Finding efficient methods for the quantum chemical treatment of protein structures: What are the effects of London-dispersion and basis-set incompleteness on peptide and water-cluster geometries?

*Lars Goerigk, and Jeffrey R. Reimers**

School of Chemistry, The University of Sydney, New South Wales 2006,

KEYWORDS peptide structures, water clusters, London-dispersion, basis-set superposition error, density functional theory

ABSTRACT We demonstrate how quantum chemical Hartree-Fock (HF) or density functional theory (DFT) optimisations with small basis sets of peptide and water cluster structures are decisively improved if London-dispersion effects, the basis-set-superposition error (BSSE) and other basis-set incompleteness errors are addressed properly. To achieve this better description, we concentrate on three empirical corrections to these problems recently advanced by Grimme and co-workers that lead to computational strategies that are both accurate and efficient. Our analysis encompasses a reoptimised version of Hobza's P26 set of tripeptide structures, a new test set of conformers of cysteine dimers, and isomers of the water hexamers. These systems reflect features commonly found in protein crystal structures. In all cases, we recommend

Grimme's DFT-D3 correction for London-dispersion. We recommend usage of large basis sets like cc-pVTZ whenever possible to reduce any BSSE effects and, if this is not possible, to use Grimme's gCP correction to account for BSSE when small basis sets are used. We demonstrate that S-S and C-S bond lengths are very prone to basis-set incompleteness and that polarisation functions have to be used on S atoms. On the double-ζ level, the PW6B95-D3-gCP DFT method combined with the SVP and 6-31G* basis sets yields accurate results. Alternatively, the HF-D3-gCP/SV method is recommended, with inclusion of polarisation functions for S atoms only. Minimal basis sets offer an intruiging route to highly efficient calculations, but due to significant basis-set incompleteness effects calculated bond lengths become prohibitively large, making applications to large proteins very difficult, but we show that Grimme's newest HF-3c correction addresses this problem and makes this computational strategy very attractive. Our results provide a useful guideline for future applications to the optimisation, quantum refinement and dynamics of large proteins.

1. Introduction

Computational studies of proteins range in nature from methods that seek to avoid structural details, such as bioinformatics,[1] to methods based on qualitative structure-activity relationships (QSAR[2]) to methods based on empirical molecular-mechanics (MM) force fields.[3] The next step in this hierarchy is formed by methods involving partial quantum-mechanical (QM) treatments (QM/MM)[4] followed by ones based solely on quantum descriptions of the electronic motions.[5,6,7] These QM/MM or QM treatments are often restricted to Hartree-Fock (HF) or density functional theory (DFT[8]) because of computational limitations.[4] Full QM treatments of big systems are usually carried out by various linear-scaling fragmentation schemes or by using computationally

fast levels of theory in combination with high parallelisation on CPU or GPU high-performance clusters.[5,6,7,9] The advantage of incorporating the quantum mechanical electronic structure of proteins is that such QM calculations are less biased towards the features used in the parameterisation of empirical methods. This, in principle, allows treating a broader range of systems, including those that contain new or unusual structural features.

QM and QM/MM treatments of proteins are often carried out using Hartree-Fock (HF) or density functional theory (DFT) approximations combined with small basis sets such as double-ζ or even minimal basis sets.[4,7] When doing this, one faces three problems: a proper treatment of London-dispersion (attractive van-der-Waals forces), the intramolecular basis-set-superposition error and other errors due to the incompleteness of small basis sets. Even extensive calculations, such as the PW91/6-31G* optimisation of the 150,000 atom photosystem-I trimer using linear-scaling DFT, suffer from these problems.[5] It is well known that HF and DFT do not correctly describe London-dispersion effects,[10] which, however, are crucial to the structural stability of biomolecular systems.[11] Nevertheless, many studies, e.g. in the field of quantum refinement of protein X-ray crystal structures, were reported for HF or DFT approximations without taking these effects into account.[4] For smaller to medium sized systems, and later also for larger van-der-Waals and protein-ligand complexes, it was demonstrated that dispersion-corrected density functional theory is an accurate remedy to overcome the London-dispersion problem of DFT.[12] Various methods for treating dispersion effects with DFT have been devised, see ref. [13] for a review. Herein, we will concentrate on Grimme's widely-used and established DFT-D3 approach,[14] which has been shown to work excellently for noncovalent interaction energies,[14,15] general thermochemistry[16] and geometries[14b,17,18] without any significant additional cost. Recently, we demonstrated that DFT-D3 also improves the structural features of the lysozyme

protein in the framework of a quantum refinement scheme that we are currently developing.[6] This study also showed that dispersion-corrected corrected HF theory is a valuable tool, which is in accordance with other recent findings for noncovalent interaction energies and geometries.[7,14b, 16]

However, previous studies on proteins – including ours – have neglected important errors induced by using small basis sets. We follow here a classification allowing these errors to be described as two separate problems: the basis-set-superposition error (BSSE) and the basis-set-incompleteness error (BSIE).[19,20] BSSE describes the well-known fact that noncovalent interaction energies are overestimated when using incomplete, atom-centred one-particle basis sets. In the picture of a non-covalently bound dimer one can imagine that the monomers "borrow" atomic basis functions from each other, when calculating the absolute energy of the dimer. Compared to a treatment of each isolated monomer, the dimer itself is therefore artificially overstabilised. This problem affects of course not only energies but also all properties derived from them such as geometries. For the intermolecular case of a noncovalently bound dimer, the counterpoise correction of Boys and Bernadi is a useful remedy.[21] Although questioned frequently in the literature,[22] it has become a popular – albeit cost-expensive – tool to estimate BSSE effects. Note that other corrections have been proposed, each with their own advantages and disadvantages.[23]

In the same spirit as for the dimer case, one can also imagine that parts of a single molecule borrow basis functions from other regions of the same system. This has been defined as intramolecular BSSE (IBSSE). In our context, it is important to note that IBSSE was found to influence the conformational energy profile of peptides,[24] and therefore it is expected that it will also significantly affect protein structures. It is rather difficult to apply these above mentioned

BSSE corrections to IBSSE, as they could possibly introduce covalent-bond breakage, unusual spin states or unchemical fragments. They can also introduce arbitrariness that makes them technically and conceptually very difficult to apply in a "black-box" generalised fashion. Finally, these corrections are also time-consuming, prohibiting their application to large systems. These shortcomings were recently addressed by Kruse and Grimme who developed a new approach called "gCP" ("geometrical counterpoise correction").[20] The name stems from the fact that gCP is only based on the atomic coordinates of the system and does not directly take into account the atomic basis functions, making it very time-efficient. In principle, gCP shares some parallels to DFT-D3. It is an additive atomic pair-wise correction that is evaluated at basically no computational cost when combined with a HF or DFT calculation. It contains three empirical parameters fitted to Boys-Bernardi counterpoise corrections for a set of intermolecular interaction energies. However, it is straightforward to apply this method also for the treatment of IBSSE. Currently, gCP is available for four small basis sets, including one minimal basis set, and it was designed for the treatment of large systems. In their original work, Kruse and Grimme demonstrated its applicability to various test cases, including a minimal-basis-set optimisation of the crambine protein comprising about 600 atoms. In a separate study, Kruse et al. enhanced the popular B3LYP/6-31G* model chemistry with the help of DFT-D3 and gCP.[25] It has been often argued[25] that this level of theory relies for its widespread success on error-compensation between missing London-dispersion and overstabilisation of BSSE. However, this error-compensation is completely unforeseeable and Kruse et al. demonstrated that actually B3LYP-D3-gCP/6-31G* is a more reliable method than the original, providing improved accuracy for general thermochemistry and, in particular, for organic reaction energies and barrier heights. Herein, we will elaborate whether the same can also be said for geometries.

The gCP approach effectively corrects for BSSE but does not include BSIE effects. Very recently, Sure and Grimme outlined how HF, even when corrected with DFT-D3 and gCP significantly overestimates the lengths of polar bonds when applied with a minimal basis set.[26] However, HF with the above-mentioned corrections is a valuable tool for the treatment of large systems and they addressed this BSIE problem by introducing a third, empirical correction. It was fitted against hybrid-DFT bond-lengths obtained with a triple-ζ basis for organic molecules. The entire approach was dubbed HF-3c, as it contains the three corrections for dispersion, BSSE and BSIE. Applications to non-covalently bound dimers, large van-der-Waals complexes and gas-phase structures of small proteins showed that indeed HF-3c is a promising, time-efficient new tool that is worthwhile to investigate further.

As mentioned before, we have recently engaged in developing a full quantum refinement scheme for protein X-ray structures.[6] In this framework, we are faced with all three problems addressed above, as the usage of high-level QM methods, which include London-dispersion effects, or large basis sets is prohibited. Particularly the BSSE and BSIE problems have not been addressed thoroughly in this context. Previous studies ignored these effects, partially because of the lack of efficient corrections, and partially also because error compensation was expected.[4] However, we argue that having a more reliable method should always be favoured over unforeseeable error-compensation effects and we address this issue herein. In this study we concentrate on the effects of method, London-dispersion, BSSE and BSIE. Before thoroughly investigating their effects on protein crystal structures, their influence on peptide model systems must first be established. Quantum chemical studies of peptides in the gas-phase provide first insights into method performance and help in the development of low-level methods for larger proteins.

Most previous model studies of calculating protein properties have concentrated on relative conformational energies of either amino acids,[27] di- and tripeptides,[28] or very recently also some biologically-relevant tetrapeptides.[29] To enable focus on the quality of optimised geometries in a systematic fashion, Hobza and co-workers have introduced a test set of 26 tripeptide structures under the name P26.[30] The tripeptides in this set contain aromatic side-chains and are therefore ideal test cases to study London-dispersion and IBSSE effects. Initially, Hobza and co-workers briefly considered the effects of London-dispersion using Grimme's older DFT-D2 correction, but effects induced by small basis sets were not considered.

