Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies on the example of glucose and α-maltose

Mateusz Marianski,* Adriana Supady,* Teresa Ingram,* Markus Schneider,* and Carsten Baldauf*

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany

E-mail: marianski@fhi-berlin.mpg.de; supady@fhi-berlin.mpg.de; ingram@fhi-berlin.mpg.de; markus.schneider@fhi-berlin.mpg.de; baldauf@fhi-berlin.mpg.de

Abstract

A big hurdle when entering the field of carbohydrate research stems from the complications in the analytical and computational treatment. In effect, this extremely important class of biomolecules remains under-investigated when compared, for example, with the maturity of genomics and proteomics research. On the theory side, the commonly used empirical methods suffer from insufficient amount of high-quality experimental data against which they can be thoroughly validated. In order to provide a pivotal point for an ascent of accurate carbohydrate simulations, we present here a structure/energy benchmark set of diverse glucose (in three isomeric forms) and α-maltose conformations at the coupled-cluster level as well as an assessment of commonly-used energy functions. We observe that empirical force fields and semi-empirical approaches, on average, do not reproduce accurately the reference energy hierarchies. While the force fields maintain accuracy for the low-energy structures, the semi-empirical methods perform unsatisfactory for the entire range. On the contrary, density-functional approximations reproduce the reference energy hierarchies with better than chemical accuracy already at the generalized gradient approximation level (GGA). Particularly, the novel meta-GGA functional SCAN provides the accuracy of more expensive hybrid functionals at the fraction of their computational cost. In conclusion, we advocate for electronic-structure theory methods to become the routine tool for simulations of carbohydrates.

Introduction

Carbohydrates are an important class of biomolecules; they can adopt multiple functions in living organisms, for example as nutrient and energy storage, as structural biopolymer in the cell walls of plants, as recognition modules in the immune system, *etc.*[1–6] Analytics of complex carbohydrates is of interest for example in the research on biopharmaceuticals, in immunology, or in food chemistry. Recent advances in analytical methodology point towards a new standard in the analytical characterization of such compounds: ion-mobility spectrometry is able to separate complex carbohydrates of very similar chemical structure.[7,8] Especially in conjunction with simulations that can predict the ion mobility of different carbohydrate species, this technique can become standard in glyco-analytics.

The importance of carbohydrates is not limited to living organism. From the viewpoint of chemical industry they represent a potentially sustainable source of energy and basic chemicals.[9–11] Due to their global abundance, especially polymeric crystalline carbohydrates are of interest as they can eventually supersede the current fossil-fuel driven energy supply and chemical industry. Key to that is an understanding of the extraordinary chemical stability of crystalline polymeric carbohydrates,[12] which will eventually lead to the development of energy-efficient decomposition reactions.

We see a possible high impact of simulation and modeling in all these fields, but what is missing at present is a systematic assessment of the accuracy of different approximation levels. Systematically generated data sets are extremely helpful in order to derive trends over parts of chemical space,[13,14] but have also, in conjunction with benchmark-level energies, proven essential in bringing theoretical methods to a common-ground to allow their evaluation[15–17] and systematic parametrization.[18,19] With the present study, we reach two important milestones on that route: (i) a benchmark data set of glucose and maltose conformers with energies at the coupled-cluster level and (ii) an evaluation of a wide range of energy functions, from empirical force fields (FFs) to semi-empirical quantum chemistry methods (SQM) and density-functional tight binding (DFTB) to density-functional theory

(DFT).

Similar PES-based comparisons have been done in the past, however they lacked generality, being limited to only one set of methods or mono-saccharides.[20–25] Hemmingsen *et al.*[20] examined the performance of 20 FFs against DFT (including corrections from coupled-cluster theory) for five monosaccharides and a water-monosaccharide complex and concluded that none of the FFs performed consistently better in all cases. In a study from 2009 Stortz *et al.*[21] compared predictions of 18 FFs and the semi-empirical PM3CARB-1 method against experimental crystal structures and DFT calculations for three disaccharides. They found that more recent parameterizations provide results more consistent with the benchmark data. Systematic evaluation of different density-functional approximations (DFAs) against MP2 energies for α,β-D-allopyranose, β-D-glucopyranose and 3,6-anhydro-4-O-methyl-D-galacticol (34 conformers in total) was done by Csonka *et al.*[22] In a test set composed of different anomers, $^{1}C_{4}$ and $^{4}C_{1}$ ring puckers and orientation of hydrogen bonds, M05-2X and PBE provided the most reliable results. In another study Csonka *et al.*[23] improved the reference MP2 energies by including corrections from the CCSD(T) level of theory. The refined benchmark energies altered previous conclusions in favor of B3PW91 and PBE0 functionals. Sameera and Pantazis[24] employed 58 structures of mono-saccharides, which varied in constitutional isomerism (open-chain isomer) and ring puckering (chairs and skew boats), to evaluate performance of a number of wave function and density-functional methods. They found that the recently highlighted[26] localized pair natural orbital-coupled electron pair approximation (LPNO-CEPA) method provided CCSD(T)-quality reference data. Among DFAs, they highlighted M06, M06-2X, B3PW91 and PBE0 functionals and discouraged using the popular LYP correlation functional for studying ring-opening reaction. Govender *et al.*[25] evaluated seven SQM approaches (including the reparametrized AM1/d-CB1 method) against DFT and CCSD(T) data. Finally, we note that, besides the aforementioned papers, vast literature on theoretical modeling of carbohydrates exists,[27–32] also including experimental support,[33–35] to name very few.

