THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

# Ab Initio Quantum Chemistry for Protein Structures

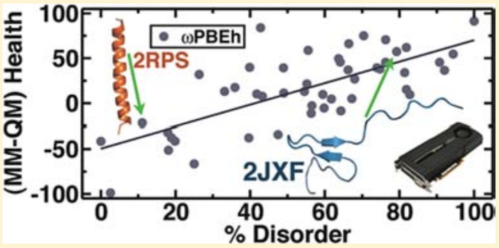
Heather J. Kulik,[†,‡] Nathan Luehr,[†,‡] Ivan S. Ufimtsev,[†,‡] and Todd J. Martinez*[,†,‡]

[†]Department of Chemistry and PULSE Institute, Stanford University, Stanford, California, 94305, United States

[‡]SLAC National Accelerator Laboratory, Menlo Park, California 94025, United States

S *Supporting Information*

**ABSTRACT:** Structural properties of over 55 small proteins have been determined using both density-based and wave-function-based electronic structure methods in order to assess the ability of ab initio "force fields" to retain the properties described by experimental structures measured with crystallography or nuclear magnetic resonance. The efficiency of the GPU-based quantum chemistry algorithms implemented in our TeraChem program enables us to carry out systematic optimization of ab initio protein structures, which we compare against experimental and molecular mechanics force field references. We show that the quality of the ab initio optimized structures, as judged by conventional protein health metrics, increases with increasing basis set size. On the other hand, there is little evidence for a significant improvement of predicted structures using density functional theory as compared to Hartree–Fock methods. Although occasional pathologies of minimal basis sets are observed, these are easily alleviated with even the smallest double-$\zeta$ basis sets.



## INTRODUCTION

Recent algorithmic advances[1−3] connected with the introduction of novel streaming architectures, such as graphical processing units (GPUs), have enabled the application of quantum chemistry methods to large molecular structures, such as polypeptides and proteins.[4,5] An open question is the expected accuracy of these ab initio quantum chemistry methods for protein structures, especially as compared to the empirical force fields that have been developed and extensively validated for this purpose.[6,7] A major advantage of ab initio methods in the biophysical context is the ability to describe phenomena that cannot be easily described with force fields, such as electronic polarization, charge transfer, and chemical reactivity. However, this advantage may be of limited utility if tractable ab initio methods are incapable of describing the structure accurately. Our goal here is, therefore, to validate ab initio methods for protein structures.

In this work, we harness the advances[8] of electronic structure on GPUs to quantify protein structure descriptions by ab initio methods on an expansive protein data set. Previously, such structural studies have been intractable without resorting to empirical force fields[9] or semiempirical methods.[10] We obtain starting protein structures by completing a systematic search of the Protein Databank (PDB).[11] We validate ab initio quantum chemistry against experimental and force field results with gas-phase and aqueous environment structural optimizations on a data set that includes a variety of secondary structure motifs and amino acid abundance comparable to the PDB. This first large-scale test of ab initio approaches for the description of protein structures does not include expansive consideration of entropic effects, focusing instead on the retention of the structure in local minima nearest the experimental structure. We limit consideration of differences between alternate conformations to select prototypical cases. Comparison between ab initio structures and protonation states on this subset of proteins with force fields and experimental structures permits us to identify fundamental differences obtained both when using an ab initio approach compared to empirical force fields and between various ab initio methods.

Most of the calculations we carry out here are based on local optimization (starting from experimental structures) for proteins in isolation. The experimental structures typically refer to either crystal or solution environments. Thus, it is important to note that the structures of proteins in the gas phase (or even in low dielectric organic solvents) could differ from those in solution, although the extent of these differences remains an open question.[12,13] For example, diminished dielectric screening will enhance electrostatic interactions, such as those present in salt bridges between charged residues.[14] A further potential difference between gas-phase and solvated protein structures lies in the charge state of the terminal residues. Although the terminal residues of peptides solvated by aqueous solutions almost invariably exist in a zwitterionic form,[15−17] this may not be the case in low dielectric environments. There is increasing evidence[18−20] that the zwitterionic charge separation is stabilized by salt bridges between residues in the gas phase. Although the determination of gas-phase protein structures is itself an important area of research, relevant, for example, in mass spectrometric analysis methods,[21−24] detailed treatment of this problem requires explicit consideration of many possible charge states for the protein and protonation states of the residues. We do not

**Received:** August 4, 2012
**Revised:** August 29, 2012
**Published:** September 13, 2012

attempt exhaustive searches over alternative protonation and charge states here, but note that ab initio methods are ideally suited to this problem, while a force-field-based approach would require extensive modification and reparameterization.[25] In fact, density functional theory (DFT) has been used to predict the most stable gas-phase charge states and structures in a few peptides with up to 20 residues.[26]

We have also carried out local optimizations in aqueous environments for the proteins that exhibited the largest structural deviations from experiment. These calculations use a hybrid quantum mechanical/molecular mechanical (QM/MM) method.[27,28] Unlike most previous applications of QM/MM, the entire protein is treated quantum mechanically and only the solvent is treated with an empirical force field. These calculations allow us to determine the extent to which environment effects are responsible for any structural discrepancies between theory and experiment.

## ■ PROTEIN SELECTION METHODS

We selected candidate protein structures from the Protein Databank (PDB)[11] by eliminating structures with 30% or more sequence identity and restricting the search to structures without ligands, nucleic acids, modified residues, or multiple chains. We further reduced the scope of our search to small polypeptides ranging from 5 to 35 residues in length. From this subset of 413 candidate polypeptides, we considered only structures with a total charge ≤ ±2, as evaluated from an empirical protonation state determination procedure,[29−31] giving a total of 58 proteins (70−590 atoms in each structure, 272 atoms on average) in our data set. For proteins in the PDB where an ensemble of experimental structures are provided (e.g., in solution NMR), the first model is always used for comparisons between experiment and theory and as a starting point for geometry optimizations. Characteristics of this protein data set, including histograms of the percent of secondary structure properties in proteins[32] and the frequency of amino acids in the primary structure of the proteins, are depicted in Figure 1. A summary of the PDB ID, type, size, and characteristics of proteins in the test set is provided in the Supporting Information.

