Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding

Jan Řezáč*

Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, 166 10 Prague, Czech Republic

E-mail: rezac@uochb.cas.cz

December 13, 2019

Abstract

The Non-Covalent Interactions Atlas project (www.nciatlas.org) aims to cover a wide range of non-covalent interactions with a new generation of benchmark data sets. This paper presents the first two data sets focused on hydrogen bonding: HB375, featuring neutral systems, and IHB100 for ionic H-bonds. Both data sets are complemented by ten-point dissociation curves (HB375x10, IHB100x10). The interaction energies are extrapolated to the CCSD(T)/CBS limit from calculations in large basis sets. The paper also summarizes the design principles that will be used to construct the subsequent data sets in the series. The testing of DFT-D methods on the HB375 set has revealed interesting, previously unnoticed issues. The application of the new data to the testing and parameterization of semiempirical QM methods is also discussed.

*To whom correspondence should be addressed

1 Introduction

The validation of the accuracy of approximate computational methods and the development of those that rely on empirical parameters depend on external reference data. In the field of non-covalent interactions, accurate experimental data for isolated non-covalent complexes are scarce, and the comparison between computed interaction energy and experimentally observed variables is not straightforward. It has thus become common to rely on accurate calculations as the benchmark. Here, the coupled-clusters singles, doubles and perturbative triples method, extrapolated to complete basis set limit (CCSD(T)/CBS) has become the "gold standard", as it offers very high accuracy while still being applicable to systems as large as several tens of atoms. Since these calculations are very expensive, the common practice is to reuse previously published results. This has been greatly facilitated by the development of benchmark data sets built specifically for this purpose. The use of predefined data sets has also brought the advantage of comparing the results calculated on them across different publications. A more detailed introduction into this area can be found in our recent reviews.[1,2]

At present, the most widely used data sets, such as S66[3,4] and X40,[5] offer very accurate interaction energies for dozens of systems. To improve the coverage of the potential energy surface, they have been accompanied by sets of non-equilibrium geometries, which has resulted in hundreds of reference points. However, the coverage of the chemical space still remains limited. More of these data sets are usually used in a single application, and they have become part of larger collections of benchmark data such as GMTKN databases.[6] Nevertheless, not even the use of a collection of data sets guarantees broad and balanced coverage of chemical space. Other large data sets created from scratch, such as the BFDb set, covering amino-acid interactions in proteins,[7] may cover a specific problem in depth while not being suitable for general-purpose applications.

It is clear that we lack data sets that would be larger, more general, and as accurate as possible. Such data sets would allow for a more robust statistical analysis of the performance

of the existing methods and facilitate the development of new general-purpose approximate and empirical methods. The most important issue to address in the future data sets is their diversity, and it is closely related to their size. The S66 set covers hydrogen bonds and London dispersion using the simplest possible examples. It is sufficient for the validation of robust *ab initio* methods, where no outliers can be expected, or for fitting few parameters in simple corrections. However, it is not diverse enough to capture specific failures of more approximate methods or for the development of more empirical approaches with a larger number of parameters.

Building a large universal database covering everything is not only impossible because of the computational demands, but also impractical for the applications. A more rational approach is to construct individual data sets covering specific areas of the chemical space, where each of them will satisfy the requirements outlined above. In the field of non-covalent interactions, this means covering the different types of interactions in different, well-defined parts of the chemical space. Each of these sets should be comprehensive and diverse, which requires including hundreds of systems instead of the dozens found in the existing data sets. With the increase of the computational power available, it is still a realistic goal to calculate such data sets at a true gold-standard level.

Such a project aimed to map non-covalent interactions across the chemical space, the Non-Covalent Interactions Atlas (NCIA, www.nciatlas.org), is being started. This paper outlines the main principles of constructing the next-generation data sets and features the first two of them covering hydrogen bonding in organic molecules. Hydrogen bonds have been selected for a very practical reason – larger and diverse data sets of H-bonds are now very important for further development of semiempirical quantum-mechanical (SQM) methods. Two data sets, HB375 and IHB100, cover hydrogen bonding in neutral and ionic systems, respectively. Featuring 375 neutral and 100 ionic complexes, they are large enough for even the most demanding applications. The equilibrium geometries are accompanied by dissociation curves (ten points per system in total) with results calculated or rescaled to

the same high benchmark level (the data sets HB375x10 and IHB100x10). This paper thus introduces 4750 data points with "gold-standard" interaction energies.

The first set comprises 375 neutral complexes, covering molecules consisting of hydrogen, carbon, nitrogen and oxygen. It includes 223 classical H-bonds between oxygen and nitrogen, 39 CH–X bonds, and 113 complexes of the same molecules not forming a hydrogen bond and their complexes with nonpolar molecules to be used as a control group. For comparison, the widely used S66 data set contains only 23 H-bonds, and the largest set of diverse H-bonds we have used so far contains 104 systems (and it was taken from a work that did not focus on true benchmark accuracy).[8,9]

The second set of 100 systems covers ionic H-bonds (where one of the molecules is charged) in the same chemical space. Here, the selection is limited to complexes stable in the gas phase with well-defined H-bonds. Not all monomer combinations satisfy this requirement, what is also a reminder that these systems are only models needed in method development and their relevance to real-world applications is limited. Here, the previously available data set covering this topic that we had built for the parameterization of the corrections for SQM methods contained only 15 systems.

The aim of the NCIA project is to provide data sets of the highest quality not only in terms of size and accuracy, but also concerning other criteria. I have used my long-term experience with constructing, distributing and using benchmark data sets to design new data sets to be as useful and as easy to use as possible. In comparison to S66 and S66x8, not only are the new data sets larger and more diverse, but they also address some weaknesses of older data sets and bring new improvements. 1) The geometries have been meticulously optimized and verified to be true minima, so that they can also serve as a benchmark for more approximate methods. 2) The dissociation curves have been built using a more straightforward, reproducible protocol (e.g. the HB375 set is only a subset of HB375x10, rather than having different geometries). 3) The dissociation curves have been calculated at the same level as the equilibrium geometries. 4) The data sets are well organized and

annotated with metadata, which simplifies their use and improves reproducibility of results obtained using these sets. 5) Because of the size of the new data sets, we have provided predefined statistically relevant subsets of different sizes that would simplify comparison of the results in cases where it would be impractical to use the complete data set. 6) The data sets are published also in a transparent machine-readable format, providing not only the composite benchmark interaction energies, but also the components used to construct them.

As this is the first paper on what will be a series of compatible data sets constructed using similar principles, special attention will be paid to the description of the construction of the data sets and the methodology of the benchmark calculations. The applicability of the new data sets is then demonstrated on the analysis of the performance of selected wavefunction, density functional theory (DFT) and semiempirical QM methods. Although it may seem as a mere repetition of what has already been done on other data sets, it is shown that the increased size and diversity of the data set make it possible to obtain interesting new information.

The data sets presented here are only the first steps in a series that will bring broader and deeper coverage of different types of non-covalent interactions. At the moment, we are working on extending the data set of hydrogen bonds to other elements, namely sulfur, phosphorus and halogens, and building data sets of dispersion-dominated complexes and repulsive contacts spanning the same chemical space.

2 Data Set Construction

2.1 Composition of the HB375 data set

All the complexes were built automatically from monomers, in which possible H-bond donor and acceptor sites were labeled. In the HB375 set, the hydrogen donors used were water, methanol, phenol and acetic acid (OH group), ammonia, methylamine, dimethylamine, aniline, 1H-pyrrole and N-methylacetamide (NH group), and methane, ethene, ethyne and

benzene (CH group). The acceptors used included acetaldehyde, acetamide, acetic acid, acetone, dimethylcarbonate, dimethylether, dimethylperoxide, furan, methanol, methylacetate, methylcyanate, nitromethane, nitrosomethane, N-methylacetamide, phenol and water (oxygen), and acetonitrile, ammonia, aniline, dimethylamine, dimethyldiazene, methylamine, methylazide, methylcyanate, methylisocyanide, nitrosomethane, N-methyl-2-propanimine, pyridine and trimethylamine (nitrogen). This yielded 378 possible combinations of molecules. At the beginning, more geometries were considered as some of the molecules feature multiple non-identical sites that can participate in a H-bond. At the end, the binding motif with the lowest energy was selected. In few systems, the donor and acceptor roles changed and the dimer became identical to another one present in the data set; the resulting set thus comprises 375 unique structures (a complete list is provided in the Supporting Information, Table S1). For the lowest energy minimum to be found for each system, the starting structure was refined with simulated annealing and optimized. Each minimum was verified by a calculation of vibrational frequencies, and if it was found to be a transition state, the optimization was repeated from a modified geometry (the details of the protocol are provided below).

The resulting structures were then automatically sorted into the following groups: OH-O, NH-O, OH-N, NH-N, CH-O and CH-N hydrogen bonds, and systems with no H-bond (the group labeled as noHB). Detailed information on the groups is provided in Table 1. If multiple H-bonds were present, the shortest one (relative to the sum of the van der Waals radii of the atoms) was used to categorize the system. For a system to be considered hydrogen-bonded, the distance between the hydrogen and its acceptor had to be less than 90% of the sum of their van der Waals radii, and the XH-Y angle had to be larger than 120°(these can be considered rather weak criteria). This procedure yielded a perfect assignment of the well-defined H-bonds, and there are only few questionable systems in the noHB group where some interactions can be considered as weak H-bonds, but with geometric features outside these arbitrary limits. When combined, the CH-X and noHB groups correspond to the "others" group of S66, which includes weaker interactions of polar molecules where electrostatic and

dispersion interactions are mixed. The figures of all the complexes are available at the project website[10] for closer inspection of the geometries.

