A fast doubly hybrid density functional method close to chemical accuracy using a local opposite spin ansatz

Igor Ying Zhang, Xin Xu, Yousung Jung, and William A. Goddard III

1Department of Chemistry, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Ministry of education (MOE) Laboratory for Computational Physical Science, Fudan University, Shanghai 200433, China; 2State Key Laboratory of Physical Chemistry of Solid Surfaces, College for Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China; 3World Class University (WCU) Professor Program at the Graduate School of Energy, Water, and Sustainability (EEWS), Korea Advanced Institute of Science and Technology, Daejeon 305 701, Republic of Korea; and 4Materials and Process Simulation Center (139 74), California Institute of Technology, Pasadena, CA 91125

Contributed by William A. Goddard III, September 21, 2011 (sent for review April 15, 2011)

We develop and validate the XYGJ-OS functional, based on the adiabatic connection formalism and Görling-Levy perturbation theory to second order and using the opposite-spin (OS) ansatz combined with locality of electron correlation. XYGJ-OS with local implementation scales as $N^2$ with an overall accuracy of 1.28 kcal/mol for thermochemistry, bond dissociation energies, reaction barrier heights, and nonbonded interactions, comparable to that of 1.06 kcal/mol for the accurate coupled-cluster based G3 method (scales as $N^7$) and much better than many popular density functional methods: B3LYP (4.98), PBE0 (4.36), and PBE (12.10).

Obtaining chemical accuracy (~1 kcal/mol) to quantify key chemical quantities (e.g., heats of formation, bond dissociation energies, and reaction barrier heights) using quantum mechanics (QM) has been a major focus in the development of the theory. This has led to, for example, the G3 method (1, 2) that approaches this chemical accuracy. Because G3 is a coupled cluster based method, it scales on the order of $N^7$, where $N$ measures the system size, limiting to fairly small species for routine use.

The desire to predict unique physiochemical phenomena (e.g., solvation, catalysis, self assembly, and drug design) in practical (large) systems has brought about a second major focus of the theoretical development, leading, for example, to divide and conquer formulations to attain more efficient scaling (3), but with much lower accuracy than G3.

Density functional theory (DFT) in the framework of Kohn-Sham (KS) scheme (4, 5) provides a “shortcut” to the many body problem. Many density functional approximations (DFAs), provide typical scaling of $N^3$ with an overall accuracy of 1.28 kcal/mol for thermochemistry, bond dissociation energies, reaction barrier heights, and nonbonded interactions, comparable to that of 1.06 kcal/mol for the accurate coupled-cluster based G3 method (scales as $N^7$) and much better than many popular density functional theory methods: B3LYP (4.98), PBE0 (4.36), and PBE (12.10).

Theory

The Holy Grail in KS DFT is to find the exact, yet unknown, exchange correlation functional $E_{xc}[ho]$ using density $\rho$ as the basic variable (4, 5). In practice, an approximate $E_{xc}$ must be adopted, which is often partitioned into the exchange and correlation parts

$$E_{xc}[ho] \approx E_{x}^{DFA}[ho] + E_{c}^{DFA}[ho].$$

$E_{x}^{DFA}$ has been extended to include a portion of nonlocal $E_{x}^{HF}$, where the superscript “HF” emphasizes that the exchange part has the same form as in Hartree Fock theory (6). The exchange correlation potential $\upsilon_{xc}$ in the nth cycle of the self consistent field (SCF) process to solve the KS equation is obtained as a functional of $\rho$ of the previous cycle.

On the other hand, Görling and Levy (GL) (7, 8) argued that the same KS scheme should work as well in terms of KS orbitals $\phi_i$ and eigenvalues $\epsilon_i$. GL proposed a formally exact KS scheme based on perturbation theory, where $E_{xc}$ was expressed as:

$$E_{xc}[ho] = E_{x}^{HF}[(\phi_i)] + \sum_{j=2}^{\infty} E_{c,j}^{GL}[(\phi_i), (\epsilon_i), (\upsilon_{\upsilon_1}(r), \upsilon_{\upsilon_2}(r), \ldots, \upsilon_{\upsilon_{j-1}}(r))].$$

$$E_{c,2}^{GL} = E_{c,2}^{GL} = \sum_{I} \frac{|\Phi_{I}^{\upsilon} - \upsilon_{\upsilon}(\Phi_{I}^{\upsilon})|^2}{E_{\upsilon,0} - E_{\upsilon}}.$$

Here $E_{c,2}^{GL}$ stands for the GL perturbation theory up to the second order, $\{\Phi_{I}^{\upsilon}, E_{\upsilon}\}$ and $\{\Phi_{I}^{\upsilon}, E_{\upsilon}\}$ are the wave function and energy for the ground state and the $\upsilon$th excited state of an $N$ electron KS system, respectively; $V_{ee}$ is the operator of electron electron repulsion, and the $\upsilon_{\upsilon}$ potential may be determined from the “exchange only” KS equation (8). With knowledge of the potentials $\upsilon_{\upsilon}(r)$. Eq. 2 gives the formally exact exchange correlation energy as functional of the KS orbitals $\phi_i$ and eigenvalues $\epsilon_i$, with which the approximate $E_{xc}^{DFA}[ho]$ defined in [1] can be compared. However, in practice, such a procedure is difficult to apply to higher than second order due to the unfavorable scaling with the system size, and in many cases perturbation theory is nonconvergent. Thus the scheme must be simplified to make it applicable for including higher order contributions.


The authors declare no conflict of interest.

1To whom correspondence may be addressed. E-mail: xchem@fudan.edu.cn, or yjin@kaist.ac.kr, or wgaw@caltech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi/10.1073/pnas.1115123108/-/DCSupplemental.
The adiabatic connection method ([ACM], ref. 9, 10), which defines a family of partially interacting N electron systems with a coupling constant for a fixed ρ, provides a powerful tool for developing and understanding E_

(6, 11 13). ACM suggests that Eq. 2 up to second order, while being more appropriate for coupling close to zero, is only exact if the potential energy of exchange correlation depends linearly on the coupling strength, whereas [1] is more associated with full electron electron coupling (13, also see SI Text for more discussion). To go beyond the linear approximation and to take advantages of both [1] and Eq. 2, we define an empirical DHDF that combines Eq. 2 (\(j = 2\)) with [1]:

\[ E_{\text{DHDF}}^{\text{1st}}[\rho] = c_i E_{\text{LDA}}^{\text{LDA}} + c_2 E_{\text{GGA}}^{\text{GGA}} + c_3 E_{\text{H}}^{\text{H}} + c_4 E_{\text{LDA}}^{\text{LDA}} + c_5 E_{\text{GGA}}^{\text{GGA}} + c_6 E_{\text{PT2}}^{\text{PT2}}. \]  

[4]

Here \(E_{\text{LDA}}^{\text{LDA}}\) and \(E_{\text{GGA}}^{\text{GGA}}\) are the exchange and correlation components within the local density approximation (LDA), and \(E_{\text{GGA}}^{\text{GGA}}\) are the corresponding correlation terms to LDA within the generalized gradient approximation (GGA). The meta GGA functionals that include kinetic energy density or the Laplacian of \(\rho\) are also used in place of GGAs. Such a combination suggests that only the energetically most important double excitation PT2 term in Eq. 2 need be calculated explicitly, the higher order correlations are embedded into the parameterized terms such as \(E_{\text{GGA}}^{\text{GGA}} - E_{\text{H}}^{\text{H}}\) (12). Other examples of DHDFs include MC3BB (14), B2PL YP (15), B2GP YP (16), and oB97X 2 (17), which are derived and constructed differently (see below and SI Text for more discussion).

