Chapter 1 Electrostatic Fields in Small Thiocyanate Molecules with Ensembles Generated using the AMOEBA Force Field

1.1 Introduction

One of the chief difficulties in protein electrostatics is obtaining adequate sampling. Throughout my research, I have operated under the assumption that the most important degree of freedom is the nitrile dipole orientation. Expanding from one-dimensional umbrella sampling to two-dimensional umbrella sampling did, in fact, yield improved correlation between calculated PB fields and experimentally measured vibrational absorption energies. However, increasing the umbrella window time from 400 ps to 2000 ps did not yield the same degree of improvement. Looking strictly at the RFM, instances of improved (Rap E30D/K31E), unchanged (Rap E30D/K31, Rap E30/K31E), and decreased (Ral monomer, Rap E30/K31) correlation were all observed. The first two cases are not unexpected—either sampling was inadequate and improved sampling improved the simulations physical veracity or the sampling was adequate enough and improved sampling does not have a significant effect.

The case where increasing the sampling actually decreases the correlation to experiment is perplexing though—what about increasing the simulation duration would actual decrease how well the *in vitro* system is represented? The answer is proteins are very complex. It's possible that increasing simulation time could allow a rare local minimum to be visited but not escaped, resulting in an over-representation of that structure. Likewise, it's also possible that it's purely coincidence that the local structure sampled happens to also well-correlate to the experimental measurement. Using napkin math, the Ral monomer has approximately 170 alkane-like sidechain dihedrals and Ral docked to Rap E30/K31 has approximately 460 sidechain dihedrals. Assuming all of them have a non-zero probability of being at approximately 60˚, 180˚, and -60˚ and counting each of those states as a single state (for three total probable states), then there

are 10^{80} possible combinations of sidechain dihedrals for Ral and 10^{220} combinations for Ral docked to Rap E30/K31. Obviously, a large number of these are so energetically unflavored that they can be neglected—there are certain combinations that are physically impossible due to steric overlap. Assuming that only residues with sidechains within 10 Å of the probe are relevant for electrostatic field calculations ($1/r^2$), which for $Y31C_{SCN}$ docked to Rap E30/K31 is 13, and using an averaging number of dihedrals per residue as 1.7, there are still 10^3 possible dihedral combinations—significantly more tractable. Yet because of energy barriers, more than 10^3 frames are needed to see all possible states; the scope of how many more frames is dependent on the size of the barriers. But again, there are orientations of these 13 residues which may only be energetically favorable given specific orientation of residues further than 10 Å from the probe. It's very easy to fall go down the rabbit hole and get lost in a sea of dihedral permutations.

The number of sidechain degrees of freedom in a protein is massive and it's impossible to be 100% certain that the entire ensemble is represented in the appropriate proportions at this stage in computational efficiency. In this regard, I have stepped back to a smaller subset of systems: methylthiocyanate, ethylthiocyanate, hexylthiocyanate, and acetyl-cyanocysteine-N-methylamine peptide-like small molecule. Furthermore, since the principles behind using the AMOEBA force field are still a concern (solute dielectric) and it was clear that sampling in Amber03 and performing field calculations in AMOEBA was unsuccessful, these molecules have also been simulated in the AMOEBA force field. The solvated protein system is currently too large for AMOEBA, but these smaller systems can easily be simulated in a reasonable amount of time (approximately 0.8-1.0 ns/day).

It has been shown that bother AMOEBA as well as GAFF can reproduce experimental Stark shifts for a given probe in a variety of solvent environments.[1] However, interesting biology involves water interacting with many different solutes. The central focus of my work has been understanding how to quantify electrostatic fields in

biologically-relevant systems. In that regard, a good reproduction of Stark shifts of different protein systems in water strictly using MD has not been reported. Here we investigate how well Stark shifts can be reproduced for various thiocyanate-containing solutes which will be easier to obtain complete ensembles. The goal here is a proof-of-concept and may be used as a springboard for future works on increasingly larger probe-containing systems.

Simulations on these same small molecules in Amber03 has also been started by an undergraduate in the lab for future comparison.

1.2 Results and Discussion

1.2.1 Sampling CN Orientations

Figure 1-1 shows the one-dimensional dihedral distributions for all non-hydrogen dihedral angles. From this it's clear that ethylthiocyanate is well-able to sample all of the expected alkane-like dihedral space in 4 ns. Hexylthiocyanate and the capped cyanocysteine are more difficult to assess from Figure 1-1 due to being unable to distinguish between dihedral permutations. Figure 1-2 therefore shows the two-dimensional dihedral distribution for these two molecules for the dihedrals involving SCN. Looking at Figure 1-2, hexylthiocyanate appears to behave alkane-like, with all alkane-like windows visited at least briefly. However, either (180˚, 60˚), (180˚, -60˚), (-60˚, -60˚), and (60˚, 60˚) are particularly favorable or the system has not yet had enough time to fully sample the other alkane-like regions. Likewise, the capped cyanocysteine appears to very much favor the (180˚, -60˚) alkane region, with a small-but-significant probability at (60˚, 60˚).

