JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Representative Amino Acid Side-Chain Interactions in Protein−DNA Complexes: A Comparison of Highly Accurate Correlated *Ab Initio* Quantum Mechanical Calculations and Efficient Approaches for Applications to Large Systems

Jiří Hostaš,[†,‡] Dávid Jakubec,[†,‡] Roman A. Laskowski,[§] Ramachandran Gnanasekaran,[†] Jan Řezáč,[†] Jiří Vondrášek,[*,†] and Pavel Hobza[*,†,▽]

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 166 10 Prague, Czech Republic

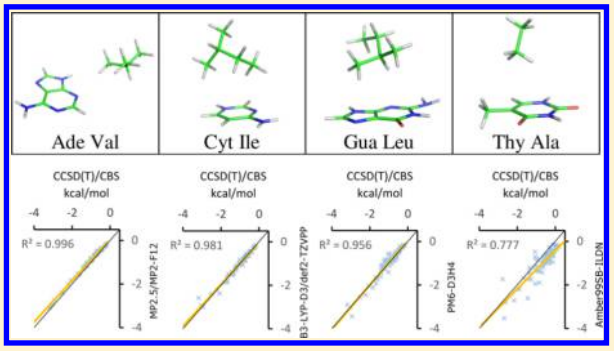
[‡]Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague, Czech Republic

[§]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[▽]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic

Supporting Information

ABSTRACT: Representative pairs of amino acid side chains and nucleic acid bases extracted from available high-quality structures of protein−DNA complexes were analyzed using a range of computational methods. CCSD(T)/CBS interaction energies were calculated for the chosen 272 pairs. These reference interaction energies were used to test the MP2.5/CBS, MP2.X/CBS, MP2-F12, DFT-D3, PM6, and Amber force field methods. Method MP2.5 provided excellent agreement with reference data (root-mean-square error (RMSE) of 0.11 kcal/mol), which is more than 1 order of magnitude faster than the CCSD(T) method. When MP2-F12 and MP2.5 were combined, the results were within reasonable accuracy (0.20 kcal/mol), with a computational savings of almost 2 orders of magnitude. Therefore, this method is a promising tool for accurate calculations of interaction energies in protein−DNA motifs of up to ∼100 atoms, for which CCSD(T)/CBS benchmark calculations are not feasible. B3-LYP-D3 calculated with def2-TZVPP and def2-QZVP basis sets yielded sufficiently good results with a reasonably small RMSE. This method provided better results for neutral systems, whereas positively charged species exhibited the worst agreement with the benchmark data. The Amber force field yielded unbalanced results—performing well for systems containing nonpolar amino acids but severely underestimating interaction energies for charged complexes. The semiempirical PM6 method with corrections for hydrogen bonding and dispersion energy (PM6-D3H4) exhibited considerably smaller error than the Amber force field, which makes it an effective tool for modeling extended protein−ligand complexes (of up to 10 000 atoms).



1. INTRODUCTION

Despite ongoing efforts,[1] an understanding of the rules that govern protein−DNA recognition is far from complete. In recent years, the growing amount of structural data has opened space for bioinformatics and computational analyses. Insights gained from such analyses into the DNA recognition process of zinc finger domains and transcription activator-like effector (TALE) proteins have led to powerful genetic engineering technologies.[2] Understanding the principles governing protein−DNA selectivity may lead to the development of new applications in biotechnology and medicine.

Recently, we quantitatively examined protein−DNA interactions by calculating the interaction energies for all 20 × 4 amino acid−DNA base combinations.[3] We obtained the geometries of the pairs from the web version of the Atlas of Protein Side Chain Interactions.[4] The structural data in the Atlas were extracted from a nonredundant subset of all available protein−DNA complexes and are not specific to any single protein family or DNA-binding motif. To evaluate the interaction energy between DNA and protein building blocks, we employed molecular mechanics (MM) utilizing empirical force fields.