Several key issues distinguish our approach from GL perturbation theory (7, 8) and other DHDFs (14 23).

i. Instead of seeking for \(\{\rho, \epsilon, u, j\}\) self consistently and order by order via Eq. 2, we use a conventional DFA defined in [1] to generate \(\{\rho, \epsilon, u\}\) and to evaluate \(E_x\) defined in Eq. 4 in a post SCF manner. In fact, the GL perturbation theory, Eq. 2, was built on KS orbitals with a local effective potential, while we have extended it to include hybrid functionals with generalized KS orbitals (12, 13).

ii. \(E_{\text{PT2}}^{\text{PT2}}\) differs from \(E_{\text{GL}}^{\text{GL}}\) in that the singles contribution is not explicitly calculated, but we argue (12, 13) that it can be reasonably absorbed into the parameterization in Eq. 4. Higher order contributions \(E_{\text{ij}}^{\text{ij}}\) (\(j > 2\)) are also implicitly included via the parameterization.

iii. Neglecting the nonlocal exchange correlation effects \(\{E_{\text{ij}}^{\text{ij}}, E_{\text{ij}}^{\text{PT2}}\}\) in Eq. 4 leads to a pure DFA, while neglecting the local exchange correlation effects, Eq. 4 gives an approximation to GL2. Hence our DHDF may be regarded as an interpolation approach between a pure DFA and GL2, while both of them are in the framework of the KS scheme. As different orbitals are used to construct the PT2 term, neglecting the local exchange correlation effects in other DHDFs (14 23) will bring back the conventional MP2, such that these DHDFs may be regarded as an interpolation approach between a pure DFA and MP2, the wave function based lowest level correlation method. These functionals go beyond the frame work of the KS scheme (23).

Assuming LDA and GGA in Eq. 4 are SVWN (24, 25) and BLYP (26, 27), respectively, we previously developed XYG3 (12), a DHDF which is remarkably accurate for a wide range of systems and important chemical properties (12, 13, 28 31). Nevertheless, the PT2 term in XYG3 and other DHDFs is evaluated in a way similar to MP2:

\[ E_{\text{ij}}^{\text{PT2}} = \frac{1}{J} \sum_{\substack{ab \, \sum \left(\langle \phi_i \phi_j | \phi_a \phi_b \rangle - \langle \phi_i \phi_j | \phi_a \phi_b \rangle \right)^2 \epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b.} \]  

[5]

where the subscripts (i, j) and (a, b) denote the occupied and unoccupied KS spin orbitals, respectively, leading to a formal scaling as \(N^3\), as opposed to a formal scaling of \(N^6\) as in B3LYP (24 27, 32, 33). This unfavorable scaling raises an issue for the practicality to apply DHDFs to large systems.

Density fitting approximations have often been used in electronic structure theories to reduce computational expense. In the so called “resolution of the identity” RI MP2 method (34, 35), the product of occupied and virtual orbitals (\(\alpha\beta\) pair) is expanded with auxiliary functions, such that numerous 4 center 2 electron integrals based on molecular orbitals (MO) are replaced by fewer 3 and 2 center integrals with cheaper transformation from atomic orbitals (AO) to MO. RI MP2 is about 5 20 times faster than conventional MP2 (36, 37). Nevertheless, RI MP2 reduces only the prefactor, it does not change the scaling. Similarly, we have RI XYG3, which retains the original accuracy but is faster for small systems with large basis sets.

Here we propose a unique OS ansatz for DHDF, that yields a balanced description of nonlocal correlation effects while considerably reducing computational time. Our OS ansatz is motivated by the observation that the most important electron correlation effects involve correlations of the OS electrons in the same orbital. The OS ansatz leads to \(N^6\) scaling (36, 37) [using auxiliary basis expansions (34, 35) and Laplace quadrature approximations (38)]. It should be emphasized that as the same spin (SS) correlation is very important in accurate description of open shell systems and magnetic properties, such contributions cannot be simply neglected. In XYG3 OS, the SS correlation effects are included within the standard DFA.

In recognition of the “nearightedness of electron correlation” as emphasized by Kohn (39), we then build upon XYG3 OS to introduce the local approximation for the OS electron correlation by utilizing the sparsity of the RI expansion coefficients, integral matrices, and Laplace transform matrices. The local implementation of XYG3 OS scales as \(N^2\) while retaining the accuracy of the original XYG3 OS (see SI Text for more discussion).

Thus our proposed functional form (XYG3 OS) is

\[ E_{\text{XYG3 OS}}^{\text{XYG3 OS}}[\rho] = E_{\text{H}}^{\text{H}} + (1 - \epsilon_{\text{ij}})E_{\text{ij}}^{\text{ij}} + (\epsilon_{\text{xyW}}E_{\text{xyW}}^{\text{xyW}} + \epsilon_{\text{lyp}}E_{\text{lyp}}^{\text{lyp}}) + E_{\text{PT2}}^{\text{PT2}}. \]  

[6]

In Eq. 6 we normalize the HF exchange and Slater exchange (24), while eliminating the \(\Delta E_{\text{GGA}}^{\text{GGA}}\) contribution. The correlation part consists of \(E_{\text{xyW}}^{\text{xyW}}\) (25), \(E_{\text{lyp}}^{\text{lyp}}\) (27), and \(E_{\text{PT2}}^{\text{PT2}}\), where the first term includes both the SS and OS effects while the second and third terms include only OS components. Our concept is that the combination of VWN, LYP, and PT2 OS yields a balanced description of both local and nonlocal spin dependent correlation terms. To determine the optimal four parameters in Eq. 6, we use the experimental heats of formation (HOF) data for the G3/99 set of 223 molecules (1, 2) as the training set, leading to \(\{\epsilon_{\text{xyW}}, \epsilon_{\text{lyp}} | E_{\text{PT2}}^{\text{PT2}}\} = \{0.7731, 0.2309, 0.2754, 0.4364\}.