Because of it is capped and therefore few possible solute-thiocyanate interactions should be present, it is hypothesized that the peptide simply has not yet had adequate time to sample all states and an enhanced method, such as umbrella sampling, should be used to generate a more-complete ensemble. The same may be said for hexylthiocyanate also,

although to a lesser degree. Looking at Figure 1-3, it can clearly be seen that ethylthiocyanate and hexylthiocyanate can rotate dihedrals much more quickly than the capped cyanocysteine, which spends nearly 6 ns in a single (χ_1, χ_2) conformation, only to escape it near the end. It's possible that more sampling could significantly alter the dihedral probability distribution as well as see the probe revert back to the prior state and remain there—more sampling or an enhanced MD technique is needed in either case.

1.2.2 Electrostatic Fields

Figure 1-4 shows the average calculated field as a function of time for fields calculated using Figure 1-4A) the induced method (IM), Figure 1-4B) the midpoint method (MPM), Figure 1-4C) MPM where the monopole, permanent dipole, induced dipole, and quadrupole contributions of the SCN atoms have been removed, and Figure 1-4D) MPM where the monopole, permanent dipole, and quadrupole contributions of the SCN atoms have been removed. The IM is trivial to calculate using built-in output from Tinker during the simulation run. The MPM method required additional work-up after the simulation was completed; both method have previously been described.

The average levels off relatively quickly for all small molecules, indicating either convergence with respect to electrostatic field or oversampling within some number of local minima and inadequate sampling within other structures. It's likely the former for methyl- and ethylthiocyanate and the latter for hexylthiocyanate and the capped cyanocysteine, based on methylthiocyanate having very few structure degrees of freedom, Figure 1-1, and Figure 1-2. Once again, it's likely a method of enhanced MD (or simply more simulation time) are necessary for the larger two molecules.

Due to the close proximity between the SCN atoms and the nitrile midpoint, SCN will be the dominate contributor to the electrostatic field. In the IM, this is significantly reduced by the damping factors in the SCF subroutines. The MPM does not use such damping terms, however, which is why the magnitudes in Figure 1-4C, Figure 1-4D,

Figure 1-5, Figure 1-6B, and Figure 1-6D are so large. Since we are not interested in the self-field of the probe, but rather the external field felt by the probe, we have looked at removing the field contributions of the SCN atoms, which is identical to what was done in the previous chapter. The SCN field is due to the charges on SCN, which should, on average, be the approximately constant for any given system—the force field parameters defining bond lengths and angles should ensure that. Because AMOEBA also has an induced dipole term, however, part of the SCN field is not constant—the field due to the induced dipoles on the S, C, and N atoms. Therefore, we have looked at both removing and keeping that term. Figure 1-5 shows the contribution of each multipole part on SCN to the electrostatic field at the bond midpoint. As expected, all of the permanent terms (monopole, permanent dipole, quadrupole, permanent total) are approximately constant for all probes. Again, this is unsurprising given all the SCN share multipole parameters. The induce dipole field contribution (and total including induced dipole contributions) vary because it is a response to the local field experienced by the probe—the exact value we are trying to vary by modulating the atoms SCN is attached to.

1.2.3 Correlating Small Molecule Fields to Experiment

Figure 1-6 shows the calculated fields plotted against the experimental absorption energies. It is important to keep in mind that, no matter how well correlated any of the data may be, there are only three data points in each fit—there are no experimental absorption frequency measurements for the capped cyanocysteine at this time and it is simply placed along the best-fit line based on its calculated field—and all interpretations of the data need to weary of this. In fact, a squared correlation coefficient of 0.97 (Figure 1-6D), still only has a p-value of 0.11 for 3 data points—there is an 11% likelihood that the most correlated data are correlated by chance.

In the induced method (Figure 1-6A), MPM total field (Figure 1-6B), and MPM field less permanent SCN field (Figure 1-6D) we see a positive correlation between

calculated fields and experimental absorption energies. However, in the MPM field less the total SCN field (Figure 1-6C), we see a strong negative correlation. The only difference between Figure 1-6(A,B,D) and Figure 1-6C is that Figure 1-6C does *not* include the probe response to its external field environment. It appears as though allowing the SCN to polarize due to its surroundings is important for accurately predicting vibrational Stark shifts in the correct direction.