Here, we set out to compare the performance of these empirical potential-based methods with reliable quantum mechanical (QM) methods. The QM methods should adequately describe the dominant hydrogen bonding interactions, as well as electrostatic and London dispersion

Received: April 29, 2015
Published: July 23, 2015

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092

interactions. Dispersion energy in particular plays an important role in biomolecular complexes, and its proper description is of primary importance.[5] In addition, further stabilization due to charge transfer can play a key role in charged complexes; however, this contribution is not covered directly at the MM level. An accurate QM approach, such as the coupled-cluster method covering single and double excitations iteratively and triple excitations perturbatively [CCSD(T)] at the complete basis set (CBS) limit, is needed to benchmark other computational chemistry methods, which can fail in certain cases (e.g., charge-transfer complexes).[6] These CCSD(T)/CBS energies are assumed to be closest to the "true" energy values.[7]

Previously, we determined interaction energies between amino acid side chains with a reasonable level of accuracy using DFT-based methods.[8,9] These methods represent a good compromise between accuracy and cost, yielding reliable characteristics for systems with several hundred atoms. In this study, we evaluated the energy of interactions between selected amino acid side chains and nucleic acid base pairs. For each pair, we used an experimentally determined representative geometry and compared the energies computed using several different *ab initio* QM and force field methods with the CCSD(T) reference method.

2. METHODS

2.1. Description of a Representative Set of Amino Acid Side Chain−DNA Base Pair Interactions. To obtain a representative set of amino acid side chain−DNA base pair interactions, we extracted data from an updated version of the "Atlas of Protein Side-Chain Interactions" (available at http://www.ebi.ac.uk/thornton-srv/databases/sidechains/) as previously described.[3] As of March 2014, the Atlas comprised 1569 unique structures of protein−DNA complexes. Amino acid−nucleotide pairs were extracted from each complex by a procedure based on the SIRIUS set of scripts described by Singh and Thornton.[10] These programs recognize a pair as "interacting" if certain distance criteria between predefined reference atoms are met. This procedure resulted in 20 × 4 sets of contacts. Each dimer, consisting of a single amino acid and a single nucleotide, was transformed to utilize the same frame of reference, with respect to the DNA base, yielding 20 distributions of amino acid residues around each DNA base. The root-mean-square deviation (RMSD) between atom positions was calculated for all pairs of amino acid side chains. The dimer with the highest number of structures within an RMSD of 1.5 Å was considered a cluster representative and set aside; "neighboring" structures and contacts were regarded as its associated cluster. This procedure was repeated up to six times, depending on the size of the cluster.

Only those with a distance of <4.5 Å between any DNA base atom and any amino acid side chain atom were further considered. We prepared a total of 272 clusters and cluster representatives. The geometries of all pairs are available at http://bioinfo.uochb.cas.cz/projects/pdna-iea/ and http://www.begdb.com.

The Cα representations of amino acids were prepared by replacing the carbonyl and amide groups with hydrogen atoms, as described by Berka et al.[9] In this procedure, each amino acid has methyl group at Cα, thus eliminating any potential nonspecific interactions between the backbone and the DNA base. Histidine was protonated on the ε-N atom; proline was modeled as a neutral tetrahydropyrrole. Only the base of each nucleotide was preserved, and the deoxyribose C1′ carbon of the N-glycosidic bond was replaced with a hydrogen atom.

Since the Atlas contains only the positions of heavy atoms, hydrogens were added to each cluster representative, using a Chimera-1.8.1[11] script, and optimized at the B3-LYP-D3/def2-TZVPP level[12−15] for *ab initio* methods; conjugate gradient optimization utilizing the Amber force field parameters was used for comparison with the MM method. The heavy atoms were kept in their original positions.

The cluster representatives were classified according to the physicochemical character of each amino acid (see Figure 1):

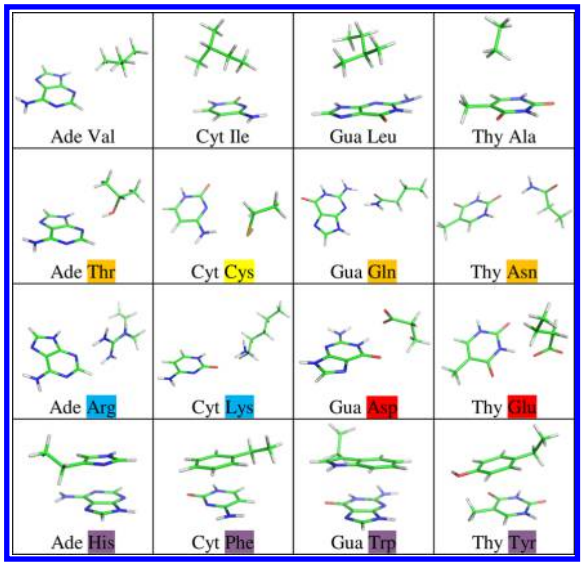


Figure 1. Illustration of the geometries of DNA bases and amino acid Cα representations of different types. The first row from the top contains nonpolar, the second row polar, the third row charged, and the fourth row aromatic amino acids in interaction with DNA bases. (aliphatic, polar, polar with sulfur, basic, acidic and aromatic types are shown in white, orange, yellow, blue, red, and violet, respectively). The cluster representatives for the most stable pair are shown.