To simplify the work with the data set, the systems were assigned unique identification numbers. These numbers consist of a digit identifying the group (see Table 1), a dot, and a three-digit number identifying the system within the group (the first system thus has the number 1.001). If the data set has been extended, this scheme makes it possible to add systems into the groups without renumbering the existing part of the data set. These numbers are the primary identifiers used in the data and geometry files provided here or at the NCIA website.

Table 1: Composition of the HB375 data set: Groups by interaction type, their size, and average interaction energy (kcal/mol) in each group.

#	Group	Size	$\langle \Delta E^{int} \rangle$
1	OH-O	60	-8.1
2	NH-O	65	-5.6
3	OH-N	45	-9.0
4	NH-N	53	-5.8
5	CH-N	20	-4.4
6	CH-O	19	-4.5
7	noHB	113	-3.3

2.2 Composition of the IHB100 data set

The data set of ionic H-bonds was constructed analogously, by combining a set of protonated H-bond donors with neutral acceptors, and of neutral donors with anionic acceptors. The hydrogen donors used were acetic acid, methanol, phenol, water (OH), oxonium (OH^+), 1H-pyrrole, ammonia, aniline, dimethylamine, methylamine, N-methylacetamide (NH), ammonium, trimethylammonium, guanidinium, imidazolium (NH^+), benzene, ethene, ethyne and methane (CH). The hydrogen bond acceptors were acetaldehyde, acetamide, acetic acid, acetone, dimethylcarbonate, dimethylether, dimethylperoxide, furan, methanol, methylacetate, methylcyanate, N-methylacetamide, nitromethane, nitrosomethane, phenol, water (O), ac-

etate, benzoate, carbonate, hydroxide, nitrate, nitrite, oxalate (O^-), acetonitrile, ammonia, aniline, cyanide, dimethylamine, dimethyldiazene, methylamine, methylazide, methylcyanate, N-methyl-2-propanimine, nitrosomethane, pyridine, trimethylamine (N), methylisocyanide (C) and cyanide (C^-). This results in 321 combinations. After geometry optimization, all the systems where the original hydrogen bond was not conserved were removed (these included systems where proton transfer occurred or where the proton was shared by the two molecules and there was no distinct XH covalent bond and H-Y hydrogen bond). From the resulting 199 systems, five more were removed because of convergence problems in the benchmark calculations. The resulting 194 systems were grouped by the elements involved in the hydrogen bond and their charge. Due to large differences in the population of these groups, the largest groups were truncated to fifteen randomly chosen systems, which resulted in the final data set of 100 complexes (listed in the Supporting Information, Table S2). In contrast to the HB375 set, the simulated annealing step was skipped in the preparation of the geometries, as it might result in proton transfer even in systems with an energy barrier. Finally, all the geometries were verified to be minima by the calculation of vibrational frequencies.

The set is divided into groups by the donor and acceptor element and its charge using the same algorithm as that used for the neutral systems. More groups are formed here. Even after the truncation of the most populated groups, the set is not completely balanced because only some combinations form stable H-bonds in the gas phase (e.g. the molecules considered do not form any stable OH^+–N H-bond). The final 13 groups are listed in Table 2. A numbering scheme analogous to HB375 is employed to identify the systems in the IHB100 data set.

2.3 Dissociation curves

Dissociation-curve scans have been performed for all the systems using the automated, reproducible protocol described here. The geometries along the curve are determined by the

Table 2: Composition of the IHB100 data set: Groups by interaction type, their size, and average interaction energy (kcal/mol) in each group. The number of systems in the groups before the reduction of the data set size is provided in parentheses.

#	Group	Size	$\langle \Delta E^{int} \rangle$
01	OH^+-O	1	-22.4
02	NH^+-O	15(59)	-23.0
03	NH^+-N	15(35)	-22.7
04	NH^+-C	4	-23.9
05	OH-O^-	15(19)	-26.2
06	OH-N^-	3	-25.5
07	OH-C^-	2	-17.8
08	NH-O^-	15(33)	-19.4
09	NH-N^-	6	-16.1
10	NH-C^-	5	-14.5
11	CH-O^-	15(23)	-9.3
12	CH-N^-	3	-8.3
13	CH-C^-	1	-2.6

vector $\vec{v}$, defining the direction, and the displacement constructed by scaling the reference distance r_{ref} by the scaling factor f. The scaling factors are the same for all systems, chosen to cover the whole curve but also to allow the precise identification of the minimum. They are the same as those used in the X40x10 and S66x10 data sets: f = 0.8, 0.85, 0.90, 0.95, 1.0, 1.05, 1.1, 1.25, 1.5 and 2.0. Unlike in S66, where the S66 set used different geometries interpolated from the S66x8 curves, here the 1.0 point is identical to the geometry from the equilibrium distance set. To construct the displaced geometry of a complex AB, the atoms in the molecule B are translated by the vector $\vec{t}$:

$$\vec{t} = \frac{\vec{v}}{|\vec{v}|} * (f - 1) * r_{ref}. \quad (1)$$

For complexes with a single dominant H-bond, the displacement vector is the axis of the H-bond – it is the vector between the hydrogen and the electron-donor atom, and the reference distance is the distance between these two atoms. In complexes without a H-bond (the noHB group), the vector is constructed from the centers of mass of the two molecules, and the reference distance is the length of the closest contact between the molecules. Finally,

in systems with multiple H-bonds of similar length (up to 102% of the length of the shortest one), the vector is calculated as the direction from the averaged position of the atoms involved in these H-bonds on one molecule and on the other. The reference distance is the length of the shortest H-bond in the system. This ensures that in symmetric systems such as the acetic acid dimer, the dissociated geometries conserve this symmetry.

3 Methods

3.1 Preparation and geometry optimization of the complexes

The protocol started with optimized geometries of the monomer molecules. In these, the hydrogen-bond donor site (on the hydrogen atom, in the extension of the X-H bond) and hydrogen-bond acceptor site (on an electronegative atom, in the most sterically accessible direction) were labeled. The dimers were then constructed automatically by pairing the donor and acceptor sites in a linear arrangement, putting the molecules at a distance equal to the sum of the van der Waals radii of the atoms.

The neutral systems were then subject to simulated annealing in order to identify the most stable minimum. This step was skipped for the ionic hydrogen bond in order to prevent an unwanted proton transfer. After an initial optimization, a molecular dynamics simulation was run at the DFTB3-D3H5 level.[11] Over a 5 ps simulation, the system was cooled from the initial temperature of 20K to zero using a Berendsen thermostat. The starting temperature was chosen so that the complexes would not dissociate, but the intermolecular degrees of freedom were sampled sufficiently.

The final geometries were obtained using a DFT-D3 optimization. Here, a hybrid DFT functional was needed to reduce the delocalization error in the ionic systems – the same setup was also used for the neutral ones for the sake of consistency. The B3LYP functional with D3 correction[12] was selected as a widely available method that yielded accurate geometries of non-covalent complexes.[13] The calculations were performed in the def2-QZVP basis set[14]

(for which the D3 correction had been parameterized). In the D3 dispersion correction, the Becke-Johnson damping was used.[15] After an initial optimization, the geometry was symmetrized and optimized again. Very fine DFT grid (100 radial / 590 spherical points) and tight convergence criteria were used to obtain high-quality geometry. In both data sets, the RMS of the residual gradient in each system was lower than 0.01 kcal/mol/Å.

The optimized geometries were then verified by vibrational analysis; if the geometry was found to be a transition state, it was modified, reoptimized and tested again. In most cases, this issue was solved by applying a small displacement along the vibrational mode with imaginary frequency, or by another short annealing at the DFT-D3 level. Only in last few systems, the geometry had to be modified manually.

The dissociation curves calculated at the benchmark level allow the verification of the DFT-D3 geometries in terms of intermolecular distance. In the HB375 data set, the DFT geometry is always the lowest point calculated on the curve; therefore, the accuracy of the intermolecular distance cannot be worse than ±2.5%. In the IHB100 data set, there are six systems where this error is larger, about 5%; all of them are CH-anion H-bonds. This indicates that the geometries are reliable for neutral and the vast majority of ionic systems, and the error in the few remaining ones is acceptably small.

The DFT-D3 optimizations and vibrational analysis calculations were performed in Psi4[16] with density fitting applied. This code had been selected because it provides a fine control of the geometry optimization of non-covalent complexes. DFTB3-D3H5 calculations were carried out in DFTB+,[17] using Cuby as a driver for both optimization and molecular dynamics.[18]

3.2 Benchmark calculations

The well-known composite CCSD(T)/CBS scheme[19–21] has been used to obtain the benchmark interaction energies. The Hartree-Fock (HF) energy in the largest basis set is used without extrapolation, MP2 correlation energy is extrapolated to the CBS limit using the

Helgaker formula,[22] and the δCCSD(T) correction calculated in a smaller basis set is added:

$$E^{CCSD(T)/CBS} = E^{HF} + E^{MP2/CBS} + \delta E^{CCSD(T)}, \tag{2}$$

$$\delta E^{CCSD(T)} = E^{CCSD(T)} - E^{MP2}. \tag{3}$$

The approach that has been used here is conservative; it is based on canonical MP2 and CCSD(T) calculations, employing as large basis sets as possible. The overall accuracy is determined by the basis set used in the CCSD(T) calculation, and this cannot be avoided even with the available explicitly correlated methods where the (T) term is not explicitly correlated and has to be scaled empirically to maintain balance with the CCSD contribution. The canonical approach avoids any empiricism, possibly at the cost of negligibly lower accuracy, and is computationally more efficient.