We also introduce a new test for method quality based on the conformers of cysteine dimers linked through a disulphide bridge. In our previous QM refinement study, we identified the description of cysteines as crucial to obtain proper agreement with measured X-ray reflections.[6] Particularly, disulphide bridges were shown to be very sensitive to the method of choice, and our new test set reveals difficult cases.

When treating protein crystal structures, one also has to deal with enclosed clusters of water molecules. Any theoretical method that works well for proteins must inherently also describe water clusters properly. This has rarely been regarded in this context and therefore we examine the influence of BSSE and BSIE on the structures of water hexamers.[31]

The test sets are briefly introduced in the next section. Section 3 describes all relevant theoretical and computational details. Section 4 addresses the tests sets under consideration of the outlined problems. We are confident that this study sheds light into the various effects of chosen method and small basis sets and that our findings can be used as guideline for method development for large proteins.

2. The test sets

2.1 The P26 set

Hobza's P26 set[30] comprises a total of 26 conformers of 5 peptides that contain aromatic sidechains. Table 1 shows each peptide's composition and number of conformers. The conformations were selected from a total of 76 structures and encompass a diversity of backbone and side-chain arrangements. Detailed information about the generation of the conformers is given in ref. 30.

The P26 structures were obtained by MP2/cc-pVTZ[32] optimisation. While this level of theory is not sufficient to describe accurately the relative conformational energies of peptides,[16,30] Hobza et al. have argued that for covalent bonds it is qualitatively acceptable. However, some of the P26 conformers also contain hydrogen bonds, and recently it was shown for tetrapeptides that IBSSE produces a sizeable overestimation of hydrogen-bond strength at the MP2/cc-pVTZ level.[29] Also it is known that diffuse functions improve the description of hydrogen bonds,[33] and the most straightforward way in which these structures could be improved would be to use aug-cc-pVTZ[32] instead of cc-pVTZ. However, such an expansion also introduces costly diffuse functions with high angular momenta that are actually not needed for the description of hydrogen bonds, with previous benchmark calculations on water clusters showing that it is sufficient to just use one set of diffuse s- and p-functions on each heavy atom.[16, 34] We follow this strategy and dub this modified basis set aug'-cc-pVTZ.

We reoptimised the P26 data set at the MP2/aug'-cc-pVTZ level, and all structures are provided in the Supporting Information (SI). Comparisons with the original structures show that covalent bond lengths are basically not affected whilst hydrogen bonds increase on average by

0.04 Å. Root-mean-square-deviations between the original and reoptimised structures are usually below 0.05 Å (see Table S1 in SI). Only in two cases the deviations are larger: for GGF05 the RMSD is 0.51 Å and for WG10 it is 0.116 Å. The average RMSD for the entire set is 0.046 Å. While this difference is small compared to the gross effects that we focus on, having the most reliable reference data is always an advantage.

Finally, we would like to comment on the accuracy of the structures. The MP2 structures should not be understood as a quantitative reference. There is still a portion of remaining IBSSE (between 5 and 10% for the chosen basis set), not all of the electron correlation is covered and MP2 itself has inherent problems with London-dispersion. However, the purpose of this study is to examine lower levels of theory and to demonstrate the effects of London-dispersion, BSSE and BSIE. The MP2 structures are only used to qualitatively support our conclusions and findings. We additionally carried out analyses with SCS-MP2[35]/cc-pVTZ and B2PLYP-D3[36]/cc-pVTZ structures and the main conclusions for the lower levels of theory are still the same (see SI). However, a thorough discussion of these additional structures is beyond the point of this manuscript.

2.2 The CYS2 set

In our previous quantum refinement study, we found that it is particularly difficult to describe disulphide bridges appropriately and that statistical analysis tools such as *R*-factors are very sensitive to errors made for cysteine residues.[6] Relative conformational energies of cysteine monomers in gas-phase have been investigated thoroughly in quantum chemical benchmark studies,[16, 27c] but no analogous test set is available for disulphide-linked cysteines. We introduce the model system shown in **Figure 1**, in which methylamide and acetyl groups cap the C- and N-

termini of the cysteine backbones, respectively, to mimic effects of a continuing peptide backbone. In fact, although this system is considered to be a model in our context, it also has been studied experimentally in the past, with a focus on reactivity with oxygen and electronic circular dichroism spectroscopy.[37]

We carried out a thorough conformational analysis of the system and will elaborate these results elsewhere. For this study we picked the three conformers shown in **Figure 1**, named CYS2a - CYS2c, based on their structural appearance and optimised them at the MP2/aug'-cc-pVTZ level of theory, to be consistent with our treatment of P26 (Cartesian coordinates of all structures are provided in the SI). CYS2a and CYS2b appear more rigid due to two hydrogen bonds connecting the adjacent backbones with each other. CYS2c only has one hydrogen bond connecting the two backbones and has overall a less rigid appearance.

2.3 The water hexamer test set

Very recently, the structures of the cage, prism and book isomers of water hexamers were resolved by broadband rotational spectroscopy.[31] The authors provided accurate experimental estimates and vibrationally averaged MP2 results for the O-O distances in those clusters. Recently, these structures were used by Hujo and Grimme for an evaluation of van-der-Waals (DFT-NL) functionals for large basis sets that do not suffer from BSSE.[18] Their analysis showed that using the experimentally obtained structures gives an almost quantitative picture of tested QM methods, even though they were not vibrationally averaged. We will follow the same procedure here in this work, with a focus on smaller basis sets, BSSE and BSIE.

3. Technical details

All calculations were carried out with TURBOMOLE 6.4.[38] The convergence criterion for each SCF step was set to 10^{-7} E_h. Geometries were optimised until the energy change between two subsequent optimisation steps was below the same energy threshold as for the SCF calculations. MP2 optimisations of the structures of the P26 and CYS2 sets were carried out using aug'-cc-pVTZ, as defined in Section 2. All MP2 calculations were sped up with the resolution-of-the-identity (RI) approximation and respective auxiliary basis functions were taken from the TURBOMOLE library.[39] Subsequent optimisations of all systems were carried out by Hartree-Fock (HF) and the following DFT methods: BLYP,[40,41] BP86,[40,42] PBE,[43] B97-D,[12d] TPSS,[44] B3LYP,[45] PBE0,[46] PW6B95,[47] and BHLYP.[48] These calculations were carried out with the cc-pVTZ triple-ζ basis set,[32] Pople's 6-31G*[49] and Ahlrichs' SVP and SV[50] double-ζ sets, and Huzinaga's minimal basis set MINIS,[51] which was taken from the EMSL basis-set exchange library.[52] Except for MINIS, the calculation of Coulomb contributions of HF and the DFT methods was carried out with the RI-J approximation, with auxiliary basis sets again being taken from the TURBOMOLE library.[53] All DFT calculations were carried out with the TURBOMOLE grid *m5*;[54] test calculations using the TURBOMOLE "reference" grid, a grid comparable to Gaussian's ultrafine grid, indicated the quality of this approach.

The HF and DFT calculations were additionally carried out with Grimme's DFT-D3 and/or gCP corrections. These are independent additive corrections of the form:

$$E^{HF/DFT-D3-gCP} = E^{HF/DFT} + E^{DFT-D3} + E^{gCP}, \quad (1)$$

where $E^{HF/DFT}$ is the original HF or DFT energy, $E^{DFT\text{-}D3}$ is the DFT-D3 energy contribution and E^{gCP} is the gCP energy contribution.

The DFT-D3 method is well established.[14] It is an additive, atomic pair-wise correction that considers the correct asymptotic R^{-6} long-range and the R^{-8} medium-range behaviours for interatomic distances. Herein, we applied DFT-D3 with the Becke-Johnson damping function, which adds a constant contribution even at the unified-atom limit.[14b] This correction is sometimes called DFT-D3(BJ) to distinguish it from the first version of DFT-D3.[6, 14b, 15-16] However, DFT-D3(BJ) has been shown to be superior to the previous version and has been recommended as a standard procedure for dispersion-corrected DFT[14b] and hence we skip the suffix in the parentheses. DFT-D3 depends on three empirical parameters that have been optimised for HF and about 50 DFT methods and implemented in Grimme's *DFT-D3* program[55] as well as in TURBOMOLE 6.4.

The gCP correction accounts for BSSE without any significant cost, making it particularly relevant for applications to large systems.[20] It contains three empirical parameters which have been fitted to Boys-Bernardi counterpoise corrections to intermolecular interaction energies for the basis sets 6-31G*, SVP, SV and MINIS. Different sets of parameters were proposed for HF and for DFT. Unlike DFT-D3 there are no functional-specific parameters and all parameters are the same for all DFT approximations. The benefit of combining DFT-D3 and gCP has been demonstrated recently.[20, 25] We use Grimme's freely available program gCP.[55]

Sure and Grimme very recently developed a correction for basis-set incompleteness, which addresses the problem of elongated bond lengths for HF combined with minimal basis sets.[26] This correction depends on three additional empirical parameters fitted using a set of B3LYP triple-ζ structures of organic molecules. The correction is specifically designed for a

basis set, called MINIX, which is a mixture of MINIS, additionally augmented with one set of p-type polarisation functions for certain elements. Elements beyond potassium are described by a double-ζ basis, but this is of no concern in our present context because the heaviest element studied is sulphur. This new correction is combined with DFT-D3 and gCP, and the entire approach is called HF-3c ("3 corrections"). These corrections were obtained with a special program obtained from the authors.

4. Results and discussion

First, we focus our discussion on London-dispersion and IBSSE, demonstrating problems that arise owing to basis-set incompleteness. The HF-3c method is then applied to remedy these problems.

4.1. Discussion of the P26 set

4.1.1 Effects of DFT-D3 on the cc-pVTZ level

It has already been outlined that is it crucial to use dispersion-corrections in DFT optimisations of biologically relevant structures. For instance, in their original study of P26, Hobza et al. recommended TPSS-D/6-311++G(3df,3pd) as promising DFT method.[30] However, this recommendation was based on the older DFT-D2 correction.[12a] Additionally, those TPSS-D results were compared with B3LYP/6-31G* calculations, and as expected that latter level of theory underperforms because of the lack of a dispersion correction. Dispersion-corrected B3LYP was not discussed at all, nor was it treated with the same basis set as TPSS-D.