Chemical and structural characteristics of carbohydrates

Most of the known biological carbohydrates are composed of approximately 20 different monosaccharide building blocks that are connected to each other by a glycosidic bond. Carbohydrates are not necessarily composed as linear chains which sets them apart from the always linear backbones of peptides and nucleic acids - the carbohydrate building blocks have one donor (the anomeric C) but multiple acceptors for glycosidic bonds. In effect, distinct glycosidic linkages exist: 1→2, 1→3, 1→4, and 1→6, where the first number refers to the carbon that donates the oxygen atom and the second links to the number of the accepting carbon at the reducing end of the respective disaccharide unit. In addition, glycosidic bonds can exist in two enantiomeric forms (α or β). From combinations of this multiplicity of linkages and stereochemistries, saccharides can arise that have different spatial structure (hence function) despite identical composition and sequence.[7,36] The rich diversity of carbohydrates surpasses the number of possible sequences in nucleic acids and peptides by orders of magnitude, even with relatively small numbers of building blocks.[37]

The significant conformational degrees of freedom for the overall structure are rotations around the single bonds of the glycosidic linkages and the conformation of the monosaccharide rings. The canonical ring puckers for pyranose rings are: chair "C", half-chair "H", envelope "E", boat "B", and skew boats "S".[30,38] Subtypes of these ring puckers differ in the relative position of distinct ring atoms indicated by subscript and superscript numbers, e.g. $^{1}C_{4}$ or $^{3}S_{1}$. The relative stability of these structures is defined by electrostatics, e.g. hydrogen bonds, or mid-range electron correlation, e.g. the anomeric effect, whereas van der Waals interactions (i.e. long-range electron correlation) should be negligible in such a compact system. In this work, we consider the open-chain glucose as well as the two anomers of D-glucopyranose (hereafter simply referred as glucose; see upper panel of Figure 1). The relative stability of these three isomers feature the energetics of the mutarotation reaction that links the open and closed-chain forms. Prediction of the relative stability of the reactant and the product of such reactions can be more problematic for theoretical methods than

describing the conformational hierarchies.[39,40] Thus we distinguish a subset of closed-ring α- and β-glucose structures from the benchmark sets. In addition to the monosaccharide glucose we investigate the disaccharide α-maltose. Here another level of conformational flexibility has to be considered, the two rotational degrees of freedom ϕ and ψ of the glycosidic bond (see lower panel of Figure 1). As a result of the increase in size and flexibility, van der Waals interactions should gain importance as various orientations of the glycosidic bond yield different distances between the two interconnected glucose building blocks.

Although the two analyzed sugars neither exhaust the list of possible linkages nor feature the possible modifications by functional groups, they do emphasize fundamental challenges characteristic to carbohydrates. Variations in the monosaccharide, i.e. substitution of glucose to xylose or mannose, or alternative glycosidic-bond connectivities will not alter the validity of the underlying physical model. Two aspects however can impact on intramolecular interactions and molecular flexibility, the introduction of further chemical modifications (e.g. acetylation or phosphorylation) or of the more flexible 1→6 linkage. These aspects should be studied in the near future to complete the the picture.

Finally, in contrast to structures that were used to parametrize carbohydrate force fields,[41,42] the designed benchmark set does not focus solely on the low-energy ring pucker ($^4C_1$) but features conformations that populate different parts of conformational space. Although $^4C_1$ puckers will dominate the structural space of monomers or short oligomers under unperturbed conditions, several reasons advocate for investigating the alternative ring puckers. First, there are experimental and theoretical evidences that non-chair conformations are featured in the enzymatic center of glycoside hydrolases.[30,43,44] Moreover, complex carbohydrates in the gas phase feature dynamics that facilitates non-chair conformations.[7,8] Consequently, there is a possibility that conformers high in energy in shorter oligosaccharides are likely to be found in central positions of larger oligomers where seemingly unfavorable conformations of a single building block can be compensated for by interactions between others. Finally, the correct assessment of kinetics of the conformational changes requires crossing

seemingly unfavourable parts of free-energy surface which modulate the slow dynamics of the molecule. This problem has been recognized for the protein force fields.[45] Consequently, we decided to consider an extended conformational space for the carbohydrates in our benchmark study.

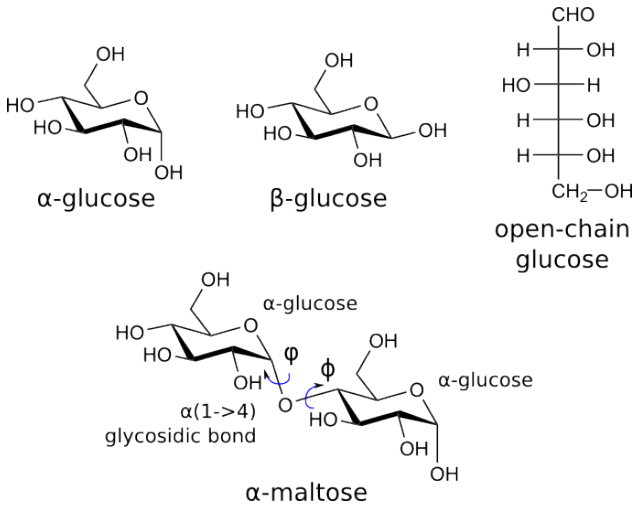

Figure 1: Chemical structures of α, β, and open-chain glucose and α maltose.