The benchmark calculations, and all the wavefunction methods tested, have employed the counterpoise (CP) correction for the basis set superposition error.[23] It has been shown that in larger basis sets, such as those used here, the CP-corrected results are more accurate than those obtained without CP correction or by averaging CP-corrected and uncorrected interaction energies.[24] The frozen-core approximation is applied to all the correlation energy calculations; its effect in systems containing only light atoms should be negligible.

The correlation-consistent basis sets of Dunning with diffuse functions[25] are used for all the wavefunction calculations presented here; the text utilizes shorthand notation, where aXZ stands for aug-cc-pVXZ. For the CCSD(T) calculation in the composite scheme to be considered a true "gold standard", it must be carried out in at least triple-zeta quality basis set.[26–28] This paper uses the heavy-aug-cc-pVTZ basis (abbreviated here as haTZ); its performance is practically the same as that of the fully augmented aTZ basis but it saves some resources. This setup was shown to yield error of about 1% with respect to a more accurate CCSD(T)/CBS estimate in the A24 data set.[27] This basis set is also close to the

limit at which the calculations of the largest systems (up to 30 atoms) can be efficiently run at the available supercomputing infrastructure. Although the smaller systems can be calculated in a larger basis set, the same scheme is used for the whole data set, as the consistency of the results is a more important objective than a non-systematic (and small) improvement of accuracy.

For the MP2/CBS term, the extrapolation from aQZ and a5Z basis sets is still affordable (and one order of magnitude faster than CCSD(T)/haTZ); therefore, it was used to construct the benchmark, although using extrapolation from aTZ and aQZ would also be acceptable. For details on the timing of the calculations and the differences from the results obtained with smaller basis sets, see Table 3. The setup described so far is denoted the "gold level" in the following text. Some calculations on the non-equilibrium geometries were performed in smaller basis sets, with MP2 extrapolated from aTZ and aQZ bases, and δCCSD(T) calculated in aDZ. This setup is referred to as the "silver level", consistently with prior literature.[28]

All the MP2 and CCSD(T) calculations were carried out using the Psi4 program.[16] Density fitting was applied in both SCF and correlation parts of the calculation using auxiliary basis sets matching the orbital basis. The SCF convergence threshold was tightened to 10^{-8} a.u.

Table 3: Timing of the MP2/CBS and CCSD(T)/CBS composite calculations in the HB375 data set. For lower-level schemes, the mean unsigned error (MUE) to the higher level is reported.

Method	CBS scheme	Avg. timing core hours/system	MUE to higher level kcal/mol
MP2	aTZ→aQZ	7	0.003
MP2	aQZ→a5Z	45	
MP2 + δCCSD(T)	aTZ→aQZ + aDZ	32	0.053
MP2 + δCCSD(T)	aQZ→a5Z + haTZ	295	

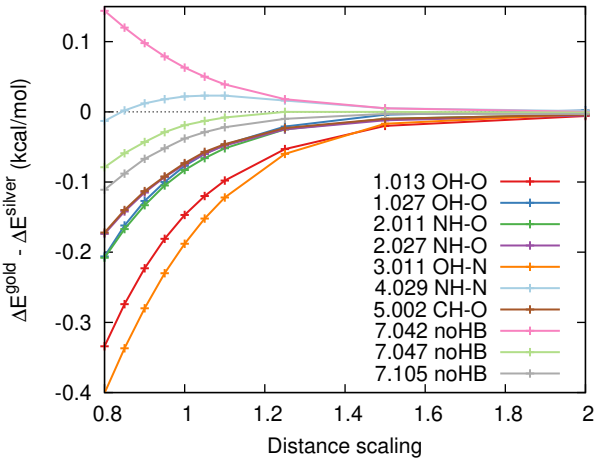
3.3 Dissociation curve scaling

The dissociation curves are smooth, well-behaved functions; their representation by ten points necessarily contains some redundant information. It would thus be a waste of computational resources calculating all the points at the gold level. The following protocol was used to avoid unnecessary calculations while conserving the gold-standard accuracy: To provide a solid baseline, all the points of the dissociation curves were calculated at the silver level. Subsequently, the minimum and the endpoints of the curve ($f = 0.8$, 1.0 and 2.0) were also calculated at the gold level. Complete curves at the gold level were calculated for ten randomly selected systems (1.013, 1.027, 2.011, 2.027, 3.011, 4.029, 5.002, 7.042, 7.047 and 7.105), and the difference between the lower and higher levels was analyzed. It is plotted in Figure 1. The average of the differences can be fitted with a power function with an exponent of approximately -4.8, but it is obvious that a more flexible function is needed to cover all the cases. Having three anchor points, a system of equations with three parameters can be solved algebraically. The difference is thus represented as:

$$\Delta_{gold-silver}\Delta E^{int}(f) = \frac{a}{f^3} + \frac{b}{f^5} + \frac{c}{f^6}, \tag{4}$$

where f is the scaling factor used to construct the dissociation curve and a, b and c are the coefficients obtained by solving the system of equations for three values of f where the result is known. The exponents were selected to provide the best fit in the ten model systems. It was found that more reliable results can be obtained when the value of $\Delta_{gold-silver}\Delta E^{int}$ in the most distant point ($f = 2.0$) is set to zero, rather than using the actual value from the calculations (which is nearly zero). When the gold-level benchmark is reconstructed using this correction (as $\Delta E^{int}_{gold} = \Delta E^{int}_{silver} + \Delta_{gold-silver}\Delta E^{int}$, the error (RMSE) to true gold-level calculations across the dissociation curves in the ten selected systems is 0.0025 kcal/mol. This is negligible in comparison with the other errors present even in the gold-level setup; it is thus safe to assume that the dissociation curves corrected using this approach are still of

Figure 1: Difference between the dissociation curves calculated at the gold and silver levels in ten systems randomly selected from the HB375x10 data set.



gold-level quality. Solving a system of equations rather than fitting the curve ensures that in the anchor points (in the minimum and at the closest distance, the most distant point was treated differently), the exact value of the gold-level calculation is conserved.

3.4 Methods tested on the new data sets

The first class of methods tested in this paper comprise correlated wavefunction methods that can be derived from the benchmark calculations: MP2 and its spin-component-scaled variants SCS-MP2[29] and SCS-MI-MP2,[30] and CCSD with analogous SCS-CCSD[31] and SCS-MI-CCSD[32] methods. These results were extrapolated using the same scheme and basis sets as the respective benchmark. Dispersion-corrected MP2[33] (MP2D) interaction energies were constructed from MP2/CBS results.

The second test included several DFT methods with multiple variants of D3 dispersion correction[12,15,34,35] and D4 dispersion correction.[36,37] The DFT interaction energies based on BLYP, B3LYP, BHLYP, BP, PBE and PBE0 functionals were calculated using Psi4; the D3 correction was added using Cuby.[18] The D4 correction was calculated using a standalone program provided by the authors of the method.[38] The double-hybrid DFT calculations with DSD-BLYP and DSD-BLYP-D3 functionals[39] were carried out in Orca 4.2.[40] All the DFT

calculations were performed in the def2-QZVP basis set,[14] as this is the basis for which the D3 dispersion correction had been parameterized.

The third series of tested methods consisted of several semiempirical QM methods, and multiple versions of corrections for non-covalent interactions applicable to these methods. From the classical SQM methods based on the NDDO approximation,[41] the author had selected the PM6 method[42] and its combination with the D3H4 correction for dispersion and H-bonding,[43] and PM7, which already included similar corrections.[44] These calculations were carried out in MOPAC 2016.[45] The remaining methods were multiple variants of the third-order self-consistent-charge density-functional tight binding (DFTB3),[46,47] and the empirical tight binding GFN2-xTB methods.[48] All of the DFTB3 calculations use the 3OB parameter set.[49] The default DFTB3 method employs the XH damping (with exponent 4.0) as a correction for hydrogen bonding.[49] DFTB3-D3 adds the D3 dispersion correction.[50] DFTB3-D3H4 employs an empirical H-bonding correction and a different parameterization of the D3 dispersion correction.[9,43] Finally, DFTB3-D3H5 uses a novel H-bonding correction embedded in the DFTB calculation.[11] All of these calculations were performed using the DFTB+ program.[17] Its latest version implements all the corrections except for D3H4, which was added using the Cuby framework.[18] The GFN2-xTB calculations were carried out in a standalone program provided by the authors of the method.[51]

3.5 Clustering analysis

In some applications, such as in the benchmarking of less empirical methods, it is not necessary to use data sets as large as these presented here. It is thus desirable to define a representative subset of systems that covers the same chemical space, but more sparsely. A random selection of the subset is, of course, a valid approach, but more advanced statistical methods can yield a subset containing more information. A very useful way how to achieve this is clustering which groups the systems according to their similarity. The most diverse subset in terms of the measure of similarity used is then formed by taking one representa-

tive from each cluster. Similar approaches had already been applied to the extraction of representative data points from larger collections of benchmark data.[52,53]

The clustering analysis relies on measuring the similarity between the systems, which can be defined in different ways. Since we are interested in the benchmarking of computational methods, an interesting choice is to characterize each of the systems by the errors of a set of diverse computational methods applied to it. A subset obtained using this approach will contain systems in which these methods fail in the most diverse ways, maximizing the chance that the subset can be used to identify these errors. Technically, each system is assigned a vector assembled from the errors of 45 computational methods used in the paper (all the wavefunction, DFT, and SQM methods introduced earlier, and also some byproducts of the benchmark calculations, such as HF and MP2 calculations in smaller basis sets). The similarity between two systems is then evaluated as the correlation between these two vectors, formally expressed by the Pearson correlation coefficient. Using the correlation coefficient instead of simple Euclidean distance between the vectors puts focus on the nature of the errors, rather than on their magnitude (which is often proportional to the interaction energy in the system).