nonpolar (G, A, V, I, L, P; 76 contacts), polar (T, S, N, Q, C, M; 69 contacts), aromatic (F, Y, W, H; 63 contacts), positively charged (K, R; 33 contacts) and negatively charged (D, E; 31 contacts). A more detailed description of the methodology is provided in our previous study,[3] in which we highlight bioinformatic aspects and features derived from distributions involving tens of thousands of contacts.

2.2. Interaction Energy Calculations. All interaction energies were calculated using fixed monomer geometries, i.e., deformation energy was not considered. The only exception was the accuracy assessment of the Amber force field, in which we took into account the deformation energy of hydrogen atoms and compared the results with those obtained at the B3-LYP-D3/def2-TZVPP level. All calculations were performed *in vacuo*, and no symmetry was assumed.

2.3. Benchmark CCSD(T)/CBS Interaction Energies. To calculate the reference interaction energies, we needed a method able to describe systems ranging from 20 atoms to 38 atoms. The natural choice is the "gold standard" of computational chemistry−the CCSD(T) method extrapolated to the CBS limit. We approximated the interaction energy as follows:

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092

$$\Delta E^{\mathrm{CCSD(T)/CBS}} = \Delta E^{\mathrm{HF}}(\text{large-size basis set}) + \Delta E^{\mathrm{MP2}}(\mathrm{CBS}) + (\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}})(\text{medium-size basis set}) \quad (1)$$

The Hartree−Fock (HF) energy converges with increasing basis set size faster than the correlation energy; therefore, a calculation in a single large basis set, such as aug-cc-pVQZ, is appropriate. The second term, ΔE^{MP2}(CBS), was determined by applying a two-point Helgaker extrapolation scheme.[16] We used Dunning's correlation-consistent series of basis sets with diffuse functions: aug-cc-pVTZ (augTZ) and aug-cc-pVQZ (augQZ).[17] We calculated the CCSD(T) correction term in a smaller aug-cc-pVDZ (augDZ) basis set, because this quantity, which is defined as the difference between CCSD(T) and MP2 energies, is less dependent on the basis set size. The augDZ was the smallest basis set that provided reliable results with errors of <0.1 kcal/mol.[18,19] This is the same setup that we used to generate our extensive datasets of benchmark interaction energies, such as S66x8 and X40x10.[20,21]

All interaction energies were corrected for BSSE using the counterpoise scheme described by Boys and Bernardi.[22] The resolution of identity was used to accelerate the MP2 calculations.[23] The frozen-core approximation was applied systematically to all calculations of correlation energy.

2.4. Post-Hartree−Fock Methods. The CCSD(T) method has been proven accurate, robust, and size-consistent, but it is limited to systems with less than several dozen atoms. For larger systems, the highest accuracy can be achieved with empirically scaled methods based on scaling of the same- and opposite-spin contributions (such as SCS-MP2 and SCS-MI-MP2)[24] and methods constructed from MP3 energy (such as MP2.5 and MP2.X).[25−27]

Explicitly correlated MP2 methods improve upon basis-set convergence toward the CBS limit and represent another important direction. Here, we employ a closed-shell (semi)-canonical variant with density fitting (RI-MP2-F12) utilizing the cc-pVDZ and cc-pVDZ-F12 basis sets as implemented in Turbomole 6.5.[28,29]

In the present work, we tested the following post-HF methods: MP2.5, SCS-MP2, SCS-MI-MP2, MP2-F12, and MP2.X. To neglect the influence of basis set selection, we used the same basis set as for calculation of the CCSD(T) correction term, augDZ. The second basis set used for MP2.5 and MP2.X calculations was the split-valence 6-31G*(0.25) basis set containing diffuse *d*-functions (α = 0.25) on heavy atoms.[30] In combination with the MP2 method, it provides reliable interaction energies comparable to those obtained with the much-larger aug-cc-pVDZ basis set.[31] There is no complementary auxiliary basis set; therefore, we used a similarly sized def2-SVPD for resolution of identity. The BSSE was eliminated using the counterpoise correction.