Benchmark calculations and chemical accuracy

An accuracy of 1 kcal/mol is often referred to as chemical accuracy, i.e. the accuracy limit accessible in experimental measurements. The reproduction and prediction of experimental results by computer simulations demands first of all an accurate treatment of the molecular potential-energy surface (PES). In simulations, ground state physico-chemical properties of molecules are derived from sampling a PES. As a consequence, the computational accuracy requirements for the PES must be much higher, maybe in the 10 cal/mol range. The calculation of the dissociation energy of the HF dimer by Řezáč and Hobza[46] may serve as an illustrative example, where the authors elegantly dissect the contributions necessary to match a highly accurate experimental estimate of the dissociation energy D_0.[47] They also restate the complication of comparing the experimental observable to a computed number: the dissociation energy consists of the interaction energy ΔE^{int} and the change of the zero-

point vibration energy ΔE^{ZPV}. Both numbers cannot be separated in an experiment. While ΔE^{int} is readily available from the PES, ΔE^{ZPV} results from sampling the PES. The extreme accuracy of the experiment by Bohac *et al.*, but also the need for sampling the PES to calculate ΔE^{ZPV} that Řezáč and Hobza refer to, illustrate the necessity of PES accuracy ideally below the 1 kcal/mol gold standard. But how does one judge the accuracy of PES with no experimental equivalent being readily available? Instead, computationally demanding high-level methods like coupled-cluster or configuration-interaction calculations are used in computational chemistry to evaluate the performance of computationally-feasible more approximate methods.

An alternative approach is the comparison of bulk/thermodynamic quantities observed in an experiment to a properties derived from simulations that sample the free-energy surface (FES) of molecular systems. This is a binding paradigm in, e.g., the parametrization of empirical force fields.[48] Thermodynamic quantities are "mapped back" from the ensemble/time averages onto detailed molecular dynamics. Methods parametrized to experimental values are usually specific to a given environment, whereas transferability for example, to the gas phase or less polar interior of proteins cannot be guaranteed. FES-based parametrization does not aim at reproducing minute details, but tries to include implicitly other contributions, for instance solvation or entropic effects.[43] These aspects should be taken into account when comparing empirical force fields to high-level first-principles methods. When FFs are used slightly outside their range of validity, it is the underlying PES that determines the appropriateness of a method. Such a "mapping forward" can only work reliably from as exact as possible potential-energy function.

Systems of the size of glucose lie within the reach of the so-called "gold standard" method, the coupled-cluster method with single, double, and perturbative triple excitations (CCSD(T)).[49] The application of CCSD(T) to larger systems is hindered by the N^7-scaling of the computational cost with the systems size N and by the slow convergence of the electronic correlation energy with the basis set size.[50] To circumvent this limitation, composite

treatments and techniques using localized molecular orbitals are commonly employed.

First, composite schemes combine the accuracy of CCSD(T) and tractability of the better-scaling MP2 method:

$$E^{\texttt{large basis}}_{\texttt{CCSD(T)}} = E^{\texttt{large basis}}_{\texttt{MP2}} + \delta^{\texttt{small basis}}_{\texttt{CCSD(T)-MP2}}. \tag{1}$$

The `large basis` set calculations performed at MP2 level are routinely extrapolated to the complete basis set limit (CBS); then the CBS energy is corrected by the difference between MP2 and CCSD(T) correlations energies computed in a `small basis` set. This focal-point extrapolation relies on the difference between MP2 and CCSD(T) correlation energies, which converges faster to a constant value than the absolute correlation energy;[51–53] although some error cancellation in small basis sets still contributes to the extrapolated energies.[54] Therefore, the approximation requires careful examination with respect to the larger basis set for producing benchmark quality data.

As an alternative approach, many efforts have been made to reduce computational costs of the CCSD(T) method itself, particularly by developing approximations which exploit the inherent locality of the electron correlation, see e.g.[55,56] and references therein. In this study we use the domain-based local pair natural orbital (DLPNO-)CCSD(T) method developed by Riplinger and Neese.[57,58] It presents the correlation energy as a sum over electron pair correlation energies of localized orbitals derived from a single determinant reference wave function. This allows for efficient screening and selection of the most significant excitations which will contribute to the correlation energy at the CCSD(T) level whereas remaining pairs enter correlation energy at MP2 level or will be neglected. Then, the method uses pair natural orbitals[59,60] to localize electron pairs into specific domains expanded in terms of projected atomic orbitals. Within the DLPNO-CCSD(T) framework, three different cut-off parameters tune the method's proximity to canonical CCSD(T). Therefore, the careful consideration of the error imposed by the parameters, which control the pair correlation

energy cutoff, the size of the virtual space, and the size of single domains, allows to calculate benchmark-quality energies for systems too large for conventional CCSD(T) calculations.[55,56] The accuracy of the DLPNO-CCSD(T) has been tested with four benchmark sets (chemical reactions,[61] S66 non-covalent[15,16] interactions and two conformational hierarchies of small biomolecules[62,63]) demonstrating that it produces reliable energies with `normal` settings.[55] Also, DLPNO-CCSD(T) performed with high accuracy for the different hydrogen-bonded conformers of butano-1,4-diol.[62] The similarity of this molecule to the carbohydrates suggests DLPNO-CCSD(T) as a suitable reference method for carbohydrates.