The observation that removing all of the SCN contribution to the field, including the induced dipole contribution, results in negative correlations merited re-visiting the previously reported results. Due to the way fields were calculated, I could not examine adding back in only the induced dipole contribution without significant (months) of repeated calculations. Furthermore, we looked at removing the entire sidechain, CH_2SCN, for the same reason that it should be approximately constant. We can, however, look at total field without any contributions removed, which would reintroduce the seemingly vital probe induced dipole field. The correlations with and without removing any atom contributions are plotted in Figure 1-7. For all the monopole methods, there is essentially zero change—the CH_2SCN is relatively constant among all simulations. For the explicit solvent AMOEBA, there is also essentially no change. For AMOEBA with implicit solve, the magnitude of the correlation increases, although it becomes more negative rather than changing signs. For future calculations, keeping the field due to the induced dipole on probe may be important, but for our previous results it's likely that the non-transferability of ensembles from Amber03 to AMOEBA is more significant and convoluting.

We also looked at field standard deviations compared to experimental FWHM, shown in Figure 1-8. Aside from the MPM total field (Figure 1-8B), the correlations are in the positive direction, although, again, with only three data points, their significant is questionable.

1.3 CONCLUSION

Sampling SCN-labeled small molecules in AMOEBA appears to be promising and merits further investigation. It appears that the polarizability of the probe itself is important for correctly quantifying the direction of vibrational shifts due to the VSE. It's also likely necessary to perform some sort of enhanced MD on the cyanocysteine and hexylthiocyanate to ensure correct ensembles.

The probe it has already shown when looking at different solvent environments,[1] combined with this preliminary study is promising. Sampling in AMOEBA has shown to correlate calculated electrostatic fields to experimental vibrational absorption energies via the VSE, rather than the observed negative correlations reported sampling in a point charge force field. Addition experiments and simulations should be examined to further investigate.

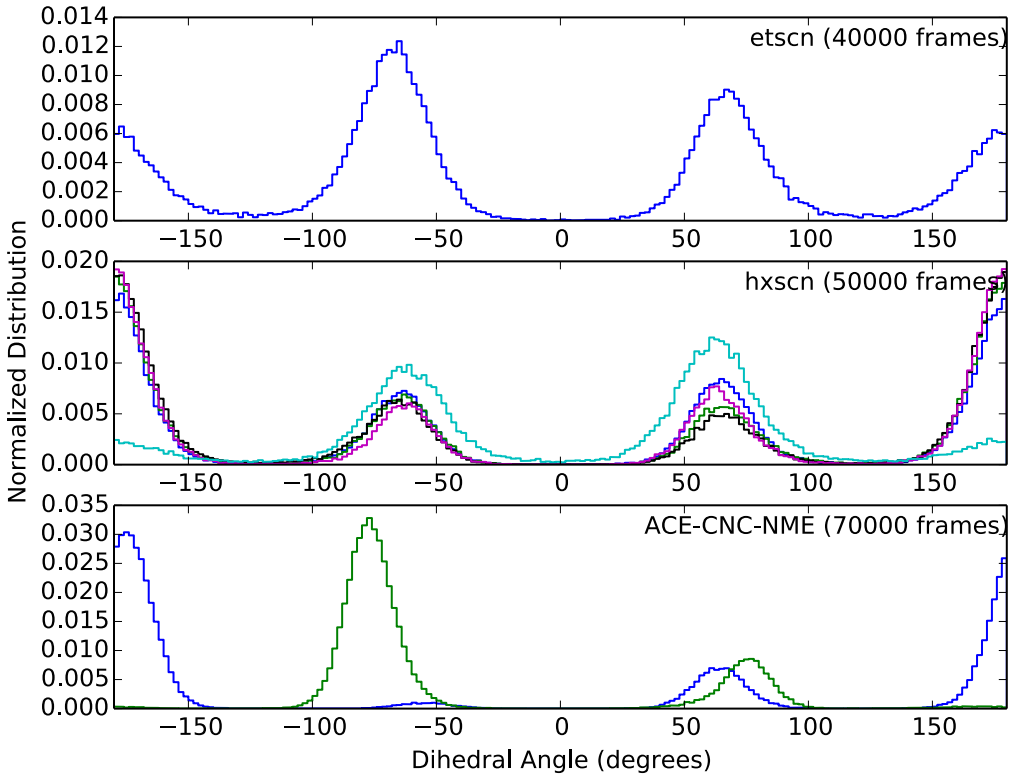

Figure 1-1: One-Dimensional Dihedral Probability Distributions

Non-Hydrogen dihedral probability distributions for ethylthiocyanate (top), hexylthiocyanate (middle), and capped cyanocysteine (bottom). Ethylthiocyanate (top): (blue) C1-C2-S-C dihedral. Hexylthiocyanate (middle): (blue) C1-C2-C3-C4 dihedral; (green) C2-C3-C4-C5 dihedral; (black) C3-C4-C5-C6 dihedral; (magenta) χ_1 analogue C4-C5-C6-S dihedral; (cyan) χ_2 analogue C5-C6-S-C dihedral. Capped cyanocysteine (bottom): (blue) χ_1 N-Cα-Cβ-S dihedral; (green) χ_2 Cα-Cβ-S-C dihedral.