As expected for a data set of small polypeptides,[33,34] the $\alpha$-helical secondary structure predominates over the $\beta$-sheet structure.[32] Thus, we sample the accuracy of electronic structure approaches to describe sheet structures with less detail. The majority of the residues in the protein data set are nonpolar (51%), with smaller contributions from polar (31%) and charged (18%) residues, largely in agreement with the relative abundance of amino acids in the human genome[35] and PDB.[11] Our set undersamples charged residues, especially histidine, because histidine is a good chelator, and we eliminated metal-containing proteins from our data set.[36] We slightly oversample polar residues, most notably cysteine, likely due to the enhanced use of disulfide bonds to restrict the structure in small, engineered polypeptides.[37,38]

## ■ COMPUTATIONAL METHODS

We assessed the ability of both restricted Hartree−Fock (RHF) and a number of density functional theory (DFT) approaches to reproduce structural properties observed in experimental X-ray crystallographic and nuclear magnetic resonance structures for several standard basis sets with our TeraChem package.[8] In the case of DFT, we attempted to study all 58 proteins with the BLYP,[39,40] B3LYP,[41] and $\omega$PBEh[42] functionals, but shortcomings of the BLYP and B3LYP approaches (specifically, convergence difficulties due to the well-known band-gap problem[43−48]) prevented electronic structure convergence and optimization of the complete protein data set. It is becoming clear that convergence difficulties are a pervasive problem when dealing with DFT for large polypeptides.[5,49] Thus, we focus here on results obtained with the range-corrected exchange-correlation functional ($\omega$PBEh) and RHF. Both of these approximations include the full strength long-range exact exchange interactions that are vital to avoid/minimize the self-interaction and delocalization errors[50] that lead to unrealistically small HOMO−LUMO gaps for some exchange-correlation functionals. Using $\omega$PBEh and RHF, we were able to optimize over 55 structures with the minimal basis sets, MINI[51] and STO-3G,[52] and also the larger double-$\zeta$ 3-21g[53] and 6-31g[54] basis sets. Calculations with MINI were motivated by previous observations that basis set superposition errors (BSSE) might be reduced with this basis set as compared to STO-3G.[55] Additional RHF calculations were carried out using the MINI basis set and including dispersion effects with the DFT-D3 empirical correction (referred to here as MINI-D).[56] Results for the semilocal BLYP and hybrid B3LYP (15 and 35 structures, respectively) using both the minimal STO-3G and the larger 3-21g and 6-31g basis sets are presented in the Supporting Information.

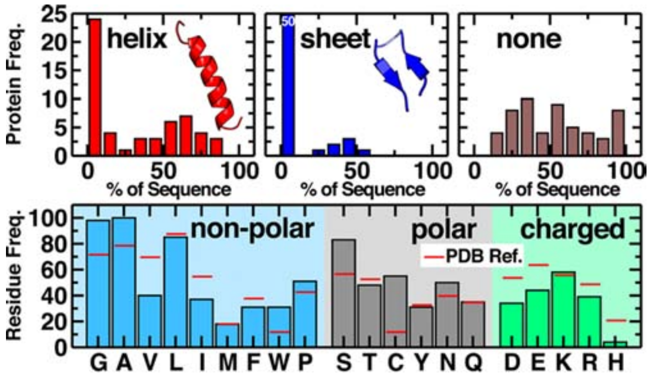


Figure 1. Histograms of primary and secondary structure properties of the 58-protein data set used for ab initio calculations. The upper three panels provide histograms of the percentage of helical, sheet, or unassigned secondary structures (note that the bin corresponding to 0% sheet in the upper middle panel is truncated, and its value is indicated in the corresponding bar). The lower panel shows the frequency of individual amino acids (labeled by a single letter code), which are nonpolar, polar, or charged, in the sequence space of the entire data set. This is compared to the expected frequency (red lines) for a data set of this size based on the entire set of proteins in the PDB.

Empirical force field optimizations for comparison with ab initio results were carried out with AMBER using the ff03 force field.[57] Quantum mechanical (QM) calculations combined with molecular mechanics (MM) treatment of a spherical shell of surrounding TIP3P[58] water molecules (QM/MM) embedded in a dielectric continuum using the self-consistent reaction field model[59] were carried out using TeraChem for 20 of the proteins in our data set. In these cases, the spherical shell of water (10 Å radius) was generated with the tleap component of AMBER. In one case discussed below (HIV-1 epitope, PDB ID: 1LB0; 226 atoms), we have used the nudged elastic band (NEB) method[60] to characterize reaction pathways between distinct structural minima. Ab initio and empirical force field

**Table 1. Average Health Checks, including $C_\alpha$ RMSD (Å), Clashes per Thousand, % Bad Side-Chain Rotamers, Number of $C_\beta$ Deviations, and % Favorable Ramachandran Values, for Several Levels of Theory Compared to the Experimentally Determined Reference Structure**

| | $C_\alpha$ RMSD | | | | | | |
|---|---|---|---|---|---|---|---|
| method | avg | min | max | clashes/1000 | poor rotamers | $C_\beta$ dev | good Ramachandran |
| AMBER | 0.57 | 0.23 | 0.94 | 3 | 9% | 0.18 | 80% |
| RHF/MINI | 0.68 | 0.31 | 1.49 | 45 | 18% | 0.48 | 80% |
| RHF/MINI-D | 0.67 | 0.27 | 1.32 | 44 | 18% | 0.52 | 78% |
| RHF/STO-3G | 0.71 | 0.22 | 1.34 | 40 | 15% | 0.22 | 75% |
| RHF/3-21g | 0.68 | 0.23 | 1.68 | 14 | 11% | 0.28 | 86% |
| RHF/6-31g | 0.74 | 0.24 | 1.44 | 8 | 9% | 0.28 | 86% |
| ωPBEh/MINI | 0.69 | 0.17 | 1.26 | 75 | 24% | 1.10 | 69% |
| ωPBEh/STO-3G | 0.70 | 0.29 | 1.38 | 72 | 19% | 0.50 | 71% |
| ωPBEh/3-21g | 0.69 | 0.32 | 1.48 | 21 | 15% | 0.46 | 81% |
| ωPBEh/6-31g | 0.65 | 0.24 | 1.41 | 9 | 13% | 0.41 | 85% |
| experiment | | | | 9 | 19% | 0.31 | 80% |

