Analysis of Ethane in Charged Environments and Electric Fields

Alexander M. Cappiello

Department of Chemistry
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The work presented here is done in conjunction with a larger project by Dr. David Yaron to apply machine learning techniques to develop new semi-empirical methods to address electronic structure calculations. The broad goal of the project, named Molecular Similarity in Quantum Chemistry (MSQC), is to exploit the fact that molecules in similar environments have similar properties. It aims to provide a bridge for accurate, computationally inexpensive calculations. Vital to this process is having data on a reference fragment under a wide range of conditions. Initially, point charges were used to achieve this result. However, analysis of the resulting data with principal component analysis (PCA) suggested that this was only exploring a narrow range of conditions. From here, new types of environment generation were explored to correct this. More elaborate systems of point charges as well as electric fields from multipoles were created as alternatives. When a trial run was done using fields, a closer look at the PCA data revealed that the previous concerns may have been unwarranted. Consequently, both fields and charges appear to be acceptable environments.

Introduction

Traditionally, computational chemistry calculations can be thought of as a compromise between the accuracy of the result and computer time required to produce it. Since this time tradeoff is significant and measurable, it has greatly limited the usefulness of electronic structure calculations in solving real-world problems, especially with larger molecules. The goal of the MSQC project is to develop new methods of calculation based on similarity between a given a fragment in changing environments. More specifically, the project seeks to take a molecular fragment where highly accurate (high level) data exists for the fragment in a variety of environments in conjunction with a less accurate calculation (low level) in a unique environment or configuration and predict more accurate values than the low level calculation can provide. Thus, from a computer science perspective, memory becomes a tradeoff factor in addition to time vs. accuracy. However, in order to create a useful model from the high level data, calculations must be done on the fragment in a sufficiently broad range of conditions. Data for the analysis is generated using the Gaussian 09 System.

The first approach to generate environments was to use point charges on the corners of a cube *(2)*. Using PCA on data generated by this method revealed that regardless of number of environments, they may be unable to sufficiently perturb the fragment. The purpose of this paper is to explore various new approaches to environment generation in hopes of increasing the diversity of calculations done on the fragment. Since using point charges in fixed positions appears to be too limited, the first alternative that was explored was using point charges placed randomly. In theory, this should increase the randomness of the environment, but it becomes difficult to regulate environment creation in that the result may be an unreasonable

environment. A second new type of environment generation was to place dipoles (represented as two nearby point charges) in the same manner. Lastly, electric fields from a multipole were added as another alternative type of environment. All three methods have been implemented, but they have not been equally tested. In the end, the project requires the method (or combination of multiple methods) that yields the most diverse environments. In essence, as many degrees of freedom as possible should be perturbed by the environments.

Method and Analysis

The entirety of this analysis is based on calculations done on ethane. The overall process is this: generate environments using a given method, choose a high level basis set and one or more low level basis sets to be used in the calculations, specify configuration parameters for the fragment, and then run the calculations. A template file is used for the fragment with placeholders for parameters and charges. This template file is used to construct a typical Gaussian input file. To be useful, the only change between trials was the environment. The high level basis is 6-31G and the low level basis sets are STO-3G and a general basis set in a narrow and diffuse configuration. Note that only the high level basis is used for PCA analysis. Next, the following configurations of ethane (shown below) were explored:

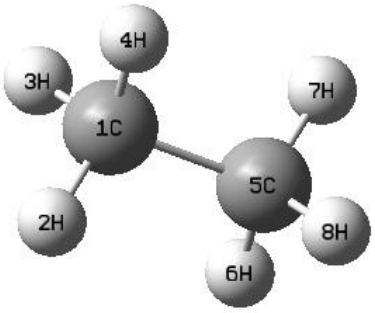

Figure 1 *(1)*

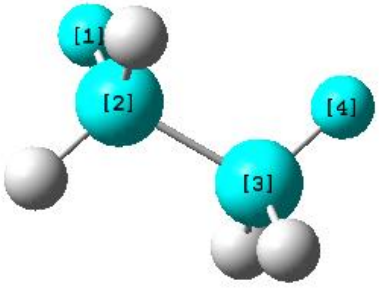

Figure 2 *(1)*

C-C bond length between 1C and 5C (Figure 1): 1.54Å (standard), 1.39 Å, and 1.69 Å
C-H bond length between 1C and 3H (Figure 1): 1.12 Å (standard), 0.97 Å, and 1.27 Å
Dihedral angle given by rotating atom 4 (Figure 2): 60° (standard), 30°, and 0°
Note that only one nonstandard parameter was used at a time. Thus, a total of 7 configurations of ethane were used for calculations.
Charges can then be added directly to the end of the Gaussian input file using the charge keyword *(3)*. For the original environments, charges were placed on the corners of a cube centered on the fragment *(2)*. In this case, a 6 Å box was used with random charge magnitudes from -5 to 5 with a total of 100 environments.
An environment based on randomly placed charges is constructed in a similar manner, instead placing charges at random coordinates within a specified bounding box. These coordinates are

then checked to ensure that a "safe" distance is maintained from the molecule, done by checking each atom. Safe distances are specified as a function of the atomic radius of the atom. Thus, we get the expression:

$$safeDist = c_1 r_{atom} + c_2$$

Using the condition of staying 5 Å away from hydrogen and 6 Å away from carbon, the equation becomes:

$$safeDist = 7.143 r_{atom} + 1.214$$

At this point, no other checks are done, but the concern is that with fairly high probability, the resulting environment is severely unbalanced. At this point, no large trials have been done using these environments.

One alternative to help mitigate the problems with random charge environments is to simulate dipoles as a pair of close point charges, with equal magnitude and opposite sign. This is done by placing a point charge randomly as above and then placing its dipole complement randomly in spherical polar coordinates relative to the first where $0 < r \leq 1$, $0 < \theta \leq \pi$, $0 < \phi \leq 2\pi$ and then converting back to absolute Cartesian coordinates in the environment. Likewise, no large trials have been done in these environments yet.

Alternatively, fields can be used using the Gaussian field keyword *(4)*. Gaussian supports multipoles in the form of dipoles, quadrapoles, octapoles, and hexadecapoles (Figure 3), which are specified by their orientation and magnitude.

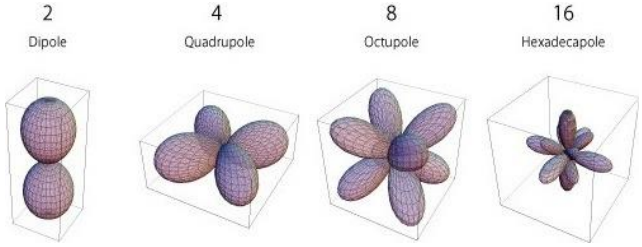

Figure 3: Multipoles *(6)*

Environments were generated using one field. With only one field, there is no interdependence between environment elements, so environments were generated systematically, instead of randomly. To explore the widest possible range of fields, all 31 orientations were used with magnitudes from 50 to 800 in steps of 50, giving a total of 496 environments. 800 was established as an upper bound because a dipole field of 800 induces a dipole moment in ethane comparable to that of isolated fluoroethane.

Generated data was initially tested by applying PCA to the density matrix of a given atom of the fragment. The density matrix is an n x n matrix where n is the number of basis functions for the given atom. This is transformed into a 1 x n^2 matrix that becomes a row in a new matrix where there is a row for each environment. For each column, the mean is calculated and the subtracted from each element in the column. PCA is run using the MATLAB function princomp() *(5)*. The result of interest is the 'latent' data, which is a vector containing the eigenvalues of the

covariance matrix of the input. The x value for which the vector converges to 0 can be interpreted as the number of degrees of freedom being explored.

Following this analysis, it is logical to apply the same methods to the full density matrix because only taking data from a single atom at a time might be misleading. This was done by simply extending the process outlined above, where each row is composed of the density matrix for each atom sequentially. More concretely, let d_a be the 1 x n^2 density matrix for atom a. Then each row of the input matrix becomes: $d_1 \mid d_2 \mid \ldots \mid d_m$ for m atoms in the fragment. Ideally, this should provide a more complete picture of the effectiveness of each environment.