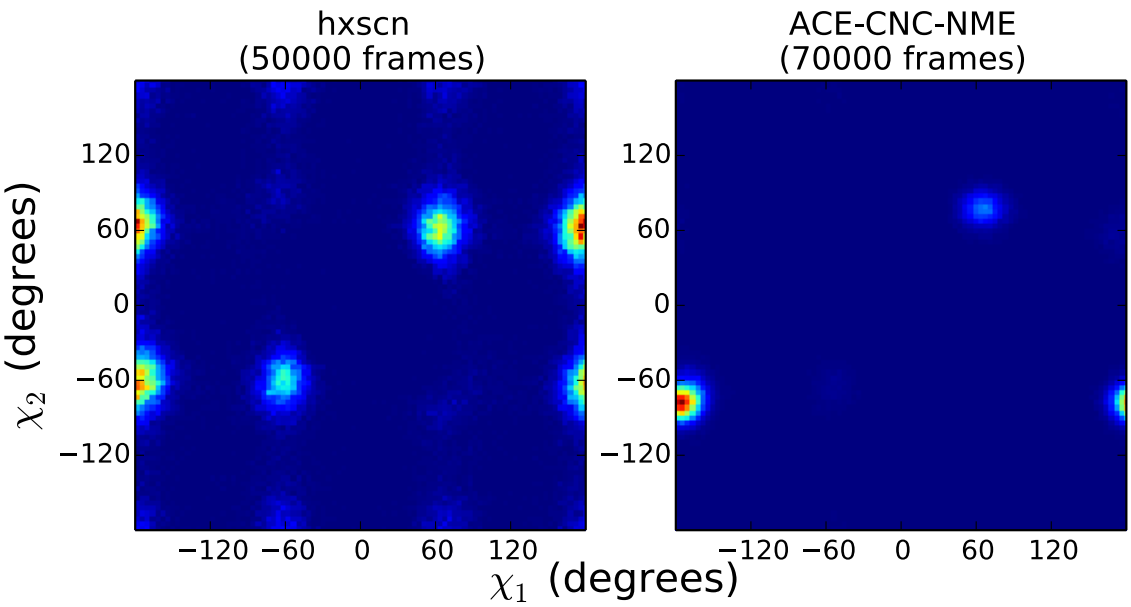

Figure 1-2: Two-Dimensional Dihedral Probability Distributions for Hexylthiocyanate and Capped Cyanocysteine

Hexylthiocyanate (left) χ_1 (C4-C5-C6-S) and χ_2 (C5-C6-S-C) analogous two-dimensional dihedral distribution after 5 ns of simulation. Capped cyanocysteine (right) χ_1 (N-Cα-Cβ-S) and χ_2 (Cα-Cβ-S-C) two-dimensional dihedral distribution after 7 ns of simulation.