Therefore, we start the discussion of the P26 set with a short analysis of London-dispersion effects based on DFT-D3 and the same basis set for each method. In this section, we restrict our discussion to the cc-pVTZ basis set and the TPSS, B3LYP and HF methods. The analysis is based on root-mean-square deviations (RMSDs) involving the Cartesian coordinates of all atoms, and they can be influenced by both short-range effects in covalent bonds, and long-range inter-residue effects stemming from London-dispersion and hydrogen bonding. A detailed list of all RMSDs of these tested methods with respect to MP2/aug'-cc-pVTZ geometries for all 26 structures is given in the SI in Table S3. As can be seen therein, for both DFT methods, the effect of the dispersion correction is large and using DFT-D3 reduces the RMSDs significantly, for example by up to 1 Å in the case of conformer FGG252. The only exception is structure FGG215. During both optimisations (with and without D3-correction) the phenyl-ring turned away from the terminating glycyl part leading to very similar RMSDs around 0.66 Å. Table 2 shows RMSDs averaged over all 26 systems. The averaged RMSD of 0.671 Å for TPSS is reduced to 0.141 Å for TPSS-D3, which is a bit better than the value of 0.151 Å for TPSS-D2 and the previously reported value of 0.16 Å for TPSS-D2 based on MP2/cc-pVTZ structures. B3LYP-D3 has a slightly lower averaged RMSD of 0.120 Å.

Using the DFT-D3 correction has a large effect on HF optimised structures and improves them in all cases, including FGG215 (Table S3), for which the phenyl-ring does not turn away any more. The final averaged RMSD of 0.115 Å compares well with that obtained using B3LYP-D3 (**Table 2**).

4.1.2 Effects of gCP on the 6-31G* level

Next, we address the effect of IBSSE on peptide structures for small basis sets. We first discuss Pople's 6-31G* set as it is very widely used. Figure 2 shows averaged RMSDs for P26 using the same three methods as discussed in the previous section as well as the BP86 functional. BP86/6-31G* without any corrections has been proposed as a reliable method for efficient peptide and protein optimisation.[4a, 4c, 4e] If none of the corrections are applied, then all RMSDs are between 0.525 Å and 0.567 Å, with BP86/6-31G* having the highest value. When the structures are optimised using the gCP correction, all RMSDs increase. The smallest increase is observed for HF, however it is still sizeable with an RMSD of 0.673 Å for HF-gCP compared to 0.559 Å for HF. The effects of BSSE for B3LYP are similar to those for HF but much larger for the meta-GGA and GGA functionals. Particularly BP86 shows the largest effect with an increase from 0.567 Å to 0.812 Å. These numbers reflect what has already been described for intermolecular BSSE: the overstabilisation leads to shorter distances between the aromatic moieties and the opposite termini of the peptide backbones. Having shown that these systems are strongly influenced by BSSE, the lower RMSDs for uncorrected methods on their own imply that BSSE corrections should not be used. We further comment on this later, after having discussed the effects of the simultaneous application of both corrections.

As **Figure 2** shows, adding just a dispersion correction to the pure QM method improves the RMSDs to between 0.109 Å and 0.144 Å. However, there is still a stabilising contribution from BSSE and for DFT-D3-gCP and HF-D3-gCP we still observe a slight increase compared to DFT-D3 and HF-D3. When both corrections are applied, the lowest RMSD is observed for HF-D3-gCP (0.125 Å), followed by B3LYP-D3-gCP (0.168 Å), BP86-D3-gCP (0.179 Å) and TPSS-D3-gCP (0.205 Å). These numbers clearly show that the popular BP86/631G* and B3LYP/6-

31G* combinations should not only be used with caution for energetic properties (as recently shown for B3LYP[25]), but also for geometries. The effects or BSSE and dispersion are also depicted for BP86 and the FGG99 conformer in Figure 3.

One argument against using gCP would be the lower RMSDs calculated for the respective methods without this correction, relying on error compensation effects to provide better results. However, in ref. 25 it has been outlined that these errors are not always foreseeable. Two examples in this study are the WG10 and WGG05 conformers. In both cases B3LYP yields a higher RMSD than B3LYP-gCP, which is contrary to the averaged RMSDs of the whole set. Therefore, we argue that is better to have a more robust approach, even though it might lead to an increase in statistical errors for some propoerties. In accordance with the findings for energetics, we also recommend to use both the DFT-D3 and gCP corrections for geometries to get a better picture of the "true" performance of a method.

So far we have concentrated mainly on how London-dispersion and BSSE effects influence the distances between the aromatic moieties and the adjacent peptide backbones. However, also hydrogen bonds are present in some systems and analysis of the effects both on those gives us further insight into the feasibility of using of small basis sets for peptide optimisations. Figure 4 shows the hydrogen-bond lengths averaged over all systems for the same four methods as discussed before. Overall, the same trends as before are observed: whenever the gCP correction is applied, the hydrogen-bond lengths increase by up to 0.2 Å for the DFT methods. Only HF seems to be less affected by BSSE with changes of under 0.1 Å. Interestingly the averaged H-bond lengths are the same for BP86 and BP86-D3-gCP. The same is also seen for B3LYP, perhaps indicating why these uncorrected methods are widely used and provide good results. This level error cancellation is not observed for TPPS and HF. Particularly HF needs the

DFT-D3 correction to shorten the bond-lengths. Thus, we have demonstrated that there is a sizeable influence of BSSE on the hydrogen-bond lengths and the distances between the two adjacent parts of the peptides. Next, we will discuss the influence of basis sets.

4.1.3 RMSDs for various basis sets and methods

We have now established that a combination of both corrections is beneficial for both DFT and HF. Finally, we compare various density functionals and HF with each other and discuss basis-set effects. Figure 5 shows averaged RMSDs for P26 for cc-pVTZ and the four basis sets for which the gCP correction was parameterised; full details are provided in the SI in Table S3. For cc-pVTZ, dispersion-corrected GGA functionals BP86-D3 and BLYP-D3 yield slightly lower RMSDs (about 0.12 Å) than TPSS-D3 and B97-D3 (0.14 Å), competing with hybrid functionals. PBE-D3 yields the worst RMSD at the cc-pVTZ level of almost 0.2 Å. The best result of this study is found for PW6B95-D3 (0.111 Å), which is in accordance with previous findings for this functional in terms of robustness for accurate energetic and geometries.[14b, 16] B3LYP-D3 follows closely, whereas PBE0-D3 and BH-LYP-D3 behave similarly to GGA functionals. HF-D3 is competitive with PW6B95-D3 with an RMSD of 0.115 Å.

For small basis sets the gCP correction is important. However, gCP does not correct for BSIE and the RMSDs for all DFT methods shown in **Figure 5** for 6-31G* are slightly higher than those for cc-pVTZ. Also the differences between the various functionals are now larger. The best two DFT methods are BLYP-D3-gCP and PW6B95-D3-gCP with RMSDs of 0.161 Å. The basis set dependence of HF is much less than that of DFT, which is why HF-D3-gCP/6-31G* has an RMSD of 0.125 Å, which is very close to that of HF-D3/cc-pVTZ. When going to the

Ahlrichs basis sets SVP and SV, the RMSDs increase again for all methods with SV having significantly higher RMSDs. The lack of polarisation functions in SV is a likely reason for that. The trends for all methods remain basically the same and HF-D3-gCP yields in all cases the lowest RMSDs (0.153 Å for SVP and 0.220 Å for SV).

The RMSDs for the MINIS minimal basis set are unproportionally large. For the DFT methods, MINIS gives different trends to those found for the other basis sets (**Figure 5**). Most DFT methods lead to RMSDs in a range between 0.48 Å and 0.64 Å, with BHLYP being the exception with an RMSD of 0.351 Å, which is clearly contrary to the trends found for larger basis sets. HF-D3-gCP is the only method that does not show such a big increase. In fact, its RMSD of 0.256 Å is similar to or better than the RMSDs for the DFT-D3-gCP/SV methods. We analysed the results further and could conclude that even for those smaller basis sets, the same basic trends in terms of BSSE and dispersion are observed as discussed for 6-31G*, which means that neither of the corrections is responsible for these high increases in the RMSDs. In fact, this only leaves one likely explanation, the basis-set incompleteness itself.

Based solely on RMSDs, we can recommend the BLYP-D3, PW6B95-D3 and HF-D3 methods, in combination with gCP for small basis sets. However, the effects of BSIE are analysed in the next section and it is addressed whether these recommendations are also valid for a reliable description of covalent bond lengths.

4.1.4 Understanding the BSIE effects on covalent bonds

The analysis of RMSDs for P26 demonstrated the benefits of including dispersion and BSSE corrections in DFT and HF optimisations. We have also mentioned that adding these corrections influences H-bond lengths and distances between aromatic moieties and the adjacent

backbone termini. However, when using very small basis sets, the RMSDs increase noticeably and one possibility for this is the BSIE. An ideal method for applications such as quantum refinement of protein X-ray structures must also yield adequate bond lengths as these do contribute to the *R*-factors that compare theoretical models with measured X-ray reflections.

We analyse four types of bonds: the carbonyl bond, the C-N peptide bond, the C-C_α bond, and N-C_α bonds. We note in passing that adding both corrections to DFT not have a sizeable effect on covalent bond lengths and only a marginal effect of HF, for which bonds are slightly shortened (see Table S6 for more for information). Therefore our analysis includes those corrections in the following. The results for the five GGA and meta-GGA functionals are averaged as well as the calculated bond lengths for the four hybrid functionals. Additionally also the results for HF-D3(-gCP) are discussed. These results are shown for all basis sets and the four different types of bonds in Figure 6. For all bond lengths and all methods we see the same basis-set dependence. The less complete the basis set, the longer the bond lengths get. Whereas 6-31G* and SVP are still similar to cc-pVTZ, stripping the basis set of polarisation functions (SV basis set) leads to elongation, particularly for the polar carbonyl group. The next largest elongation is seen when going from the SV to the minimal basis set. For QM refinement purposes the bonds for MINIS are unacceptably long. Note, that this is not only true for HF but for DFT, as well.