This similarity measure is then used in the complete-linkage clustering algorithm,[54] which yields a specified number of clusters. This clustering algorithm is well suited for this application as it distributes the systems into compact clusters of similar size. A single system representing each cluster is selected as the one that is the most similar to all the other members of the cluster.

In addition to generating the subsets, the analysis of the results of the clustering provides important insights into the composition of the data set and the redundancy of the information it contains. This is discussed in detail in Section 5.6.

3.6 Processing of the results

The results of the methods tested on the new data sets are analyzed using the statistical measures commonly used in this field. The root-mean-square error (RMSE) is used as the primary (and often the only) error measure. The systematic part of the error is separated by using mean signed error (MSE). In the HB375 data set, the errors are usually reported also for the individual groups, while the smaller IHB100 is not divided further in this work.

The processing of the results has been automated using the Cuby framework.[18,55] The data sets are provided in a format used by Cuby and will be bundled with the framework in the future. To perform a calculation on the whole data set or a user-defined selection thereof, only a single input file has to be written. The data set definition file contains all the metadata necessary for selecting the predefined groups, etc. Cuby prepares and runs all the calculations, and reports the statistical analysis of the results. In the data sets of dissociation curves, the errors are also evaluated as a function of the distance, and the curves can be plotted automatically.

4 Data Availability

The geometries of all the complexes as well as the tables of the benchmark interaction energies and associated metadata are provided in the Supporting Information. The metadata describe the classification of the system and its assignment to the predefined groups used in the analysis of the data set. More detailed information, including visualizations of all the geometries, can be browsed at the NCIA website.[10] Since the core HB375 and IHB100 data sets are just subsets of the HB375x10 and IHB100x10 sets, the data are provided only for the complete sets of dissociation curves, from which the sets of equilibrium geometries can be easily extracted.

Machine-readable (but human-friendly) data files containing the data sets have been prepared in the format used by the Cuby framework developed by the author. These are

YAML files[56] containing a description of the data set, the list of systems with the benchmark results, metadata and reference to the geometry files, and additional information used e.g. for the analysis and visualization of the dissociation curves. These files contain also all the interaction energies and their components used to construct the composite benchmark results, and some results of the methods tested on the data sets. These files are included in the SI for reference, but their up-to-date version will be included in the Cuby package once this paper is published. As the structure of the files is simple and YAML parsers are available in all common programming languages, it would be straightforward to use these files outside the Cuby framework as well.

The author of this paper is also preparing an archive of Psi4 output files from all the calculations used to construct the benchmark. These will be made available at the NCIA website or another repository in the future.

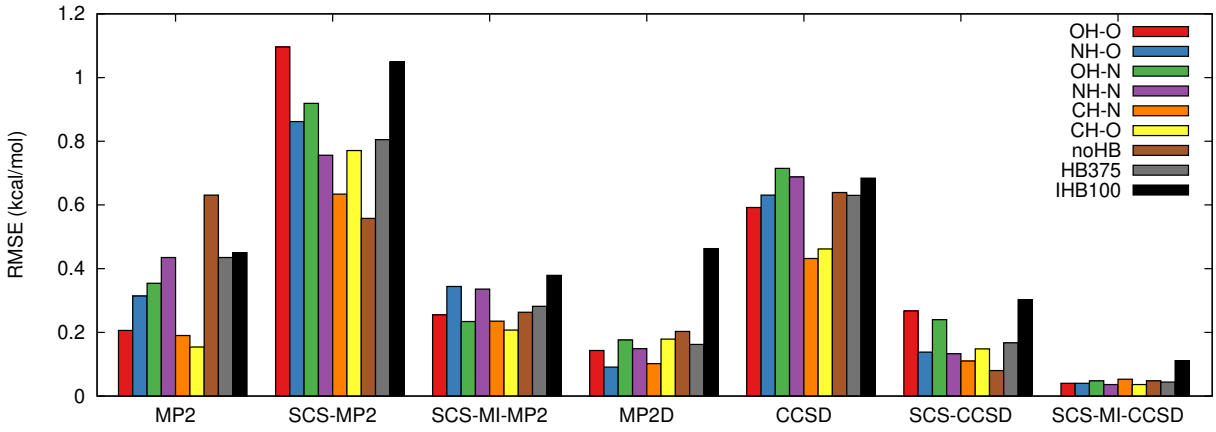
5 Results and Discussion

5.1 Correlated wavefunction methods

The text below briefly analyzes the performance of several correlated wavefunction methods applicable to non-covalent interactions. Here, one cannot expect any new discoveries, but these results form a baseline to which the more approximate methods described below should be compared. These results are a byproduct of the benchmark calculations; they were extrapolated to the CBS limit using the same composite scheme and basis sets as the benchmark, namely MP2 extrapolation from aQZ and a5Z basis sets, and the CCSD correction calculated in the haTZ basis in the methods beyond MP2. In the HB375 data set, the groups are analyzed separately, and the IHB100 set is not divided further. The results are summarized in the form of a plot in Figure 2, and the data used to construct the plot are available in the Supporting Information Table S5.

At the MP2 level, the author first tested the two spin-component-scaled approaches SCS-

Figure 2: Errors of correlated calculations in the HB375 data set by group, in the whole set, and in the IHB100 data set. All results have been extrapolated to the CBS limit using the same scheme and basis sets as those used in the benchmark calculations.



MP2 and SCS-MI-MP2, where the latter is intended for non-covalent interactions (MI stands for molecular interactions). Since MP2 is known to perform well for hydrogen bonds, none of these methods provides any significant advantage, although SCS-MI-MP2 slightly lowers the overall error and provides better balance between the groups. A more interesting approach is MP2D, MP2 with a dispersion correction that replaces the overestimated uncoupled HF dispersion with the more appropriate coupled Kohn-Sham dispersion used in DFT-D3. This correction is empirical, based on precalculated C_6 coefficients, and adds practically no cost to the MP2 calculation. In the neutral H-bonds, the results are only slightly better than the best DFT-D3 methods, but the errors are distributed more evenly. In the ionic H-bonds, dispersion does not play such an important role, and the error is almost the same as in uncorrected MP2. This is not a trivial finding because all DFT approaches fail in the ionic systems (with about four times larger errors than MP2D; the details are discussed in the following section), and MP2D might be the cheapest method to achieve this level of accuracy for all non-covalent interactions including ionic systems.

CCSD yields worse results than MP2, as it underestimates the dispersion correction and there is no error compensation as in the case of MP2. The excellent accuracy of SCS-MI-MP2 is confirmed again; the RMSE is 0.05 kcal/mol in the neutral data set and 0.11 kcal/mol

in the ionic systems (this difference can be attributed to the different magnitude of the interaction energies in these data sets, rather than to the different behavior of the method).

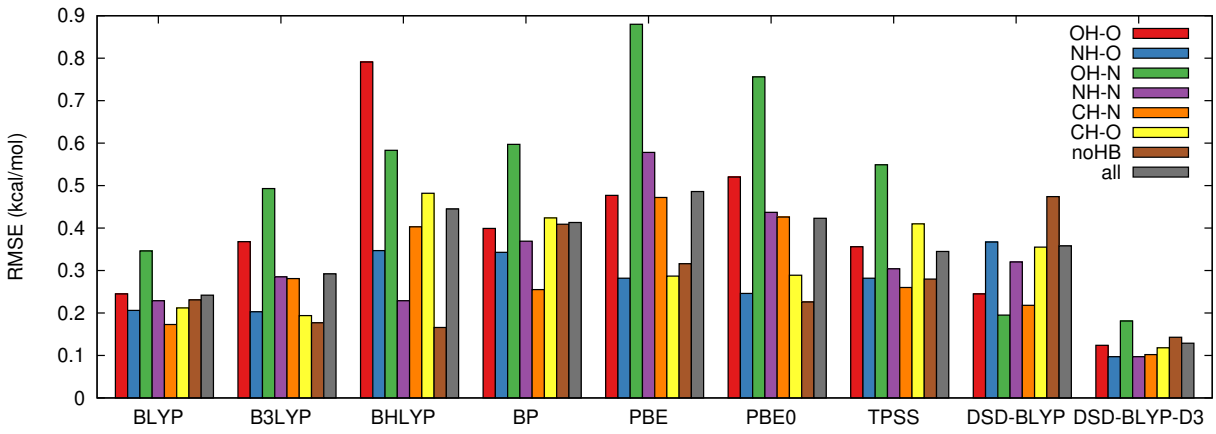
Overall, the correlated wavefunction methods that are free from obvious systematic errors (such as MP2D and all CCSD-based methods) provide similar results across all the groups – in both neutral and ionic systems. This should be highlighted, as the DFT methods discussed below fail in both of these aspects.

5.2 DFT-D3 calculations

The author has briefly tested some DFT-D3 methods on the new data sets. This is by no means exhaustive analysis of the performance of DFT methods, only a preview of new information that can be obtained with the new, larger data sets. Nevertheless, interesting, previously undescribed phenomena have been found. For most of the calculations, the default DFT-D3 setup with the recommended def2-QZVP basis set and Becke-Johnson damping has been used.[15] The results of different functionals across the groups of the HB375 data set are plotted in Figure 3 and listed in the SI, Table S6. Overall, these DFT functionals perform similarly to what has already been reported in other data sets. There is, however, one outstanding feature that deserves special attention. In practically all the cases, the error in the OH-N group is significantly larger than in other groups. This group has, on average, the strongest H-bonds (see the average interaction energies in Table 1), but they are only marginally stronger than in the OH-O group, where the errors are significantly smaller. It is clearly a DFT-related issue, as there is no similar trend visible in the results of correlated wavefunction methods discussed above. When expressed as the difference between the errors in OH-N and OH-O groups, this effect ranges from 0.1 kcal/mol in BLYP-D3 to 0.4 kcal/mol in PBE (when BHLYP is excluded because it yields very large errors in both cases).