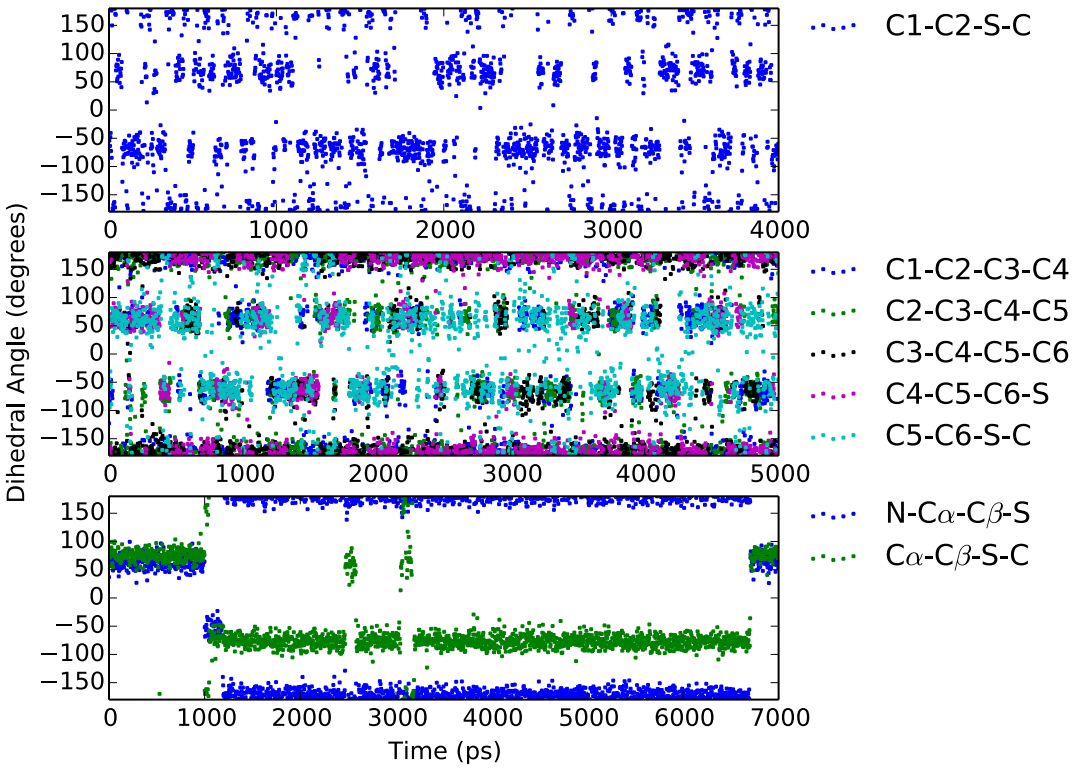

Figure 1-3: Dihedral Angles at Each Time Step

Non-Hydrogen dihedral angles as a function of time for ethylthiocyanate (top), hexylthiocyanate (middle), and capped cyanocysteine (bottom). Ethylthiocyanate (top): (blue) C1-C2-S-C dihedral. Hexylthiocyanate (middle): (blue) C1-C2-C3-C4 dihedral; (green) C2-C3-C4-C5 dihedral; (black) C3-C4-C5-C6 dihedral; (magenta) χ_1 analogue C4-C5-C6-S dihedral; (cyan) χ_2 analogue C5-C6-S-C dihedral. Capped cyanocysteine (bottom): (blue) χ_1 N-Cα-Cβ-S dihedral; (green) χ_2 Cα-Cβ-S-C dihedral.

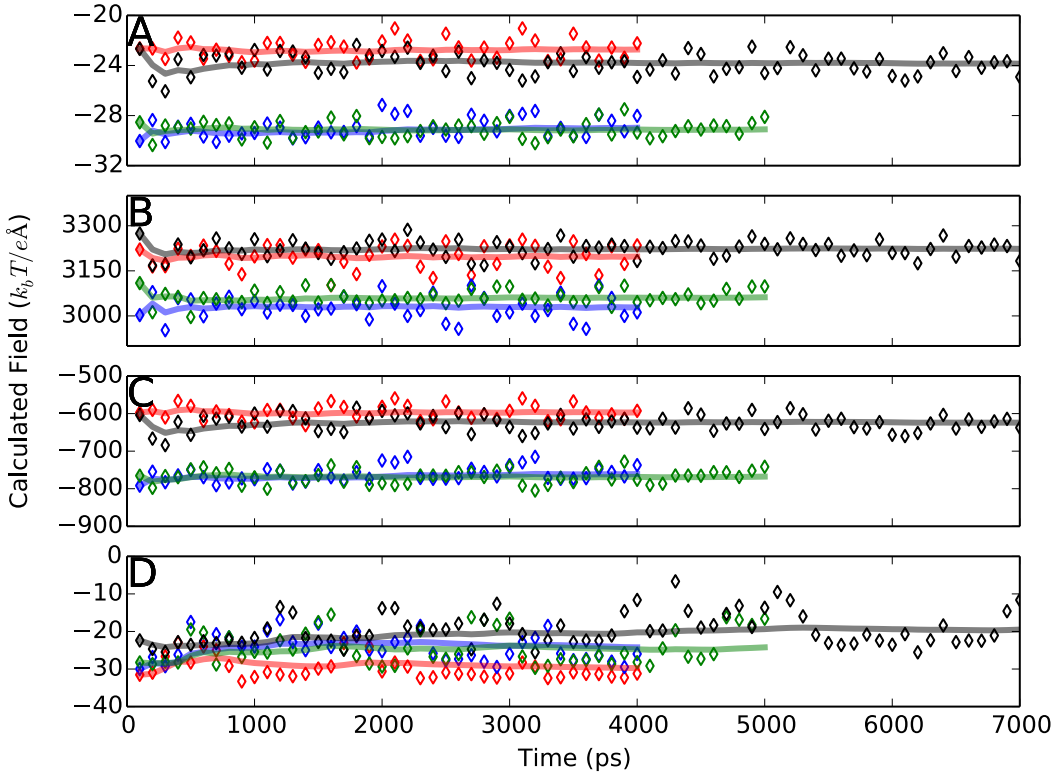

Figure 1-4: Average Electrostatic Field as a Function of Simulation Time

Average electrostatic field for methylthiocyanate (red), ethylthiocyanate (blue), hexylthiocyanate (green), and capped cyanocysteine (black) after some amount of frames, indicated on the x-axis. Diamonds are the average for the previous 1000 frames (100 ps).

