595	-0.236	-0.216	-0.106	-2.285	-1.343
$\max_+$	$+25.947$	$+12.731$	$+0.729$	$+0.193$	$+0.083$	$+2.282$	$+1.343$

1-*e* integrals					
	NDDO-retained			NDDO-neglected	
notation	$\bar{U}_{\mu\mu}$	$\bar{V}_{\mu\nu,\mathrm{B}}$	$\bar{\beta}_{\mu\lambda}$	$\bar{U}_{\mu\nu}$	$\bar{V}_{\mu\lambda,\mathrm{C}}$
r^2	0.983	0.994	0.740	$-^b$	0.969
σ	6.954	0.774	11.977	0.911	0.196
$\max_-$	-146.816	-79.351	-15.201	-6.535	-4.760
$\max_+$	$-^a$	$+16.409$	$+17.133$	$+6.546$	$+4.678$

[a]: $\bar{U}_{\mu\mu}$ is always negative.

[b]: r^2 is undefined for $\bar{U}_{\mu\nu}$ because $U_{\mu\nu}$ is zero by symmetry in the atomic basis.

Table 2: Statistical measures of molecular integrals in the Löwdin basis. The valence-only minimal CEP-4G basis function is employed for integral evaluation. The integral notation is specified in Table 1. The correlation coefficient r^2 and the standard deviation σ (in eV) for the NDDO-retained integrals and the NDDO-neglected integrals refer to the corresponding integrals in the atomic basis and to zero, respectively. $\max_-$ and $\max_+$ denote the most negative and most positive values of these integrals, respectively.

GTO	$\bar{\Delta}E^{\mathrm{NDDO}}$	$\bar{\Delta}E^{\bar{U}_{\mu\nu}}$	$\bar{\Delta}E^{\bar{g}_{2\mathrm{ch}}}$	$\bar{\Delta}E^{\bar{g}_{2\mathrm{cx}}}$	$\bar{\Delta}E^{\bar{V}_{\mu\lambda,\mathrm{C}}}$	$\bar{\Delta}E^{\bar{g}_{3\mathrm{ch}}}$	$\bar{\Delta}E^{\bar{g}_{3\mathrm{cx}}}$
CEP-4G	+79.06	+78.18	+19.79	+18.56	+48.85	−3.90	−0.92
STO-3G	+175.33	+146.60	+14.64	+14.50	+36.83	−2.23	−0.61
STO-6G	+173.50	+150.99	+14.71	+14.31	+36.52	−2.42	−0.60
3-21G	−42.69	−385.57	−60.76	−56.02	−112.61	−2.10	−0.36
6-31G	−169.99	−379.59	−64.90	−60.59	−124.33	−2.41	−0.58
6-311G	+3143.04	−433.87	−79.39	−73.65	−157.05	−5.83	−1.34

Table 3: Mean induced error ($\bar{\Delta}E^i$, in eV) relative to the standard HF-SCF molecular electronic energy. See the text for the seven distinct schemes of neglecting molecular integrals.