Results and Discussion We emphasize that we use the fully optimized B3LYP orbitals to generate the density and to calculate each term in Eqs. 4 or 6 (12, 13, 29 31). But the choice of LDA = SVWN and GGA = BLYP in Eqs. 4 or 6, as well as using B3LYP orbitals as input, is not unique. We find that any conventional DFTs defined in [1] can also serve the same purpose, leading to similar performance, albeit with a reoptimized set of mixing parameters. In our method, we assume that our KS wave function is the zeroth order wave function in the GL perturbation theory that gives the correct ground state electron density. Instead, MC3BB is a multicoefficient method, which mixes the total energies from the conventional MP2 and a conventional DFA calculation. Hence there are two independent SCF calculations in the MC3BB type of...
DHDFs, which lead to two sets of different orbitals, yielding two different densities. In MC3BB, the SCF HF orbitals are used for MP2 evaluation (14), while it is well known that HF wave function is the one determinantal wave function which gives the lowest expectation value with the fully interacting Hamiltonian. It has been shown recently (23) that the B2PLYP family of functionals works in a way which is technically similar to MP2. These functionals adopt orbitals that minimize the one determinantal wave function based on the so called density scaled one parameter hybrid approximation. Hence, the density from the B2PLYP type of functionals is by construction not meant to be the true density. In contrast to what was originally proposed (15), there is no singles’ contribution by construction in the B2PLYP family of functionals, and its theoretical basis is provided by the multideterminant extension of the Kohn Sham scheme (23). On the other hand, the key idea of the XYG3 type of functionals is to combine the GL perturbation theory and the standard KS scheme with DHDFs which lead to two sets of different orbitals, yielding two different energies. Only the energetically most important double excitation PT2 term in Eq. 2 is calculated explicitly using orbitals generated from a conventional DFA in [1]. Our present OS ansatz further reduces the computational cost by only calculating the most important electron correlation effects contributed by the OS electrons in the same orbital. XYGJ OS presents a unique combination of speed and accuracy. Hence, we propose that the current DHDFs shall be categorized into three types (13), as represented by MC3BB (14), B2PLYP (15), and XYG3 (12). The former two go beyond the KS scheme.

It is possible to carry out a SCF calculation for any orbital dependent $E_{xc}$ using, for example, the direct optimization approach to compute the optimized effective potential (OEP) as proposed by Yang and coworkers (40). It was found that such a SCF procedure can often lead to unphysically unbound state with too low total energy mainly due to the near degeneracy of the orbitals in the single excitation terms. This unbound issue can be largely remedied by calculating only the double excitation terms in the post SCF manner (40). This procedure is indeed the calculation approach adopted by the XYG3 type of functionals. Furthermore, we note that extra terms will appear when formulating the analytical gradients for doing geometry optimization with our functionals, which, however, will not impose any practical difficulty in implementation as compared to other types of DHDFs.

We find that XYGJ OS is remarkably accurate for a broad range of systems and important chemical properties (1, 2, 41 45) other than HOE which are not used in fitting the parameters. In Table 1 (more details are in SI Text) compares the overall performance of some representative DFT methods, showing that XYGJ OS is the best or nearly best for essentially all properties, leading to chemical accuracy (1.28 kcal/mol) comparable to the G3 theory (which contains four empirical parameters involving the number of electron pairs) (1.06) and much better than MP2 (7.49 kcal/mol).

### Heats of Formation

The 223 molecule G3/99 set (1, 2) provides a test of accuracy for the main group covalent systems. XYGJ OS gives mean absolute deviation (MAD) of 1.65 kcal/mol, lying between those of G2 (1.89) and G3 (1.06) theories. Note that the MADs for HOE listed in Table 1 associated with XYG3, B2PLYP, B2PLYP D, and oB97X2 (LP) are taken from the original ref. 12, 17, 22, which were obtained by using the way these functionals were parameterized. It was shown that HOE calculations with DHDFs are most prone to the basis set effects (29). The results with the G3Large basis set (1) used for optimization of the XYGJ OS functional can be found in SI Text, where severe basis set dependences are clearly shown.

### Charged Species

Charged species are not included in our training set. The G2 I set (2) for ionization potential (IP) contains 14 atoms and 24 molecules. XYGJ OS gives a MAD of 1.23 kcal/mol, being one of the best DFT methods for calculating IP. Over the 25 cases in the G2 I set for electron affinity (EA) (2), XYGJ OS leads to MAD = 1.97 kcal/mol. Generally, increasing the size of the basis set will increase the accuracy for EA calculation with DHDFs. For the eight systems for proton affinity in the G2 and G3 sets (1), XYGJ OS leads to MAD = 1.68 kcal/mol, comparable to the performance of conventional DFAs.

### Table 1. MAD, (in kcal/mol) for various benchmarks

<table>
<thead>
<tr>
<th>Methods</th>
<th>HOF (223)</th>
<th>IP (38)</th>
<th>EA (25)</th>
<th>PA (8)</th>
<th>BDE (92)</th>
<th>NHTBH (38)</th>
<th>HTBH (38)</th>
<th>NICE (31)</th>
<th>All (493)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVWN (LDA)</td>
<td>130.88</td>
<td>15.14</td>
<td>17.30</td>
<td>5.68</td>
<td>18.14</td>
<td>12.53</td>
<td>17.95</td>
<td>3.29</td>
<td>67.28</td>
</tr>
<tr>
<td>BLYP</td>
<td>10.16</td>
<td>6.02</td>
<td>2.47</td>
<td>1.75</td>
<td>7.00</td>
<td>8.29</td>
<td>7.68</td>
<td>1.49</td>
<td>7.84</td>
</tr>
<tr>
<td>PBE</td>
<td>20.71</td>
<td>5.13</td>
<td>2.40</td>
<td>1.56</td>
<td>3.91</td>
<td>8.57</td>
<td>9.48</td>
<td>1.17</td>
<td>12.10</td>
</tr>
<tr>
<td>TPSS</td>
<td>5.01</td>
<td>5.36</td>
<td>2.41</td>
<td>1.66</td>
<td>5.88</td>
<td>9.04</td>
<td>8.26</td>
<td>1.14</td>
<td>5.33</td>
</tr>
<tr>
<td>B3LYP</td>
<td>6.08</td>
<td>3.74</td>
<td>2.45</td>
<td>1.40</td>
<td>5.51</td>
<td>4.84</td>
<td>4.26</td>
<td>0.98</td>
<td>4.98</td>
</tr>
<tr>
<td>B3LYP D3</td>
<td>4.15</td>
<td>3.77</td>
<td>2.47</td>
<td>1.18</td>
<td>4.29</td>
<td>5.17</td>
<td>4.97</td>
<td>0.64</td>
<td>3.93</td>
</tr>
<tr>
<td>PBE0</td>
<td>5.64</td>
<td>3.84</td>
<td>2.97</td>
<td>1.25</td>
<td>3.67</td>
<td>3.56</td>
<td>4.38</td>
<td>0.71</td>
<td>4.36</td>
</tr>
<tr>
<td>M06 2X</td>
<td>2.26</td>
<td>2.72</td>
<td>2.37</td>
<td>1.94</td>
<td>1.40</td>
<td>1.26</td>
<td>1.25</td>
<td>0.28</td>
<td>1.86</td>
</tr>
<tr>
<td>XYG3†</td>
<td>1.81</td>
<td>1.31</td>
<td>1.84</td>
<td>1.61</td>
<td>1.57</td>
<td>1.29</td>
<td>0.75</td>
<td>0.32</td>
<td>1.51</td>
</tr>
<tr>
<td>XYGJ OS</td>
<td>1.65</td>
<td>1.23</td>
<td>1.97</td>
<td>1.68</td>
<td>0.71</td>
<td>1.18</td>
<td>0.88</td>
<td>0.35</td>
<td>1.28</td>
</tr>
<tr>
<td>MC3BB</td>
<td>3.28</td>
<td>2.78</td>
<td>4.01</td>
<td>1.09</td>
<td>2.43</td>
<td>1.44</td>
<td>0.80</td>
<td>0.58</td>
<td>2.58</td>
</tr>
<tr>
<td>B2PLYP</td>
<td>2.73</td>
<td>2.48</td>
<td>2.15</td>
<td>1.52</td>
<td>2.95</td>
<td>2.23</td>
<td>1.73</td>
<td>0.55</td>
<td>2.45</td>
</tr>
<tr>
<td>B2PLYP D†</td>
<td>1.67</td>
<td>2.48</td>
<td>2.15</td>
<td>1.34</td>
<td>2.27</td>
<td>2.47</td>
<td>2.11</td>
<td>0.45</td>
<td>1.88</td>
</tr>
<tr>
<td>oB97X2 (LP)‡</td>
<td>1.52</td>
<td>1.73</td>
<td>1.56</td>
<td>1.09</td>
<td>1.62</td>
<td>1.67</td>
<td>0.74</td>
<td>0.47</td>
<td>1.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Time*</th>
</tr>
</thead>
<tbody>
<tr>
<td>C100H202</td>
<td>C100H100</td>
</tr>
<tr>
<td>HOF</td>
<td>213.23</td>
</tr>
<tr>
<td>MP2</td>
<td>10.63</td>
</tr>
<tr>
<td>G2</td>
<td>1.89</td>
</tr>
<tr>
<td>G3</td>
<td>1.06</td>
</tr>
</tbody>
</table>