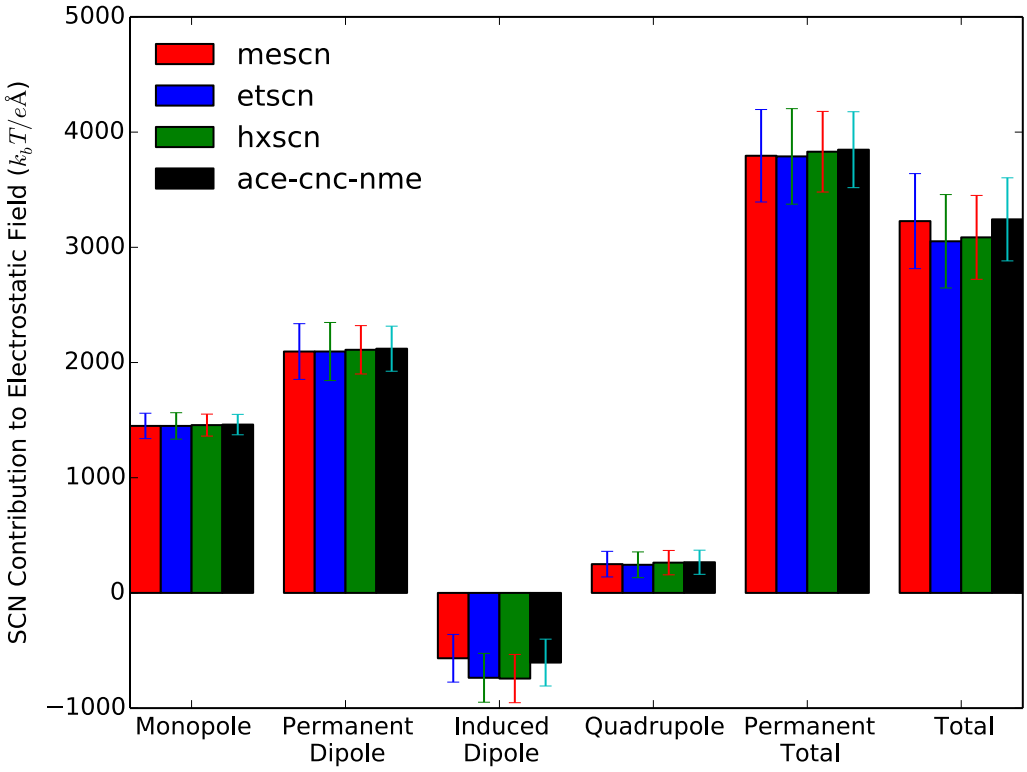

Figure 1-5: SCN Field Contributions are Constant

Contributions to the electrostatic field at the nitrile bond midpoint due to different multipole contributions from the SCN atoms. SCN is the dominating contributor to the electrostatic field due to the close proximity between the location of interest (field midpoint) and the SCN atoms.