NEB calculations were carried out using the implementations in TeraChem and AMBER, respectively. Here, we also used a dielectric continuum model to represent the effect of the surrounding aqueous solvent ($\varepsilon = 78.39$) on geometries along the reaction path (optimized in the absence of solvent). This protein (PDB ID: 1LB0) was solved experimentally by NMR in aqueous solution, justifying the choice of dielectric constant used here. These continuum solvation (conductor-like screening model, COSMO[61]) calculations were carried out with the Q-Chem quantum chemistry package[62] using the switching/Gaussian approach[63,64] for the discretization of the solvent-accessible surface with 110 Lebedev points/atom. Optimized protein structures were evaluated using health criteria that were computed by the MolProbity suite.[65] Reported root-mean-square deviations (RMSDs) are given with respect to the corresponding experimental structure from the PDB, after placing structures in maximal alignment. We used a cutoff of 1.4 Å to define covalent (N/O)−H bonds, while bond lengths in the range of 1.4−2.0 Å were considered as (N/O)−H hydrogen bonds.[66]

## RESULTS

Structural properties of the initial experimental protein data set were evaluated and compared against the same properties for the ab initio and force field optimized structures. The aggregated results are summarized in Table 1 (detailed information for the individual structures is available in the Supporting Information). The average backbone root-mean-squared deviation (RMSD) for these polypeptides with respect to the starting experimental structure is between 0.57 and 0.77 Å for all theoretical methods, and a comparison of the individual RMSD values is provided in the Supporting Information. For reference, the average RMSD between different models of solution NMR structures in our data set is 1.73 Å, likely owing to the lower secondary structure content in these small proteins compared to larger, more globular proteins.[34] The other "health checks" we use are as follows: (1) clashes, or steric overlaps > 0.4 Å, per 1000 atoms, (2) the percentage of poor side-chain dihedrals or rotamers, (3) the number of $\beta$-carbon deviations > 0.25 Å from the predicted position expected based on the backbone coordinates,[67] and (4) the percentage of backbone dihedrals that fall into a favored region on a Ramachandran plot. All of these health checks are given as reported by MolProbity, and the detailed definitions of these quantities can be found in previous literature.[65] Comparison of the average of these criteria over all proteins in each data set and for each method reveals that residue clashes vary the most significantly across methods and basis set size. Minimal basis sets, regardless of the use of RHF or DFT methods, produce an order of magnitude higher clashing values (40−75 per 1000) with respect to the starting value in the experimental structures (9 per 1000), force field results (3 per 1000), and the 6-31g basis set results (8−9 per 1000).

Comparison of the RHF/MINI and RHF/MINI-D results shows that the inclusion of empirical dispersion effects does not improve the results significantly (by any of the metrics in Table 1), indicating that the errors in the small basis set results are likely dominated by the limited electronic flexibility and/or BSSE. Furthermore, comparison of MINI and STO-3G results suggests little difference between these methods. Given previous suggestions[55] that MINI was less sensitive to BSSE, it is likely that the limited electronic flexibility is the major error in the small basis set results. Indeed, the results with the smallest of the double-$\zeta$ basis sets (3-21g) are already a significant improvement over the minimal basis set results with respect to clashing (14−21 per 1000) and other health checks. Since these calculations only require about 50% more computational time than STO-3G calculations, it seems prudent to use double-$\zeta$ basis sets, at least until some correction for the limited flexibility of minimal basis sets is introduced. We note that an empirical geometry-dependent BSSE correction scheme for minimal basis sets has recently been proposed,[68] and the accuracy of this scheme will be investigated in future work.

One should expect the "health checks" to produce the best values for force field approaches, since the energy functions of force fields are optimized to reproduce the properties of "healthy" folded experimental structures.[69,70] Nevertheless, both RHF and ωPBEh results with the 6-31g basis set exhibit health checks nearly as good as the force field approach in most categories and improve upon force fields with respect to the percent of dihedrals lying in favorable regions on Ramachandran plots. Importantly, there is evidence that the force field approach may produce "healthier" structures even when the experimental results suggest "unhealthy" anomalies in the structure. For example, the $\beta$-carbon deviation computed from empirical force field optimizations is anomalously low compared with *both* the experimental structures and the ab initio optimizations. This is strongly suggestive of a bias toward "healthy" structures in the empirical force fields, indicating that ab initio approaches may be more reliable in the prediction and assessment of structures for proteins with unusual and/or disordered structural characteristics.

We highlight one specific example where the empirical force field predicts a "healthier" structure than both ab initio optimization and experiment. The failures of the empirical force field in this example are representative of other cases in the data set. The complete structures and health characteristics of all the proteins in the data set are provided in the Supporting Information for the reader interested in further details. The example we choose here is the crustacean cardioactive peptide (CCAP) of the fruit fly (PDB ID: 1Y49),[71] a highly conserved nonapeptide that is made cyclic by a disulfide bridge between the third and ninth residues. As shown in Figure 2, the residues close to the disulfide bridge have unusual backbone dihedrals and side-chain rotamers in both ab initio and experimental structures. In contrast, force field optimization enforces a more prototypically favorable structure around the disulfide bridge, at variance with the experimental structure.

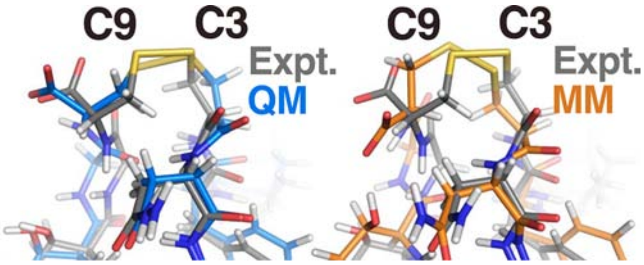


**Figure 2.** Comparison of experimental (gray, Expt.) and ωPBEh/6-31g (blue, QM) optimized structures (left panel) and comparison of experiment (gray, Expt.) to force field (orange, MM) optimized structures (right panel) around the Cys9−Cys3 disulfide bridge of CCAP (PDB ID: 1Y49).

General trends in our data set are also evident: the low secondary structure content (e.g., steric zippers, PDB ID: 3FTR)[72] or perturbed secondary structure (e.g., helical inhibitor of papillomavirus, PDB ID: 1RIJ)[73] in proteins is more consistently described between experiment and ab initio methods than with force fields. In some cases, the less-ordered proteins (e.g., ubiquitin fragments, PDB ID: 2JTA;[74] hemocyte migrating peptide, PDB ID: 2RPS;[75] metallothionein, PDB ID: 1T2Y[76]) exhibit "healthier" characteristics in ab initio results compared with the experimental structures, likely because these structures are more challenging to refine. On the other hand, prototypical secondary structures, particularly $\alpha$-helices (e.g., a human serum apolipoprotein component, PDB ID: 1ODP[77]), seem to be better described by force fields than ab initio methods, as judged by RMSD to experimental structures.