MP3 calculations take advantage of the resolution of identity, which greatly accelerates the calculations (see Table 1 in the Results section). MP2.5 and MP2.X, which were developed in our laboratory,[25−27] are known to provide highly accurate interaction energies for different types of molecular clusters.[31] The method is more demanding, with regard to CPU time, than various versions of the MP2 method, but it is much faster than the CCSD(T)/CBS method used in the present work for benchmark calculations. Because the DFT-D method might fail for charged complexes with significant charge transfer, we used MP2.5 as the benchmark method for complexes for which CCSD(T)/CBS calculations are not feasible.

Table 1. Computational Times (Adenine−Tryptophan System, 37 Atoms, Intel Xeon E5630 2.53 GHz, 8 Cores, 5.8 GB RAM per Core)

method and basis set	time [h]
CCSD(T)/aug-cc-pVDZ	273
RI-MP2/aug-cc-pVQZ	117
RI-DFT/def2-QZVP	31.3
RI-MP2/aug-cc-pVTZ	12.8
RI-MP3/aug-cc-pVDZ	7.3
RI-MP2-F12/cc-pVDZ	6.5
RI-DFT/def2-TZVPP	3.3
RI-MP3/6-31G*(0.25)	0.6
Amber force field	0.001

2.5. Density Functional Interaction Energies. DFT is the method of choice for large complexes with hundreds of atoms, because of its favorable balance between accuracy and computational cost. Here, we included DFT-D3 interaction energies, in which the DFT energies were calculated with a B3-LYP functional in def2-TZVPP and def2-QZVP basis sets.[15] Both were augmented with the D3 empirical dispersion term, thus covering one of the most severe deficiencies of the DFT methodology while keeping essentially the same CPU-time requirements.[14] Calculations utilized the Becke-Johnson damping function.[32] Three-body nonadditive terms were not considered[14] (for complexes presented here, the average absolute value is <0.05 kcal/mol). All calculations were carried out by means of the resolution of identity.

The present combination of the DFT functional, empirical dispersion correction, and basis sets provided accurate interaction energies for various types of noncovalently bound molecular clusters.[33] In this study, we also included results for B-LYP and TPSS functionals.[34−37]

2.6. Semiempirical Quantum Chemical Methods. We investigated the performance of PM6[38] augmented with empirical corrections for hydrogen bonding and dispersion interactions (PM6-D3H4).[39] For comparison, we also included the PM6 method as implemented in MOPAC software.[40]

2.7. Empirical Force-Field Interaction Energies. The empirical force-field calculations were carried out with the Gromacs-4.5.5 package[41] with the Amber99SB-ILDN protein force field[42] combined with Amber94 nucleic-acid parameters.[43] We selected this force field because it was the most commonly used among those tested in our previous study in which B3-LYP/def2-TZVPP was the reference method.[3]

Parameters of the Cα representations of amino acids were added manually. They were based on the existing amino acid topologies, with the Cα carbonyl and amide groups replaced with hydrogen atoms. The charges of these added Cα hydrogen atoms were symmetrically distributed to keep the overall integral charge of the amino acid as follows: −1 for aspartate and glutamate, +1 for lysine and arginine, and 0 for others. The atom types were HC for all amino acids except proline, where the atom type of the tetrahydropyrrole N hydrogen was H. This procedure introduced a symmetry in four amino acids: alanine, valine, proline, and glycine. We did not reflect this by recalculating hydrogen atom charges for any amino acid except glycine, where all four hydrogen atoms were made equivalent to cancel detrimental effects—their orientation, with respect to the nucleotide, would have violated the resulting interaction energies.

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092

Similarly, we derived the DNA-base parameters in the force field based on the parameters of free nucleotides by stripping them of the sugar−phosphate atoms. The purine H9 and pyrimidine H1 hydrogen atoms were added to the topologies, their charges were assigned to keep the overall charge of the base at 0, and their atom types were set to H. All energy calculations and gradient optimization of hydrogens were performed in the gas phase using a double-precision version of the GROMACS-4.5.5 package.[41]

2.8. Error Analysis. There are multiple statistical tools that highlight different information about error measurements. We considered the root-mean-square error (RMSE), the mean signed error (MSE), and the mean unsigned error (MUE) to be the most robust quantities showing a method's overall performance. For a relative comparison between the different interaction types, we also present the RMSE as the percentage of the average interaction energy in the group (rRMSE), and we do likewise for the relative mean signed error (rMSE) and the relative mean unsigned error (rMUE).