This error is systematic; the strength of these H-bonds is clearly overestimated (as indicated by the negative values of MSE listed the in SI, Table S7). A closer inspection of the results reveals that this error is not caused by any specific outliers but by a trend observable

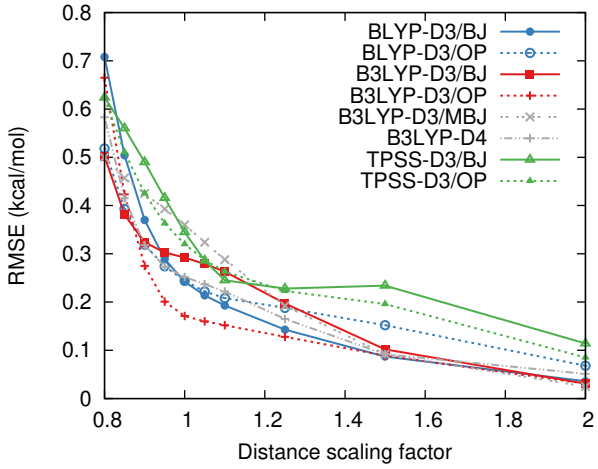
Figure 3: Errors of DFT-D3(BJ)/def2-QZVP and double-hybrid DFT calculations in the HB375 data set



in the whole group. Clearly, there is something special about strong H-bonds with nitrogen serving as an electron donor, which troubles DFT. In order to find the physical origins of this error, the author has attempted to correlate the errors in the OH-N group with the interaction energy components from the SAPT0 decomposition[57] – they correlate best with the polarization energy (which covers both intramolecular induction and charge transfer in these calculations). This supports the hypothesis that these H-bonds are overstabilized because of the delocalization error in DFT, which leads to an exaggerated contribution of charge transfer. Another supporting argument is that the difference between OH-N and OH-O groups decreases when we pass from GGA to a hybrid functional, such as from PBE to PBE0. Unfortunately, the BHLYP functional, where this trend should be very strong because it uses 50% of exact exchange, yields large errors across all the groups that hide this feature. This issue is, however, very small in the DSD-BLYP and DSD-BLYP-D3 double-hybrid functionals that use an even larger fraction of HF exchange, 70%.

The HB375x10 set was used to analyze the distance-dependence of the errors of different DFT-D3 approaches. The BHLYP, BP, PBE and PBE0 functionals yield larger errors that monotonously decrease with intermolecular distance. This is expected behavior, which does not have to be discussed further. The comparison of the remaining functionals is much more interesting. The author of this work has tested both the original Becke-Johnson (BJ)

Figure 4: Errors of DFT-D3/def2-QZVP calculations as a function of intermolecular distance scaling in the HB375x10 data set



damping and the optimized power (OP) damping with one additional parameter[35] (the parameters are not available for the other functionals tested here). The zero damping does not have to be discussed, as it consistently yields larger errors. For the B3LYP functional, the author has also tested a modified version of the BJ damping published recently[34] (MBJ), and the D4 correction, which is based on the same formalism but uses charge-dependent C_6 coefficients. The results are plotted in Figure 4, and the source data are listed in SI, Table S8. It is immediately clear that B3LYP with the BJ damping does not perform well around the equilibrium distance, which can be attributed to the damping function. There is a negative systematic error (the MSE at equilibrium is -0.21 kcal/mol), which implies that the dispersion correction is not damped strongly enough. When OP damping is used, the results are significantly better and the hump on the curve disappears. This also explains why BLYP-D3/BJ has yielded better results than B3LYP-D3/BJ in other data sets as well – it is caused solely by the damping function, and when a better damping function is used, B3LYP becomes more accurate, as can be expected. In TPSS-D3/BJ, a similar issue is observed at longer distances.

A similar, albeit weaker, issue can be found in B3LYP-D4 (see Fig. 4), which uses the same BJ damping function that differs only in the parameters used. These results suggest

that the most widespread BJ damping, as it is parameterized for some functionals, is not optimal for the description of H-bonds (and for other interactions at short distances probably either), and this issue should be paid more attention. The independently parameterized MBJ version yields a smooth curve, but the errors are consistently larger. It is thus not clear whether this issue can be solved by mere reparameterization or a different form of the damping function is needed. The present data set makes it possible to analyze this in detail and will be useful in any future reparameterizations provided that it is complemented by an analogously constructed large set of dispersion-bound complexes (which is now under development).

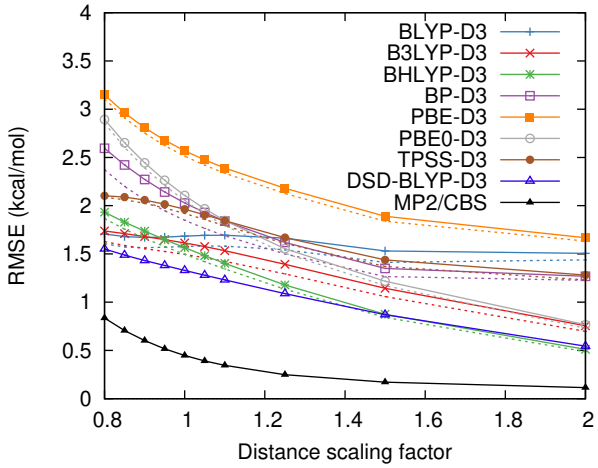
These findings raise a question about the use of B3LYP-D3/BJ for the optimization of the complexes. However, previously published results[13] as well as our own testing against reference CCSD(T)/CBS geometries in the A24 data set[58] show that this method yields very accurate geometries superior to BLYP-D3/BJ, and that the benchmark interaction energies calculated on these geometries are close to the ones computed on the reference geometries.

The double-hybrid functional DSD-BLYP achieves accuracy similar to the dispersion-corrected GGA and hybrid functionals. When the D3 correction is included in the DSD-BLYP-D3, the results improve to a level comparable to correlated wavefunction methods (the RMSE in the HB375 data set is 0.13 kcal/mol). The MP2D method, whose computational complexity is comparable, yields the RMSE of 0.16 kcal/mol.

In the IHB100x10 data set of ionic systems, the author has analyzed the distance dependence of the errors of DFT with D3 and D4 corrections. The results are summarized in Supporting Information Tables S9 and S10, and plotted in Figure 5. Here, D4 systematically performs better than D3, but the improvement is very small. The text below thus discusses only the differences between the functionals, which are much more significant.

The errors here are significantly higher, which cannot be attributed solely to the larger magnitude of the interaction energies in this data set. The main source of the problems here is the delocalization error of DFT, which affects the calculation even at longer distances. In

Figure 5: Errors of DFT-D3 (solid line) and DFT-D4 (dotted) calculations as a function of intermolecular distance scaling in the IHB100x10 data set. MP2/CBS results (black line) are provided for comparison.



the pure GGA and meta-GGA functionals, the error decays only slowly (if at all) with the intermolecular distance; all of these functionals converge to the RMSE of about 1.5 kcal/mol at twice the equilibrium separation. This problem is slightly alleviated in hybrid functionals, where the errors at the longest distance drop below 1 kcal/mol. As expected, the functional with the largest amount of exact exchange, BHLYP, performs best at the longest distances. Nevertheless, it is comparable to B3LYP at equilibrium, and it is not a functional that can be recommended for neutral systems. The double-hybrid DSD-BLYP-D3 with 70% of exact exchange yields similar results here, and is very accurate in the neutral systems, but it is of course more expensive. Even the simplest correlated wavefunction methods outperform DFT significantly because they are not affected by this error. For comparison, MP2/CBS results are included in Figure 5; the error in equilibrium geometries is lower than 0.5 kcal/mol, further decaying at longer distances.

Calculations of ionic systems in the gas phase, and especially of those where the ionized and ionizable sites are separated by an insulating region, are a weak point for DFT. It should be, however, mentioned that these model systems are far from the reality – they were designed to test and develop methods; in most applications, however, there is an environment such as solvent that shields strong electrostatic interactions and DFT results can be expected

to be better.

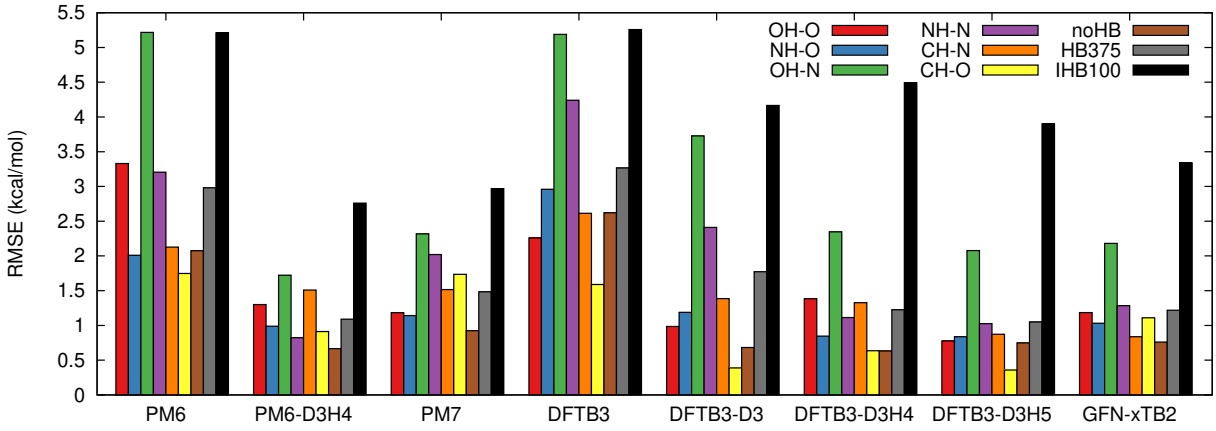
5.3 Semiempirical QM calculations

The description of non-covalent interactions in semiempirical QM methods is not very good; these problems stem directly from the approximations on which these methods are based. Besides London dispersion (which is missing in all mean field methods), also hydrogen bonding is affected very strongly. All of these methods use the minimal valence basis, which limits the description of polarization, and the use of only s-functions on hydrogen atoms does not allow any polarization at all. Additionally, DFTB uses monopole approximation for all electrostatic interactions, which prevents any on-atom polarization. Multiple corrections aimed to improve the description of H-bonds have been devised, and some of them are tested here. The results for the HB375 and IHB100 data sets are plotted in Figure 6 (the original results are available in the SI, Table S11).