Overall, it appears that proteins with regions of disorder are better described by ab initio methods, while prototypically folded structures are better described by MM. We have quantified a measure of disorder for each protein structure in our data set as

$$\text{Disorder} = \frac{1}{2}\frac{N_{\text{res}}^{\text{unassigned-SS}}}{N_{\text{res}}} + \frac{1}{4}\frac{N_{\text{SS-int}}}{N_{\text{SS}}} + \frac{1}{4}\frac{N_{\text{res}}^{\text{atypical}}}{N_{\text{res}}} \quad (1)$$

where $N_{\text{res}}$ is the number of residues in the protein, $N_{\text{res}}^{\text{unassigned-SS}}$ is the number of residues with an unassigned secondary structure, $N_{\text{SS-int}}$ is the number of 1−2 residue insertions without a secondary structure or distinct secondary structure regions interrupting a run of the same secondary sequence type,[78] $N_{\text{SS}}$ is the number of distinct secondary sequence runs, and $N_{\text{res}}^{\text{atypical}}$ is the number of residues identified as having bad properties by MolProbity (i.e., bad side-chain rotamer or bad backbone). This disorder metric is evaluated using the experimental structures for each protein. Next, we define the relative health for a given protein structure as

$$\text{Relative Health} = \frac{\text{Health}_{\text{MM}} - \text{Health}_{\text{QM}}}{\text{Health}_{\text{Expt}}} \quad (2)$$

where $\text{Health}_{\text{MM}}$, $\text{Health}_{\text{QM}}$, and $\text{Health}_{\text{Expt}}$ are the MolProbity scores assigned to the force field optimized, ab initio optimized, and experimental structures, respectively. Note that large MolProbity scores correspond to less healthy protein structures, and thus negative values of Relative Health imply that a better description is provided by the empirical force field. In Figure 3, we plot the relative health versus the disorder for all structures in the data set, considering both the RHF/6-31g and the ωPBEh/6-31g ab initio methods. There is a clear correlation in that disordered proteins are treated better with QM, while more ordered proteins are treated equally well or better by the empirical force field. This correlation holds for both RHF and ωPBEh, and there is little detectable structural improvement from the inclusion of electron correlation in ωPBEh.

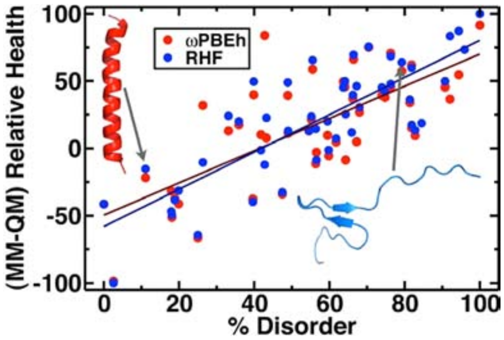


**Figure 3.** Plot of normalized relative health from QM (blue dots, RHF/6-31g; red dots, ωPBEh/6-31g) and MM optimizations as a function of protein disorder for a 51-protein data set. Negative values of relative health indicate that the MM-optimized structure is "healthier" as judged by MolProbity analysis. Trend lines are indicated in red and blue for ωPBEh and RHF, respectively. Structures of proteins with more disorder are better described by ab initio methods (both ωPBEh and RHF) than by empirical force fields. Structures representative of low (PDB ID: 2JXF) and high (PDB ID: 2RPS) disorder from the full data set are overlaid on the graph along with arrows pointing to the respective data points.

We now examine the steric overlaps between residues more closely for experimental, force field, and ab initio approaches (Figure 4). To eliminate any effects introduced by a nonuniform sampling of residues in our protein data set, we renormalized the clashing frequency of each residue by the abundance of that amino acid in our data set and further scaled these for each method so that the amino acid with the highest clashing frequency is listed as 100%. The relative frequency of steric overlap by residue was equivalent for both HF and DFT (ωPBEh) approaches, so we focus here on comparison of RHF/STO-3G, RHF/3-21g, and RHF/6-31g ab initio data. The greatest absolute amount of clashing is observed for minimal basis set ab initio results ($N$ = 172 for RHF/STO-3G,

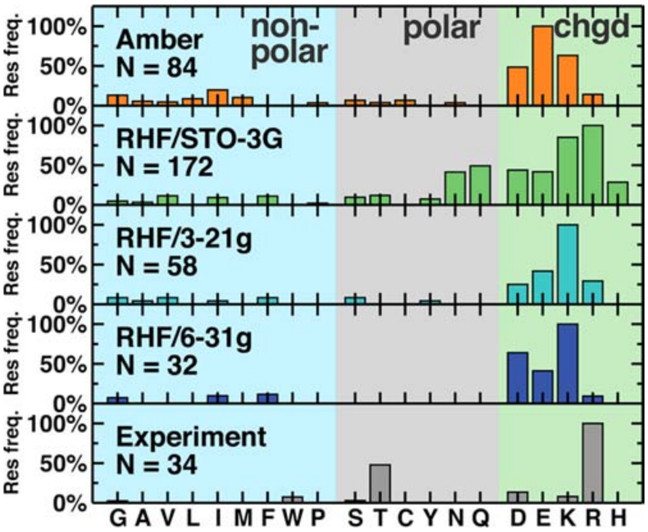


**Figure 4.** Relative clashing rate for each amino acid (for AMBER, RHF/STO-3G, RHF/3-21g, RHF/6-31g, and experimental structures from top to bottom) per occurrence in the protein data set. The residue with the highest frequency of clashing has been renormalized to 100%, and the total number of clashes (*N*) is indicated for each method.

where *N* is the total number of clashes observed in the optimized structures), while the largest basis set reduces the clashing to the lowest value ($N$ = 32 for RHF/6-31g). In all cases, clashing is most prominent for charged residues (D: aspartate; E: glutamate; K: lysine; and R: arginine), likely indicating that the steric overlap that arises is due to the formation of salt bridges or strong hydrogen bonds. The significant clashing occurring on polar residues (N: asparagine and Q: glutamine) for minimal basis set RHF cannot be explained in this way, however, and this steric overlap is likely indicative of a more fundamental difference in the protein structure evaluated at this basis set.

Bond length distributions for key peptide bond distances provide a further useful point of comparison for ab initio, experimental, and force field results, as shown in Figure 5.