2.9. Computational Details. All DFT, MP2, MP3, and CCSD(T) interaction energy calculations were carried out as implemented in Turbomole 6.5. We used the GROMACS-4.5.5 package[41] for all force-field calculations and the MOPAC 2012 package[40] for PM6 calculations. The cuby framework (http://cuby.molecular.cz), which was developed by one of the current authors (Jan Řezáč), was used to automate the calculations.

3. RESULTS AND DISCUSSION

In this study we concentrated on comparison of selected computational chemistry methods and evaluation of their accuracy and efficiency. Table 1 lists the wall time spent on computations by methods in the standard computational cluster. The relative RMSE for the four groups of DNA base−amino acid residue complexes shown in Figures 2 and 3 illustrates the correlation between the benchmark CCSD(T)/CBS results and the results obtained by other methods. Complete results are summarized in Tables S1−S5 (see the Supporting Information).

3.1. Performance of MP2.5/CBS. As shown in our previous work, large errors can be expected in complexes with dominant dispersion interaction, because of the strongly overbinding MP2/CBS term.[25] In this study, we used two basis sets, 6-31G*(0.25) and aug-cc-pVDZ, to calculate the MP2.5 correction terms (the difference between MP2.5 and MP2 energies). The smaller basis set, 6-31G*(0.25), performed comparably well (see Table S1) to the considerably larger (and thus more time-consuming) aug-cc-pVDZ basis set (see Table 1). By analyzing different types of molecular clusters, we found that the smaller basis set performed worse only for systems with aromatic amino acids, in which it overestimated the interaction energies (MSE of −0.07). In contrast, the larger basis set underestimated them (MSE of 0.06 kcal/mol). With the smaller basis set, the RMSE for aromatic systems was twice as large as for the other neutral complexes (0.18 kcal/mol) (see Table S1). We found comparable absolute errors for charged systems (RMSEs of 0.16 and 0.15 kcal/mol for small and large basis sets, respectively).

For all types of interactions, the correlation between the CCSD(T) and MP2.5 correction terms (one of the highly important parameters) was more than 95%, except for charged systems, in which it was only 72% and 67% for the larger and smaller basis sets, respectively. On the other hand, we observed the opposite trend when considering relative measures (the rRMSE for charged systems is <2%, compared to 4% for the much more weakly bound neutral complexes). This inconsistent behavior is probably caused by the relatively small basis set size and the highest absolute interaction energies among all interaction types.

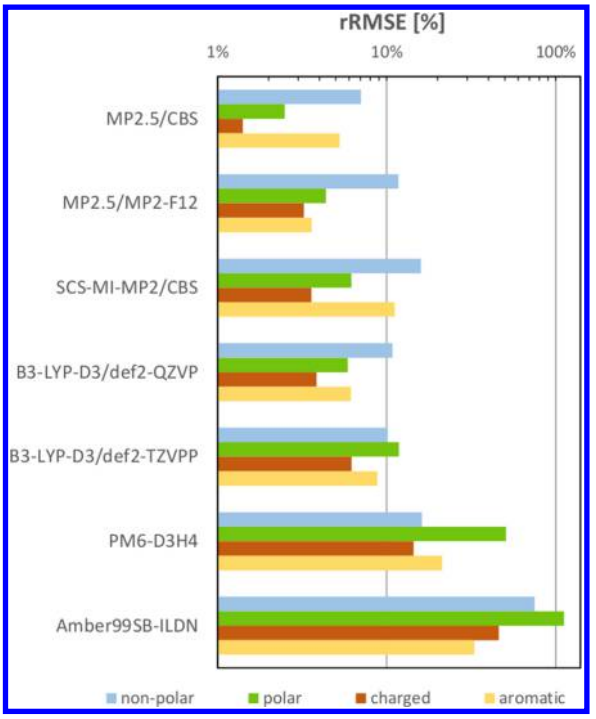


Figure 2. Relative errors (%) for the four groups of complexes between a DNA base and the following types of amino acid residues: nonpolar, polar, charged, and aromatic. Error was determined as RMSE, relative to the average absolute interaction energy in the group. MP2.5/MP2-F12 was calculated using 6-31G*(0.25)/cc-pVDZ-F12 basis sets. The aug-cc-pVDZ basis set was used for calculations of correction terms.

These findings indicate that one must pay close attention when dealing with charged systems; the appropriate methods should be applied only after performing tests on similar complexes. Furthermore, the present findings support our previous conclusions about the good performance of the 6-31G*(0.25) basis set, which makes it a promising tool for future use on extended protein−DNA molecular clusters.

