Frontiers in electronic structure theory

C. David Sherrill

Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry and School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, USA

(Received 26 February 2010; accepted 2 March 2010; published online 18 March 2010)

Current and emerging research areas in electronic structure theory promise to greatly extend the scope and quality of quantum chemical computations. Two particularly challenging problems are the accurate description of electronic near-degeneracies (as occur in bond-breaking reactions, first-row transition elements, etc.) and the description of long-range dispersion interactions in density functional theory. Additionally, even with the emergence of reduced-scaling electronic structure methods and basis set extrapolation techniques, quantum chemical computations remain very time-consuming for large molecules or large basis sets. A variety of techniques, including density fitting and explicit correlation methods, are making rapid progress toward solving these challenges. © 2010 American Institute of Physics. [doi:10.1063/1.3369628]

The past 15 years have seen major advances in electronic structure theory. Coupled-cluster theory or multireference methods, when joined with basis set extrapolation techniques, allow very accurate computations of small molecules. Indeed, for the smallest molecules, numerical approximations and better algorithms. Together, these advances have made electronic structure theory an increasingly important part of modern scientific research. For example, chemists proposing reaction mechanisms are now generally expected to provide supporting evidence from quantum chemistry computations.

Despite tremendous recent progress, there remains ample room for innovation and improvement in electronic structure theory. In this spotlight article, I will discuss a few of the areas which promise significant advances in the near future. Because this is intended to be a survey of selected current and future research areas, and not a retrospective review article, the discussion of prior work is incomplete.

One of the continuing challenges of electronic structure theory is the tremendous computational cost required to accurately solve the electronic Schrödinger equation. Linear-scaling self-consistent-field (SCF) methods in the mid-1990s represented a major step forward, and local correlation methods have greatly reduced the computational costs of post-Hartree–Fock methods. Alternatively, there is also an increasing interest in methods which break up a large molecule into multiple pieces, perform computations on each piece separately, and then combine the results. Such approaches include divide-and-conquer density functional theory (DFT) methods, fragment molecular orbital (MO) approaches, and recent molecular cluster schemes explored in the context of coupled-cluster theory.

Numerical approximations. The reduced-scaling advances discussed above are indispensable in making routine computations possible for systems with hundreds of atoms or more. However, even though the computational scaling has been reduced dramatically, the prefactor in the computational cost remains high (particularly if one desires to use such methods for dynamical simulations, where a very large number of computations need to be performed). Thus, in recent years there has been a growing interest in numerical techniques which can speed up the computations.

Various methods are available to approximate the four-index electron repulsion integrals as products of three-index intermediates. These approaches include resolution of the identity (RI, which is also called density fitting or DF), pseudospectral, and Cholesky decomposition (CD) techniques. A general comparison of DF and CD methods has recently been published by Weigend, Kattanek, and Ahlrichs. These methods have been extremely successful and appear destined for adoption into every major electronic structure program package within the next few years. These techniques often reduce computational costs so much that they can make some impractical computations become practical, and recent work continues to push these techniques into new contexts.

In the RI/DF approach, the four-index integrals are approximated as

\[ (\mu \nu | \rho \sigma) \approx \sum_{PQ} (\mu \nu | P)(J^{-1})_{PQ}(Q | \rho \sigma). \] (1)

The two index quantity, \( (J^{-1})_{PQ} \), is the inverse of the Coulomb metric evaluated in an auxiliary basis set.
\[
[J]_{PQ} = \int P(\mathbf{r}_1) \frac{1}{r_{12}} Q(\mathbf{r}_2) d^3 r_1 d^3 r_2.
\]

The three-index quantity \((\mu \nu | P)\) serves to cast the product \((\mu \nu |\) onto the auxiliary basis via the Coulomb metric

\[
(\mu \nu | P) = \int \mu(\mathbf{r}_1) \nu(\mathbf{r}_1) \frac{1}{r_{12}} P(\mathbf{r}_2) d^3 r_1 d^3 r_2.
\]

In the RI approach, the auxiliary basis typically comprises atom-centered Gaussian functions. In the related pseudospectral approach, a real-space grid is used instead. In the CD approach, one uses a similar equation

\[
(\mu \nu | \rho \sigma) = \sum_{Q} t^{Q}_{\mu \nu} \rho^{Q}_{\sigma},
\]

where the three-index quantities \(t^{Q}_{\mu \nu}\) are now referred to as “Cholesky vectors” and they are constructed using a recursive algorithm.

