

## **Vinay Ramasesh**

### **Getting Involved with Research**

Before I entered the Texas Academy of Mathematics and Science as a junior, I had very little experience with actual scientific research. As a freshman, I'd messed around with fractals and music, finding a few patterns and proving some interesting results, but overall, my scientific endeavors consisted mainly of building motors and simple mag-lev devices, and doing math problems.

Luckily, that all changed when I came to TAMS. One of the great things about this school was, and still is, the amount of research opportunities that it provides. As a residential high school located on the University of North Texas campus, TAMS makes it very easy for its students to conduct research under UNT faculty members. Not only does living on campus make going in to the lab very convenient, but the University professors are eager to have bright high school students working in their labs.

The summer before I entered TAMS, I knew I wanted to do research while I was there, but I had no idea what field to pursue. My main fear was that I wouldn't know nearly enough to contribute to any scientific research. As a newly-graduated sophomore, the only science classes I had taken were biology and chemistry, and I wasn't too comfortable with biology. My strength, I felt, was math, but again, I didn't know enough to conduct pure math research. I began to wonder how any high school student could expect to make a meaningful contribution to any professor's research. So, I looked at previous TAMS students and their work, to get a sense of what research would be like. Many projects caught my eye, and most were conducted in the lab of Dr. Angela Wilson, Professor of Chemistry at UNT. Her work was in computational quantum chemistry, and it combined computational work with high-level math. Not only did her

research seem interesting, but the projects TAMS students had done in the past seemed both challenging and doable. I sent Dr. Wilson an e-mail, requesting more information about her work, and her reply was encouraging. She gave me several links to websites with information about computational chemistry, and told me to contact her when the school year started if I was still interested.

I set to work reading about computational chemistry. A lot of the sites she pointed me to had information that went well above my head, but I was able to understand the gist. *Ab initio* computational chemistry attempts to predict thermodynamic properties of molecules from first principles. In the molecular realm, “first principles” translates to quantum mechanics, in the form of the Schrödinger equation:

$$E\varphi(x) = -\frac{\hbar^2}{2m}\nabla^2\varphi(x) + V(x)\varphi(x)$$

This famous equation relates the wavefunction,  $\varphi(x)$ , of an electron to the potential energy,  $E$ , of the molecular system. The wavefunction loosely corresponds to the probability distribution of an electron within a molecule, and the potential energy value is used as a starting point for other, more practical, calculations, such as those involving molecular enthalpies of formation. Unfortunately, this equation is so complex that it cannot be solved analytically for systems with more than one electron! This poses a bit of a problem, since the universe isn't composed solely of hydrogen.

That's where computational chemistry comes in. It uses numerical approximation methods, similar to Euler's method or the Runge-Kutta algorithm for solving differential equations numerically. These approximation schemes, known as computational methods, are used in conjunction with basis sets, or sets of functions which approximate the electronic wavefunction. Currently, a variety of different computational methods and basis sets are used, with each

enjoying varied success with different types of molecules. One general trend is evident, however: as the size of the molecule we want to calculate the properties of increases, the computational expense of performing the calculation—in terms of computer processor time, memory resources, and hard disk space—increases very steeply. For example, the simplest computational method which takes into account electronic interactions, Second-Order Møller-Plesset Perturbation Theory (MP2), scales as  $N^5$ , where  $N$  is the number of basis functions in the system. This steep scaling of computational cost with system size is a major reason that the insights of computational chemistry cannot be applied to systems of biological interest, such as proteins or nucleotides.

When I met with Dr. Wilson in the fall, she told me that this problem was one her group was looking to work on. They had recently created a new computational method, the Correlation-Consistent Composite Approach (ccCA) which employs several existing methods working in conjunction to accurately predict molecular energies. One of their current goals was to extend the scope of the method to work on larger molecules, but to do this they needed to overcome the scaling problem.