3.2. Use of MP2-F12 Instead of Extrapolation to the CBS Limit. One attractive option to calculate interaction energy values close to the complete basis set limit is the use of explicitly correlated methods.[44] We tested RI-MP2-F12, which can be utilized instead of extrapolating results from two separate MP2 calculations with a systematically increasing size of correlation-consistent basis sets.

The MP2-F12 method, together with the cc-pVDZ(-F12) basis set, performed well for complexes with nonpolar and polar amino acids, when compared to MP2/CBS (RMSE of <0.10 kcal/mol). Larger discrepancies were found for complexes containing aromatic amino acids and especially for charged systems (RMSE values of 0.17 and 0.31 kcal/mol, respectively).

Interestingly, the MP2.5/6-31G*(0.25) correction term has a mean signed error for systems with aromatic amino acids with the opposite sign; therefore, a partial error cancellation takes place when the two methods are compared. The resulting RMSE for these complexes is roughly 0.12 kcal/mol (see Tables S1 and S2 in the Supporting Information). The overall

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092

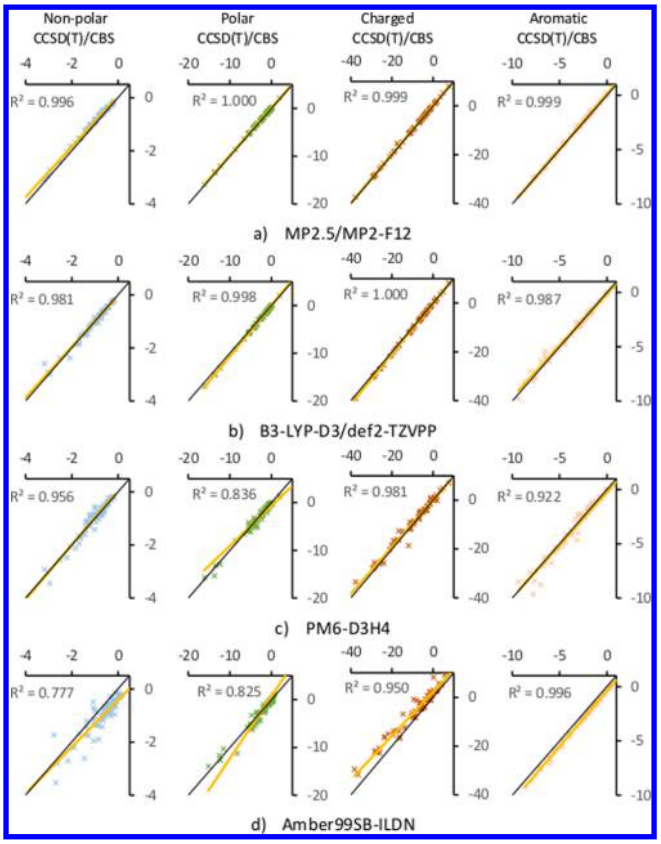


Figure 3. Correlation plots of the CCSD(T)/CBS and MP2.5/MP2-F12, B3-LYP-D3/def2-TZVPP, PM6-D3H4, and Amber99SB-ILDN methods. The yellow line represents the linear regression, and the black line has a slope of 1. All energies are given in units of kcal/mol.

RMSE of this parameter-free approach was only 0.20 kcal/mol with computational savings of almost 2 orders of magnitude, compared to the CCSD(T)/CBS calculation.

3.3. Comparison of DFT-D3, PM6, and Amber Force Field with CCSD(T)/CBS Methods. The DFT-D method systematically overestimated the strength of interactions. The use of a larger basis set reduced both the absolute and relative errors by roughly one-third (RMSE value of 0.25 kcal/mol). This difference stemmed mainly from the effect of the basis set in polar and negatively charged systems, in which the larger basis set yielded significantly better results. On the other hand, we found that both basis sets performed well for nonpolar and aromatic systems. Charged amino acid complexes yielded the worst agreement with benchmark values, and surprisingly, positively charged complexes provided systematically worse results. Based on these findings, we recommend the use of both small and large basis sets for applications involving large complexes of a similar type. The other two functionals, TPSS and B-LYP, exhibited slightly larger errors (RMSEs of 0.34 and 0.35 kcal/mol, respectively) with significant computational savings.