Using big basis sets is prohibited when treating big biomolecules and a compromise has to be found between computational time and basis-set incompleteness. For this purpose, the averaged calculated bond lengths for the MP2 structures are shown as guidelines in Figure 6. It is textbook knowledge that HF theory yields too short bong lengths,[56] which can be seen for the cc-

pVTZ, 6-31G* and SVP results. Due to basis-set incompleteness, HF-D3-gCP/SV, however, agrees very well with the MP2 results. For (meta-)GGA DFT it is known that bond lengths are slightly too long,[56] which is also observed herein, in particular for the SV and MINIS basis sets. Hybrid DFT results lie in between GGA and HF as they contain portions of both methods. Also here cc-pVTZ and the double-ζ sets with polarisation functions can be recommended.

We therefore extend our recommendations from the previous section by saying that, based on the findings for P26, a hybrid functional like PW6B95-D3 empirically gives the best results when used with cc-pVTZ or with GCP and 6-31G* or SVP. When computational restrictions limit the basis-set size to at most SV, we recommend using the HF-D3-gCP method instead.

4.2 Results for the CYS2 set

4.2.1 Discussion of RMSDs

RMSDs for the three CYS2 conformers (see **Figure 1** and Section 2.2) are calculated without taking into account the terminating methyl groups, as they are rather floppy and rotation of these might affect the RMSDs and hence lead to artefacts in their interpretation. The RMSDs for all tested levels of theory are listed in the SI. Herein, we only restrict ourselves to a discussion of BP86, B3LYP and HF combined with the 6-31G* basis set.

The RMSDs for DFT-D3 and gCP corrected and uncorrected calculations are shown in Table 3. The trends for the RMSDs of the CYS2a and CYSb conformers are very similar to each other (this is also found for the other levels of theory shown in Table S8). The likely reason for this is the rigidity induced by the two hydrogen bonds connecting the adjacent backbones with

each other. For all methods we observe the same trends discussed for P26. Dispersion contributions are necessary and adding DFT-D3 halves the RMSDs. Applying the gCP correction leads to structures with slightly higher RMSDs, which means again that BSSE-uncorrected methods overestimate the non-covalent interactions. The only exception is HF-D3-gCP for CYS2a, which yields a lower RMSD than HF-D3 (0.046 Å vs. 0.067Å). HF-D3-gCP provides lower RMSDs than the two density functionals for CYS2a and the second lowest of all the methods for CYS2b.

This picture is slightly different for CYS2c. The RMSDs show a larger range for different methods. BP86-D3-gCP has the highest RMSD of the three methods discussed in **Table 3** (0.413 Å) and it is followed by B3LYP-D3-gCP with 0.326 Å. HF-D3-gCP is the only method to predict results close to the MP2 structure with an RMSD of just 0.103 Å. The trends discussed for DFT-D3 and gCP are the same as for the other two conformers.

The trends for averaged RMSDs (see Figure S1) are very similar to P26. Smaller basis sets yield higher RMSDs, and in general the same methods as for P26 can be recommended.

4.2.2 BSIE effects on S-S and C-S bonds

The basis set and method dependence of bond-lengths involving carbon, oxygen and nitrogen is the same as discussed for P26. Figure 7 therefore only shows S_γ-S_γ and C_β-S_γ bond lengths averaged over the three conformers, for the (meta-)GGAs, the hybrid functionals and HF. In principle it is again observed that bond lengths get shorter when increasing the amount of Fock exchange. Hybrid functionals give very similar results to MP2 using cc-pVTZ but again the smaller 6-31G* and SVP empirically give better values for HF. The necessity of polarisation

functions is more dramatic for S-S and C-S bonds than for the bonds discussed for P26. Indeed, the SV basis set is no longer useful for S-S and C-S bonds, with the averaged S-S bond lengths ranging between 2.24 Å and 2.32 Å, compared to 2.06 Å for MP2 or the crystallographic value of around 2.02 – 2.03 Å.[57] Note, however, that the effect is also seen for the C_β-S_γ bond is less pronounced than that for the disulphide bridges. Nevertheless, any QM refinement of protein structures without at least polarisation functions on the S-atoms is unlikely to yield acceptable *R*-factors. The findings also confirm our previous observations that cysteines are difficult to treat and here we have a likely explanation for this behaviour.[6] However, the advantage of using basis sets such as SV, is of course the computational efficiency. One strategy with low computational cost and a practical level of accuracy may be to use SVP on S-atoms only and SV for all other elements. This in combination with HF-D3-gCP is a possible practicable level of theory that can be tested in future QM refinement applications. Note that the gCP correction would need to be applied with the parameterisation for SV and not SVP, as sulphur atoms were not included in the fit set used to determine the gCP parameters.

4.3 Results for the water hexamers

For the water clusters the averaged nearest O-O distances are discussed for each of the three systems (the *book*, *cage* and *prism* isomers, as discussed in Section 2.3). The experimental values are shown in Table 4 together with results for BP86, B3LYP and HF used with the 6-31G* basis set. The results demonstrate again the established effects of both corrections. Adding the DFT-D3 correction leads to a shortening of the O-O distances, as discussed by Hujo and Grimme for the same systems.[18] For BP86 and B3LYP, the distances usually get shorter by 0.01 Å to 0.02 Å. For HF this effect is more pronounced and distances decrease by 0.06 Å to 0.07 Å. BSSE has a

large influence on the distances and uncorrected structures have distances that are too short compared to the experimental values shown in the table. The differences between corrected and uncorrected structures lie between 0.06 Å and 0.10 Å. HF-D3-gCP/6-31G* yields the best agreement with experiment. For the *book* isomer, the averaged distance is by 0.01Å too short, for the *cage* isomer by 0.02 Å and for the *prism* 0.05 Å. B3LYP-D3-gCP distances are between 0.02 and 0.06 Å too short, BP86-D3-gCP distance are too short between 0.05 and 0.1 Å.

When testing various basis sets, we made an important observation about using the minimal basis set MINIS. For all DFT methods other than BHLYP, all hexamer structures were qualitatively inconsistent with experiment, containing hydronium and hydroxyl ions. This effect depends on the amount of Fock exchange, with BHLYP and HF not suffering from this effect. Examples of these autoionised structures are shown in Figure 8. Note that this effect has nothing to do with the dispersion or BSSE corrections. Additionally we also checked whether the grid could be an error source. In all cases the same effect was observed, making it likely that basis-set incompleteness in combination with an inadequate description of exchange effects lead to autoionisation. Additionally, we also carried out optimisations with the COSMO continuum solvation model[58] and a log-range dielectric constant of 4, a value that is often used to mimic protein environments. Again, we found the same effect. When increasing the basis set size, the same was seen for some structures and levels of theory for the SV basis set. BP86 and PBE give autoionised structures for the *book* isomer, and BP86, B97-D and PBE0 struggle with the description of the *prism* isomer. However, in contrast to MINIS, these problems were overcome by using the COSMO model. Moreover, just using the gCP correction did not show this effect either, whereas the pure method, the method corrected with DFT-3, or with DFT-D3 and gCP showed these problems with the SV basis.

Figure 9 shows the averaged errors for the O-O distances for (meta-)GGAs, hybrid functionals and HF the different basis sets. In all cases the results were corrected for BSSE and dispersion effects. All levels of theory underestimate the O-O distance. The largest underestimation is seen for the SV basis set, while the smallest effects are observed for SVP and 6-31G*. The fact that cc-pVTZ underestimates the O-O distances more than the double-ζ basis sets is an indication that still some error compensation effects play a role. HF-D3-gCP is for all basis sets the best option, although also hybrid functionals in combination with SVP show very similar errors. The results for HF-D3-gCP/MINIS lie in between those for SV and SVP.

4.4 Results for the HF-3c method

We have identified the BSIE as a likely source for the elongated bond-lengths in the examples discussed in this paper. Next, we test Grimme's new HF-3c, which combines the DFT-D3 and gCP corrections with a minimal basis set and a third empirical correction to overcome BSIE effects. It can also be seen as a complementing method to HF-D3-gCP/MINIS, and we therefore compare HF-3c with HF-D3-gCP/MINIS for our three test sets. As shown in the SI, the effects on RMSDs are not very big, when comparing both methods with each other. For P26 the averaged RMSD improves from 0.263 Å for HF-D3-gCP/MINIS to 0.223Å for HF-3c. The averaged RMSD for CYS2 increases slightly from 0.177 Å for HF-D3-gCP/MINIS to 0.198Å for HF-3c, main cause here is conformer CYS2c, for which the RMSD increases from 0.236 Å to 0.300 Å, a value that is close to hybrid DFT for cc-pVTZ, to which the BSIE correction in HF-3c had been fitted.

The effect of HF-3c on bond lengths is significant though and shown in Table 5 for all bond lengths discussed. In all cases, the bond lengths get shorter and resemble DFT results for

double- and triple-ζ basis sets. HF-3c does not show the strong overbinding tendency of HF-D3/cc-pVTZ. The biggest effect is seen for the S-S and C-S bonds in the CYS2 test set. The averaged S-S bond length drops from 2.253 Å to 2.101 Å. C-S bond lengths are similarly described as GGA-DFT-D3/cc-pVTZ with 0.184 Å.

The O-O distances in the water clusters are insignificantly shorter for HF-3c than for HF-D3-gCP/MINIS, because the third correction in HF-3c is designed for closer interatomic distances and does not have any significant effect on hydrogen bonds. The average error is –0.09 Å for HF-3c compared to –0.07 Å for HF-D3-gCP/MINIS.