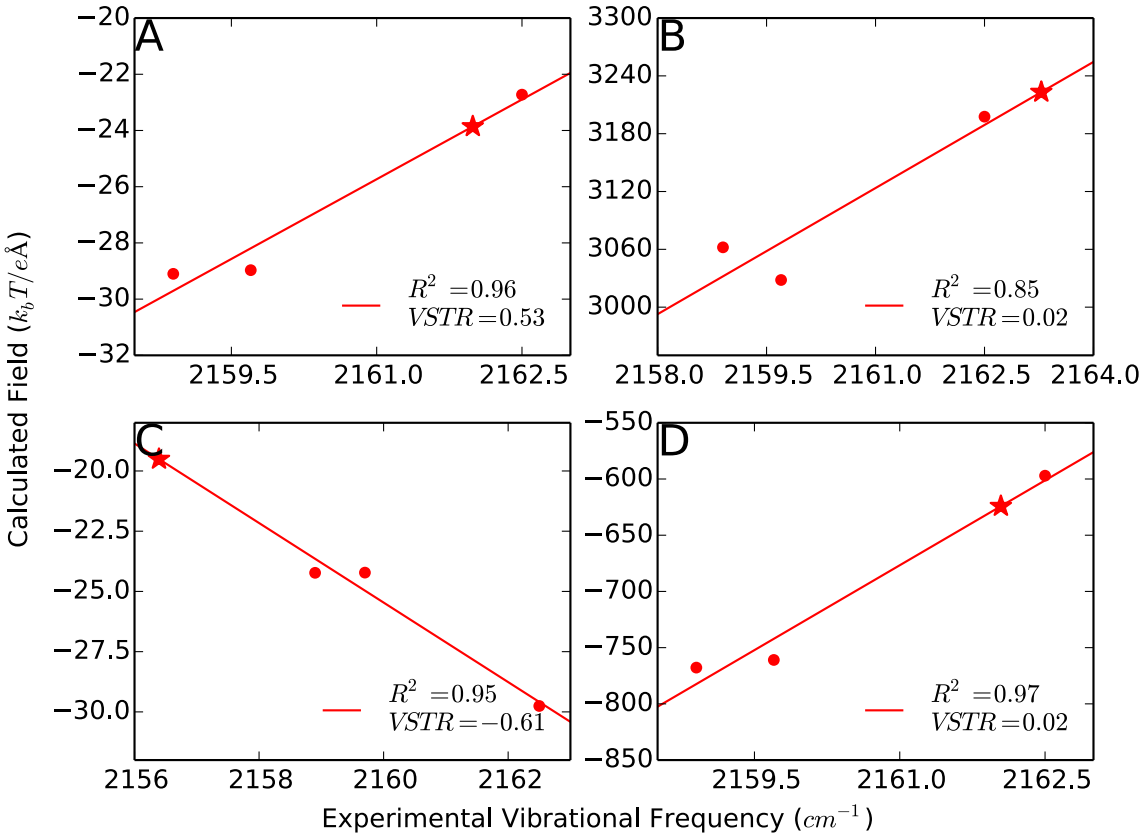

Figure 1-6: Small Molecule AMOEBA Fields Plotted Against Experimental Vibrational Absorption Energies

Calculated fields plotted against experimentally measured vibrational absorption energies. Squared correlation coefficients and VSTR are indicated in the bottom right corner of each plot. The capped cyanocysteine does not have an experimental absorption energy at this time and has therefore been speculated based on the best-fit equation obtain from the other three molecules and indicated with a star data point. A) Fields calculated with the IM. B) Total fields calculated using MPM. C) Fields calculated using MPM where the monopole, permanent dipole, *induced* dipole, and quadrupole fields due to SCN atoms have been removed. D) Fields calculated using MPM where the monopole, *permanent* dipole, and quadrupole fields due to SCN atoms have been removed.

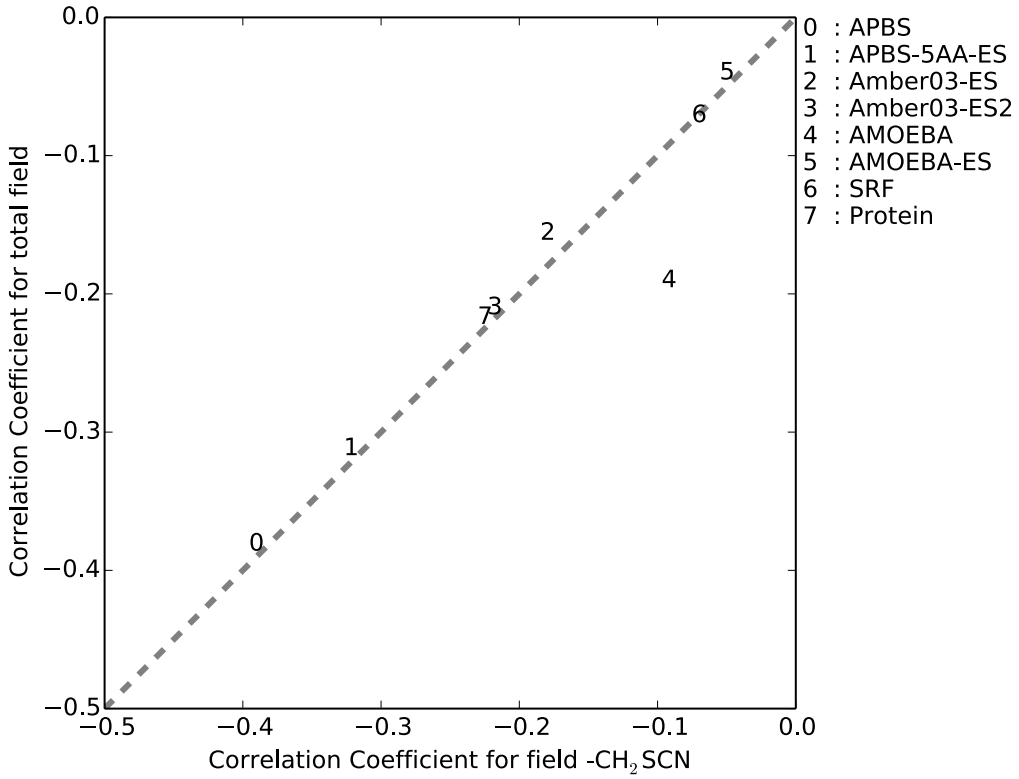

Figure 1-7: Comparing Total Field Correlations to Field Less CH_2SCN Correlations

The correlation between the total field and experimental vibrational absorption energies was compared to the field less CH_2SCN and experimental vibrational absorption energies. The dashed line is the line y=x, not a best-fit line. 0) APBS; 1) APBS 5 Å water sphere; 2) GROMACS explicit TIP3P reaction field electrostatics; 3) hybrid TIP3P reaction field electrostatics; 4) AMOEBA (CP and CPf have been excluded due to being nearly identical to without); 5) AMOEBA with explicit solvent (CP and CPf have been excluded due to being nearly identical to without); 6) the PB solvent reaction field; 5) the analytic Coulomb solute field.