As Dr. Wilson explained to me, researchers had recently developed computational methods known as “local methods,” which made certain approximations to decrease the expense of working with larger molecules. In the description of space employed by local methods, electrons interact mainly with other electrons that occupy spatially close orbitals; with conventional methods, all electrons interact with each other. This approximation reduces the number of integrals the computational program has to evaluate, and thus reduces the computational cost. However, this also reduces the accuracy of local methods. A study by Werner, *et al.*, [1], showed that the calculations made using Local Second-Order Møller-Plesset Perturbation

Theory, when compared to the conventional MP2 calculations, were only 97-98% accurate. For most molecules, this corresponds to an error in energy of about 10-15 kcal/mol. This error bound is largely unacceptable for high-accuracy studies. So, a first step in harnessing local methods for use with high-level studies is to find a way to increase their accuracy without simultaneously increasing their expense. This was the driving idea behind my work in Dr. Wilson's lab, although I was more generally involved in identifying aspects of the localization process that required further effort.

## **The Main Work**

### **1 Background**

#### **1.1 Local Methods**

Conventional methods employ a description of space in which electrons occupy molecular orbitals delocalized over the entire molecule. These orbitals increase in size with the system, leading to the high scaling of computational effort with system size. To circumvent this problem and reduce computational expense, local electron correlation methods, which use a localized description of electronic interactions, have been developed.

The implementation scheme for the local methods considered here, introduced by Pulay [2] and implemented by Pulay and Saebø, employs local orbital spaces (called domains) rather than delocalized molecular orbitals (MOs). This orbital localization restricts the number of possible substituted configurations in the electronic wavefunction, only considering those substitutions which contribute significantly to the molecular energy.

#### **1.2 Complete One-electron Basis Set Extrapolations**

Errors in quantum calculations arise from two sources—intrinsic error of the method (or level of theory), and error due to imperfect basis sets. To eliminate basis set errors, it is desirable

to calculate the value of a property with a complete basis set. However, since such calculations are impossible, an approximation to this value must be used. Dunning has introduced the correlation-consistent polarized valence (cc-pVnZ with  $n=D, T, Q, 5 \dots$ ) basis sets [3], which are now available for a large portion of the periodic table. These sets are constructed by adding shells of functions to a core set of orbital functions. As the zeta level ( $n$ ) of the basis set increases, the value of the property being calculated converges towards the one-electron complete basis set (CBS) limit. By extrapolating to the CBS limit, the error coupling of the one-particle basis set expansion and the  $n$ -particle electron correlation is eliminated, leaving only the intrinsic error of the trial wave function. Current CBS extrapolation schemes are not exact functions, but only approximations to an as-yet unknown formula.

## **2 Goal**

I sought to discover whether local second second-order Møller-Plesset perturbation theory [LMP2] energies systematically converged to a CBS limit and to discover how the accuracy of LMP2 with respect to canonical MP2 varied across basis set levels, including at the CBS limit. I explored the convergence to the CBS limit of LMP2 single-point energies versus the convergence of standard MP2 single-point energies, applying both methods to smaller hydrocarbons as well as a test set of large organic and organometallic molecules. I also tested three CBS extrapolation schemes to see which yielded the best results when used with local methods. Finally, I investigated changes in the accuracy of LMP2 CBS limit energies versus system size.

## **3 Methodology**

In order to study convergence trends of local and conventional MP2 methods for a wide variety of molecules, I selected a set of 24 relatively large systems from the 600-member training

set of Cioslowski and co-workers [4] that contain 18 or more atoms. These molecules include cyclic, poly-cyclic and non-cyclic alkanes, conjugated molecules, cage structures, and organosulfur and organometallic compounds. To this I added a set of 7 smaller hydrocarbons, to give a total test set of 31 molecules.

I optimized the molecular structures with the B3LYP computational method at the cc-pVTZ basis set. Then, I ran single-point energy calculations on these optimized structures at cc-pVDZ, cc-pVTZ, and cc-pVQZ with both conventional MP2 and LMP2.