5 Summary and recommendations

For computational efficiency, the quantum chemical study of proteins is often carried out with Hartree-Fock (HF) or density functional theory (DFT) methods combined with small basis sets. However, these approaches face three major problems for the determination of structures to the quality required for say protein X-ray structure refinement applications. HF and current DFT approximations do not adequately describe London-dispersion effects, while small basis sets induce the basis-set superposition (BSSE) and basis-set incompleteness errors (BSIE); the focus of this work was assessing correction schemes for these problems. For HF and DFT, Grimme and co-workers established two corrections called DFT-D3 and gCP. These corrections have in common that they are time-efficient and easy to combine with standard efficient HF or DFT procedures. Sure and Grimme very recently addressed the BSIE problem for HF with minimal basis sets, developing the HF-3c method, includes contains DFT-D3, gCP and a third empirical correction to BSIE and is combined with a minimal basis set.

Our assessment of these methods utilised reoptimised gas-phase tripeptide structures of Hobza's P26 set, comprising 26 conformers of 5 tripeptides with aromatic sidechains, a new set of three cysteine dimers connected by disulphide bridges, as well as water hexamer configurations. These test sets embody the most important feature in protein crystal strctures. One part of our analysis of these methods was based on root-mean-square deviations from reference structures that assess intermolecular dispersion and hydrogen bonding interactions. Additionally, all methods were also assessed based on calculated covalent bond lengths, as the success treatments of proteins, such as quantum refinement, also depends on calculating these accurately.

In **Figure 10** all of our findings are combined and used to make recommendations of practical, accurate computational strategies:

1) Regardless of the chosen basis set, dispersion corrections are crucial for both DFT and HF. We recommend Grimme's DFT-D3 correction in its latest version with Becke-Johnson damping. This correction makes HF a valuable and competitive alternative to dispersion-corrected DFT.

2) If the system size allows it, a basis set of at least triple-ζ quality, such as cc-pVTZ, should be chosen, as this reduces intramolecular BSSE effects. For a basis set of this quality, we favour the PW6B95-D3 method over other DFT approximations, but an alternative is also HF-D3 due to its low RMSDs.

3) If only a basis set of double-ζ quality can be used, it is crucial to correct for intramolecular BSSE and Grimme's new gCP correction is shown to be ideal. Although occasionally a method without this correction seems to yield better results, we demonstrate that

this arises owing to unforeseeable error cancellation and that the DFT-D3-gCP and HF-D3-gCP approaches are more robust. Using DFTD3 and gCP minimises error compensatsion to allow focussing on the quality of other aspects of the computational methods. Depending on the type of basis set we conclude that PW6B95-D3-gCP is recommended when used with the 6-31G* and SVP basis sets, both double-ζ basis sets including polarisation functions. If the system size prohibits usage of polarisation functions, the SV basis set provides a method combined HF-D3-gCP, but polarisation functions must be included on any sulphur atom.

4) If the system size does not allow using a double-ζ basis set, minimal basis sets provide the only option. Basi-set incompleteness effects are severe for minimal basis sets and are not accounted for when using the gCP correction. Indeed, for the MINIS basis set, we show that the DFT-D3-gCP and HF-D3-gCP methods yield unacceptably large bond lengths, making applications such as quantum refinement very difficult. Moreover, for most tested DFT methods, usage of MINIS lead to autoionisation of water clusters. However, we show that the HF-3c method efficiently corrects for this problem to a large extent and provides a valuable new approach. We recommend extending this approach to include DFT methods.

The methods outlined in Figure 10 provide a wide range of efficient and accurate computational schemes that allow performing quantum chemical calculations on proteins and polypeptides for structures optimisations, molecular dynamics and quantum refinement of protein X-ray structures.

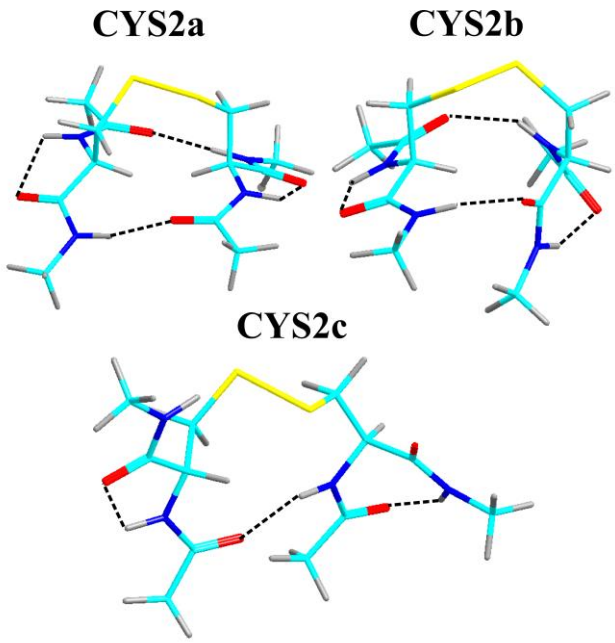

Figure 1. The three conformers of the cysteine-dimer model. Hydrogen bonds are marked with dashed lines.

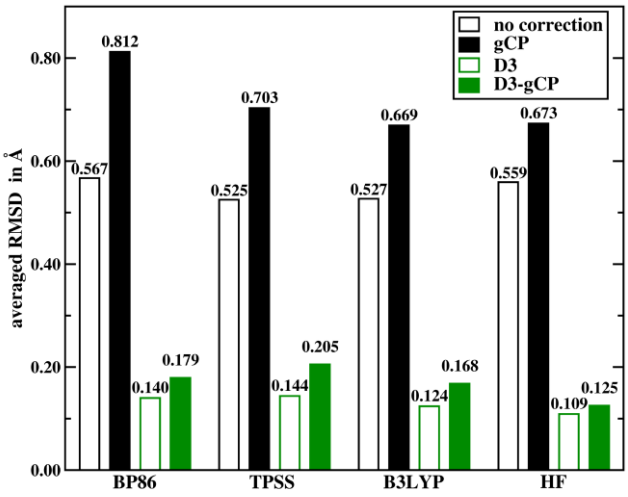

Figure 2. Root-mean-square deviations (RMSDs) for all atoms in Å averaged over the P26 set for four different methods with and without DFT-D3 and gCP corrections. The 6-31G* was used in all cases.

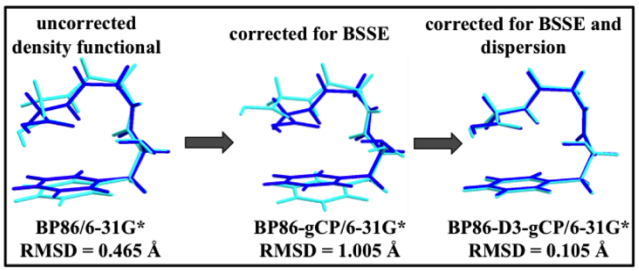

Figure 3. BP86/6-31G* (light blue) structures of the FGG99 conformer compared with the MP2 geometry (dark blue). The effects of adding the DFT-D3 and gCP corrections are shown. Root-mean-square deviations (RMSDs) with respect to MP2 are also shown.

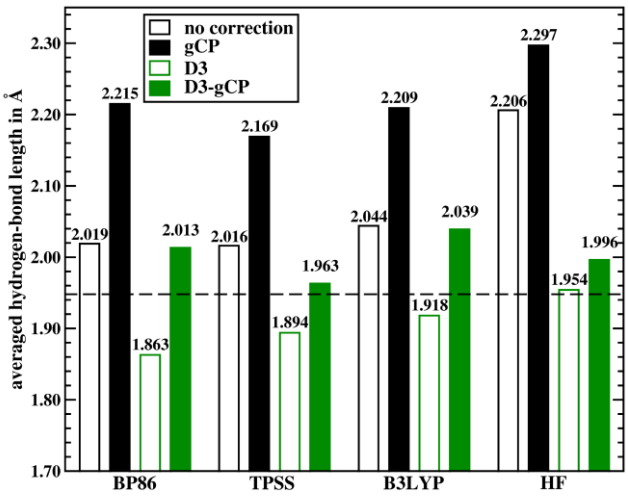

Figure 4. Averaged hydrogen-bond lengths in Å for the P26 set for four different methods with and without DFT-D3 and gCP corrections. The dashed line shows the averaged value for MP2/aug'-cc-pVTZ as a qualitative guideline.

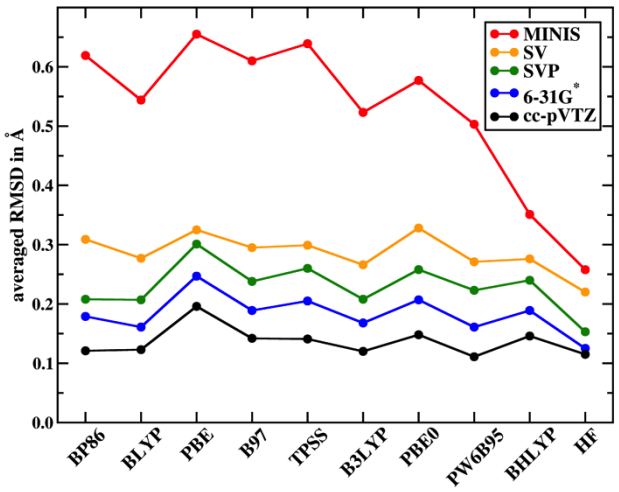

Figure 5. Root-mean square deviations (RMSDs) for all atoms in Å averaged over the P26 set for DFT-D3-gCP methods and HF-D3-gCP for five different basis sets.