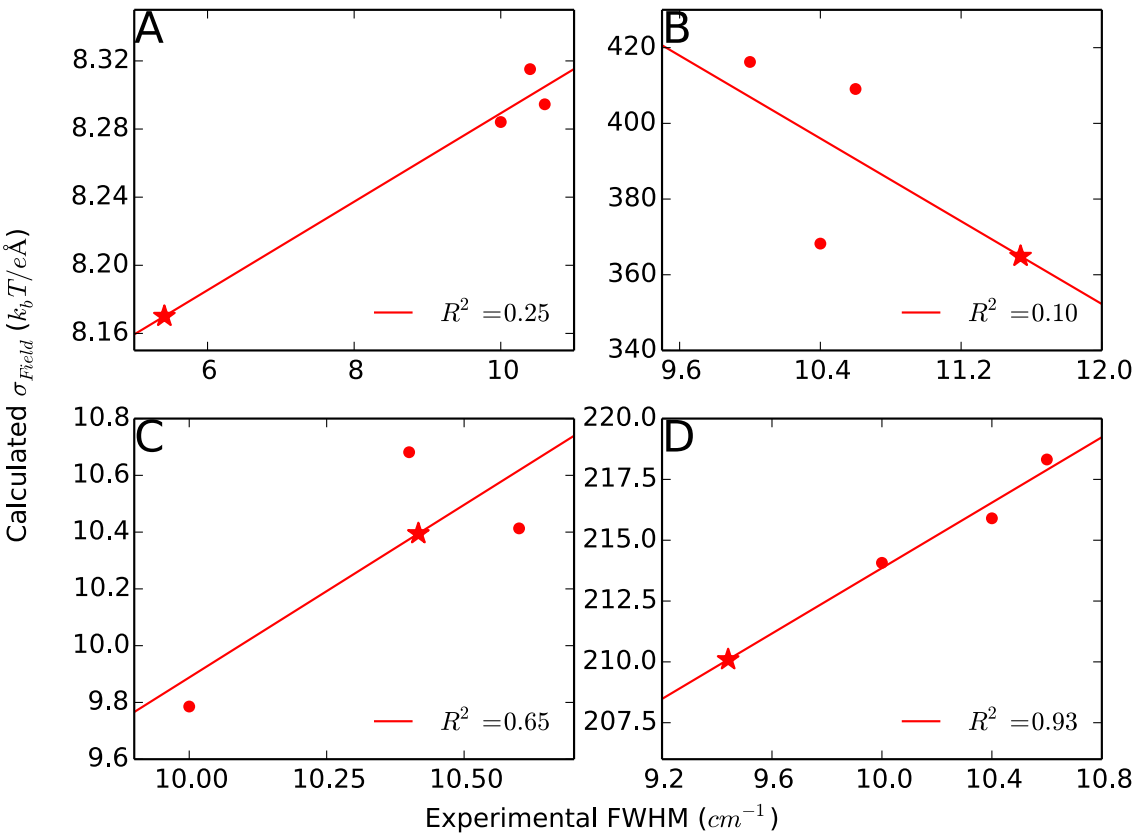

Figure 1-8: Small Molecule AMOEBA Field Standard Deviations Plotted Against Experimental FWHM

Calculated field standard deviations plotted against experimentally measured full width at half peak maximum. Squared correlation coefficients are indicated in the bottom right corner of each plot. The capped cyanocysteine does not have an experimental absorption energy at this time and has therefore been speculated based on the best-fit equation obtained from the other three molecules and indicated with a star data point. A) Fields calculated with the IM. B) Total fields calculated using MPM. C) Fields calculated using MPM where the monopole, permanent dipole, *induced* dipole, and quadrupole fields due to SCN atoms have been removed. D) Fields calculated using MPM where the monopole, *permanent* dipole, and quadrupole fields due to SCN atoms have been removed.

References

1. Fried, S. D.; Wang, L. P.; Boxer, S. G.; Ren, P. Y.; Pande, V. S., Calculations of the Electric Fields in Liquid Solutions. *J Phys Chem B* **2013,** *117* (50), 16236-16248.