Results and Discussion

The primary challenge with this data is properly interpreting the PCA results. As mentioned earlier, MATLAB's princomp() function returns a vector 'latent' containing the eigenvalues of the covariance matrix for the input *(5)*. PCA is done on each set of parameters for the fragment individually. When plotted together, the result looks like this:

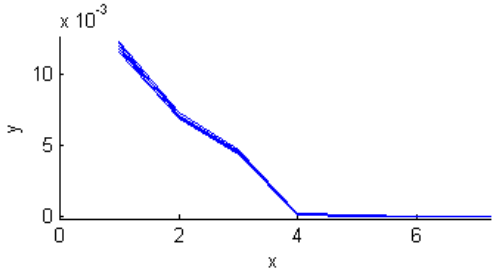

Figure 4: PCA on carbon-1 charge data

The sharp drop present in the plot suggested that the environments are only exploring a limited range of states of the fragment. In other words, using 100 environments may be no different than 4. After analyzing the data in a different light, I propose that this initial conclusion was invalid. A point of contention is over when to say that the vector converges to zero. From this plot, it appears to converge at x = 4. However, when looking at the actual values, this does not happen until much later. The vector continues to decline at a reasonable rate and then suddenly drops to effectively zero. It may actually go to zero or an arbitrary small value due to floating point rounding errors (indicated by a drop of a large order or magnitude). Looking at the log plot provides a clearer picture of this trend:

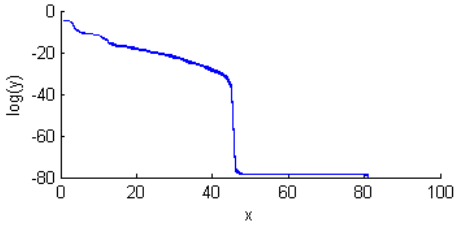

Figure 5: Log Plot of PCA on carbon-1 charge data

The decline is almost linear until the drop between x = 45 and x = 46. Moving forward, only the first value of $0 < y < 10^{-25}$ will not be included on the graph and values where y = 0 must be dropped to make the log plot. The conclusion that 100 environments is no better than 45 is very reassuring following this change in the analysis. Furthermore, drawing a conclusion after looking at the numbers is sounder than basing it off of a graph. However, two big questions remain: how does this compare to using electric fields and does applying PCA to the whole fragment change the result?

The easier question to answer is how charges and fields compare to one another. Plotting the two together gives a rather interesting result:

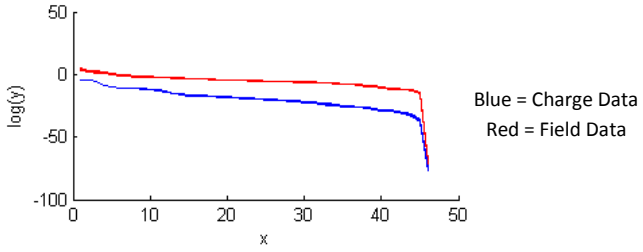

Figure 6: Log Plot of PCA on carbon-1

The most obvious feature is that both curves plummet at the same place, between x = 45 and x = 46. The other notable feature is that field data lies above the charge data throughout. The most likely explanation is that the data was not normalized before PCA was run. Thus, the preliminary conclusion is that charges and fields perturb the fragment in an equally effective manner.

Another interesting comparison is using PCA on a hydrogen atom instead of a carbon atom. Comparatively, the results are vastly different.

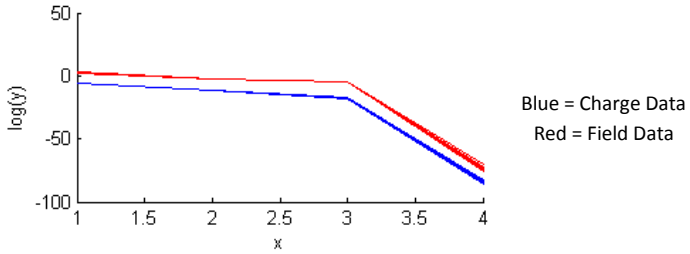

Figure 7: Log Plot of PCA on hydrogen-3

In this case, the plot actually does drop to 0 between x = 3 and x = 4. This similarity to the original plot shown is most likely a coincidence and is presumably the effect of hydrogen being such a simple system. More importantly, since the behavior is different, it builds a strong case that doing PCA on the whole fragment is a more appropriate analysis.

Intuitively, it seems reasonable that PCA on the whole fragment will yield a different result. The argument parallels describing degrees of freedom in a molecule. Namely, the number of degrees of freedom increases with the number of molecules and degrees of freedom of each atom are essentially independent. Similar, it stands to reason that only looking at the density matrix of a single atom ignores the variance in the other atoms. The result appears to confirm this hypothesis.