HOF is heat of formation (1, 2); IP is ionization potential (2); EA is electron affinity (2); PA is proton affinity (2); BDE is bond dissociation energy (40); NHTBH and HTBH are barrier heights for reactions (42, 43); NICE is the binding in molecular clusters (42, 43). The basis sets used for final energetics are G3Large unless otherwise stated. See SI Text for computational details.

*The time is single CPU hours on 2.5 GHz Xeon for linear alkane C100H202 and the C100H100 diamond structure. The basis sets used are cc pVDZ.
†Taken from ref. 12 with 6 311+G(3df,2p).
‡HOF are taken from ref. 22 with very large basis set of QCZV3P. BDEs are calculated using the corresponding HOE.
Bond Dissociation Enthalpy (BDE). BDE is arguably the most important property for chemistry, carrying additional information other than HOFs (13). XYGJ OS leads to MAD = 0.71 kcal/mol for the 92 reactions in the BDE92/07 set (40)! This accuracy surpasses the MAD of the most accurate ab initio methods, G2 (1.80) and G3 (1.09 kcal/mol). Indeed XYGJ OS leads to half the MAD of XYG3 (1.57 kcal/mol), providing strong support for the OS ansatz and the local approximation.

Reaction Barrier Height (RBH). We use the NHTBH and HTBH datasets of Zhao and Truhlar as testing sets (42, 43) of RBHs. NHTBH contains six heavy atom transfer reactions, eight unimolecular and association reactions, while HTBH comprises solely 19 hydrogen transfer reactions. XYGJ OS leads to MAD = 1.03 kcal/mol for the two sets together, a significant improvement over B3LYP (4.55 kcal/mol) and MP2 (4.67 kcal/mol). Note that the dispersion corrected methods [e.g., B3LYP D3 (46)] may deteriorate the performance for RBH calculations.

We test various methods in describing the whole H + CH4 → H2 + CH3 reaction path using the coupled cluster method with single and double as well as perturbative triple excitations (CCSD(T)/6-31+G(3df,2p) data (43) as the reference. The results are depicted in Fig. 1A. XYGJ OS results are nearly identical to the XYG3 and CCSD(T) results before the barrier. But XYGJ OS overestimates the reaction endothermicity by 1.21 kcal/mol.

Nonbonded Interaction (NBI). We use the NCIE dataset of Zhao and Truhlar (42, 43) to test the XYGJ OS performance for the description of NBIs. The NCIE set includes six hydrogen bond complexes, seven charge transfer complexes, six dipole interaction complexes, seven weak interaction complexes, and five π stacking complexes. XYGJ OS (MAD = 0.35 kcal/mol) does quite well for NBI, including the London dispersion dominant systems in the NCIE dataset. For the S22 set (47), which contains larger molecules (e.g., uracil dimer, adenine thymine complexes) that are more biologically related, XYGJ OS (MAD = 0.36 kcal/mol) leads to similar accuracy as the dispersion corrected methods (22, 46). On average, B3LYP D3 improves B3LYP by 0.34 kcal/mol for the NCIE set. While significant improvements occur for the weak interaction complexes, and the π stacking complexes, B3LYP D3 significantly overbinds the hydrogen bond complexes, and the charge transfer complexes (See SI Text for more discussion).

We test various methods in describing the intermolecular potentials of the CH4 C6H6 complex calculated by various methods. The CCSD(T) results at the complete basis set (CBS) limit are used as reference (44). As shown in Fig. 1B, XYGJ OS data are nearly on top of those of XYG3.

Scaling. Equally important is that XYG3 and other DHDFs (15 19) scale formally with the fifth power of the size, while XYGJ OS scales formally with the fourth power, and for the local implementation it scales formally with the third power. In contrast, the most accurate ab initio methods, CCSD(T), and G3, scale as the seventh power. Indeed, Table 1 shows that the total computational time for c100 chains and diamondoids with local XYGJ OS is just 3 to 4 times that for B3LYP (a part of XYGJ OS), with a scaling of N^3 for n alkanes (See SI Text for more details).

Summary

We have developed and validated here a unique doubly hybrid density functional, XYGJ OS, using Walter Kohn’s insight about “nearsightedness” of electron correlation by including explicitly only the correlation between electrons of OS and then only the parts that are in the same region of space. We show that XYGJ OS achieves nearly chemical accuracy (1.28 kcal/mol) with computational costs scaling as N^3. This unique combination of high accuracy and speed leads to a practical level of calculation while attaining chemical accuracy for large molecular systems.

ACKNOWLEDGMENTS. X.X. acknowledges the support of NSF of China (91027044, 21133004, 20923004), and the Ministry of Science and Technology of China (2007CB815206, 2011CB808505). Y.J. acknowledges the support of National Research Foundation (NRF) of Korea (2010 0023018, 2010 0029034, and 2010 0029728) by the Ministry of Education, Science and Technology (MEST) of Korea. W.A.G. acknowledges the support of USA National Science Foundation (NSF) (ECF 0609128, CTS 0668899), the Center for Catalytic Hydrocarbon Functionalization (Department of Energy, Basic Energy Sciences Award DE SC0001298), and Office of Naval Research Defense Advanced Research Projects Agency [ONR DARPA (PROM N00014 06 1 0938 and N00014 05 1 0778)]. Y.J. and W.A.G. have also been supported by the World Class University (WCU) (NRF R 31 2008 000 10055 0) program funded by the Ministry of Education, Science and Technology.