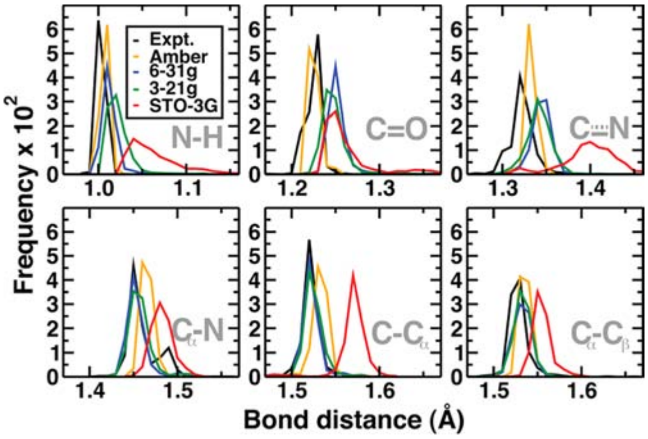


**Figure 5.** Bond length distributions for experimental structures (black line) compared to distributions obtained from optimization with AMBER (orange line), $\omega$PBEh/STO-3G (red line), $\omega$PBEh/3-21g (green line), and $\omega$PBEh/6-31g (blue line). Each panel has the same range of bond distances (0.2 Å) but is centered around the appropriate distance for the labeled bond type.

Unsurprisingly, parametrized force fields produce exceedingly narrow distributions with means that are in reasonable agreement with experimental values. However, the force field approach underestimates the width of the bond length distribution for carbon−oxygen and carbon−nitrogen bonds of the peptide bond, and it does not distinguish between C−C bonds formed between the amide carbon and $\alpha$-carbon or $\alpha$-carbon and $\beta$-carbon. In experimental structures, a broad distribution of C−O and C−N backbone bonds likely stems from the underlying resonance representations of this bond, which are adequately captured by the ab initio approaches, especially with a larger basis set. The larger basis set ab initio results also correctly predict C−C$_\alpha$ bonds to be slightly shorter than C$_\alpha$−C$_\beta$ bonds. Importantly, minimal basis set ab initio results have overly broad distributions for N−H and C−N bonds, where the former suggests deprotonation of amide nitrogens, while the latter suggests a decrease in C−N bond order. However, even a relatively small double-$\zeta$ basis set such as 3-21g recovers the correct bonding distributions and bond order for the amide bond.

Increased steric overlap and broadened bond distance distributions found in both ab initio and experimental structures (compared to the force field optimized structures) could be closely tied to the presence of strong hydrogen bonds or changes in protonation states. Focusing on the protein backbone, we considered whether shifting protonation states and character could be unearthed through examination of the number of protons forming covalent or hydrogen bonds to amide oxygen or nitrogen (Figure 6). Nearly all electronic structure methods (both RHF and $\omega$PBEh) and basis sets predict a higher number of hydrogen bonds with respect to experimental structures. This may be partly a consequence of the fact that the ab initio optimizations discussed so far were carried out without any representation of the solvent environment. Thus, the enhanced hydrogen bonding compared

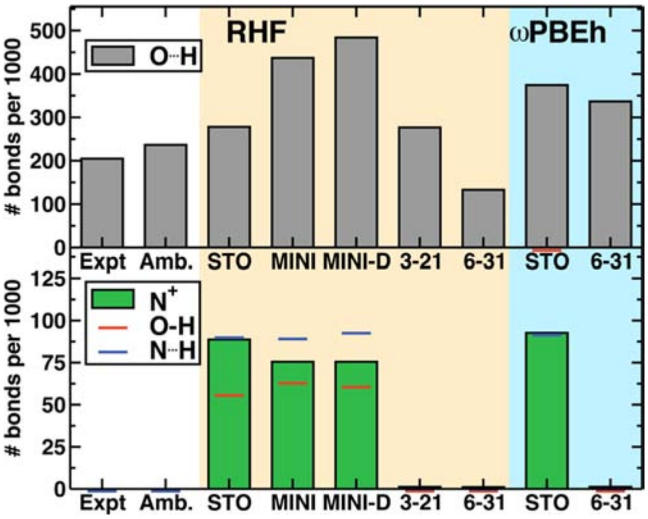


**Figure 6.** Character of protonation state and hydrogen bond of backbone amide species compared at several levels of theory against experiment. (Top) Number of carbonyl oxygens per thousand forming hydrogen bonds with neighboring species. (Bottom) Number of amide nitrogens that are deprotonated per thousand (green bar) referenced against the number of amide nitrogens that form a hydrogen bond with a neighboring species (blue line) and the number of backbone carbonyl oxygens that have been protonated (red line).

to experimental structures (which correspond to aqueous or hydrated crystal environments) may be partially due to the fact that hydrogen bonds are strengthened in the gas phase. However, detailed investigation of the minimal basis set results shows that the enhanced presence of hydrogen bonding also coincides with an unusual deprotonation of amide nitrogens. In some cases, a majority of the protons are shared between peptide bond nitrogens and oxygens, forming a covalent bond with the oxygen and a weaker hydrogen bond with the nitrogen. This result suggests that charge separation is disproportionately unfavorable in the gas phase using ab initio methods with minimal basis sets. An example of this behavior is shown for a protein fibril (PDB ID: 3FTR) in Figure 7. This anomalous internal proton transfer does not occur spontaneously when larger basis sets are used; that is, the larger basis sets do exhibit a local minimum with the expected protonation states of the peptide backbone. However, reoptimization of the anomalous minimal basis set structures with a larger basis set reveals that these alternative structures are also stationary points at the larger basis set. The calculations therefore predict that the anomalous protonation states are true metastable configurations and they should not be excluded when considering protein structures in the gas phase, as may be important in mass spectrometry experiments. Definitive results on the relative stability of these protonation states would require free energy calculations and possibly more extensive treatments of electron correlation, which is beyond the scope of the present work. Nevertheless, the anticipated coordination number of the amide nitrogen is recovered when slightly increasing the basis set size.