Second-order Möller–Plesset perturbation theory (MP2) benefits greatly from techniques such as these. The evaluation of the MP2 correlation energy, scaling as \(O(o^2 v^2)\) (where \(o\) and \(v\) are the number of occupied and virtual molecular orbitals, respectively), is actually less expensive computationally than the transformation of the four-index integrals from the atomic orbital (AO) to the MO basis, which scales as \(O(N^2_{AO}N_{MO})\), where \(N_{AO}\) and \(N_{MO}\) are the number of AOs and MOs, respectively. In the RI-MP2 approach, the \(O(N^3)\) transformation step drops to \(O(N^4)\), with the most expensive step being \(O(oN_{AUX}N^2_{AO})\), where \(N_{AUX}\) is the size of the auxiliary basis. There is an associated increase in the computational cost of evaluating the MP2 correlation energy, which goes from \(O(o^2v^2)\) to \(O(o^2v^2N_{AUX})\). However, this increase is more than compensated for by the decrease in the transformation time, and in applications performed by our group, we often see speedups due to the RI approximation of a factor of 2–5. Moreover, using basis sets which have been optimized for RI-MP2,\(^{42}\) the errors introduced are negligible. For example, the MP2/aug-cc-pVTZ binding energy of the benzene dimer changes by 0.001 kcal mol\(^{-1}\) upon invoking the RI approximation with a triple-\(\xi\) auxiliary basis set.\(^{43}\)

DF is also very helpful in linear-scaling local MP2 methods\(^{18}\) and Laplace-based MP2 methods.\(^{44}\) DF has recently allowed a very significant increase in the size of systems which may be studied by symmetry-adapted perturbation theory (SAPT).\(^{45}\) DF has been applied\(^{46,47}\) to SAPT(DFT), which employs a DFT description of the intramonomer electron correlation;\(^{48–50}\) this approach has been used to study systems as large as the 2(H\(_2\))\(_2\)-C\(_{60}\) complex with a TZVPP basis set.\(^{51}\) DF has also been added to wave function-based SAPT, allowing routine energy component analysis of systems with 1500 basis functions or more\(^{52}\) (see Fig. 1). Even in more sophisticated electron correlation methods such as multireference configuration interaction, once local correlation approximations are employed, the four-index integral processing can remain a significant fraction of the computational cost, and CD techniques have been shown to be very useful.\(^{53}\)

For small to medium-sized molecules where linear-scaling techniques do not yet become efficient, DF can provide significant speedups for Hartree–Fock and DFT computations. The evaluation and repeated processing of \(O(N^3)\) two-electron integrals dominates the iterative Hartree–Fock procedure. By casting the product \((\mu \nu |\) onto the auxiliary basis, the number of required integral computations drops to \(O(N^2_{AO}N_{AUX})\). Additionally, the \(O(N^4)\) memory footprint of the three-index integrals is often small enough to allow for storage on main memory or disk in situations, where the \(O(N^4)\) requirement of the two-electron integrals is prohibitively large or inefficient. For sufficiently large systems, the computational scaling for conventional Hartree–Fock or DF-HF drops to \(O(N^3)\); however, DF still provides a significant speedup even for fairly large systems. Of course, DF can also be applied in the evaluation of the Coulomb matrix in DFT, and the exact exchange matrix in hybrid DFT.

The “dual-basis” methods are also very promising for reducing the computational prefactor for large molecules or large basis sets. Recent advances in reduced-scaling MP2 algorithms have been so successful that often the rate-determining step in such computations is the SCF to obtain the molecular orbitals. Because the basis set required to converge the SCF is typically smaller than the basis set required to converge the correlation energy, a dual-basis approach, which uses a smaller basis for the SCF and a larger one for the correlation energy computation, suggests itself.\(^{54–57}\) These methods project the small-basis density matrix into the larger basis. Then, one either rediagonalizes the occupied-occupied and virtual-virtual blocks of the Fock matrix separately,\(^{53,55}\) or else takes a single additional SCF iteration in the larger basis.\(^{56,57}\) A similar approach by Gill and co-workers,\(^{58}\) which adds some higher-order corrections, has also been reported recently. Head-Gordon and co-workers have published reduced versions of several standard basis sets for use in dual-basis SCF.\(^{57,59}\) Results to date indicate that dual-basis approximations yield reliable energies at significantly reduced computational cost.