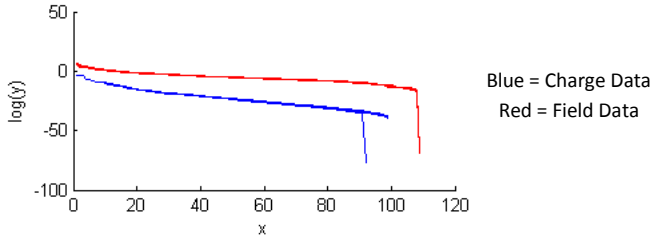

Figure 8: Log Plot of PCA on full fragment

There are several notable features of this graph. The most prominent is that the curves do not fall to 0 at the same x-value. In looking at the charge data, one line clearly breaks from the other 6, falling between x = 91 and x = 92. Presently, the cause has not been analyzed, but is not expected to pose a problem with the remaining analysis. However, it is worth noting that the line originates from the fragment with all the standard parameters. Also noteworthy is that the 6 lines that stay together do not show a drop at the end. This is because they immediately fell to 0 as opposed to an intermediate step or arbitrary small number, thus the drop cannot be expressed on this plot. This happens between x = 99 and x = 100. On the other hand, the plot of the field data drops between x = 108 and x = 109, which suggests that additional environments are beneficial for up to 108 environments. At first glance, this would appear to a significant finding. However, recall that the charge data was generated from 100 environments and the field data from 496 environments. Thus, it is entirely expected that the charge data cannot pass 100 on this plot. While an upper bound cannot be placed on the charge data because of this, it could be easily solved by generating a larger set of data. Again, the field data lies above the charge data, but this is likewise expected to be an error due to lack of normalization. Overall, the results of this analysis are far more encouraging than originally anticipated.

Summary and Conclusion

To recap, this analysis of environment generation for the MSQC project was initiated after PCA on data using environments containing point charges on the corners of cube was brought into question over whether they explore the fragment to an acceptable extent. However, the analysis here brought into question that original premise, disproved it, and found both fields and point charges on a cube to be acceptable environments. However, one of the larger questions that remain is confirming that the PCA result is actually measuring what it was

intended to. Fortunately, the results are very encouraging. First, PCA on carbon-1 not only gives a reasonable value, but gives the same value for both charge environments and field environments. This suggests that either both are equally effective or both perturb carbon to its fullest potential. Also, the fact that PCA on the charge data for the full fragment is bounded by the number of environments shows that there is a direct link between the two. Most importantly, there does not appear to be a cause for concern over the environments that are being used.

However, going forward there are many other aspects of environment generation to be explored. Among the more straightforward is finding a true upper bound for the charge on a cube environments as well as generating and analyzing data using the random charge and dipole environments that have not been explored yet. There are also more directions to be explored with regard to the field environments. Firstly, analyzing dipoles, quadripoles, octapoles, and hexadecapoles individually can show whether they make equally useful environments. Another test would be using environments with more than one field present. Also, while it appears that these individual environment types are suitable individually, it may be worthwhile combining environment types to see if it increases the diversity of environments. Lastly, the correlation explored here with PCA can be used to establish an optimal number of environments for a given fragment and configuration. The usefulness of this is twofold: to ensure that fittings actually represent the broadest possible range on the data while avoiding redundant calculations. It would be an interesting side project to explore why these limits exist and what the limiting factors are. It is possibly that it is affected by the nature of the environment, the atom/fragment itself, and/or the chosen basis set for calculations. Since the environments play a major role in the overall picture of the MSQC project, it will be important to understand how to use the environments to maximize the accuracy of the fitting process once that point is reached.

References

[1] *AMPAC 9*, version 9.2.1. Semichem, Inc: Shawnee, KS 2008; (accessed 12/16/2011).

[2] David Yaron. Overview [Word Document]. 2011.

[3] Gaussian 09 User's Reference. Charge. http://gaussian.com/g_tech/g_ur/k_charge.htm (accessed 12/16/2011).

[4] Gaussian 09 User's Reference. Field. http://gaussian.com/g_tech/g_ur/k_field.htm (accessed 12/16/2011).

[5] *MATLAB*, version 7.12.0.635 (R2011a). MathWorks: Natick, MA 2011; (accessed 12/16/2011).

[6] Physics of Correlated Systems: Yanagisawa Lab. Multipoles. http://www.cris.hokudai.ac.jp/yanagisawa/entries/research/images/multipoles.jpg (accessed 12/16/2011).