Methods

Benchmark set

Conformational sampling

The conformational space of glucose isomers and α-maltose was investigated on the first-principles level employing the search tool Fafoom[64,65] interfaced with the electronic structure code FHI-aims.[66] The genetic algorithm (GA) based search samples the significant molecular degrees of freedom (DOFs) over a number of generations in order to minimize a fitness function, here the electronic potential energy. By resorting to the electronic potential energy, we avoid a plausible bias of empirical FFs towards chair structures. We have implemented ring puckering as an additional DOF available for six-membered rings. A list of internal degrees of freedom (angles and dihedrals) of 38 idealized canonical puckers has been used as a template.[67] A ring mutation in GA sampling occurs stepwise: first, the anomeric carbon - ring oxygen bond is dissected; then, the respective dihedrals and angles are adjusted to values in the selected canonical pucker (plus one improper dihedral to maintain stereochemistry at the anomeric carbon) and the dissected oxygen-carbon bond is restored. Five C-O and one C-C (between carbons C5 and C6) rotatable bonds, and a ring pucker conformation were

selected as DOFs for α- and β-glucose (see Figure 2). In open-chain glucose we selected 10 (five C-C and five C-O bonds) torsional angles. For α-maltose the list of available DOFs was extended by ϕ/ψ angles of α-Glc(1→4) glycosidic bond. The detailed settings of the GA runs can be found in the Supporting Information.

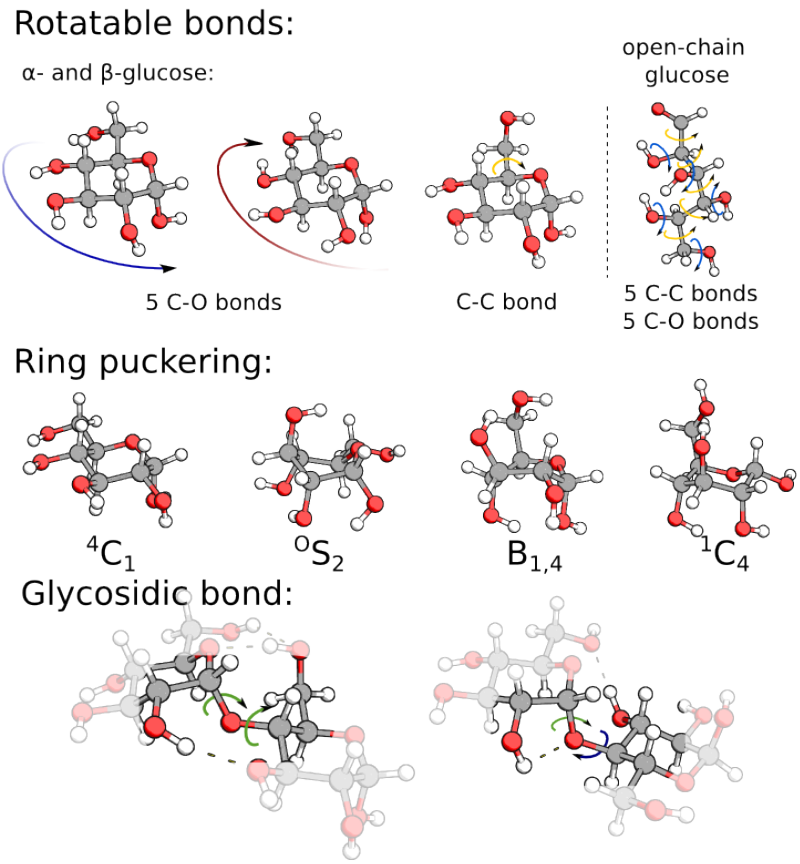

Figure 2: Selected degrees of freedom: five C-O and one C-C rotatable bonds, few examples of ring puckering and two different orientation of the glycosidic bond. The C-O and C-C bonds govern the hydrogen bonding, the ring puckering arranges oxygen atoms around the ring and two dihedral angles in the gycosidic bond orients two rings relative to each other.

Selection of reference structures

The GA search yielded 429 and 476 unique conformers of α and β-glucose respectively. Each conformer has been assigned to a respective ring pucker according to its Cremer-Pople coordinates[38] using a script adapted from Reilly *et al.*[67] The sampling yielded multiple

chair, boat and skew-boat conformations. Envelopes and half-chairs, although included in the list of possible ring mutations, were not found among the stable minima. This absence of envelopes and half-chairs agrees with the previous study by Beckham *et al.*[30] From the available minima, we selected a subset of 230 structures that were refined on a higher level of theory. We chose 15 lowest energy conformers of each $^4C_1$ and $^1C_4$ ring puckers (30 structures) and the 5 lowest energy conformers of each of six boat and six skew boat ring puckers (60 structures). If fewer than 5 conformations in boat or skew boat were found, extra structures were selected from the neighbors following the Cremer-Pople ϕ coordinate.[67] Furthermore, we selected 50 lowest energy conformers out of 570 unique structures for the open-chain glucose. The structures were subsequently optimized at MP2/def2-SVP and MP2/def2-TZVPP levels of theory employing ORCA program package.[68] Duplicate removal with a strict similarity criterion, root mean square deviation (RMSD) of less or equal 0.1 Å, yielded 80, 76, and 49 unique minima of α, β and open-chain glucose, respectively. The resulting 205 structures span an energy window of roughly 15 kcal/mol above the global minimum.