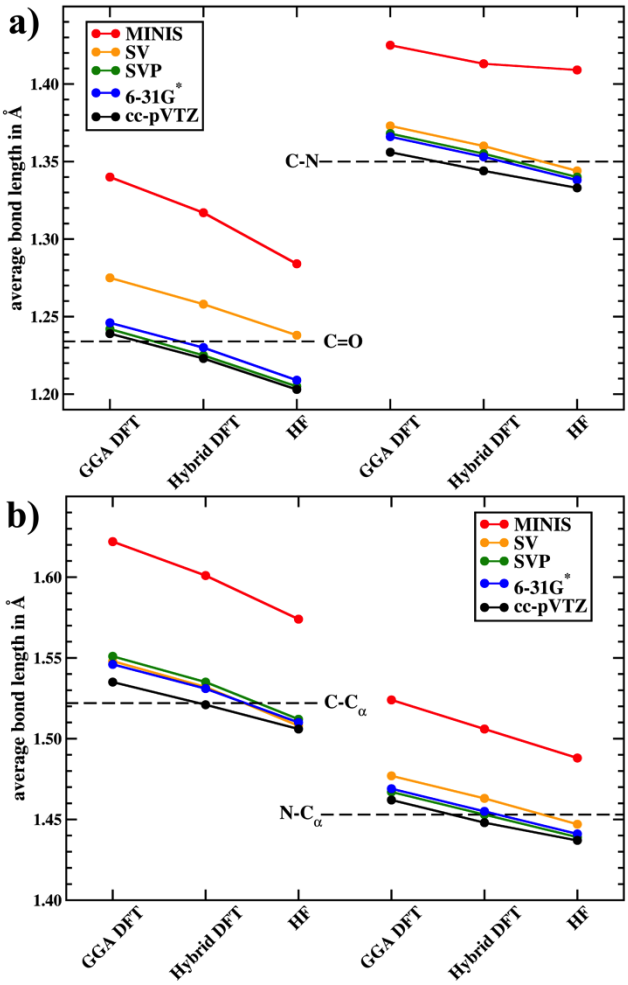

Figure 6. Averaged bond-lengths for four different types of bonds in Å. Carbonyl and peptide bonds are shown in part **a**, lengths between C_α and the carbonyl C-atoms or the peptide N-atoms are shown in part **b**. The values are averaged over all (meta)-GGA or hybrid DFT methods. Also values for HF are shown. Five different basis sets are discussed. The dashed lines show the average bond-lengths for MP2/aug'-cc-pVTZ.

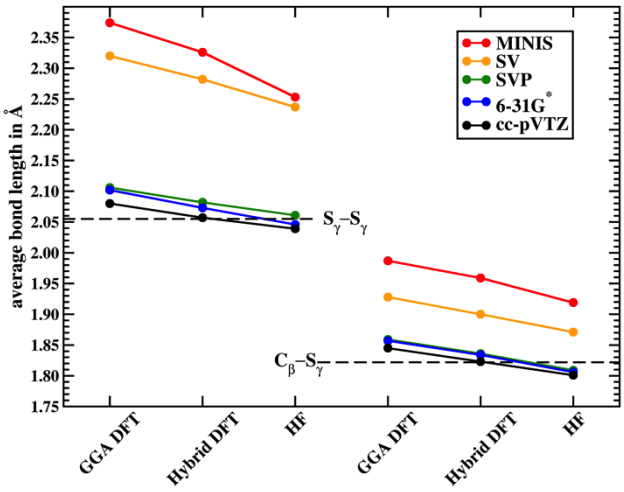

Figure 7. Averaged bond-lengths for the disulphide bridge and the C-S bonds in the cysteine-dimers in Å. The values are averaged over all (meta)-GGA or hybrid DFT methods. Also values for HF are shown. Five different basis sets are discussed. The dashed lines show the average bond-lengths for MP2/aug'-cc-pVTZ.

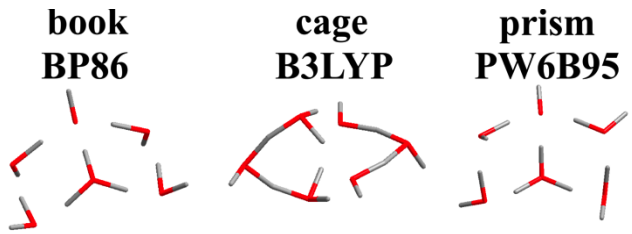

Figure 8. Example of wrong water cluster geometries obtained with density functional methods and the MINIS basis set. The DFT-D3 and gCP corrections are employed in all cases, but not source of the errors.

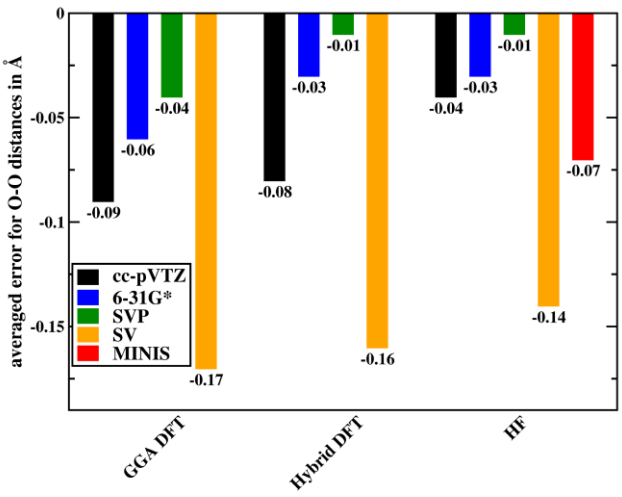

Figure 9. Averaged error for O-O distances in water hexamers for (meta-)GGA DFT, hybrid DFT and HF. All methods were corrected with DFT-D3 and gCP and results for various basis sets are shown.

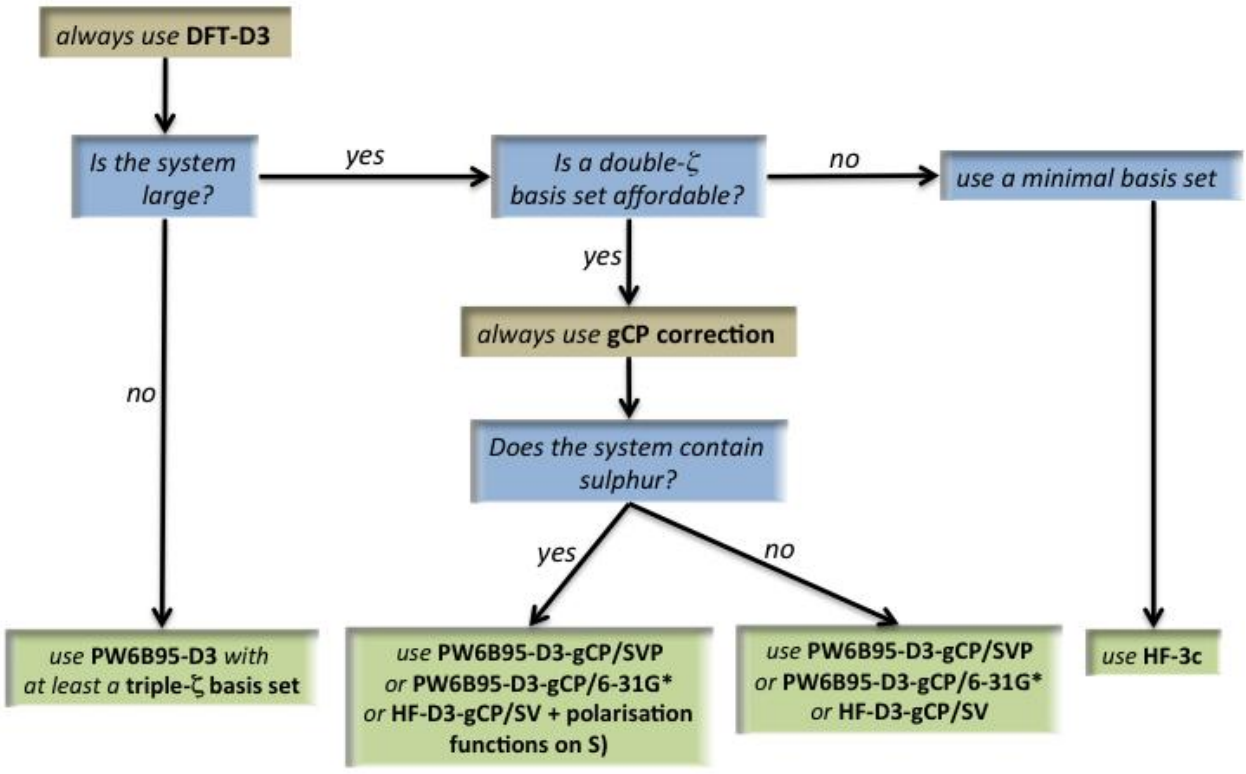

Figure 10. Conclusions and recommendations for the optimisation of polypeptides.

Table 1. Composition of Hobza's P26 benchmark set for peptide geometries.[30]

acronym	sequence	number of conformers
FGG	phenylalanyl-glycyl-glycine	7
GFA	glycyl-phenylalanyl-alanine	4
GGF	glycyl-glycyl-phenylalanine	4
WG	tryptophyl-glycine	5
WGG	tryptophyl-glycyl-glycine	6

Table 2. Averaged root-mean-square deviations (RMSDs) for all atoms in Å for P26 with respect to MP2/aug'-cc-pVTZ geometries. RMSDs are shown for TPSS, B3LYP and HF at the cc-pVTZ level with and without dispersion-corrections.

	uncorrected	with DFT-D3
TPSS	0.671	0.141
B3LYP	0.684	0.120
HF	0.630	0.115

Table 3. Root-mean-square deviations (RMSDs) for all atoms except the methyl groups in Å between MP2/aug'-cc-pVTZ reference geometries and geometries of BP86, B3LYP and HF at the cc-pVTZ level with and without dispersion- and BSSE-corrections. RMSDs are shown for the three conformers of the CYS2 set.