We tested various methods in describing the whole \( H + CH_4 \rightarrow H_2 + CH_3 \) reaction path using CCSD(T)/6-311+ +G(3df,2pd) data (32) as the reference. The results are depicted in Fig. S1. XYGJ OS results are nearly identical to the XYG3 and CCSD(T) results before the barrier. But XYGJ OS overestimates the reaction endothermicty by 1.21 kcal/mol. Note that RBH are not included in the training set of XYG3 and XYGJ OS.

We tested various methods in describing the intermolecular potentials of the \( CH_2 \cdot CH_2 \) complex calculated by various methods as shown in Fig. S2. The CCSD(T) results at the complete basis set (CBS) limit are used as reference (33). An unpruned (250,590) grid is used in calculations to avoid spurious oscillations on potential energy curves for dispersion bound complexes with MC3BB and the M06 family of functionals (34). As shown in Fig. S2, XYGJ OS data are nearly on top of those of XYG3. Note that nonbonded interaction is not included in the training set of XYG3 and XYGJ OS. However we do not include BSSE corrections with XYG3 and XYGJ OS in comparing to the G3 and M06 data bases.

The total CPU timings are compared for XYG3 RI, XYGJ OS, and local XYGJ OS (see Fig. S3). We used n alkanes as examples. The calculations were performed with cc-pVDZ basis sets with the same basis set as the auxiliary basis. The numbers of basis functions are 250 for C10, 1,210 for C50, and up to 2,410 for C100. For local XYGJ OS, a cutoff criterion of \( 5 \times 10^{-6} \) for local screening and the \( \omega.2 \) a.u. for the attenuated Coulomb fitting metric (Eq. S17 below) were used (35). As compared to XYG3 OS, the local algorithm errors for total energies of n alkanes are less than 0.01 eV (0.25 kcal/mol) up to C100. For the atomization energies of G2 set, we obtained the same statistical results (less than 0.001 eV of difference) for both XYG3 OS and its local version.

The \( oB97X\ 2\ (LP) \) results are taken from ref. 21. The \( oB97X\ 2\ (LP) \) yields a very high accuracy comparable to XYGJ OS for HOF, IP, EA, PA, NTHBH, and HTBH, where the key difference in terms of accuracy is that, in \( oB97X\ 2\ (LP) \) all the properties were part of the training set used in the fitting process while in XYGJ OS only the HOF was used in a training set. XYGJ OS contains four fitting parameters while parameters in Becke88 exchange and LYP correlation functionals are fixed at their original values. The \( oB97X\ 2\ (LP) \) contains 16 fitting parameters. Another difference is that, the parameters in the \( oB97X\ 2\ (LP) \) were fitted assuming the large Pople basis 6 311++G(3df,3pd). There is also a computational difference between \( oB97X\ 2\ (N^2) \) and XYGJ OS (N5 for the default local implementation and N8 for canonical).

From Table S1, it is clear that HOF calculations are very basis set dependent, where the calculations with the basis sets used for DHDF functional parameterization lead to the best results. Changing basis sets from 6 311+G(3df,2p) to G3Large increases XYG3 MAD for HOF by 1.63 kcal/mol. Changing basis sets from CQZV3P (quadruple basis sets plus three set of polarization and core polarization functions) to G3Large increases B2PLYP MAD by 4.85 kcal/mol and B2PLYP D MAD by 4.22 kcal/mol. Changing basis sets from 6 311++G(3df,3pd) to G3Large in creases \( oB97X\ 2\ (LP) \) MAD for HOF by 3.82 kcal/mol. For other properties, basis set dependence associated with DHDFs is relatively mild.

Dispersion correction methods could increase the errors for RBH and hydrogen bond interactions, but they should significantly improve the performance for systems (clusters) dominated by weak (vdw) interactions. Thus for RBH (Truhlar datasets
XYG3 vs. XYGJ-OS. In this section of SI, we summarize some basic physics associated with XYG3 and XYGJ OS. The way of deviation of XYG3 and XYGJ OS presented here complements that in the main text.

The Holy Grail in Kohn Sham (KS) DFT is to find the exact, yet unknown, exchange correlation functional in the main text. From perturbation theory and relations based on uniform coordinates, the agreement is worse for the dispersion differences in the basis sets. For nonbonded interactions (Truhlar transfer complexes, the agreement is worse for the dispersion error. On the other hand, for the hydrogen bond and charge transfer complexes, the agreement is worse for the dispersion corrected methods.

All calculations reported in this article were performed using a development version of Q Chem (36).

\[ E_{xc}^\text{GL} = \sum \sum \frac{\left| \langle \phi_i | \Delta \hat{H}_\lambda | \phi_j \rangle \right|^2}{e_i - e_j} + \frac{1}{4} \sum \sum \frac{\left| \langle \phi_i | \Delta \hat{H}_\lambda | \phi_j \rangle - \langle \phi_j | \Delta \hat{H}_\lambda | \phi_i \rangle \right|^2}{e_i + e_j - e_a - e_b}. \]  

where subscripts \((i, j)\) denote occupied KS orbitals, and \((a, b)\) denote virtual orbitals. Here \(\nu\) is the local exchange operator defined by the exchange part of Eq. S8.

\[ \nu_{xc}(r) = \frac{\delta E_{xc}[\rho]}{\delta \rho}. \]

while \(\hat{k}_\lambda\) in Eq. S7 is the Fock like nonlocal exchange operator, leading to \(E_{xc}^{\text{HF}}\):

\[ E_{xc}^{\text{HF}} = -\frac{1}{2} \int \int d^3r d^3r' \frac{\rho_i(r) \rho_j(r') \rho_i(r')}{|r' - r|}. \]

The superscript “\(\text{HF}\)” emphasizes that it has the same form as in Hartree Fock (HF) theory, Eq. S9 is exact if the KS orbitals give the true density. Inserting Eq. S6 into Eq. S3 leads to:

\[ E_{xc}[\rho] \approx E_{xc}^{\text{HF}}[\{\phi_i(\rho)\}] + E_{xc}^{\text{GL2}}[\{\phi_i(\rho)\}]. \]

If \(U_{xc,\lambda}\) depends linearly on \(\lambda\), the higher order term \(O(\lambda^2)\) in Eq. S6 is zero, such that the higher order term \(O(\lambda^3)\) in [S10] vanishes; i.e., [S10] provides an exact expression in this condition. The \(\{\phi_i(\rho)\}\) indicates an orbital dependent functional. For practical uses, one needs an \(E_{xc}^{\text{DFA}}[\rho]\) that generates good KS orbitals (17, 44, 45, 46).