α-Maltose, which consists of two pyranose rings, comprises a much larger conformational space due to the 36×36 possible ring pucker combinations as well as due to the flexible glycosidic bond. The GA search (details can be found in the Supporting Information) yields 2,092 unique conformers: 592 $^4C_1$-$^4C_1$, 259 $^1C_4$-$^4C_1$, 181 $^4C_1$-$^1C_4$, and 43 $^1C_4$-$^1C_4$ ring puckers, whereas the remaining 1,017 conformers featured at least one building block in a non-chair (neither $^4C_1$ nor $^1C_4$) conformation. The GA also sampled the α-Glc(1-4) glycosidic bond in diverse geometries. Among all structures, we selected a subset according to the following workflow:

1. We set a limit of 230 in total, 25 conformers for chair-pucker pairs, and 10 conformers if one of the rings is a non-chair conformation.

2. The conformers were sorted according to their relative energy.

3. Starting from the lowest-energy structure, the next higher-in-energy conformer was added to the benchmark set if: (1) the count of the respective ring pucker combination is below the limit and (2) the torsional RMSD of the glycosidic bond differs by more than 5 degrees from other conformers with the same ring puckering combination already included in the benchmark set.

This workflow generated a set composed of highly diversified structures of maltose that span an energy window of about 15 kcal/mol above the global minimum. The structures consist of 72-two chair puckers and only 7 structures in which both pyranose rings assumed non-chair conformation. The remaining 151 conformations represent combinations of chair and non-chair puckers. The 230 structures were optimized according to the same procedure as in glucose and yielded 223 unique minima at (resolution of identity) MP2 in def2-TZVPP basis set.

Energy functions

Wave-function theory

Benchmark calculations are based on wave-function theory (WFT) methods, MP2,[69] CCSD(T),[49] and DLPNO-CCSD(T),[57,58] that were employed via the ORCA code.[68] Hartree-Fock (HF) energies were extracted from DLPNO-CCSD(T) calculations. Refinement of the glucose geometries for the benchmark calculations (mentioned above) were carried out at the MP2 level with Ahlrichs' def2-SVP and def2-TZVPP basis sets.[70] Resolution of the identity (RI) MP2[71] relaxations for α-maltose employed auxiliary basis sets (def2-SVP/C, def2-TZVPP/C, def2-QZVPP/C). We tested that RI-MP2 and MP2 produce essentially indistinguishable results for glucose. The DLPNO-CCSD(T) calculations were carried out with `normal` cutoff settings.

We employed two basis set extrapolation schemes to reach the complete basis set (CBS) limit[72–75] as implemented in the ORCA program.[76] CBS(2,3) refers to the extrapolation

using smaller def2-SVP/def2-TZVP basis sets, while CBS(3,4) refers to the extrapolation using larger def2-TZVPP/def2-QZVPP basis sets.

Empirical force fields

The single-point energy evaluation for GLYCAM06[41] and CHARMM36[42] FFs have been performed with Gromacs-5.0.5.[77] We limit our evaluations to these two out-of-the-box force fields as representative cases of commonly used empirical energy functions. No distance cutoffs were used for evaluating the long range contributions. Because connectivity of atoms between open and closed chain forms of glucose differs, their total energies are incomparable on the FF level. Consequently, we use only a subset of α-glucose and β-glucose to benchmark the FF energies. It should be also noted that FFs use the same set of parameters to describe anomeric carbon in α- and β-glucose, which adds a source of possible errors as it denies the anomeric effect to impact relative stability of two enantiomers.

Semi-empirical quantum mechanics

Semi-empirical electronic structure methods follow a simplification strategy, many of the computationally demanding terms of the underlying first principles formalisms are replaced by empirical formulations. The accuracy is then tuned by fitting parameters to experimental data or high-level calculations.[78] SQMs that we test here are all implemented in the MOPAC code.[79,80] We test approximations/parametrizations that belong to the family of neglect of diatomic differential overlap (NDDO) methods. AM1,[81] PM3,[82] and PM6[83] were demonstrated to fail to reproduce inter- and intra-molecular molecular non-bonded interactions like van der Waals effects or hydrogen bonding.[84] As a consequence, different correction schemes for PM6 or the new parametrization PM7[85] were developed. Empirical dispersion corrections are discussed below. In addition to potential energy, SQMs yield heats of formation of molecules which we use for the comparison. By definition, the value is composed of electronic energy, nuclear-nuclear repulsion energy, ionization energy for the valence electrons and the

heat of atomization for all atoms in the system.[86] Where applicable, also the dispersion corrections are included at this stage. The use of heats of formation instead of total energies introduces a systematic shift of the energies, which is accounted for in calculations of MAE and ME values.