	CYS2a	CYS2b	CYS2c
BP86-D3-gCP	0.075	0.143	0.413
BP86-D3	0.069	0.070	0.428
BP86-gCP	0.165	0.425	0.432
BP86	0.122	0.341	0.426
B3LYP-D3-gCP	0.073	0.201	0.326
B3LYP-D3	0.047	0.055	0.300
B3LYP-gCP	0.166	0.451	0.383
B3LYP	0.122	0.388	0.368
HF-D3-gCP	0.046	0.171	0.103
HF-D3	0.067	0.089	0.088
HF-gCP	0.160	0.498	0.360
HF	0.128	0.450	0.295

Table 4. Averaged shortest O-O distances for the book, cage, and prism isomers of water clusters in Å. Experimental values[31] are compared with three methods with and without DFT-D3 and gCP corrections. The QM results were obtained with the 6-31G* basis set.

	book	cage	prism
exp.	2.80	2.85	2.89
BP86-D3-gCP	2.75	2.78	2.79
BP86-D3	2.67	2.69	2.70
BP86-gCP	2.76	2.80	2.81
BP86	2.68	2.71	2.72
B3LYP-D3-gCP	2.78	2.82	2.83
B3LYP-D3	2.71	2.73	2.74
B3LYP-gCP	2.80	2.84	2.86
B3LYP	2.72	2.75	2.76
HF-D3-gCP	2.81	2.82	2.84
HF-D3	2.77	2.78	2.79
HF-gCP	2.91	2.95	2.96
HF	2.85	2.88	2.90

Table 5. Bond lengths in Å for HF-3c compared to HF-D3-gCP/MINIS and reference values.

	HF-3c	HF-D3-gCP/MINIS	Reference
C-O[b]	1.222	1.284	1.234
C-N[b]	1.379	1.409	1.350
C-C_α[b]	1.568	1.574	1.522
N-C_α[b]	1.456	1.488	1.453
S_γ-S_γ[c]	2.101	2.253	2.055
C_β-S_γ[c]	1.844	1.919	1.822
O-O (book)[d]	2.73	2.75	2.80
O-O (cage)[d]	2.77	2.78	2.85
O-O (prism)[d]	2.79	2.81	2.89
av. error for O-O[e]	–0.09	–0.07	——

[a]MP2/aug'-cc-pVTZ for P26 and CYS2; experimental values for the water hexamers. [b]Averaged bond lengths for the P26 set. [c]Averaged bond lengths for the CYS2 set. [d]Averaged O-O distances for the water hexamers. [e]Averaged error for O-O distances in water hexamers.

Supporting Information. The Supporting Information contains all Cartesian coordinates for the P26 and CYS2 test sets, all root-mean-square deviataions for all tested methods, averaged covalent and hydrogen-bond lengths for P26 and CYS2, and all results for the water hexamers. This material is available free of charge via the Internet at http://pubs.acs.org.

Corresponding Author

*E-mail: jeffrey.reimers@sydney.edu.au

Acknowledgments

Lars Goerigk is supported by a postdoctoral scholarship by the German Academy of Sciences "Leopoldina" under the grant number LPDS 2011-11. This project was also supported by the Australian Research Council under the grant number DP110102932. We are grateful for allocation of computer time from the NCI National Facility in Canberra, Australia, and from Intersect Australia Ltd. We thank Dr Holger Kruse for technical support with the gCP code and Prof. Stefan Grimme for providing us with a pre-print of the HF-3c manuscript and a version of the related program.

References

1. (a) Lahti, J. L.; Tang, G. W.; Capriotti, E.; Liu, T. Y.; Altman, R. B. *Journal of the Royal Society Interface* **2012,** *9*, 1409-1437; (b) Nugent, T.; Jones, D. T. *Journal of Structural Biology* **2012,** *179*, 327-337.
2. (a) Audie, J.; Swanson, J. *Chemical Biology & Drug Design* **2013,** *81*, 50-60; (b) Du, Q. S.; Huang, R. B.; Chou, K. C. *Current Protein & Peptide Science* **2008,** *9*, 248-259; (c) Concu, R.; Podda, G.; Ubeira, F. M.; Gonzalez-Diaz, H. *Current Pharmaceutical Design* **2010,** *16*, 2710-2723.
3. (a) L. Avila, C. L. *Current Protein & Peptide Science* **2011,** *12*, 221-234; (b) Falklöf, O.; Collyer, C. A.; Reimers, J. R. *Theor. Chem. Acc.* **2012,** *131*, 1076-1091; (c) Schmid, N.; Eichenberger, A.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A.; Gunsteren, W. *European Biophysics Journal* **2011,** *40*, 843-856; (d) Zhu, X.; Lopes, P. E. M.; MacKerell, A. D. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012,** *2*, 167-185.
4. (a) Ryde, U.; Nilsson, K. *Journal of the American Chemical Society* **2003,** *125*, 14232-14233; (b) Ryde, U.; Nilsson, K. *Journal of Molecular Structure: THEOCHEM* **2003,** *632*, 259-275; (c) Ryde, U. *Dalton Trans.* **2007**, 607-625; (d) Genheden, S.; Ryde, U. *Journal of Computational Chemistry* **2011,** *32*, 187-195; (e) Hsiao, Y. W.; Sanchez-Garcia, E.; Doerr, M.; Thiel, W. *Journal of Physical Chemistry B* **2010,** *114*, 15413-15423; (f) Yu, N.; Hayik, S. A.; Wang, B.; Liao, N.; Reynolds, C. H.; Merz Jr, K. M. *Journal of Chemical Theory and Computation* **2006,** *2*, 1057-1069; (g) Li, X.; Hayik, S. A.; Merz, K. M. *Journal of Inorganic Biochemistry* **2010,** *104*, 512-522; (h) Altun, A.; Shaik, S.; Thiel, W. *Journal of Computational Chemistry* **2006,** *27*, 1324-1337; (i) Hsiao, Y. W.; Thiel, W. *Journal of Physical Chemistry B* **2011,** *115*, 2097-2106; (j) Sanchez-Garcia, E.; Doerr, M.; Thiel, W. *Journal of Computational Chemistry* **2010,** *31*, 1603-1612; (k) Schoneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. *Journal of the American Chemical Society* **2002,** *124*, 8142-8151; (l) Sun, Q.; Li, Z.; Lan, Z. G.; Pfisterer, C.; Doerr, M.; Fischer, S.; Smith, S. C.; Thiel, W. *Physical Chemistry Chemical Physics* **2012,** *14*, 11413-11424.

5. Canfield, P.; Dahlbom, M. G.; Reimers, J. R.; Hush, N. S. *J. Chem. Phys.* **2006,** *124*, 024301.
6. Goerigk, L.; Falklöf, O.; Collyer, C. A.; Reimers Jeffrey, R., First Steps Towards Quantum Refinement of Protein X-Ray Structures. In *Quantum Simulations of Materials and Biological Systems*, Zeng, J.; Zhang, R.-Q.; Treutlein, H. R., Eds. Springer: Dordrecht, 2012; pp 87-120.
7. Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. *Journal of Physical Chemistry B* **2012,** *116*, 12501-12509.
8. Kohn, W.; Sham, L. J. *Physical Review* **1965,** *140*, 1133-1138.
9. (a) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chemical Physics Letters* **1996,** *253*, 268-278; (b) Lee, T. S.; Lewis, J. P.; Yang, W. *Computational Materials Science* **1998,** *12*, 259-277; (c) Wada, M.; Sakurai, M. *Journal of Computational Chemistry* **2005,** *26*, 160-168; (d) Fedorov, D. G.; Alexeev, Y.; Kitaura, K. *Journal of Physical Chemistry Letters* **2010,** *2*, 282-288; (e) Fedorov, D. G.; Ishida, T.; Uebayasi, M.; Kitaura, K. *Journal of Physical Chemistry A* **2007,** *111*, 2722-2732; (f) Fedorov, D. G.; Kitaura, K. *Journal of Physical Chemistry A* **2007,** *111*, 6904-6914; (g) Nagata, T.; Brorsen, K.; Fedorov, D. G.; Kitaura, K.; Gordon, M. S. *Journal of Chemical Physics* **2011,** *134*, 124115; (h) Nagata, T.; Fedorov, D. G.; Sawada, T.; Kitaura, K.; Gordon, M. S. *Journal of Chemical Physics* **2011,** *134*, 034110; (i) Ohta, K.; Yoshioka, Y.; Morokuma, K.; Kitaura, K. *Chemical Physics Letters* **1983,** *101*, 12-17; (j) Gale, J. D., SIESTA: A linear-scaling method for density functional calculations. In *Computational Methods for Large Systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*, Reimers, J. R., Ed. Wiley: Hoboken NJ, 2011; pp 45-74; (k) Mayhall, N. J.; Raghavachari, K. *Journal of Chemical Theory and Computation* **2010,** *7*, 1336-1343; (l) He, X.; Merz, K. M. *Journal of Chemical Theory and Computation* **2010,** *6*, 405-411; (m) Kussmann, J.; Beer, M.; Ochsenfeld, C. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, n/a-n/a.
10. (a) Kristyán, S.; Pulay, P. *Chemical Physics Letters* **1994,** *229*, 175-180; (b) Hobza, P.; Šponer, J.; Reschel, T. *Journal of Computational Chemistry* **1995,** *16*, 1315-1325; (c) Šponer, J.; Leszczynski, J.; Hobza, P. *Journal of Computational Chemistry* **1996,** *17*, 841-850.
11. Kolar, M.; Kubar, T.; Hobza, P. *Journal of Physical Chemistry B* **2011,** *115*, 8038-8046.
12. (a) Antony, J.; Grimme, S. *Physical Chemistry Chemical Physics* **2006,** *8*, 5287-5293; (b) Antony, J.; Grimme, S. *Journal of Computational Chemistry* **2012,** *33*, 1730-1739; (c) Antony, J.; Grimme, S.; Liakos, D. G.; Neese, F. *Journal of Physical Chemistry A* **2011,** *115*, 11210-11220; (d) Grimme, S. *Journal of Computational Chemistry* **2006,** *27*, 1787-1799; (e) Grimme, S. *Journal of Computational Chemistry* **2004,** *25*, 1463-1473; (f) Grimme, S.; Antony, J.; Schwabe, T.; MÃ¼ck-Lichtenfeld, C. *Organic and Biomolecular Chemistry* **2007,** *5*, 741-758; (g) Schwabe, T.; Grimme, S. *Physical chemistry chemical physics : PCCP* **2007,** *9*, 3397-3406; (h) Waller, M. P.; Kruse, H.; Muck-Lichtenfeld, C.; Grimme, S. *Chemical Society Reviews* **2012,** *41*, 3119-3128.
13. Grimme, S. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2011,** *1*, 211-228.
14. (a) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *Journal of Chemical Physics* **2010,** *132*, 154104; (b) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of Computational Chemistry* **2011,** *32*, 1456-1465.
15. Goerigk, L.; Kruse, H.; Grimme, S. *Chemphyschem* **2011,** *12*, 3421-3433.
16. Goerigk, L.; Grimme, S. *Physical Chemistry Chemical Physics* **2011,** *13*, 6670-6688.