To improve the linear approximation of \(U_{xc,\lambda}\), we define a doubly hybrid density functional (DHDF) that combines [S10] and [S11]:

\[ E_{xc}^{\text{DHDF}}[\rho] = c_1 E_{xc}^{\text{LDA}} + c_2 \Delta E_{xc}^{\text{GGA}} + c_3 E_{xc}^{\text{HF}} + c_4 E_{xc}^{\text{LDA}} + c_5 \Delta E_{xc}^{\text{GGA}} + c_6 E_{xc}^{\text{GL2}}. \]

Here \(\{E_{xc}^{\text{LDA}}, E_{xc}^{\text{GGA}}\}\) are the exchange and correlation components within the local density approximation (LDA), and \(\{\Delta E_{xc}^{\text{GGA}}, \Delta E_{xc}^{\text{GGA}}\}\) are the corresponding correction terms to LDA within the generalized gradient approximation (GGA). The meta GGA functionals that include kinetic energy density or the Laplacian of density can also be used in place of GGAs. [S10] is more appropriate near \(\lambda = 0\), while [S11] is more appropriate near \(\lambda = 1\). Therefore Eq. S11 combines both to embrace local and nonlocal parts of both exchange and correlation contributions, which we expect to provide a good form for general applications. MC3BB (18), B2PLYP (19) and \(\omega B97X 2\) (21) are derived and constructed differently, are examples of other doubly hybrid functionals (18 21, 47 49). Neglecting the nonlocal correlation effects, Eq. S11 leads to such conventional hybrid functionals as Becke’s half and half (50) and B3LYP (12).

Based on Eq. S11, we have developed a DHDF, namely XYG3 (17), which is remarkably accurate for a wide range of systems and important chemical properties (17, 44, 45, 46). Nevertheless, including the PT2 term in DHDFs (17 21) leads to a formal scaling as \(N^3\), as opposed to a formal scaling of \(N^4\) as in B3LYP. This unfavorable scaling raises an issue for the practicality to apply XYG3 to large systems.

Here we propose a new opposite spin (OS) ansatz for DHDF, namely XYGJ OS, that yields a balanced description of nonlocal correlation effects while considerably reducing computational time. Our OS ansatz is motivated by the observation that the most important electron correlation effects involves correlation of the
OS electrons in the same orbital. The OS ansatz leads to \( N^4 \) scaling [using auxiliary basis expansions and Laplace quadrature approximations (31, 35)]. In XYGJ OS, the same spin (SS) correlation effects are included within the standard DFA. A local version of XYGJ OS is further developed that introduces the local approximation for the OS electron correlation. Local XYGJ OS scales as \( N^4 \) by recognizing the "nearsightedness of electron correlation" as emphasized by Kohn (51), while retaining the accuracy of the original XYGJ OS (described below).

Thus our proposed functional form (XYGJ OS) is

\[
E_{\text{XYGJ OS}}[\rho] = e_h E_h + (1 - e_s) E_s + (e_{\text{LYP}} E_{\text{LYP}} + e_{\text{LYP}} E_{\text{LYP}}) + e_{\text{PT2}} e_{\text{res}}^\text{PT2}.
\]

In Eq. S12 we normalize the HF exchange and Slater exchange, while eliminating the \( \Delta E_{\text{GGA}} \) contribution. The superscript "PT2" indicates that only the MP2 like perturbation part in Eq. S7 is evaluated as also done in the B2PYLP (19) and other DHDFs (17, 20, 21). The correlation part consists of \( E_{\text{LYP}}^\text{LYP} \), and \( E_{\text{LYP}}^\text{LYP} \), where both terms include the SS and OS effects while the second and third terms include only OS components. Our concept is that the combination of VWN, LYP, and PT2 OS yields a balanced description of both local and nonlocal spin dependent perturbation terms. To determine the optimal four parameters in Eq. S12, we use the experimental HOF data for the G3/99 set of 223 molecules (24 26) as the training set, leading to \( e_h, e_{\text{LYP}} E_{\text{LYP}}^\text{LYP} = 0.7731, 0.2309, 0.2754, 0.4364 \).

We would like to emphasize that we use optimized B3LYP orbital to generate the density and to calculate each term in Eq. S12 (17, 44, 45, 46). Other DHDFs have either used HF or integrals for PT2 evaluation (18), or the density and orbitals from a truncated functional with incomplete correlation (19 21, 47 49).

**Summary of local OS PT2 approximations.** In this section of SI, we summarize the approximations associated with the local OS PT2 ansatz, and a brief introduction of how the locality is exploited in the present algorithm. That is, we here summarize ref. 31 for completeness.

In order to be able to use locality of electron correlation, one has to transform the canonical orbital based PT2 expression to the local orbital based expression. To this end, we eliminate the energy denominator via the Laplace transform 1/\( \lambda = \int \frac{d\tilde{\lambda}}{\tilde{\lambda}} e^{-\tilde{\lambda}u} \) (56):

\[
E_{\text{OS PT2}} = - \sum_i \sum_a \sum_b \frac{\langle i a | j b \rangle^2}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} = - \int_0^\infty dt \sum_i \sum_a \sum_b e^{-t(\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j)} \langle i a | j b \rangle^2.
\]  

where \( \epsilon_i, \epsilon_j, \epsilon_a, \epsilon_b \) are the orbital energies of the occupied levels \( i, j \) and the virtual levels \( a, b \). The two electron MO repulsion integrals are given by:

\[
\langle i a | j b \rangle = \langle i a | j b \rangle = \int \frac{t_1}{t_2} \phi_i(r_1) \phi_j(r_1) \phi_i(r_2) \phi_j(r_2).
\]

Numerical integration over \( t \) is performed by introducing a discrete quadrature (over \( N \) points, which is typically 6 8):

\[
E_{\text{OS PT2}} = - \sum_i \sum_a \sum_b \langle i a | j b \rangle^2.
\]

\[
E_{\text{OS PT2}} = - \sum_i \sum_a \sum_b \langle i a | j b \rangle^2.
\]

In Eqs. S16, S17 and below, we use a notation for simplicity that \( \phi_i \) implies the transformed orbital at the quadrature point \( q \). We next introduce an auxiliary basis \( \{ \phi_Q \} \) expansion for the product density (Eq. S18) to reduce the number of expensive four center electron integrals into much more affordable two- and three-center integrals.

\[
\rho = \langle \phi_i \phi_j \rangle \approx \rho_{\text{integ}} = \sum_Q C_{\phi Q}^i C_{\phi Q}^j.
\]

This approximation is also known as Resolution of Identity (RI) or density fitting technique (53, 54). The auxiliary basis expansion coefficients, \( C \), are determined by minimizing the self-interaction of the difference between the fitted distribution and the actual distribution, \( \rho_{\text{diff}} = \rho_{\text{integ}} - \rho_{\text{actual}} \) defined as,

\[
\rho_{\text{diff}}(\rho_{\text{integ}}) = \int \rho_{\text{integ}}(\rho_{\text{actual}}) = \int rdr_1 \int rdr_2 \rho_{\text{integ}}(r_1, r_2) \rho_{\text{diff}}(r_2).
\]

where \( g(r_1, r_2) = 1/|r_1 - r_2| \). Differentiation of Eq. S19 with respect to \( C \) and setting it to zero leads to the following solution for \( C \), which involves only the two index two electron and three index two electron integrals:

\[
C_{\phi Q}^i = \sum_p \langle \tilde{q} \tilde{p} | P \rangle P^{-1}(Q).
\]

Inserting Eqs. S18 and S20 into Eq. S14 results in the following integral approximation:

\[
\langle i a | j b \rangle \approx \sum_Q B_{\phi Q}^i B_{\phi Q}^j,
\]

where \( B \) is the matrix defined as:

\[
B_{\phi Q}^i = \sum_p \langle \tilde{q} | P \rangle P^{-1}(Q)^{-1/2}.
\]

Substituting Eq. S21 into Eq. S15 finally leads to the following working expression for the OS PT2 correlation energy:

\[
E_{\text{OS PT2}} = \sum_{q} w_q \sum_Q X_{\phi Q} \tilde{X}_{\phi Q}.
\]

\[
X_{\phi Q} = \sum_a B_{\phi Q}^a B_{\phi Q}^a.
\]

\[
\tilde{X}_{\phi Q} = \sum_b B_{\phi Q}^b B_{\phi Q}^b.
\]
Formation of $X$ from the $B$ matrices, Eqs. S24 and S25, is the most expensive fourth order step in OS PT2.