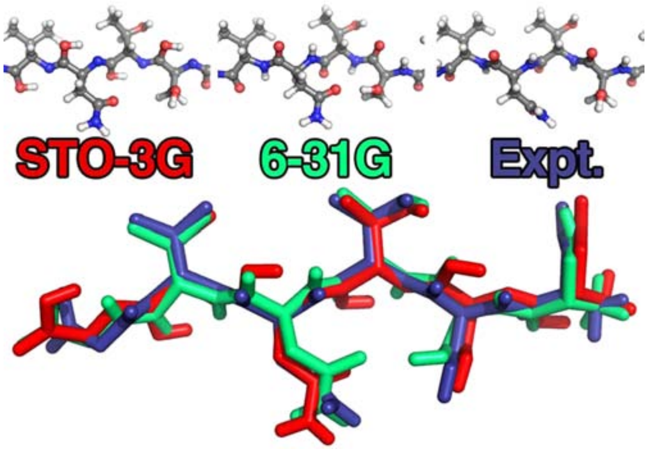


**Figure 7.** Comparison of experimental structure (upper right) for a protein fibril (PDB ID: 3FTR) with optimized structures from minimal basis RHF/STO-3G (upper left) and larger basis RHF/6-31g (upper middle). Note the anomalous protonation state in the minimal basis optimized structure that is absent in the structure optimized with a larger basis set. An overlay of all three structures is shown in the lower panel, with red, green, and blue structures corresponding to RHF/STO-3G optimization, RHF/6-31g optimization, and experimental coordinates, respectively.

Following the identification of unusual results from the electronic structure with minimal basis sets on zwitterionic, gas-phase protein structures, we now consider whether the minimal basis set structures can be improved either by neutralizing the terminal $NH_3^+$ and $COO^-$ residues or by explicit representation of the surrounding solvent. These calculations have been carried out for a subset (20 proteins) of the original data set that exhibited the largest deviations in RMSD compared with the experimental structures (details in Supporting Information). Calculations in explicit solvent were carried out by embedding the protein in a spherical shell of water molecules with a radius of 10 Å (positions of water molecules in this shell were determined by the tleap component of AMBER). Almost all (17 out of 20) of these protein structures were determined by solution NMR under aqueous conditions (see the Supporting Information). The water molecules were represented with the TIP3P force field (i.e., a QM/MM calculation). This spherical shell of protein and surrounding TIP3P water was further embedded in a dielectric continuum ($\varepsilon$ = 78.39) using an Onsager self-consistent reaction field model[59] (with the reaction field coupled to the solute through its charge and dipole moment). Solvation with explicit water molecules and a surrounding continuum eliminates the spontaneous amide backbone deprotonation described in the minimal basis set calculations above. It further improves the RMSD compared to the experimental structures and other health characteristics, as shown in Figure 8. Neutralizing the terminal residues in this subset also leads to improvement in the protein structure as judged by health metrics. This result suggests that some of the observed deviations between experimental and ab initio structures arises because of the interaction between the charged $NH_3^+$ and $COO^-$ terminal residues. This electrostatic attraction is screened in solution in the experiments and when the ab initio calculations include surrounding solvent. If the protein is modeled in isolation, this attraction can modify the structure. Such structural modifications are likely to occur in gas-phase proteins with charged terminal residues but may be minimized if the terminal residues are neutralized.

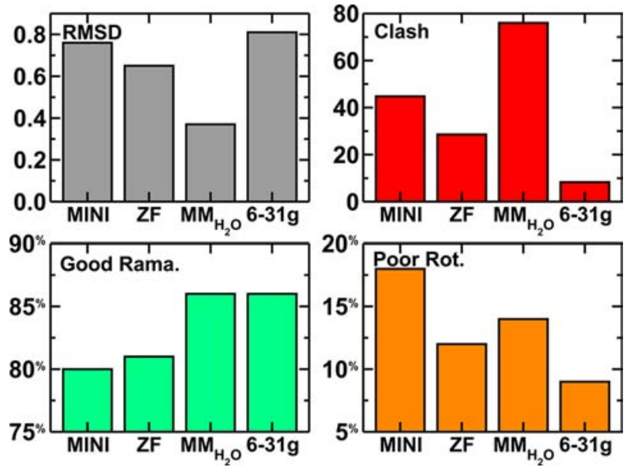


**Figure 8.** Health check metrics, including $C_\alpha$ RMSD, clash score, favorable Ramachandran values, and poor side-chain rotamers, for a minimal basis set (RHF/MINI) and larger basis set (RHF/6-31g) compared with RHF/MINI results where the zwitterionic termini are neutralized (ZF) or the protein is solvated with MM water molecules ($MM_{H_2O}$) and a surrounding polarizable continuum.

Finally, we consider the ability of electronic structure approaches to characterize the multiple local minima of an HIV-1 epitope (PDB ID: 1LB0) that adopts a $3_{10}$-helical structure rather than an $\alpha$-helical structure in an aqueous environment.[79] From the peptide sequence, we generated helices with dihedral angles ranging from a canonical $3_{10}$- ($\phi$ = 49°, $\psi$ = 26°) to an $\alpha$- ($\phi$ = 57°, $\psi$ = 47°) helical secondary structure. The path along these dihedral values was optimized with nudged elastic band[60] (NEB) for both the empirical force field and $\omega$PBEh/6-31g methods (Figure 9). In the gas phase, the ab initio results show that the $3_{10}$ helix is favored by about 6 kcal/mol over the $\alpha$ helix, with a significant barrier to interconversion occurring around ($\phi$ = 55°, $\psi$ = 40°). Since relative stabilities likely vary between gas-phase and solvated environments, we repeated COSMO calculations on the NEB images. We find that the barrier to interconversion is reduced, but the relative stability of the end point intermediates remains unchanged. Examination of the transition-state structure shows that unfavorable positioning of the central residues, particularly Lys7 and Trp8, and mismatching of backbone hydrogen bonding contributes to the barrier height. In contrast to both ab initio and experimental results, empirical force field NEB calculations predict that the $\alpha$-helical structure is favored over the $3_{10}$-helical structure by about 8 kcal/mol.