I used the energies obtained with cc-pVDZ, cc-pVTZ, and cc-pVQZ to calculate the CBS limit energy for each molecule and method. I chose three widely-used CBS extrapolation schemes to test: the simple exponential fit (CBS-F) proposed by Xantheas and Dunning [5] and Feller [6], which uses the formula

$$E(n) = E_{\text{CBS}} + B \exp(-Cn),$$

where  $n$  is the zeta level of the basis set; the mixed exponential/Gaussian functional (CBS-P) devised by Peterson *et al.* [7], which uses the formula

$$E(n) = E_{\text{CBS}} + B \exp[-(n-1)] + C \exp[-(n-1)^2];$$

and the two-point Schwartz (CBS-S4)  $1/(l_{\text{max}}^4)$  scheme [8],

$$E(l_{\text{max}}) = E_{\text{CBS}} + \frac{B}{(l_{\text{max}} + \frac{1}{2})^4},$$

where  $l_{\text{max}}$  is the value of the highest angular momentum function in the basis set (equal to the zeta-level of the basis set for the atoms used in this study).

## 4 Results

(1) As I increased the basis set level in the calculation, the errors in the LMP2 calculations decreased. On average, the LMP2/cc-pVQZ error was 42.68% lower than the LMP2/cc-pVTZ error, which was 24.69% lower than the LMP2/cc-pVDZ error. Stated otherwise, LMP2 energies are closer to the canonical MP2 values—and thus more accurate—at higher basis set levels.

To confirm the reduction in correlation error with increasing basis set size (for most molecules), I ran single-point calculations on the smaller hydrocarbons (through butane) with cc-pV5Z. As expected, errors in LMP2 correlation energy are less with cc-pV5Z than they are with cc-pVQZ. On average, LMP2/cc-pV5Z correlation errors are 52.82% of the LMP2/cc-pVQZ errors, with similar values occurring for the domain-extended variants of LMP2. The cc-pV5Z errors were never greater than either the cc-pVQZ or cc-pVTZ errors.

This leads directly to the conclusion regarding CBS limits below.

(2) At the CBS limit, LMP2 energy errors were lower than those at each individual basis-set level. LMP2/CBS-P always recovered more than 99.2% of the canonical MP2/CBS-P energy, a vast improvement over the 97.1% seen at the cc-pVDZ basis-set level. This implies that extrapolating LMP2 values to the CBS limit lowers the energy differences between LMP2 and MP2, a conclusion consistent with the reduction of errors with increasing basis-set levels. Thus, CBS extrapolations can significantly increase the accuracy of this local method. In the future, researchers can use this technique to obtain accurate potential energies of molecules for calculating enthalpies of formation and stable structures.

LMP2/CBS recovers an average of 99.59% of the conventional CBS limit correlation energy. The worst case occurred for bullvalene, in which only 99.21% was recovered. This is a vast improvement over the average of 97.53% seen at the cc-pVDZ level, with the worst case of

96.40% occurring at this level for bicyclo[2.2.2]octane. As table 2 shows, by extrapolating to the CBS limit, LMP2 correlation errors are reduced from an average of 15.81 kcal/mol at the cc-pVDZ level to an average of 3.59 kcal/mol at the CBS limit.

It is interesting to see a mathematical reason behind the reduction in errors at the CBS limit. Solving the CBS-P equations reveals that the CBS-P energy is a linear combination of the cc-pVDZ, cc-pVTZ, and cc-pVQZ energies:

$$E_{CBS} = \frac{(e^{-11} - e^{-7})E_D + (e^{-4} - e^{-10})E_T + (e^{-5} - e^{-3})E_Q}{e^{-11} - e^{-7} + e^{-4} - e^{-10} + e^{-5} - e^{-3}}$$

So, the LMP2/CBS-P error is a linear combination of the cc-pVDZ, cc-pVTZ, and cc-pVQZ errors:

$$\Delta E_{CBS} = \frac{(e^{-11} - e^{-7})\Delta E_D + (e^{-4} - e^{-10})\Delta E_T + (e^{-5} - e^{-3})\Delta E_Q}{e^{-11} - e^{-7} + e^{-4} - e^{-10} + e^{-5} - e^{-3}}$$