Density-functional tight-binding methods

Besides approximate WFT methods like the right above described SQM methods, we also test here an approximate method based on DFT. Self-consistent-charge density-functional tight binding (SCC-DFTB) alleviates some short-comings of standard DFTB for the description of molecular systems with intermediate intramolecular charge transfer and is as such well applicable to bio-organic molecules. Here we test the most recent version, DFTB3,[87] with the parametrization for organic and biological systems (3OB)[88] by using the software DFTB+.[89] In addition, also Grimme's D3 dispersion correction was used.[90,91]

Density-functional approximations

We compare a representative set of density-functional approximations (DFAs) that span several rungs of what John Perdew calls the Jacob's ladder of DFT:[92]

- Generalized gradient approximation (GGA) and meta-GGA DFAs are of particular interest as they are computationally feasible, which allows their application for *ab initio* molecular dynamics simulations, large-scale hybrid quantum mechanics/molecular mechanics simulations, or high-throughput screenings. We here test the GGA exchange-correlation functionals PBE[93] and BLYP[94,95] and the meta-GGAs M06-L,[96] M11-L,[97] and SCAN.[98,99]

- Hybrid functionals incorporate exact Hartree-Fock exchange and by that, by the means of adiabatic connection,[100] promise higher accuracy at a higher computational cost. We selected B3LYP,[18] PBE0,[101] M06,[102] M06-2X,[102] M06-HF,[103] M08-SO, M08-HX,[104] and M11.[105]

- The double-hybrid functionals include, in a manner similar to exact exchange in hybrids, a fraction of the correlation energy usually calculated using MP2 or random phase approximation (RPA) to improve deficiencies in the correlation functional.[106] As an example of this group, we chose XYG3 functional[107] as it has shown excellent performance for non-bonded biomolecular interactions.[108]

Density-functional calculations were performed with the all-electron numeric atom-centered orbitals code FHI-aims.[66] The initial GA-based search with Fafoom was carried out using the PBE functional augmented with MBD dispersion model in a *tier-1* basis set,[66] and *light* computational settings. Single-point energy evaluations of the benchmark geometries, except for the double-hybrid XYG3, were carried out using `really_tight` settings for basis set and integration grids. The resolution of identity was used for evaluation of the four-center Coulomb integrals.[106] For the XYG3 functional, the numeric correlation-consistent triple-ζ basis set NAO-VCC-3Z was used.[109,110] The NAO-VCC family of basis sets behaves like the numerical counterpart of Dunning's correlation-consistent family of basis sets.[111] The correlation energy calculations were performed using orbitals from the B3LYP functional.

Van der Waals correction schemes

Van der Waals (vdW) interactions are long-range correlation phenomena and require non-local treatment, whereas most of commonly used correlation functionals are intrinsically (semi-) local, i.e. they include local electron density (and its gradients) but lack explicit dependence on the non-local contributions.[112] Some functionals, like double hybrids or vdW functionals by Vydrov and Van Voorhis,[113] incorporate such non-local correlation in their design, but they are significantly more computationally demanding. More feasible solutions rely on the *a posteriori* addition of dispersion-correction schemes.[112,114]

In general, dispersion-corrections are crucial for dispersion-dominated systems and systematically improve the performance of GGA and hybrid functionals for biomolecules.[13,115–119] Nevertheless in some cases they might overestimate diaxial interactions in di-substituted cy-

clohexane derivatives,[120] similar to these present in some conformations of sugars.

In this study, we examine the performance of three dispersion-correction models:

- Grimme's pairwise additive D3[90] model combined with zero-damping function and corrections from three-body interactions. It includes wide parametrization for different class of atoms depending on the local environment.

- Pairwise additive dispersion-correction by Tkatchenko and Scheffler (vdW^{TS})[121] which includes dependence of C_6 coefficients on the electron density. It uses tabulated van der Waals radii, polarizabilities and C_6 coefficients of free atoms which are then rescaled according to the effective atomic volumes derived from Hirshfeld partitioning in the molecular environment.

- Many-body dispersion (MBD) by Tkatchenko *et al.*[122] (sometimes called MBD@rsSCS) computes dispersion interactions using a coupled quantum harmonic oscillator model, which goes beyond two-body contributions. As a result, predictions with the MBD model can differ substantially from pairwise-additive schemes.[118,123]

The dispersion corrections vdW^{TS} and MBD are implemented in FHI-aims.[66] The D3 dispersion corrections were calculated with the DFT-D3 program from Grimme's group webpage.[91] Only corrections combined with functionals for which the parametrization was readily available were used in this study.

The problem of dispersion interactions is not unique to density-functional approximations, but also plagues other methods and requires separate treatment. For the SQM method PM6, we consider Grimme's D3 dispersion and hydrogen bonding correction (keyword `PM6-D3`).[90] Hobza and co-workers started to develop empirical corrections for semi-empirical methods. Already its first incarnation (PM6-DH) reduced the MAE w.r.t. the S22 data set from 3.17 kcal/mol for PM6 to 0.54 kcal/mol for PM6-DH.[84] The advanced version `PM6-D3H4` further adds a hydrogen bond correction developed by Řezáč and Hobza[124,125] that

was successfully used, for example in the context of predicting the energetics of protein-ligand complexes.[126]

Mean-absolute and maximum error

We refrain from using least-square fit to quantify an error between two data sets due to its overemphasis of distant data points. Instead, we resort to mean absolute error (MAE) and maximum error (ME):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\Delta E(A)_i - \Delta E(B)_i + b| \tag{2}$$

$$\mathrm{ME} = \max_{i \in N} |\Delta E(A)_i - \Delta E(B)_i + b| \tag{3}$$

where $\Delta E(A)_i$ stands for relative energies calculated by the reference method, while $\Delta E(B)_i$ is calculated by the method being evaluated. N is the overall number of conformers and i represents a specific conformer in the data set. We uniformly shift the energies evaluated by method B by a constant value b to minimize the MAE. This shift removes spurious dependencies on arbitrary reference points, for instance the lowest energy conformer. An exemplary correlation plot is shown in Figure 3.