Finally, wavelet analysis or multiresolution analysis is a well-established technique in mathematics whose application to quantum chemistry remains in its early stages.\(^{50–65}\) The
The basic idea of this approach is to use higher-resolution descriptions in local regions where the wave function or density changes rapidly, and lower-resolution descriptions elsewhere. Such approaches are promising for their potential to provide very compact representations of the target functions.

**Electronic near-degeneracies.** Electronic near-degeneracies, which occur for bond-breaking processes, diradicals, first-row transition metals, etc., have been a persistent problem for quantum chemistry. Most standard quantum chemistry models have difficulty describing such situations, in which more than one electron configuration contributes significantly to the wave function. Potential energy curves for bond-breaking reactions can be qualitatively incorrect for many of the standard quantum chemistry methods, including even CCSD(T). In principle, adding successively higher excitation levels to the wave function can solve these problems, but this is not computationally feasible except for the smallest molecules. Alternatively, one can employ “multireference” methods, which describe the zeroth-order wave function as a combination of leading electron configurations. Progress in multireference methods has been slow because they can become very costly computationally and because the theory can also become much more complex than for standard single-reference methods. However, at present there are perhaps more groups actively addressing this problem than at any time in the past, and the diverse array of solutions being investigated seems very promising for significant progress.

For systems with only a handful of atoms, multireference configuration interaction (MRCI) (or related methods corrected for size extensivity) are the methods of choice. With modern programs, it is possible to perform extensive MRCI computations with large, correlation-consistent basis sets to achieve definitive results that compare favorably with experiment, even for spectroscopic properties. For slightly larger molecules, complete-active-space second-order perturbation theory (CASPT2) has become the preferred approach, and it has been widely used for reliable estimates of intermediates and transition states along a reaction path. Unfortunately, standard MRCI and CASPT2 are too computationally expensive to apply to large molecules. One solution has been the implementation of parallel algorithms, which allow somewhat larger systems to be studied. Another has been a reduction in computational cost through pseudospectral and local correlation approximations. The most recent work by Lindh and co-workers on linear-scaling MRCI using local correlation and integral screening techniques is quite impressive. Similarly, although multireference coupled-cluster methods have been explored for quite a few years, recent work is beginning to make these high-accuracy computations feasible for systems with a dozen atoms or more.

Alternatively, a variety of rather different techniques to describe near-degeneracies have emerged in the past few years. One approach is to use coupled-cluster-type methods to replace the traditional linear expansion of leading configurations, as in the active-space coupled-cluster methods of Head-Gordon and co-workers. Recent work along these lines has introduced a model including limited quadruple excitations which provides near-CASSCF results for difficult cases but which scales only as $O(N^4)$.

The spin-flip method of Krylov and co-workers is a simple but effective way to describe the most important leading electron configurations as spin-flipped excitations from a high-spin, single-reference electron configuration. This technique has been applied in configuration interaction, coupled-cluster, and DFT (Ref. 93) contexts. The method of moments and completely renormalized coupled-cluster approaches of Piecuch and co-workers attempts to correct single-reference coupled-cluster methods for near-degeneracy effects by adding energy corrections obtained from projection onto determinants not included in the coupled-cluster procedure. This approach appears to significantly improve results for challenging cases such as bond-breaking reactions with only modest additional computational cost.

The density matrix renormalization group (DMRG) theory bypasses the typical linear expansion of Slater determinants to use an alternative set of iteratively constructed $N$-electron basis functions (which are eigenvectors of reduced density matrices). This approach has been used to mimic what would be the largest full configuration interaction computation performed. Moreover, DMRG was used to perform complete active space computations of unprecedented size in a study of the acenes through dodecagone (correlating the $\pi$ valence electrons); these computations surprisingly indicate that the longer acenes have polyradical ground electronic states. DMRG handles the strong nondynamical correlations and may be combined with techniques such as canonical transformation theory to include dynamical correlation.

Another approach is to avoid wave functions and solve directly for the two-electron reduced density matrix (2-RDM). This idea is very appealing, but the mathematical challenge is that not all 2-RDMs correspond to possible $N$-electron wave functions; thus, constraints must be placed on the possible 2-RDMs considered (the “$N$-representability conditions.”) Fortunately, the discovery of additional $N$-representability constraints is now allowing accurate computations of molecular energies directly from linear functionals of the 2-RDM.

Together, these new techniques are allowing reliable computations on the most troublesome of chemical systems, and extending our reach to larger molecules.