While Eq. S15 is specific to the canonical orbital basis, expressions in a generalized orthogonal basis can be given as follows:

$$E_{OS\ PT2} = - \sum_{q}^{n} w_{q} \sum_{IAJ}^{a} \langle \tilde{J} A | \tilde{J} B \rangle q_{j}^{2}$$  \[S26\]

$$\tilde{\phi}_{I} = \sum_{K}^{n} \phi_{K} \langle \phi_{K} | \epsilon_{q} / 2 \rangle | \phi_{I} \rangle = \sum_{K}^{n} \phi_{K} W_{KI}$$  \[S27\]

$$\tilde{\phi}_{A} = \sum_{C}^{n} \phi_{C} \langle \phi_{C} | \epsilon_{q} / 2 \rangle | \phi_{A} \rangle = \sum_{C}^{n} \phi_{C} W_{CA}$$  \[S28\]

where indices $I, J, K, \ldots$ and $A, B, C, \ldots$ refer to localized functions that span occupied and virtual spaces. In Eqs. S27 and S28, we again assume that $\tilde{\phi}_{I}$ implies the transformed orbital at the quadrature point $q$. We chose to use orthogonal Boy_ns functions for the localized occupied orbitals, and the generalized orthogonal projected atomic orbitals (AO)’s for the localized virtual orbitals, defined as $| \phi \rangle = Q \langle \psi^{-1/2} | \mu \rangle$, where $Q$ is the virtual projection operator, and $\mu_1, \ldots$ are the AO indices. We could have used Boys virtual orbitals instead of the projected AOs since the projected AOs partly destroy the locality of AOs by applying $s^{-1/2}$. But Boys localization using the Jacobi sweeps becomes computationally very demanding due to poor convergence for large systems. Therefore we used the projected AOs for the localization of virtual orbitals.

We next introduce the auxiliary basis expansion, as before, for the product distribution $\tilde{\phi}_{I} \tilde{\phi}_{A}$, and obtain the same solution for $C$, except that here we choose to employ the short range two electron operator $g(r_{1}, r_{2}) = \text{erfc}(c(|r_{1} - r_{2}|)/|r_{1} - r_{2}|)$, which is denoted with a subscript $\omega$, to remove the long range artifact of Coulomb metric (S5).

$$C_{Q}^{\omega} = \sum_{P} \langle \tilde{I} A | P \rangle \langle P | Q \rangle_{\omega}^{-1}$$  \[S29\]

Finally, we rewrite Eq. S14 as the following working expression for the local OS PT2 algorithm:

$$E_{OS\ PT2} = - \sum_{q}^{n} w_{q} \sum_{PQRS}^{a} \langle \phi_{P} | \phi_{Q} \rangle \tilde{Y}_{QS} (R | S)$$  \[S30\]

$$Y_{PR} = \sum_{IA}^{a} C_{I}^{\omega A} C_{R}^{\omega}$$  \[S31\]

$$\tilde{Y}_{QS} = \sum_{JB}^{b} \beta B_{Q} C_{S}^{\omega}$$  \[S32\]

To exploit the locality and sparsities of the various matrices, we use a scheme called “natural blocking,” where the entire rows or columns of large matrices can be removed for processing due to the physical nature (structure) of given matrices.

Eq. S31 is the most time consuming step in a straightforward 4th order scaling implementation of OS PT2 method (like Eqs. S24 and S25). But because the occupied and virtual functions ($I, A$) are localized and the auxiliary expansion functions $Q$ are also local, the fitting coefficients $C_{Q}^{\omega}$ (where the quadrature point $q$ is implied as in Eq. S27 and S28) must also be local. In other words, for a given occupied orbital $I$, only those virtual and auxiliary functions (matrix elements) that are local to $I$ will be nonzero, and so $Y_{PQ} = \sum_{B} \beta_{B} C_{Q}^{\omega} = (\beta_{B} Q_{sig})/(B_{sig} Q_{sig})$, where $B_{sig}, Q_{sig}$ mean that $P, B,$ and $Q$ are local to $I$. Furthermore, $C_{Q}^{\omega}$ is arranged optimally for this operation, which is performed as an efficient dense matrix multiply only over the significant virtual ($B_{sig}$) and auxiliary ($P_{sig}$) indices, using the natural blocking scheme. In deleting the rows and columns, we have used a numerical cutoff of $5 \times 10^{-6}$ and the error associated with it is insignificant as shown below.

In another step, the AO to occupied MO transformation ($\lambda_{P} | \psi_{\mu} \rangle = \sum_{\nu} C_{\mu \nu} | \psi_{\nu} \rangle$) is $O(M^{3})$ work. For a large enough system, it asymptotically approaches $O(M^{2})$ because the magnitude of the $\mu \nu$ function product decays as a Gaussian with the interfunction distance. Because the integral operator is short ranged, as denoted by a subscript $\omega$, only those integrals that involve basis functions ($\mu, \nu, \nu$) that are close each other in real space will have nonzero values. A significant number of arithmetic operations can be saved by using this integral sparsity, namely inserting an IF statement in which the first MO transformation is skipped if the integral value is zero against some numerical threshold. We have used $10^{-6}$ as an integral threshold criterion, which is sufficiently tight that it adds no additional error numerically as shown below. This utilization of sparsity makes the cost of occupied index AO to MO transformation approximately $O(M^{2})$, in stead of $O(M^{3})$.

Other steps involved in the transformations and evaluations of various quantities that lead to Eqs. S31 and S32 have also been accelerated by using the similar blocking scheme and integral screening (see ref. 35 for more details). The resulting algorithm is formally cubic scaling (localization, local transformation, and inspection of the matrices), but due to a relatively small prefactor of the latter cubic steps it shows a quadratic scaling for favorable cases before these cubic steps become dominant.

As seen in Fig. 3, linear alkane chains show a quadratic scaling and a significant speedup as compared to the calculation with out using local algorithm. On the other hand, the fully three dimensional (3D) dense systems like the diamondoid do not show a significant speedup up to C100 due to the small length scale of this system even with 100 carbon atoms, $\approx 10$ Å. In contrast, the length scale of linear alkane with 100 carbon atoms is about $\approx 125$ Å, thus showing a substantial speedup. This relatively modest speedup means that, for fully 3D structures like the diamondoid structure, we have simply not yet reached the quadratic or cubic regime where sparsity can make a significant difference. There will be a significant speedup using the local algorithm even for dense 3D systems, however, if one considers large enough systems. The memory and disk requirement for the local PT2 calculation are also quadratically scaling, moderate to be applicable for large systems.