The Amber99SB-ILDN force field yielded results that were unbalanced in describing neutral and charged systems. It performed well for nonpolar and aromatic systems but significantly underestimated charged pairs. The overall performance of the Amber force field was slightly better when we included the deformation energy (RMSE decreased from 2.6 kcal/mol to 2.3 kcal/mol). Similar to the DFT results, charged systems had the worst agreement with the benchmark data, and positively charged species gave systematically worse results.

The semiempirical PM6 method without corrections yielded slightly worse results than force field calculations (RMSE value of 2.5 kcal/mol). On the other hand, we observed significant improvement after including the dispersion correction. The most illustrative cases are aromatic systems, for which the RMSE decreased from 2.9 kcal/mol to 0.8 kcal/mol. We observed the same trend in nonpolar systems (a decrease in RMSE from 1.0 kcal/mol to 0.2 kcal/mol). Upon inclusion of hydrogen bonding, we observed the greatest improvement in charged complexes (the RMSE value decreased from 3.5 kcal/mol to 1.5 kcal/mol). Positively and negatively charged complexes exhibited similar errors. Both corrections played an important role in polar systems, decreasing the RMSE value from 2.2 kcal/mol to 1.5 kcal/mol. In summary, the PM6-D3H4 method exhibited less than half of the error of force-field calculations, which makes it an effective tool for the modeling of similar extended protein−ligand complexes with up to 10 000 atoms (see Table S5 in the Supporting Information).

Tables S4 and S5 in the Supporting Information show the performance of the Amber99SB-ILDN force field and the semiempirical PM6 and B3-LYP-D3 methods, the latter with medium-size (def2-TZVPP) and large-size basis sets (def2-QZVP).

4. CONCLUSION

We have analyzed 272 representative pairs of amino acid side chains and nucleic acid bases, using CCSD(T)/CBS interaction energies to test various methods. We have reached the following conclusions:

(i) MP2.5 provides balanced and highly accurate stabilization energies with a root mean square error (RMSE) value of 0.11 kcal/mol and a relative error of 2%. The method successfully described all types of clusters investigated, which included both neutral and charged species.[27] We obtained comparable results for positively and negatively charged complexes. Charged complexes exhibited the largest absolute errors with an RMSE value of 0.15 kcal/mol.

(ii) For larger complexes, for which CCSD(T)/CBS calculations are not feasible, we found that a combination of the MP2-F12/cc-pVDZ and MP2.5/6-31G*(0.25) correction terms achieved an accuracy of 0.20 kcal/mol. This technique is a promising tool for calculation of accurate interaction energies for extended protein−DNA complexes.

(iii) B3-LYP-D3 systematically overestimated the strength of interactions. This effect was most pronounced for positively charged complexes (by 0.31 kcal/mol, on average, for the def2-QZVP basis set). Both the def2-TZVPP and def2-QZVP basis sets yielded mean unsigned errors of <0.25 kcal/mol with RMSE values of <0.40 kcal/mol; the latter basis set gave more-accurate results.

(iv)The semiempirical PM6-D3H4 method performed well, with an RMSE value of 1.1 kcal/mol.

(v) The average performance of the Amber99SB-ILDN force field was in reasonable agreement with the benchmark method as long as the amino acid−DNA base pairs were practically close to the equilibrium geometries. (i.e., when the interaction energy difference outliers were excluded).

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00398.

Detailed error analysis (Tables S1−S5) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*Tel.: +420 220 183267. Fax: +420 220 410 321. E-mail: jiri.vondrasek@uochb.cas.cz.

*Tel.: +420 220 410311. Fax: +420 220 410 321. E-mail: pavel.hobza@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was part of Research Project RVO (No. 61388963) of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. This work was also supported by the Czech Science Foundation (No. P208/12/G016) and the operational program Research and Development for Innovations of the European Social Fund (No. CZ 1.05/2.1.00/03/0058).

■ REFERENCES

(1) Laskowski, R. A.; Thornton, J. M. Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet.* **2008**, *9*, 141−151.

(2) Gaj, T.; Gersbach, C. A.; Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **2013**, *31*, 397−405.

(3) Jakubec, D.; Hostaš, J.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Large-Scale Quantitative Assessment of Binding Preferences in Protein−Nucleic Acid Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 1939−1948.

(4) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid-base interactions: A three-dimensional analysis of protein−DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860−2874.

(5) Černý, J.; Hobza, P. Non-covalent interactions in biomacromolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291−5303.