**Dispersion in DFT.** DFT (Ref. 111) has enjoyed widespread success over the past two decades. The B3LYP functional, in particular, has become a favorite tool of computational chemists for molecular problems because of its general reliability and its relatively inexpensive computational cost. However, in the past few years, theorists have begun seriously reassessing the strengths and weaknesses of DFT, with the goal of better understanding those cases when it is not reliable. One area which has gained substantial attention has been the failure of standard, Kohn–Sham DFT methods to describe long-range dispersion interactions. As electronic structure methods in general, and DFT methods in particular, are made more efficient for larger molecules, the chemistry of supramolecular (nonbonded) inter-
actions becomes an increasingly attractive target for application studies. However, it is inappropriate to study these systems with theoretical models which fail to incorporate a description of long-range dispersion.

Incorporation of dispersion effects into DFT is an active current frontier area of electronic structure theory. The simplest possible approach is the addition of empirical, pairwise atomic dispersion corrections of the form $-C_r/r^6$, which are used in force field methods. To avoid double-counting electron correlation effects at short range, these contributions can be damped for small internuclear distances. This general approach is most commonly referred to as DFT plus dispersion (DFT-D). Various tests suggest that DFT-D with triple-$\zeta$ basis sets provides results within 10%–30% of benchmark values for the binding energies of van der Waals (vdW) clusters. This is a tremendous improvement over either uncorrected density functionals (DFs) or Hartree–Fock theory (which lack long-range dispersion and often fail to predict any binding at all). It is also an improvement over MP2 theory, which tends to greatly overbind vdW clusters. At present, there are several live research questions regarding these approaches. Can their accuracy be further improved to the point where they are good enough for nearly all chemical applications (short of benchmark or spectroscopic-quality work)? How much improvement can be afforded by higher-order terms ($C_6$, $C_{10}$) or anisotropic terms? How many “atom types” (e.g., which differentiate $sp^2$ from $sp^3$ hybridization) are necessary for high accuracy?

One disadvantage to current DFT-D methods is that the dispersion correction is formulated to be independent of the chemical environment. While this appears to be a good approximation, one wonders whether this will create a “glass ceiling” to hinder further improvements of the method, and whether this will make it unreliable in unusual molecular environments. Fortunately, less empirical approaches for incorporating dispersion effects are also being actively pursued. Röthlisberger and co-workers have added effective atom-centered nonlocal potentials to describe dispersion (with the parameters determined by fitting to high-quality ab initio data). Similarly, Langreth, Lundqvist, and co-workers have proposed a vdW-DF, which includes nonlocal terms in the correlation energy functional. Becke and Johnson have introduced an “exchange dipole moment” model which obtains dispersion coefficients $C_6$, $C_8$, $C_{10}$ from the “dipole moments” which would be obtained from an electron and its associated exchange hole. These works, and additional tests, provide promising results, although the reliability of these approaches is not necessarily better than that of DFT-D at this stage. Another interesting approach is the combination of DFT with symmetry-adapted perturbation theory, called DFT-SAPT or SAPT(DFT). In these approaches, the dispersion term is obtained via the frequency-dependent density susceptibility functions of time-dependent DFT.

Alternatively, one may also include the long-range correlation contributions through “double hybrid” DFT, which mixes not only exact exchange, but also second-order perturbation theory corrections to the correlation energy. Grimme’s B2PLYP functional was the first double-hybrid.

Its performance for nonbonded interactions is not as good as one might hope, and so there is also an empirical dispersion corrected variant, B2PLYP-D. Other double hybrids have appeared, most recently a reparameterization of B3LYP to incorporate MP2 correlation by Zhang, Xu, and Goddard (denoted XYZ3), and a double-hybrid version of Chai and Head-Gordon’s oB97X functional. XYG3 was recently compared with B97-D for potential energy curves of several small vdW dimers. Both provided mean absolute errors of less than 0.2 kcal mol$^{-1}$ across the potential curves; however, one should also note that the computational cost of double-hybrid methods will be significantly greater because of the need to evaluate the MP2 correlation correction.

Of course, there have also been attempts to parameterize new functionals within the current standard framework of local or “semilocal” functionals, up through forms as complex as the hybrid, meta generalized gradient approximation functionals which depend on the local density, its gradient, and the kinetic energy density, and which mix in Hartree–Fock exchange (see, e.g., Refs. 149–151). Some of these functionals provide improved performance for noncovalent interactions, so long as the nonbonded contacts are at short or “medium” range (up to perhaps 4–5 Å). However, longer-range nonbonded contacts (which may become very numerous for larger molecules) would appear to require explicit incorporation of nonlocal terms.