Results

Validation of DLPNO-CCSD(T) as reference method for carbohydrates

First, we test the applicability of DLPNO-CCSD(T) as a benchmark method for carbohydrates. To that end we performed two tests on the 205 glucose structures:

- The comparison of DLPNO-CCSD(T) (`normal` settings) against canonical CCSD(T) calculations using CBS(2,3) extrapolation yields MAE of 0.25 kcal/mol and ME of

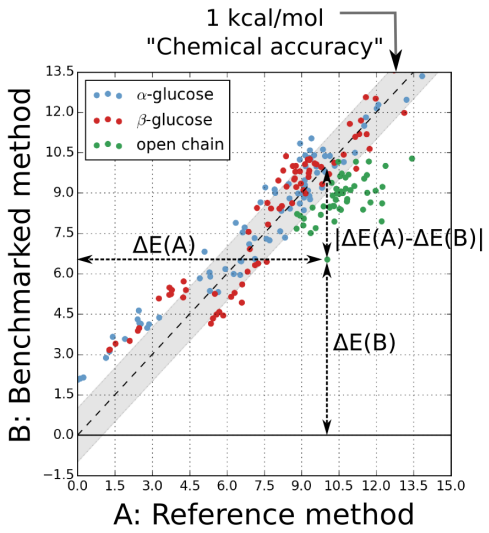

Figure 3: Example of a correlation plot used in this study. The energy scale is given in kcal/mol. On the x axis we show conformational hierarchy using reference method, on the y axis the method that is benchmarked. All ΔE(B)'s are shifted by a constant value b (see Equations 2 and 3) to minimize MAE between two data sets. If the match between reference and benchmarked method is perfect, the points should align on the dashed diagonal line. The gray shading highlights the region of so-called 'chemical accuracy' of 1 kcal/mol. Each color codes one isomer of glucose.

1.06 kcal/mol (see Figure 4a). The error compares to MAE of 0.07 kcal/mol which was found for the roughly three times smaller 1,4-butanediol.[55] Tighter DLPNO settings would further decrease the MAE, but are intractable for longer saccharides.

- We also compare the conformational energy hierarchies of DLPNO-CCSD(T)/CBS(3,4) against the focal-point extrapolated MP2/CBS(3,4)+$\delta^{\texttt{def2-TZVP}}_{\Delta E(\texttt{CCSD(T)-MP2})}$ energies. The observed deviations (MAE of 0.23 kcal/mol and ME of 1.06 kcal/mol, Figure 4b) are again small.

These consistency checks confirm the applicability of DLPNO-CCSD(T)/CBS(3,4) in `normal` settings as a benchmark-level method for carbohydrates. Hence, hereafter we will refer to DLPNO-CCSD(T)/CBS(3,4) energies as our benchmark values for glucose and α-maltose conformational hierarchies.

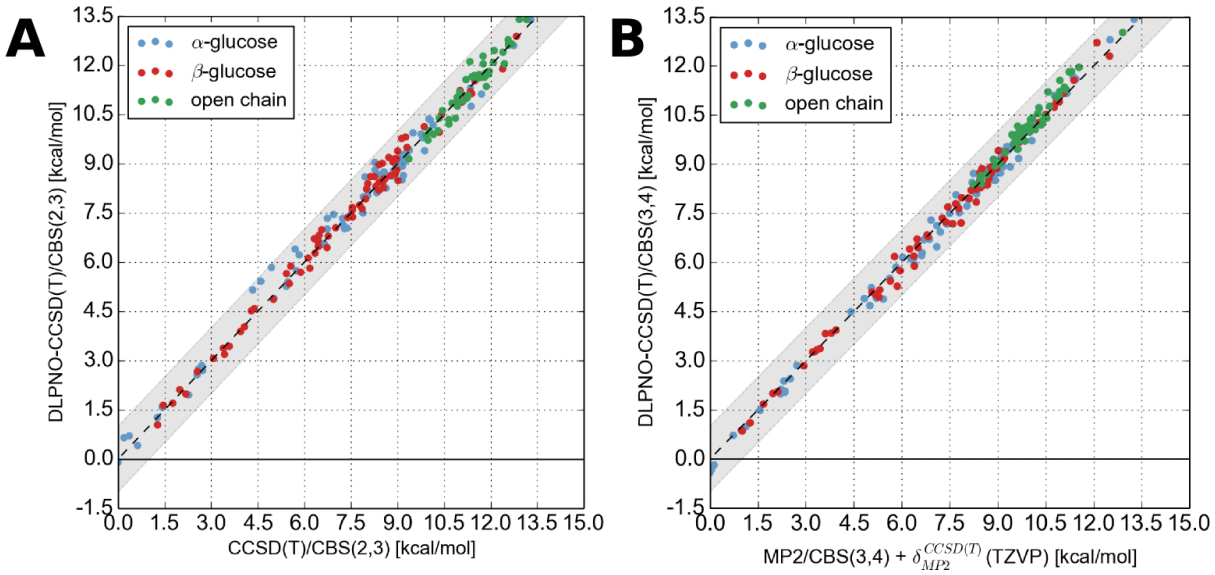

Figure 4: (**A**) The correlation between canonical CCSD(T)/CBS(2,3) and DLPNO-CCSD(T)/CBS(2,3) energies. (**B**) The correlation between MP2/CBS(3,4)+$\delta^{\texttt{def2-TZVP}}_{\Delta E(\texttt{CCSD(T)-MP2})}$ and DLPNO-CCSD(T)/CBS(3,4) energies.