The neutral systems only are discussed here first. The uncorrected PM6 method[42] yields rather large errors, but with the D3H4 corrections,[43] it becomes one of the most accurate SQM methods with the RMSE of 1.1 kcal/mol in the HB375 data set, and the results are similar for all the groups of this set. This is, however, also a result of the use of a separate parameter for each combination of elements in the H-bond in the H4 correction. PM7 includes embedded dispersion and H-bonding corrections,[44] but they are simpler and less extensively parameterized.

The DFTB3 method includes a simple H-bond correction (so-called XH damping or γ-damping),[49] but it is not strong enough to compensate for the serious underestimation of the strength of hydrogen bonds. The dispersion correction added in the DFTB3-D3 method[50] improves the description of weakly interacting systems, but the errors in strong H-bonds (especially those with a nitrogen serving as the electron donor) are still rather large. A standalone D3H4 correction[9,43] (with pairwise parameters for all combinations of elements in the H-bonds) provides better results. The best DFTB result, and the best result in

Figure 6: Errors of semiempirical QM calculations in HB375 (individual groups and the whole data set) and IHB100 (the whole set only).



this class of methods, is achieved with the DFTB3-D3H5 method[11] (the RMSE of 1.05 kcal/mol). The H5 correction is integrated into the DFTB3 calculation, and its connection to the electron density on the interacting atoms makes it possible to achieve high accuracy with a smaller number of parameters than in the more empirical H4 correction. The GFN2-xTB method[48] is also competitive with the RMSE of 1.2 kcal/mol; this method has been developed specifically for the description of non-covalent interactions.

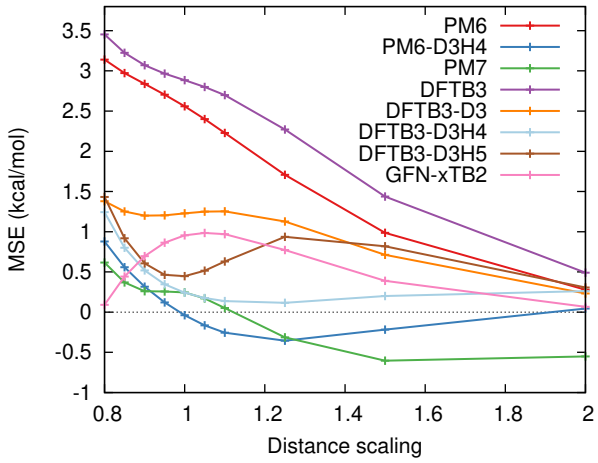
One important trend visible across all the methods tested seems to be similar to what has been observed in DFT – among the groups of the HB375 data set, the OH-N H-bonds have the error significantly larger than the remaining groups. However, the error here is of a different nature – the strength of these H-bonds is underestimated, while the opposite is true in DFT. These H-bonds are very strong, with a large contribution of polarization and possibly also of charge transfer, and the SQM methods simply fail to describe such a strong interaction. All of the methods yield a positive systematic error (MSE) in this group, although it is remarkably small in DFTB3-D3H4 (where it is accounted for by the pairwise parameters). The OH-O H-bonds are almost as strong, but their description by these methods is significantly better. This observation can be linked to other problems with the parameterization of nitrogen in DFTB.[49]

Next, the distance dependence of the errors in the HB375x10 set will be examined. The

analysis here does not focus on the overall RMSE, but primarily on the systematic part of the error represented by the MSE. The results are plotted in Figure 7 using data provided in the SI Table S13 (an analogous plot of the RMSE is provided in the SI Figure S1 and Table S12). PM6 and DFTB3 without further corrections exhibit large positive errors at short distances, which slowly decay with the intermolecular distance. The behavior of the corrected methods is more interesting. PM6-D3H4 exhibits practically no systematic error at the equilibrium distances, which implies that the parameterization is transferred well even to this large data set. At longer distances, there is a small negative error, which indicates that the correction can probably be improved by making it more short-ranged. DFTB3-D3H4 results exhibit similar distance dependence, although the curve is shifted towards positive values. The MSE of DFTB3-D3H5 is best in the equilibrium, and it grows again at longer distances. This implies that the correction is too localized; nevertheless, this was done on purpose in order to minimize the effect of the correction on the atom pairs not forming a H-bond. Finally, GFN2-xTB has the maximum MSE around the equilibrium, and the error is smaller at both shorter and longer distances. This issue is not so prominent at the level of the RMSE, which grows at short distances (see Fig. S1 in the SI), but there is a hump on the RMSE curve, which corresponds to this maximum of the MSE. This anomaly deserves a more detailed investigation in the future.

The interactions of charged species are stronger and induce larger changes in the electronic structure of the interacting molecules. The description of these interactions is thus even more challenging for the SQM methods, and the errors are larger. The tight-binding based methods have serious problems with convergence in some systems, and despite all the effort, some systems had to be removed from the set for the purpose of this analysis. They are listed in Table S11 in the SI. The total errors in the IHB100 data set are plotted in Figure 7, where they can be compared with the results in the HB375 set. For more information on the distribution of the errors, they are plotted in the form of a box plot in Figure 8 (in the box plot, the box contains 50% of the errors, the whiskers 95%, and the remaining points

Figure 7: Systematic error (represented by MSE, mean signed error) of semiempirical QM calculations as a function of intermolecular distance scaling in the HB375x10 data set



are plotted as circles.).

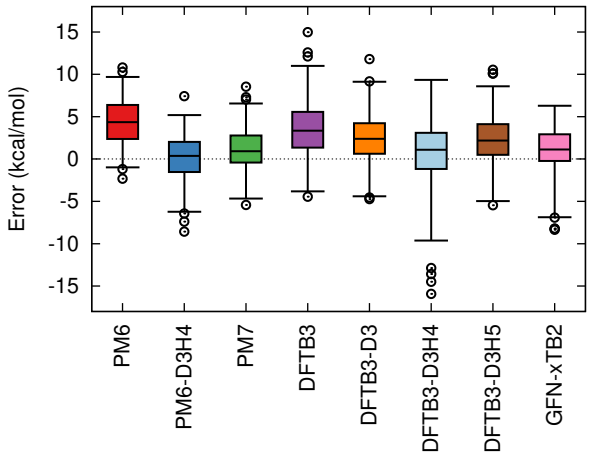
All of the SQM methods underestimate the strength of these H-bonds ; these systematic errors are best minimized by empirical D3H4 corrections. Overall, PM6-D3H4 and PM7 perform best, although the errors are more than two times larger than in neutral systems (this, however, reflects well the increased strength of the ionic H-bonds).

DFTB-based methods suffer from larger errors that are difficult to correct. Here, the usual positive errors (due to underestimated polarization) are mixed in some cases with the large negative error caused by the the delocalization error inherited from DFT. This is most pronounced in DFTB3-D3H4, where the correction independent of the DFTB calculation only strengthens the interaction even in the cases where DFTB3 already overestimates the strength of the H-bond. DFTB3-D3H5 and GFN2-xTB behave slightly better.

5.4 An overlap of HB375 and S66 data sets

The present HB375 data set contains some complexes included in the S66 set; splitting the data set using these criteria may provide useful analysis of the methods that have been parameterized on S66. More specifically, focus will only be placed on the true hydrogen bonds of nitrogen and oxygen, neglecting the systems with a more significant contribution of

Figure 8: Box plot depicting the distribution of the errors of semiempirical QM methods in the IHB100 data set.



dispersion because they form a larger part of S66 while being underrepresented here. This subset of the HB375 data set can be divided into three groups – dimers present in S66, dimers containing one monomer also used in S66, and completely different systems. The following methods will be examined: PM6-D3H4 and DFTB3-D3H5, both using empirical corrections for dispersion and hydrogen bonding (where the latter is based on a very different principle), developed by the author and parameterized using S66x8. This is complemented by similar methods developed by others, PM6-DH+ and PM7, which were also fitted using S66 or S66x8. The last methods added are B3LYP-D3, where the dispersion correction has been parameterized using S66x8, and MP2/CBS, which is free of any parameters. The results are reported in Table 4 as relative errors in the three groups defined above (using the RMSE divided by the average magnitude of the interaction energy in the group). An analogous table of the actual values of the RMSE is provided in the Supporting Information as Table S14.

As expected, all of the semiempirical methods with corrections fitted to S66 perform better in the complexes found in S66 and worse in the others. However, these results are still rather good (if we omit PM7); in the systems completely different from S66, the best of these methods, DFTB3-D3H5, yields the RMSE of 1.17 kcal/mol. It is interesting to

Table 4: Relative errors (RMSE/$|\langle \Delta E^{int} \rangle|$, in percent) of selected methods, grouped by the overlap of the system with the S66 data set. The average benchmark interaction energies in these groups are also listed for reference.