Explicit correlation methods. One disappointing result from numerous systematic studies of convergence to the complete basis set limit is that double-$\zeta$ basis sets are often insufficient to provide molecular properties to the desired accuracy. Triple-$\zeta$ basis sets are much more reliable, and for some applications, quadruple-$\zeta$ or even larger basis sets may be required. Unfortunately, computational costs can grow dramatically with these increases in the basis set. For this reason, computations on hundreds of atoms or more have been restricted to double-$\zeta$ basis sets and the moderate accuracy they allow.

“Explicitly correlated” methods provide one approach to accelerate convergence toward the complete basis set limit. These methods modify the usual wave function expansion to include terms depending explicitly upon the distances between pairs of electrons ($r_{ij}$). Ultimately, such methods have their origin in the pioneering 1929 work of Hylleraas on the helium atom. The “linear R12” approach of Kutzelnigg and Klopper introduces into the wave function expansion terms that are linear in the interelectronic distance and prescribes “standard approximations” for the elimination of some integrals and the estimation of others using a resolution of the identity. This work was very influential in stimulating recent research in explicitly correlated methods and led to the MP2-R12 and CCSD-R12 approaches. Building upon these accomplishments, theorists have improved the efficiency of integral evaluation, allowed for the use of a different basis set for the RI versus the other components of the computation, and reformulated the RI to make it more stable numerically, among other achievements.

The next major advance was Ten-no’s 2004 demonstration that replacing linear R12 terms with more complex func-
tions of the interelectronic distance can yield even faster convergence to the complete basis set limit.\textsuperscript{165} These new approaches are called F12 methods,\textsuperscript{153,165,166} and they appear very promising for solving the problem of large basis set requirements for accurate molecular properties. F12 versions of perturbation theory and coupled-cluster theory have been recently introduced.\textsuperscript{167} On the basis of recent work, it appears possible to achieve results of approximately quadruple-\(\zeta\) quality (or better) with only double-\(\zeta\) basis sets when using F12 methods.\textsuperscript{167,168} The development of special basis sets specifically for R12/F12 methods will accelerate progress in this area.\textsuperscript{169–171} Explicitly correlated methods are more complex than standard electronic structure methods; this increases the computational prefactor compared with the analogous standard method, and also makes the method harder to derive and implement computationally. However, elegant recent work by Torheyden and Valeev\textsuperscript{168} has shown how to evaluate a “universal” second-order perturbative correction for basis set completeness using R12/F12 techniques and the 2-RDM, which may be obtained from any standard electronic structure method.

Other frontiers. This article has discussed only a few of the promising and important research areas in electronic structure theory. In the last several years, many molecular scientists have begun shifting their attention from small, gas-phase systems toward larger, condensed-phase systems. Recent and current work in electronic structure theory is beginning to enable accurate computations on these larger, more complex systems. However, continued theoretical advances will be necessary to accurately account for the effects of solvent or chemical environment, and additional improvements are necessary in approximations or algorithms to allow dynamics simulations of large molecules using accurate electronic structure methods. Mixed quantum-mechanics/molecular-mechanics approaches,\textsuperscript{172–177} are probably a necessary but not sufficient tool to enable the most challenging of these simulations. Recent work on effective fragment potential methods\textsuperscript{178–181} may be helpful in this regard.

In conjunction with the improvements in electronic structure theory, computer hardware continues to advance in speed and affordability. However, computer processors are now being improved primarily through the addition of cores rather than by increasing clock speeds. It is hardly possible to purchase a single-core processor anymore, and Advanced Micro Devices’ “Magny-Cours” processor, with twelve on-die cores, is slated to ship by the end of March, 2010. Systems will be available with up to four of these chips per motherboard, for a total of 48 cores per node. This trend toward many cores seems certain to continue for the next several years. Graphical processing units are even more extreme examples of hardware parallelism. This is both an opportunity and a challenge for electronic structure theory. The multiplication of cores in affordable systems provides much greater computer power, but only if the software is set to handle it. The development of parallel algorithms, once considered by many as the domain of the government laboratories, is becoming increasingly necessary just to utilize the power of a desktop workstation. For this reason, new methods and algorithms that are more naturally parallel may out-compete more traditional algorithms that are harder to adapt to the emerging hardware.

The author is grateful to Professor Edward Valeev and Professor Daniel Crawford (Virginia Tech) and to Robert Parrish (Georgia Tech) for helpful discussions.