17. Grimme, S.; Schreiner, P. R. *Angewandte Chemie-International Edition* **2011,** *50*, 12639-12642.
18. Hujo, W.; Grimme, S. *J. Chem. Theory Comput.* **2012** *9*, 308-315.
19. (a) Kestner, N. R. *Journal of Chemical Physics* **1968,** *48*, 252; (b) Jansen, H. B.; Ros, P. *Chemical Physics Letters* **1969,** *3*, 140-143; (c) Liu, B.; McLean, A. D. *Journal of Chemical Physics* **1973,** *59*, 4557-4558; (d) van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. *Chemical Reviews* **1994,** *94*, 1873-1885.
20. Kruse, H.; Grimme, S. *Journal of Chemical Physics* **2012,** *136*, 154101.
21. Boys, S. F.; Bernardi, F. *Molecular Physics* **1970,** *19*, 553.
22. (a) Mayer, I.; Turi, L. *Journal of Molecular Structure-Theochem* **1991,** *73*, 43-65; (b) Cook, D. B.; Sordo, J. A.; Sordo, T. L. *International Journal of Quantum Chemistry* **1993,** *48*, 375-384; (c) Gutowski, M.; Chalasinski, G. *Journal of Chemical Physics* **1993,** *98*, 5540-5554; (d) Wieczorek, R.; Haskamp, L.; Dannenberg, J. J. *Journal of Physical Chemistry A* **2004,** *108*, 6713-6723.
23. (a) Mayer, I. *International Journal of Quantum Chemistry* **1983,** *23*, 341-363; (b) Asturiol, D.; Duran, M.; Salvador, P. *Journal of Chemical Physics* **2008,** *128*, 144108 ; (c) Balabin, R. M. *Journal of Chemical Physics* **2010,** *132*, 231101 ; (d) Jensen, F. *Journal of Chemical Theory and Computation* **2010,** *6*, 100-106; (e) Deng, J.; Gilbert, A. T. B.; Gill, P. M. W. *Journal of Chemical Physics* **2011,** *135*.
24. (a) Holroyd, L. F.; van Mourik, T. *Chemical Physics Letters* **2007,** *442*, 42-46; (b) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *Journal of Physical Chemistry A* **2006,** *110*, 8-12; (c) Valdes, H.; Klusak, V.; Pitonak, M.; Exner, O.; Stary, I.; Hobza, P.; Rulisek, L. *Journal of Computational Chemistry* **2008,** *29*, 861-870.
25. Kruse, H.; Goerigk, L.; Grimme, S. *Journal of Organic Chemistry* **2012,** *77*, 10824-10834.
26. Sure, R.; Grimme, S. *Journal of Computational Chemistry* **2013**, submitted.
27. (a) Moreno, J. R. A.; Moreno, M. D. Q.; Urena, F. P.; Gonzalez, J. J. L. *Tetrahedron-Asymmetry* **2012,** *23*, 1084-1092; (b) Bohórquez, H. J.; Cárdenas, C.; Matta, C. F.; Boyd, R. J.; Patarroyo, M. E., Methods in Biocomputational Chemistry: A Lesson from the Amino Acids , . In *Quantum Biochemistry*, Matta, C. F., Ed. Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany. , 2010; pp 403-421; (c) Wilke, J. J.; Lind, M. C.; Schaefer, H. F.; Csaszar, A. G.; Allen, W. D. *Journal of Chemical Theory and Computation* **2009,** *5*, 1511-1523.
28. (a) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. *Journal of Physical Chemistry A* **2007,** *111*, 1146-1154; (b) Gloaguen, E.; Valdes, H.; Pagliarulo, F.; Pollet, R.; Tardivel, B.; Hobza, P.; Piuzzi, F.; Mons, M. *Journal of Physical Chemistry A* **2010,** *114*, 2973-2982; (c) Reha, D.; Valdes, H.; Vondrasek, J.; Hobza, P.; Abu-Riziq, A.; Crews, B.; de Vries, M. S. *Chemistry-a European Journal* **2005,** *11*, 6803-6817; (d) Valdes, H.; Pluhackova, K.; Hobza, P. *Journal of Chemical Theory and Computation* **2009,** *5*, 2248-2256; (e) Valdes, H.; Reha, D.; Hobza, P. *Journal of Physical Chemistry B* **2006,** *110*, 6385-6396; (f) Valdes, H.; Spiwok, V.; Rezac, J.; Reha, D.; Abo-Riziq, A. G.; de Vries, M. S.; Hobza, P. *Chemistry-a European Journal* **2008,** *14*, 4886-4898; (g) Toroz, D.; Van Mourik, T. *Molecular Physics* **2006,** *104*, 559-570.
29. Goerigk, L.; Karton, A.; Martin, J. M. L.; Radom, L. *Physical Chemistry Chemical Physics* **2013**, published online, DOI: 10.1039/c1033cp00057e.
30. Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Physical Chemistry Chemical Physics* **2008,** *10*, 2747-2757.

31. Perez, C.; Muckle, M. T.; Zaleski, D. P.; Seifert, N. A.; Temelso, B.; Shields, G. C.; Kisiel, Z.; Pate, B. H. *Science* **2012,** *336*, 897-901.
32. Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *Journal of Chemical Physics* **1992,** *96*, 6796-6806.
33. (a) Halkier, A.; Koch, H.; Jorgensen, P.; Christiansen, O.; Beck Nielsen, I. M.; Helgaker, T. *Theoretical Chemistry Accounts* **1997,** *97*, 150; (b) Jure@v cka, P.; Cern@'y, J.; Hobza, P.; Salahub, D. *Journal of Computational Chemistry* **2007,** *28*, 555.
34. (a) Goerigk, L.; Grimme, S. *Journal of Chemical Theory and Computation* **2010,** *6*, 107-126; (b) Goerigk, L.; Grimme, S. *Journal of Chemical Theory and Computation* **2011,** *7*, 291-309.
35. Grimme, S. *Journal of Chemical Physics* **2003,** *118*, 9095-9102.
36. (a) Grimme, S. *Journal of Chemical Physics* **2006,** *124*, 034108; (b) Neese, F.; Schwabe, T.; Grimme, S. *Journal of Chemical Physics* **2007,** *126*, 124115.
37. (a) Gustus, E. L. *Journal of Organic Chemistry* **1967,** *32*, 3425-3430; (b) Takagi, T.; Okano, R.; Miyazawa, T. *Biochimica Et Biophysica Acta* **1973,** *310*, 11-19; (c) Hill, R. R.; Ghadimi, M. *Journal of the Society of Dyers and Colourists* **1996,** *112*, 148–152.
38. (a) *TURBOMOLE V6.4 2012, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com;* (b) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chemical Physics Letters* **1989,** *162*, 165-169.
39. Weigend, F.; Haser, M. *Theoretical Chemistry Accounts* **1997,** *97*, 331.
40. Becke, A. D. *Physical Review A* **1988,** *38*, 3098-3100.
41. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988,** *37*, 785-789.
42. Perdew, J. P.; Wang, Y. *Physical Review B* **1986,** *33*, 8800-8802.
43. Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996,** *77*, 3865-3868.
44. Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Physical Review Letters* **2003,** *91*, 146401.
45. (a) Becke, A. D. *J. Chem. Phys.* **1993,** *98*, 5648-5652; (b) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994,** *98*, 11623-11627.
46. (a) Ernzerhof, M.; Scuseria, G. E. *Journal of Chemical Physics* **1999,** *110*, 5029-5036; (b) Adamo, C.; Barone, V. *Journal of Chemical Physics* **1999,** *110*, 6158-6170.
47. Zhao, Y.; Truhlar, D. G. *Journal of Physical Chemistry A* **2005,** *109*, 5656-5667.
48. Becke, A. D. *Journal of Chemical Physics* **1993,** *98*, 1372-1377.
49. Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972,** *56*, 2257.
50. Horn, H.; Ahlrichs, R. *Journal of Chemical Physics* **1992,** *97*, 2571.
51. Tatewaki, H.; Huzinaga, S. *Journal of Computational Chemistry* **1980,** *1*, 205-228.
52. EMSL basis-set exchange, https://bse.pnl.gov/bse/portal.
53. Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chemical Physics Letters* **1995,** *240*, 283-289.
54. Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theoretical Chemistry Accounts* **1997,** *97*, 119.
55. Website of the Grimme group, http://www.thch.uni-bonn.de/tc/index.php?section=downloads&lang=english.
56. Jensen, F., *Introduction to Computational Chemistry, 2nd Edition*. John Wiley @& Sons: 2007.
57. Petersen, M. T. N.; Jonson, P. H.; Petersen, S. B. *Protein Engineering* **1999,** *12*, 535-548.

58. Klamt, A.; Schuurmann, G. *Journal of the Chemical Society-Perkin Transactions 2* **1993**, 799-805.