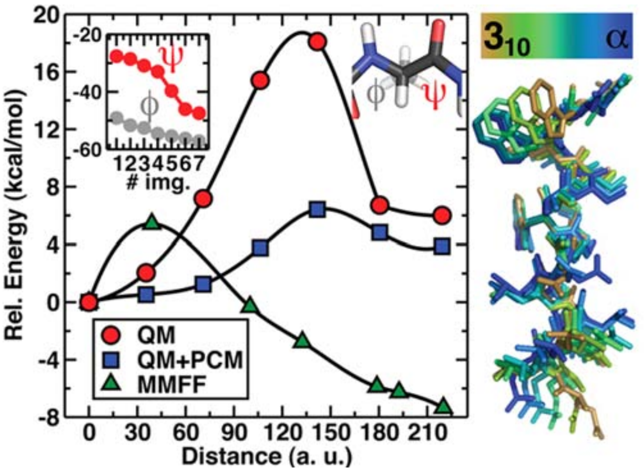


**Figure 9.** Left panel: minimum energy path (MEP) connecting $3_{10}$ and $\alpha$ helices for an HIV epitope (PDB ID: 1LB0), computed with AMBER (green triangles, MMFF) and ab initio $\omega$PBEh/6-31g (red circles, QM). Energetics including continuum solvation along the ab initio optimized MEP determined in isolation are also shown (blue squares, QM + PCM). Average backbone dihedrals $\phi$ and $\psi$ for each of the NEB images along the MEP are provided in the inset graph. Right panel: overlay of conformers along the AMBER-optimized MEP color-coded from orange to blue.

## ■ CONCLUSIONS

Ab initio methods are uniquely poised to answer the open questions of both the relative stability of protonation states and how these influence the protein structure. In this work, we have leveraged recent developments in both algorithms and computer architectures that bring ab initio calculations of optimized protein structures within reach. These advances allowed us to optimize the structures of more than 50 proteins ranging in size up to 35 residues. Each of these structures was optimized using at least four different basis sets (STO-3G, MINI, 3-21g, and 6-31g) and two different levels of theory (RHF and DFT with the range-corrected $\omega$PBEh functional).

Our results show that unbiased ab initio methods are able to retain the protein structure at the same level as empirical force fields that have been extensively parametrized for this purpose. Furthermore, there is evidence that the ab initio methods are consistently better than force fields at reproducing experimental structures for proteins with disordered regions. We highlighted a few examples from our data set where unusual structural motifs observed in the experimental data are reproduced by ab initio methods, but not by the empirical force field.

We also uncovered unusual internal proton transfer for some of the structures when using minimal basis sets. Further investigation showed that larger basis sets predicted the existence of a minimum at the expected structure but also that the unexpected structure remained a local minimum that might be accessible at elevated temperatures. This result encourages the use of larger basis sets that do not undergo spontaneous internal proton transfer. However, it also highlights the ability of ab initio methods to describe chemical phenomena that cannot be described easily with empirical force fields.

Finally, for a subset of the proteins that exhibited the largest deviations from experimental structures when optimized in isolation, we used a QM/MM approach (where the entire protein is modeled with QM) to carry out optimizations in an aqueous environment. This led to improvements in the structure compared to experimental data, suggesting that the relatively small deviations observed when optimizing protein structures in isolation are due to true differences in the structure of the isolated and hydrated protein. Empirical force fields are less sensitive to the immediate environment of the protein, which is perhaps expected since force fields are parametrized to reproduce structural properties in solution.

Our results here show that even relatively low-level ab initio methods provide a useful description of protein structures. There is little discernible difference between RHF and DFT with respect to structural features, although it is likely that relative energies of different conformations are significantly improved with DFT since it includes electron correlation effects. Further investigations including entropic effects and sampling alternate conformations can be expected to be possible in the near future, and this will be required to determine the accuracy of ab initio methods in predicting the free energy landscapes of proteins.

## ■ ASSOCIATED CONTENT

### Ⓢ Supporting Information

Details of proteins included in the data set used, including PDB IDs and secondary structure assignment; details of health checks for all optimized structures; and coordinates for all optimized structures. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: todd.martinez@stanford.edu.

### Notes

The authors declare the following competing financial interest(s): T.J.M. and I.S.U. are co-founders of PetaChem, LLC.

## ■ ACKNOWLEDGMENTS

We thank Prof. S. Grimme for providing optimized DFT-D3 dispersion parameters for RHF with the MINI basis set. This work was supported by the NSF (CHE-10-47577) and the Department of Defense (Office of the Director of Defense Research and Engineering) through a National Security Science and Engineering Faculty Fellowship. This research used computational resources provided through DOE Contract

No. DE-AC02-7600515 and the Keeneland Computing Facility at the Georgia Institute of Technology, which is supported by the National Science Foundation under Contract No. OCI-0910735. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. T.J.M. and I.S.U. are cofounders of PetaChem, LLC.

## ■ REFERENCES

(1) Ufimtsev, I. S.; Martínez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222−231.

(2) Ufimtsev, I. S.; Martínez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619−2628.

(3) Ufimtsev, I. S.; Martínez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004−1015.

(4) Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. *J. Phys. Chem. Lett.* **2011**, *2*, 1789−1793.

(5) Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2011**, *7*, 1814−1823.

(6) MacKerrell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(7) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; et al. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(8) Petachem. http://www.petachem.com.

(9) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.

(10) Stewart, J. J. P. *J. Mol. Model.* **2009**, *15*, 765−805.

(11) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535−542.

(12) Breuker, K.; Brueschweiler, S.; Tollinger, M. *Angew. Chem., Int. Ed.* **2011**, *50*, 873−877.

(13) Mattos, C.; Ringe, D. *Curr. Opin. Struct. Biol.* **2001**, *11*, 761−764.

(14) Marchese, R.; Grandori, R.; Carloni, P.; Raugei, S. *PLoS Comput. Biol.* **2010**, *6*, e1000775.

(15) Xu, S. J.; Nilles, M.; Bowen, K. H. *J. Chem. Phys.* **2003**, *119*, 10696−10701.

(16) Wyttenbach, T.; Liu, D. F.; Bowers, M. T. *Int. J. Mass Spectrom.* **2005**, *240*, 221−232.

(17) Bush, M. F.; Prell, J. S.; Saykally, R. J.; Williams, E. R. *J. Am. Chem. Soc.* **2007**, *129*, 13544−13553.

(18) Rodgers, M. T.; Campbell, S.; Marzluff, E. M.; Beauchamp, J. L. *Int. J. Mass Spectrom. Ion Processes* **1995**, *148*, 1−23.

(19) Summerfield, S. G.; Whiting, A.; Gaskell, S. J. *Int. J. Mass Spectrom. Ion Processes* **1997**, *162*, 149−161.

(20) Cerda, B. A.; Wesdemiotis, C. *Analyst* **2000**, *125*, 657−660.

(21) Patriksson, A.; Marklund, E.; Van Der Spoel, D. *Biochemistry* **2007**, *46*, 933−945.

(22) Benesch, J. L. P.; Ruotolo, B. T.; Simmons, D. A.; Robinson, C. V. *Chem. Rev.* **2007**, *107*, 3544−3567.