S66 overlap:	none	monomer	dimer
PM6-D3H4	19.1	17.6	13.9
DFTB-D3H5	18.6	18.0	12.9
PM6-DH+	20.5	19.4	13.8
PM7	26.8	23.6	19.5
B3LYP-D3	5.0	5.6	6.1
MP2/CBS	7.4	3.5	1.8
	$\langle \Delta E^{int} \rangle$, kcal/mol		
	-6.31	-7.15	-8.29

note that the same trend has been observed for completely non-empirical MP2/CBS. Here, the source of this issue is different. S66 contained mostly prototypical H-bonds, which are stronger than the average in a broader set (as documented by the average interaction energy in each group listed in the table). A stronger H-bond means a smaller relative contribution of dispersion, and MP2 is known to overestimate it severely. Finally, DFT-D, represented here by B3LYP-D3, provides very balanced results. Although there may be some bias for S66 in dispersion-dominated complexes (as the correction was parameterized on S66x8), in hydrogen bonds this effect is negligible and the major part of the error comes from the DFT itself.

5.5 Reparameterization of semiempirical methods

It should also be tested how the use of the new dataset in method parameterization changes the results. The H-bonding correction in the DFTB3-D3H5 and PM6-D3H4 methods has been chosen as the test case. The H5 correction has two global parameters that define the spatial extent of the correction (which will not be parameterized) and an element-specific parameter k_{YH} for the atom Y in a XH-Y hydrogen bond that determines the strength of the correction. Here, this parameter for O and N will be optimized. The H4 correction uses four pairwise parameters determining the strength of the correction for the combinations of the

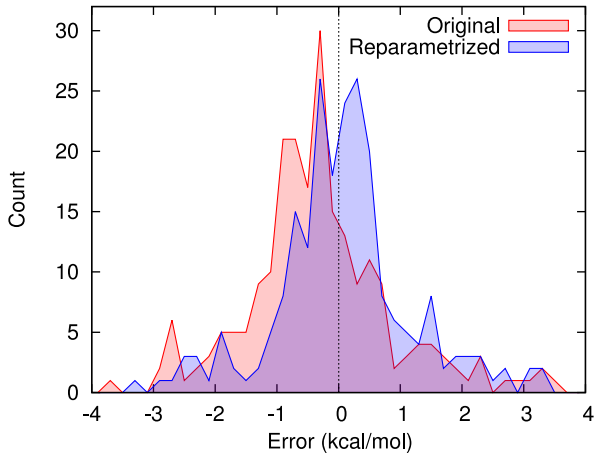
donor and acceptor atoms X and Y, with X and Y being oxygen or nitrogen, and an extra parameter for water. Only these five parameters will be modified, whereas the remaining global parameters will be kept fixed. The original corrections have been fitted to the H-bonded complexes of the S66x8 data set. To use a corresponding set of hydrogen bonds, only the first four groups of the HB375 data set (the H-bonds of nitrogen and oxygen) have been used for the reparameterization. Finally, since only the scaling factors determining the strength of the corrections is being fitted while the geometric part is kept fixed, only the equilibrium geometries are used. The parameterization has been performed as a gradient optimization based on the finite-difference gradient, minimizing the RMSE in the training set.

The reparameterization of DFTB3-D3H5 has only had a small effect. The original values of the parameters $k_{OH} = 0.06$ and $k_{NH} = 0.18$ have changed only slightly to 0.065 and 1.95. The RMSE (in the training set) has decreased only negligibly from 1.22 to 1.20 kcal/mol. A more important result is the change of the systematic error; the MSE has decreased from 0.47 to 0.27 kcal/mol. This implies that the original parameterization is already at the limit of what can be achieved with such a simple correction, but the reparameterization has allowed it to adapt to the different distribution of the errors in the larger and more diverse data set.

The results for PM6-D3H4 are very similar. The values of the parameters have changed only slightly, and the overall error has dropped from 1.23 to 1.11 kcal/mol. However, the improvement in the systematic error is more significant here. The MSE has changed from -0.32 kcal/mol to 0.07 kcal/mol. This effect can be clearly seen in the plot of the distribution of the errors in Figure 9. It is obvious that the correction has been fitted to a limited set of strong H-bonds, which has led to the overcorrection of weaker ones. When the correction is fitted to a larger and more diverse data set, this bias is removed.

The improvements achieved by the reparameterization are too small to warrant the release of a new version of the corrections. Nevertheless, the results show that the original parameterization on S66 has led to a bias that can be observed in the larger HB375 data set

Figure 9: Distribution of the errors of PM6-D3H4 in groups 1–4 of the HB375 data set before and after the reparameterization of the H4 correction.



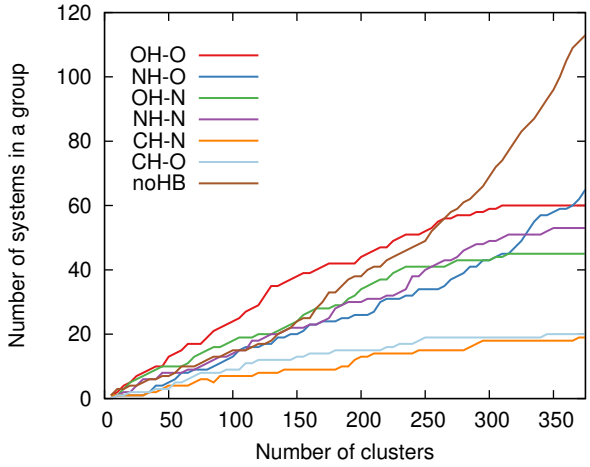
and is removed when this data set is used for the fitting.

5.6 Clustering

The clustering algorithm described above has been applied to the HB375 and IHB100 data sets. In the HB375 set, predefined subsets of 20, 50, 100 and 200 systems had been built that can be used for applications requiring different amounts of data. The members of each subset are listed in the SI Table S3 and are labelled in the dataset definition file. When it is processed using the Cuby framework, a single keyword can be used for selecting the desired subset. Users are encouraged to use these predefined subsets, so that their results are easily reproducible and comparable across publications.

Additionally, the analysis of the results of the clustering provides valuable information on the composition and character of the whole data set. An important validation of this approach lies in the analysis of the relationship between the clusters defined in this way and in the chemical classification of the systems. The HB375 data set is divided into seven groups by the nature of the interaction and the elements involved in the hydrogen bond. Interestingly, when it is split into seven clusters, each of the representatives of these clusters belongs to a different group. As the number of clusters is increased, the assignment of

Figure 10: Population of the HB375 groups in subsets obtained by clustering, plotted as a function of the number of the systems in the subset.



their representatives to these groups reflects well the overall composition of the data set (see Fig. 10). The two largest groups in the set, NH-O and, more importantly, noHB, seem to be the most redundant ones, as their participation in the clusters grows even after other groups become saturated. This leads to two important conclusions: First, the similarity measure used is not biased towards some chemical properties of the systems, which makes the subsets generated by the clustering well balanced and safe for general use. Second, it is a piece of evidence suggesting that the whole data set is well balanced, as different errors are distributed evenly across it.

The same procedure has been used to obtain subsets of 20 and 50 systems from the IHB100 data sets. Here, the population of the groups is less balanced, which is reflected by the composition of the subsets obtained by clustering. Some groups with only a few members are not included in the subsets, and larger groups are represented by more systems.

The results of the clustering also provide insight into the redundancy of the complete HB375 data set at the more coarse-grained level. Inspection of the relationship between the number of clusters and their sizes (plotted in Fig. 10) reveals that when only a few clusters are formed, the systems are not distributed evenly, but a large part of them falls into several most important clusters. These large clusters correspond well to the physical groups in the

data set, and the smaller clusters collect the outliers. Once about 230 clusters are formed, the distribution becomes more even and there is no cluster significantly larger than others. This may be the point where the maximum information can be obtained from a limited number of systems.

This analysis highlights the advantage of the clustering approach – in contrast to a random selection of the subset, selecting one representative from each cluster eliminates the redundant information and maximizes the diversity.

For the use of the subset generated by the clustering, it is important to know how the results obtained in the subset relate to those obtained in the whole data set. This has been tested for three methods from different categories, MP2/CBS, DFT-D3 and PM6-D3H4, on the HB375 data set. The results are summarized in Table 5. In DFT-D3 and PM6-D3H4, the errors are distributed more randomly (as these are only the errors that cannot be removed by the corrections applied), so there is no significant dependence of the average error on the subset size. However, MP2/CBS has a well-known systematic error – it overestimates dispersion-bound complexes while hydrogen bonds are described comparatively better. The error is thus the lowest in the smallest subset, where the ratio between strong H-bonds to the weak interactions from the noHB group is large, and it grows with the subset size as more systems where dispersion is important are introduced.

Table 5: Error (RMSE, in kcal/mol) of selected methods in subsets of different size generated by clustering, and in the whole data set.

selection	MP2/CBS	B3LYP-D3	PM6-D3H4
20 clusters	0.167	0.373	1.228
50 clusters	0.293	0.380	1.420
100 clusters	0.301	0.373	1.269
200 clusters	0.384	0.362	1.215
all	0.435	0.333	1.091

6 Conclusions

The Non-Covalent Interactions Atlas project (www.nciatlas.org) aims to provide high-quality benchmark data for a wide range of non-covalent interactions across the chemical space. This paper presents the first two data sets covering hydrogen bonds in organic molecules (more generally, in HCNO chemical space): HB375, featuring 375 neutral complexes, and IHB100 with 100 ionic systems. The benchmark interaction energies have been calculated using a conventional composite CCSD(T)/CBS scheme employing large basis sets. Both sets have been extended with ten-point dissociation curves, with interaction energies accurately rescaled to the same level (the sets HB375x10 and IHB100x10). In total, this study brings 4,750 data points computed at a true "gold standard" level. This is an increase of an order of magnitude compared to the previous state of the art. Moreover, not only the size but also other properties of the data sets have been improved, addressing many shortcomings of currently used data sets such as S66x8. The data sets, metadata and additional results are available in several forms including data files ready for machine processing. Predefined subsets obtained by clustering analysis are also provided for applications that do not need so much data.