Fig. S1. Accuracy of various theoretical methods for calculating the potential energy curve of the H + CH₄ → H₂ + CH₃ reaction. Reaction coordinate is defined as [R(CH) R(HH)] (in Å). CCSD(T)/6 311++G(3df,2pd) results (32) are used as reference. Note that RBH are not included in the training set of XYG3 and XYGJ OS.
Fig. S2. The intermolecular potentials for the CH₄ C₆H₆ complexes from various methods. R is defined as carbon of CH₄ to the ring center of C₆H₆ (in Å). Data in red are the coupled cluster method with single and double as well as perturbative triple excitations (CCSD(T)) at the complete basis set limit from ref. 33. XYGJ OS data are nearly on top of those of XYG3. Note that nonbonded interaction is not included in the training set of XYG3 and XYGJ OS. But BSSE corrections are NOT included in the calculations. XYG3 and XYGJ OS are making use of BSSE to have this good agreement.
Fig. S3. Scaling of the XYG3, XYGJ OS, and its local version method as measured by the log of total CPU time vs the alkane chain length.
Table S1. Mean absolute deviations (MADs, in kcal/mol) for various benchmark sets with the G3Large basis set. BSSE correction is not included unless otherwise stated.

<table>
<thead>
<tr>
<th>Methods</th>
<th>HOF</th>
<th>IP</th>
<th>EA</th>
<th>PA</th>
<th>BDE</th>
<th>NHTBH</th>
<th>HTBH</th>
<th>NCIE</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(223)</td>
<td>(38)</td>
<td>(25)</td>
<td>(8)</td>
<td>(92)</td>
<td>(38)</td>
<td>(38)</td>
<td>(31)</td>
<td>(493)</td>
</tr>
<tr>
<td>1st rung</td>
<td>SVWN (SPL)</td>
<td>130.88</td>
<td>15.14</td>
<td>17.30</td>
<td>5.68</td>
<td>18.14</td>
<td>12.53</td>
<td>17.95</td>
<td>3.29</td>
</tr>
<tr>
<td>2nd</td>
<td>BLYP</td>
<td>10.16</td>
<td>6.02</td>
<td>2.47</td>
<td>1.75</td>
<td>7.00</td>
<td>8.29</td>
<td>7.68</td>
<td>1.49</td>
</tr>
<tr>
<td>Rung</td>
<td>PW91</td>
<td>22.04</td>
<td>5.19</td>
<td>3.19</td>
<td>1.53</td>
<td>3.75</td>
<td>8.95</td>
<td>9.76</td>
<td>1.31</td>
</tr>
<tr>
<td>3rd</td>
<td>M06 L</td>
<td>5.20</td>
<td>5.31</td>
<td>3.56</td>
<td>1.60</td>
<td>4.17</td>
<td>3.81</td>
<td>4.33</td>
<td>0.59</td>
</tr>
<tr>
<td>Rung</td>
<td>TPSS</td>
<td>5.01</td>
<td>5.36</td>
<td>2.41</td>
<td>1.66</td>
<td>5.88</td>
<td>9.04</td>
<td>8.26</td>
<td>1.14</td>
</tr>
<tr>
<td>4th</td>
<td>B3LYP D</td>
<td>6.08</td>
<td>3.74</td>
<td>2.45</td>
<td>1.40</td>
<td>5.51</td>
<td>4.84</td>
<td>4.26</td>
<td>0.98</td>
</tr>
<tr>
<td>Rung</td>
<td>B3LYP D3</td>
<td>4.15</td>
<td>3.77</td>
<td>2.47</td>
<td>1.18</td>
<td>4.22</td>
<td>3.87</td>
<td>3.48</td>
<td>0.71</td>
</tr>
<tr>
<td>5th</td>
<td>M06 2X</td>
<td>2.26</td>
<td>2.72</td>
<td>2.37</td>
<td>1.94</td>
<td>4.00</td>
<td>3.12</td>
<td>3.04</td>
<td>0.69</td>
</tr>
<tr>
<td>Rung</td>
<td>M06</td>
<td>3.37</td>
<td>3.63</td>
<td>2.04</td>
<td>1.78</td>
<td>5.70</td>
<td>4.21</td>
<td>3.16</td>
<td>0.93</td>
</tr>
</tbody>
</table>

*BSSE corrections are included for the NCIE set.

Table S2. Mean absolute deviation for RBH of Truhlar’s NHTBH38 and HTBH38 sets (basis sets: 6-311+G(3df,2p))

<table>
<thead>
<tr>
<th>RBH</th>
<th>UM10</th>
<th>NS16</th>
<th>HAT12</th>
<th>HT38</th>
<th>NHT38</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3LYP</td>
<td>2.02</td>
<td>3.38</td>
<td>8.50</td>
<td>4.31</td>
<td>4.64</td>
<td>4.47</td>
</tr>
<tr>
<td>B3LYP D</td>
<td>2.50</td>
<td>4.03</td>
<td>8.96</td>
<td>5.06</td>
<td>5.18</td>
<td>5.12</td>
</tr>
<tr>
<td>B3LYP D3</td>
<td>2.37</td>
<td>4.01</td>
<td>8.93</td>
<td>4.83</td>
<td>5.13</td>
<td>4.98</td>
</tr>
</tbody>
</table>

Here HT38 refers to the forward and reverse barrier heights for 19 hydrogen transfer reactions; HAT12 refers to the forward and reverse barrier heights for six heavy atom transfer reactions, NS16 refers to the forward and reverse barrier heights for eight nucleophilic substitution reactions and UM10 refers to the forward and reverse barrier heights for five association and unimolecular reactions.

Table S3. Mean absolute deviation for Truhlar’s NCIE31 set of nonbonded interactions (basis sets: 6-311+G(3df,2p))

<table>
<thead>
<tr>
<th>RBH</th>
<th>HB6</th>
<th>CT7</th>
<th>DI6</th>
<th>WI7</th>
<th>PP55</th>
<th>Total(31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3LYP</td>
<td>0.63</td>
<td>0.76</td>
<td>0.61</td>
<td>0.27</td>
<td>2.93</td>
<td>0.95</td>
</tr>
<tr>
<td>B3LYP D</td>
<td>0.86</td>
<td>1.49</td>
<td>0.60</td>
<td>0.14</td>
<td>0.34</td>
<td>0.71</td>
</tr>
<tr>
<td>B3LYP D3</td>
<td>0.70</td>
<td>1.60</td>
<td>0.78</td>
<td>0.11</td>
<td>0.21</td>
<td>0.70</td>
</tr>
</tbody>
</table>

NCIE31 consists of six hydrogen bond (HB) complexes, seven charge transfer (CT) complexes, six dipole interaction (DI) complexes, seven weak interaction (WI) complexes, and five π-π stacking (PPS) complexes. The WI and PPS are dominated by London dispersion.