(23) Jarrold, M. F. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1659−1671.

(24) Konermann, L. *J. Phys. Chem. B* **2007**, *111*, 6534−6543.

(25) Kirschner, K. N.; Lewin, A. H.; Bowen, J. P. *J. Comput. Chem.* **2003**, *24*, 111−128.

(26) Dal Peraro, M.; Raugei, S.; Carloni, P.; Klein, M. L. *ChemPhysChem* **2005**, *6*, 1715−1718.

(27) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.

(28) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389−427.

(29) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Int. J. Mass Spectrom. Ion Processes* **2005**, *33*, W368−W371.

(30) Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. *Proteins* **2006**, *63*, 928−938.

(31) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(32) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.

(33) Gellman, S. H. *Curr. Opin. Chem. Biol.* **1998**, *2*, 717−725.

(34) Banerjee, A.; Datta, S.; Pramanik, A.; Shamala, N.; Balaram, P. *J. Am. Chem. Soc.* **1996**, *118*, 9477−9483.

(35) Echols, N.; Harrison, P.; Balasubramanian, S.; Luscombe, N. M.; Bertone, P.; Zhang, Z. L.; Gerstein, M. *Nucleic Acids Res.* **2002**, *30*, 2515−2523.

(36) Berg, J. M. *J. Biol. Chem.* **1990**, *265*, 6513−6516.

(37) Wells, J. A.; Estell, D. A. *Trends Biochem. Sci.* **1988**, *13*, 291−297.

(38) Ainavarapu, R. K.; Brujic, J.; Huang, H. H.; Wiita, A. P.; Lu, H.; Li, L.; Walther, K. A.; Carrion-Vazquez, M.; Li, H.; Fernandez, J. M. *Biophys. J.* **2007**, *92*, 225−233.

(39) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(40) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(41) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(42) Henderson, T. M.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108.

(43) Godby, R.; Schluter, M.; Sham, L. *Phys. Rev. B* **1988**, *37*, 10159−10175.

(44) Lany, S.; Zunger, A. *Phys. Rev. B* **2009**, *80*, 085202.

(45) Perdew, J.; Levy, M. *Phys. Rev. Lett.* **1983**, *51*, 1884−1887.

(46) Seidl, A.; Gorling, A.; Vogl, P.; Majewski, J. A.; Levy, M. *Phys. Rev. B* **1996**, *53*, 3764−3774.

(47) Cohen, A.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 792−794.

(48) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2008**, *100*, 146401.

(49) Rudberg, E. *J. Phys.: Condens. Matter* **2012**, *24*, 072202.

(50) Stein, T.; Eisenberg, H.; Kronik, L.; Baer, R. *Phys. Rev. Lett.* **2010**, *105*, 266802.

(51) Huzinaga, S.; Andzelm, J.; Klobukowski, M.; Radzio-Andzelm, E.; Sakai, Y.; Tatewaki, H. *Gaussian Basis Sets for Molecular Calculations*; Elsevier: Amsterdam, 1984.

(52) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657−2664.

(53) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939−947.

(54) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.

(55) Remko, M.; Scheiner, S. *J. Pharm. Sci.* **1988**, *77*, 304−308.

(56) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

(57) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M. et al. *Amber 11*; University of California: San Francisco, CA, 2010.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(59) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486−1493.

(60) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9901−9904.

(61) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, *2*, 799−805.

(62) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172−3191.

(63) Lange, A. W.; Herbert, J. M. *J. Chem. Phys.* **2010**, *133*, 244111.

(64) Lange, A. W.; Herbert, J. M. *Chem. Phys. Lett.* **2011**, *509*, 77−87.

(65) Chen, V. B.; Arendall, W. B., III; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. *Acta Crystallogr., Sect. D* **2010**, *66*, 12−21.

(66) Benco, L.; Tunega, D.; Hafner, J.; Lischka, H. *J. Phys. Chem. B* **2001**, *105*, 10812−10817.

(67) Lovell, S. C.; Davis, I. W.; Arendall, W. B., III; de Bakker, P. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437−450.

(68) Kruse, H.; Grimme, S. *J. Chem. Phys.* **2012**, *136*, 154101.

(69) Gordon, D. B.; Marshall, S. A.; Mayo, S. L. *Curr. Opin. Struct. Biol.* **1999**, *9*, 509−513.

(70) Mittal, J.; Best, R. B. *Biophys. J.* **2010**, *99*, L26–L28.

(71) Nagata, K.; Masaru, T. *Pept. Sci.* **2005**, *44*, 441.

(72) Wiltzius, J. J. W.; Landau, M.; Nelson, R.; Sawaya, M. R.; Apostol, M. I.; Goldschmidt, L.; Soriaga, A. B.; Cascio, D.; Rajashankar, K.; Eisenberg, D. *Nat. Struct. Mol. Biol.* **2009**, *16*, 973–978.

(73) Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J. D. *Biochemistry* **2004**, *43*, 7421–7431.

(74) Jaremko, L.; Jaremko, M.; Pasikowski, P.; Cebrat, M.; Stefanowicz, P.; Lisowski, M.; Artym, J.; Zimecki, M.; Zhukov, I.; Szewczuk, Z. *Biopolymers* **2009**, *91*, 423–431.

(75) Nakatogawa, S.-i.; Oda, Y.; Kamiya, M.; Kamijima, T.; Aizawa, T.; Clark, K. D.; Demura, M.; Kawano, K.; Strand, M. R.; Hayakawa, Y. *Curr. Biol.* **2009**, *19*, 779–785.

(76) Cobine, P. A.; McKay, R. T.; Zangger, K.; Dameron, C. T.; Armitage, I. M. *Eur. J. Biochem.* **2004**, *271*, 4213–4221.

(77) Wang, G.; Treleaven, W. D.; Cushley, R. J. *Biochim. Biophys. Acta* **1996**, *1301*, 174–184.

(78) The minimum number of residues for a sequence to count as a "run" of a secondary structure is four. If there is no discernible secondary structure or if the number of insertions is larger than the number of secondary structure runs, the ratio $N_{\text{SS-int}}/N_{\text{SS}}$ is set to unity. $N_{\text{SS-int}}$ is zero only when the protein is typed by a single ordered SS type and any disordered regions occur only at the tail ends of the protein.

(79) Biron, Z.; Khare, S.; Samson, A. O.; Hayek, Y.; Naider, F.; Anglister, J. *Biochemistry* **2002**, *41*, 12687–12696.