Although the main purpose of these data sets is their application in the development of novel approximate methods, even a brief analysis of the results of common DFT-D3 methods has revealed interesting issues. First, it has shown that the delocalization error is surprisingly large even in neutral systems. Specifically, in the group of OH-N bonds, it can be as large as 0.4 kcal/mol for some commonly used GGA functionals. Second, in some functionals with the D3 dispersion using Becke-Johnson damping, the damping function is not optimal and introduces an unnecessarily large error specifically around the equilibrium distance of hydrogen bonds. The surprising superiority of BLYP-D3 interaction energies over B3LYP-D3, which has already been noted in the literature, can be attributed to this problem in B3LYP-D3. The new data set is not suitable as a training set for dispersion correction on its own; nevertheless, when more data for dispersion-bound complexes are available, the

reoptimization of the damping may improve the accuracy of some DFT-D3 methods.

The main application of the new data sets is in the development of approximate, empirical computational methods such as semiempirical QM methods. This paper has used the HB375 and IHB100 data sets for testing several SQM methods with corrections for H-bonding, and for an experimental reparameterization of this correction in two of them, PM6-D3H4 and DFTB3-D3H5. In the neutral systems, the SQM methods with the latest corrections perform rather well, with the RMSE only slightly above 1 kcal/mol. This is an important validation, because these corrections have been developed using significantly smaller training sets. The result of the limited reparameterization described here is even more important. In both methods, the overall error (RMSE) cannot be reduced much, which indicates that the present parameterization of the corrections is sufficient and there is no room for improvement without changing the form of the corrections. However, the systematic component of the error in the HB375 data set could be reduced, because the original parameterization of the method on the H-bonds from the S66 data set has introduced a bias towards strong H-bonds. The main advantages of the new data sets, their large size and increased diversity, could be fully exploited in the development of the next generation of approximate methods, where more complex corrections could be applied.

The data sets presented here are just the first part of a series that aims to cover also other classes of interactions in broader chemical space. At present, a complementary data set of halogen bonds involving additional elements (S, P and halogens) is almost finished, and a large data set of dispersion-bound complexes is being developed.

7 Acknowledgements

We acknowledge the support from the Czech Science Foundation, Grant No. 19-13905S, and from the European Regional Development Fund; OP RDE; Project: Chemical Biology for Drugging Undruggable Targets (Chem-BioDrug, No. CZ.02.1.01/0.0/0.0/16_019/0000729).

This work is part of the Research Project RVO 61388963 of the Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences. It has also been supported by the Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project IT4Innovations National Supercomputing Center, LM2015070.

8 Supporting Information

The Supporting Information contains: 1) additional tables and figures supporting the main text, including tables of the data used in the plots featured here, 2) geometries of all the systems and tables of benchmark interaction energies needed for the reproduction of the results presented here, and 3) a machine-readable definition of the data sets containing the metadata describing the categorization of the systems.

References

(1) Řezáč, J.; Hobza, P. Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications. *Chem. Rev.* **2016**, *116*, 5038–5071.

(2) Řezáč, J.; Hobza, P. *Encyclopedia of Physical Organic Chemistry*; John Wiley & Sons, Inc., 2017.

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

(4) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.

(5) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark calculations of noncovalent interactions of halogenated molecules. *J. Chem. Theory Comput.* **2012**, *8*, 4285–4292.

(6) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(7) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *The Journal of Chemical Physics* **2017**, *147*, 161727.

(8) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

(9) Miriyala, V. M.; Řezáč, J. Description of non-covalent interactions in SCC-DFTB methods. *J. Comput. Chem.* **2017**, *38*, 688–697.

(10) Non-Covalent Interactions Atlas website. 2019; `http://www.nciatlas.org`.

(11) Řezáč, J. Empirical Self-Consistent Correction for the Description of Hydrogen Bonds in DFTB3. *J. Chem. Theory Comput.* **2017**, *13*, 4804–4817.

(12) 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.

(13) Witte, J.; Goldey, M.; Neaton, J. B.; Head-Gordon, M. Beyond Energies: Geome-

tries of Nonbonded Molecular Complexes as Metrics for Assessing Electronic Structure Approaches. *J. Chem. Theory Comput.* **2015**, *11*, 1481–1492.

(14) 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.

(15) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(16) Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D. Psi4: an open-source ab initio electronic structure program. *WIREs Comput Mol Sci* **2012**, *2*, 556–565.

(17) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method†. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(18) Řezáč, J. Cuby: An integrative framework for computational chemistry. *J. Comput. Chem.* **2016**, *37*, 1230–1237.

(19) Koch, H.; Fernández, B.; Christiansen, O. The benzene–argon complex: A ground and excited state ab initio study. *J. Chem. Phys.* **1998**, *108*, 2784.

(20) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. Estimates of the Ab Initio Limit for π-π Interactions: The Benzene Dimer. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

(21) Hobza, P.; Šponer, J. Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

(22) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *The Journal of Chemical Physics* **1997**, *106*, 9639–9646.

(23) Boys, S.; Bernardi, F. Calculation of Small Molecular Interactions by Differences of Separate Total Energies - Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(24) Burns, L. A.; Marshall, M. S.; Sherrill, C. D. Comparing Counterpoise-Corrected, Uncorrected, and Averaged Binding Energies for Benchmarking Noncovalent Interactions. *J. Chem. Theory Comput.* **2014**, *10*, 49–57.

(25) Woon, D. E.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994**, *100*, 2975.

(26) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, δMP2CCSD(T): Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. *The Journal of Chemical Physics* **2011**, *135*, 194102.

(27) Řezáč, J.; Dubecký, M.; Jurečka, P.; Hobza, P. Extensions and applications of the A24 data set of accurate interaction energies. *Phys. Chem. Chem. Phys.* **2015**, *17*, 19268–19277.

(28) Kodrycka, M.; Patkowski, K. Platinum, gold, and silver standards of intermolecular interaction energy calculations. *J. Chem. Phys.* **2019**, *151*, 070901.

(29) Grimme, S. Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* **2003**, *118*, 9095.

(30) DiStasio, R.; Head-Gordon, M. Optimized spin-component scaled second-order Moller-Plesset perturbation theory for intermolecular interaction energies. *Mol. Phys.* **2007**, *105*, 1073–1083.

(31) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. Improvement of the coupled-cluster singles and doubles method via scaling same- and opposite-spin components of the double excitation correlation energy. *J. Chem. Phys.* **2008**, *128*, 124111.

(32) Pitoňák, M.; Řezáč, J.; Hobza, P. Spin-component scaled coupled-clusters singles and doubles optimized towards calculation of noncovalent interactions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9611.

(33) Řezáč, J.; Greenwell, C.; Beran, G. J. O. Accurate Noncovalent Interactions via Dispersion-Corrected Second-Order Møller–Plesset Perturbation Theory. *J. Chem. Theory Comput.* **2018**, *14*, 4711–4721.

(34) Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7*, 2197–2203.

(35) Witte, J.; Mardirossian, N.; Neaton, J. B.; Head-Gordon, M. Assessing DFT-D3 Damping Functions Across Widely Used Density Functionals: Can We Do Better? *J. Chem. Theory Comput.* **2017**, *13*, 2043–2052.

(36) Caldeweyher, E.; Bannwarth, C.; Grimme, S. Extension of the D3 dispersion coefficient model. *The Journal of Chemical Physics* **2017**, *147*, 034112.

(37) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(38) DFTD4 program. 2019; `https://github.com/dftd4/dftd4`, original-date: 2019-02-28T15:45:22Z.

(39) Kozuch, S.; Gruzman, D.; Martin, J. M. L. DSD-BLYP: A General Purpose Double Hy-

brid Density Functional Including Spin Component Scaling and Dispersion Correction. *J. Phys. Chem. C* **2010**, *114*, 20801–20808.

(40) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Computational Molecular Science* **2018**, *8*, e1327.

(41) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

(42) 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.

(43) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

(44) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model* **2013**, *19*, 1–32.

(45) Stewart, J. J. P. MOPAC 2016. 2016; `http://openmopac.net/`.

(46) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260.

(47) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(48) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multi-

pole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(49) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(50) Grimme, S. Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angew. Chem. Int. Ed.* **2013**, *52*, 6306–6312.

(51) XTB, Semiempirical Extended Tight-Binding Program Package. 2019; `https://github.com/grimme-lab/xtb`, original-date: 2019-09-30T12:40:09Z.

(52) Gould, T. 'Diet GMTKN55' offers accelerated benchmarking through a representative subset approach. *Phys. Chem. Chem. Phys.* **2018**,

(53) Morgante, P.; Peverati, R. Statistically representative databases for density functional theory via data science. *Phys. Chem. Chem. Phys.* **2019**, *21*, 19092–19103.

(54) Gordon, A. D. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General)* **1987**, *150*, 119–137.

(55) Řezáč, J. Cuby 4, software framework for computational chemistry. 2015; `http://cuby4.molecular.cz/`.

(56) The Official YAML Web Site. `http://yaml.org/`.

(57) Hohenstein, E. G.; Sherrill, C. D. Density fitting of intramonomer correlation effects in symmetry-adapted perturbation theory. *J. Chem. Phys.* **2010**, *133*, 014101.

(58) 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.