Notice that the coefficient of  $\Delta E_T$  is negative, while those of  $\Delta E_D$  and  $\Delta E_Q$  are positive. Thus, the LMP2/cc-pVTZ error cancels, to some degree, the cc-pVDZ and cc-pVQZ errors. We can use the percentages given in the section 4.1 to express the average errors at the cc-pVDZ and cc-pVTZ levels as multiples of the cc-pVQZ error. So on average,

$$\Delta E_{CBS} \approx \frac{2.317(e^{-11} - e^{-7})\Delta E_Q + 1.745(e^{-4} - e^{-10})\Delta E_Q + (e^{-5} - e^{-3})\Delta E_Q}{e^{-11} - e^{-7} + e^{-4} - e^{-10} + e^{-5} - e^{-3}},$$

which simplifies to

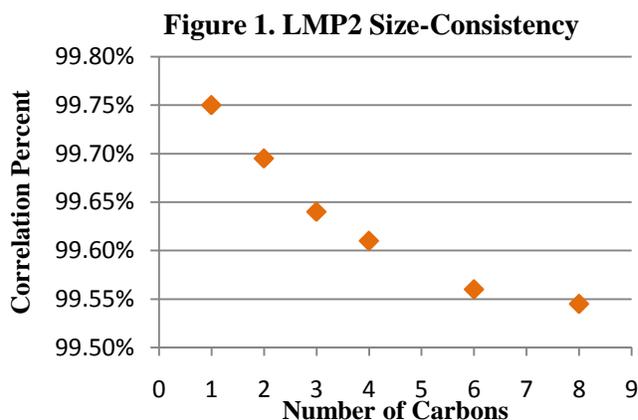
$$\Delta E_{CBS} \approx 0.516\Delta E_Q,$$

showing why the CBS-P energy is more accurate than the cc-pVQZ energy, and therefore more accurate than the energies obtained with other basis sets.

(3) Comparison of CBS schemes: Now that we know that extrapolating LMP2 calculations to the CBS limit is an inexpensive way to increase their accuracy, we need to know

which CBS scheme to use to calculate such limits. The CBS-F scheme initially appeared to perform best with local methods; with it, the local energy errors were lowest. However, as I increased the sizes of domains used in the LMP2 calculation, the magnitude of the LMP2/CBS-F energies grew larger than the MP2/CBS-F values. This is physically incorrect, as local energies should always be lower in magnitude than the corresponding non-local values. This flaw, along with the observation that the MP2/CBS-F values were inconsistent with the MP2/CBS-P and MP2/CBS-S4 energies, shows the weakness of the CBS-F scheme. LMP2/CBS-P errors were, on average, 0.35 kcal/mol less than LMP2/CBS-S4 errors. This implies that the CBS-P scheme will yield the most accurate results in studies with local methods, a fact useful in any study involving large molecules.

(4) Size-consistency of LMP2: Figure 1 shows a plot of the percent of the MP2/CBS-P energy recovered with LMP2/CBS-P for alkanes versus molecular size. The decreasing trend, best modeled with a power function, implies that LMP2/CBS-P values will be within 1% of the



non-local values for alkanes with up to 1045 carbons. Thus, low-cost LMP2/CBS-P calculations will likely be accurate for systems well into the range of biomolecules.

My main contribution was to make accurate quantum calculations on important molecules more feasible by determining that LMP2 energy values can be made more accurate by extrapolating to the CBS limit and by establishing that the CBS-P scheme should be used for such an extrapolation. My research may, therefore,

one day aid vital advances in studies that depend on knowing thermodynamic properties of large molecules.

## 5 References

- [1] Werner, H.-J.; Pflüger, K. *Ann. Rep. in Comp. Chem.* **2007**, 2, 53.
- [2] Pulay, P. *Chem Phys. Lett.* **1983**, 100, 154.
- [3] Dunning Jr., T. H. *J. Chem. Phys.* **1989**, 90, 1007.
- [4] Cioslowski, J.; Schimeczek, M.; Liu, G.; Stoyanov, V. *J. Chem. Phys.* **2000**, 113, 9377.
- [5] Xantheas, S. S.; Dunning Jr., T. H. *J. Phys. Chem.* **1993**, 97, 18.
- [6] Feller, D. *J. Chem. Phys.* **1992**, 96, 6104.
- [7] Peterson, K. A.; Woon, D. E.; Dunning Jr., T. H. *J. Chem. Phys.* **1994**, 100, 7410.
- [8] Martin, J. M. L. *Chem. Phys. Lett.* **1996**, 259, 669.