(6) Černý, J.; Pitoňák, M.; Riley, K. E.; Hobza, P. Complete basis set extrapolation and hybrid schemes for geometry gradients of non-covalent complexes. *J. Chem. Theory Comput.* **2011**, *7*, 3924−3934.

(7) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T), at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151−2155.

(8) Černý, J.; Pitoňák, M.; Riley, K. E.; Hobza, P. Complete basis set extrapolation and hybrid schemes for geometry gradients of non-covalent complexes. *J. Chem. Theory Comput.* **2011**, *7*, 3924−3934.

(9) Berka, K.; Laskowski, R. A.; Riley, K. E.; Hobza, P.; Vondrášek, J. Representative amino acid side chain interactions in proteins. A comparison of highly accurate correlated *ab initio* quantum chemical and empirical potential procedures. *J. Chem. Theory Comput.* **2009**, *5*, 982−992.

(10) Singh, J.; Thornton, J. M. SIRIUS. An Automated Method for the Analysis of the Preferred Packing Arrangements between Protein Groups. *J. Mol. Biol.* **1990**, *211*, 595−615.

(11) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−12.

(12) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *Ab Initio* Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(13) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

(14) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(15) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.

(16) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis set convergence in correlated calculations on Ne, N_2 and H_2O. *Chem. Phys. Lett.* **1998**, *286*, 243−252.

(17) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(18) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. Highly Accurate CCSD(T) and DFT−SAPT stabilization Energies of H-Bonded and Stacked Structures of the Uracil Dimer. *ChemPhysChem* **2008**, *9*, 1636−1644.

(19) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. An Assessment of Theoretical Methods for Nonbonded Interactions: Comparison to Complete Basis Set Limit Coupled-Cluster Potential Energy Curves for the Benzene Dimer, the Methane Dimer, Benzene−Methane, and Benzene−H_2S. *J. Phys. Chem. A* **2009**, *113*, 10146−10159.

(20) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structure. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.

(21) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 4285−4292.

(22) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553−566.

(23) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of Approximate Integrals in *Ab Initio* Theory—An Application in Mp2 Energy Calculations. *Chem. Phys. Lett.* **1993**, *208*, 359−363.

(24) Distasio, R. A., Jr.; Head-Gordon, M. Optimized spin-component scaled second-order Møller-Plesset perturbation theory for intermolecular interaction energies. *Mol. Phys.* **2007**, *105*, 1073−1083.

(25) Sedlák, R.; Riley, K. E.; Řezáč, J.; Pitoňák, M.; Hobza, P. MP2.5 and MP2.X: Approaching CCSD(T) Quality Description of Non-covalent Interaction at the Cost of Single CCSD Iteration. *ChemPhysChem* **2013**, *14*, 698−707.

(26) Riley, K. E.; Řezáč, J.; Hobza, P. MP2.X: A generalized MP2.5 method that produces improved binding energies with smaller basis sets. *Phys. Chem. Chem. Phys.* **2011**, *13*, 21121−21125.

(27) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. Scaled MP3 Non-Covalent Interaction Energies Agree Closely with Accurate CCSD(T) Benchmark Data. *ChemPhysChem* **2009**, *10*, 282−289.

(28) Peterson, K. A.; Adler, T. B.; Werner, H. J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al-Ar. *J. Chem. Phys.* **2008**, *128*, 084102.

(29) TURBOMOLE v6.5 2013, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

(30) van Lenthe, J. H.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. Weakly Bonded Systems. *Adv. Chem. Phys.* **1987**, *69*, 521−566.

(31) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110*, 5023−5063.

(32) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(33) Goerigk, L.; Kruse, H.; Grimme, S. Benchmarking Density Functional Methods against the S66 and S66x8 Datasets for Non-Covalent Interactions. *ChemPhysChem* **2011**, *12*, 3421−3433.
(34) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098−3100.
(35) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.
(36) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 146401−146404.
(37) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation-energy density functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157*, 200−206.
(38) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(39) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.
(40) Stewart, J. J. P. *MOPAC2012*; Stewart Computational Chemistry: Colorado Springs, CO, USA, 2012 (http://OpenMOPAC.net).
(41) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(42) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(43) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(44) Liakos, D. G.; Izsák, R.; Valeev, E. F.; Neese, F. What is the most efficient way to reach the canonical MP2 basis set limit? *Mol. Phys.* **2013**, *111*, 2653−2662.

DOI: 10.1021/acs.jctc.5b00398
J. Chem. Theory Comput. 2015, 11, 4086−4092