Benchmarks for glucose

Exemplary correlation plots for a selection of methods that belong to different approximations are shown in Figure 5 and MAE and ME errors for all tested methods are summarized in Figure 6. The remaining correlation plots are supplied in the Supporting Information. In addition, we evaluate the accuracy using only a subset of closed-ring structures, namely conformers of α and β-glucose (156 conformations) to remove the impact of the mutarotation reaction on the calculated MAEs.

Force fields

Two tested force fields, CHARMM36 and GLYCAM06, demonstrate rather similar behavior. CHARMM36, yields MAE of 2.08 kcal/mol, whereas GLYCAM06 has higher MAE of 2.58 kcal/mol. The MEs are equal 5.88 kcal/mol for CHARMM36 and 8.14 kcal/mol for GLYCAM06. The inspection of the correlation plots (see Figure 5) indicates that both FFs are somewhat accurate in the low-energy regime, populated solely by $^4C_1$ ring puckers, but

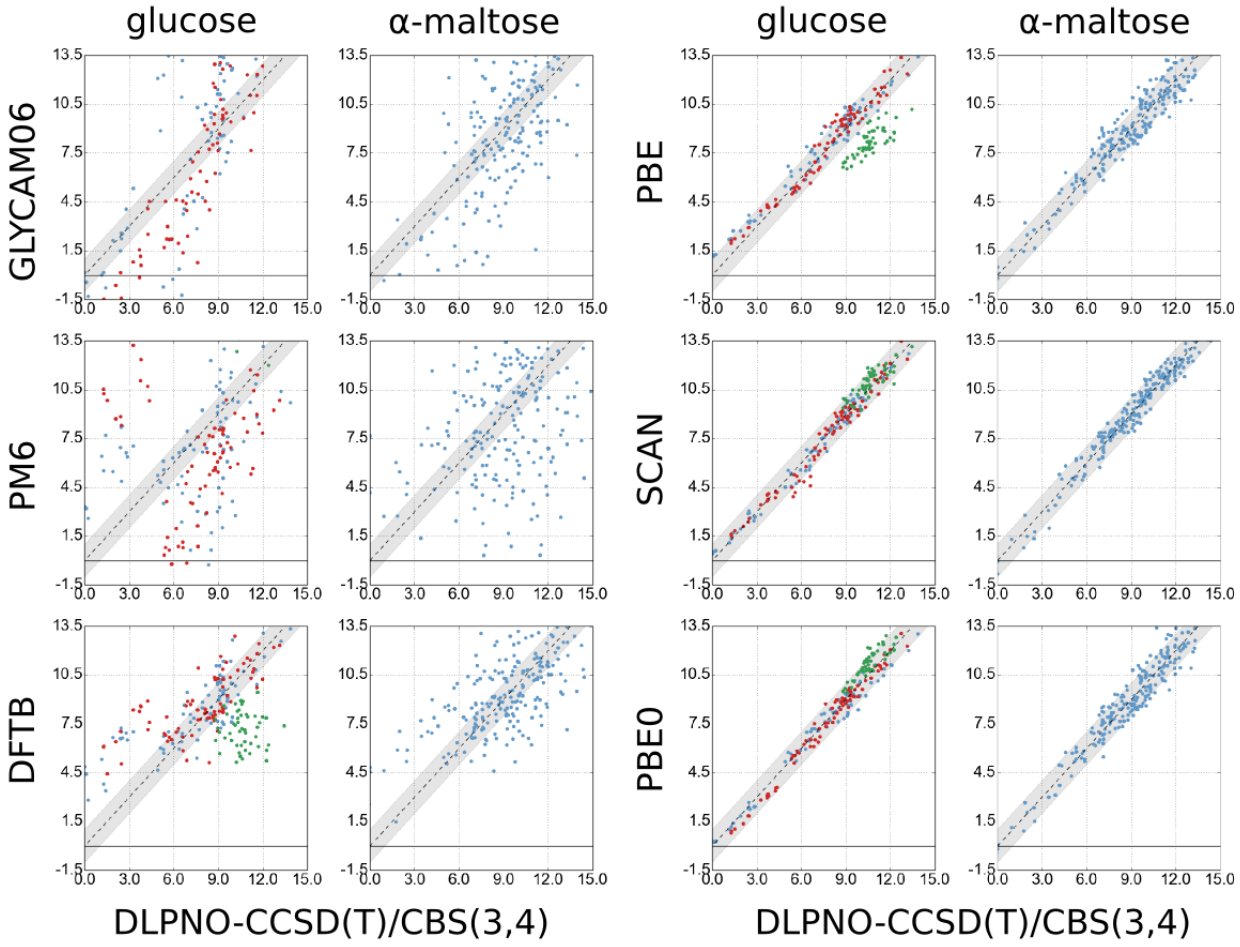

Figure 5: Correlation between conformational hierarchies of DLPNO-CCSD(T)/CBS(3,4) (x axis) and different approximate methods (y axis). In the glucose plots, blue dots represent α-glucose, red dots the β anomer and green dots belong to the open-chain isomer. The relative energies (in kcal/mol) are calculated to the with respect to DLPNO-CCSD(T) lowest energy conformer. Please note that some dots might lay outside of depicted range on y axis.

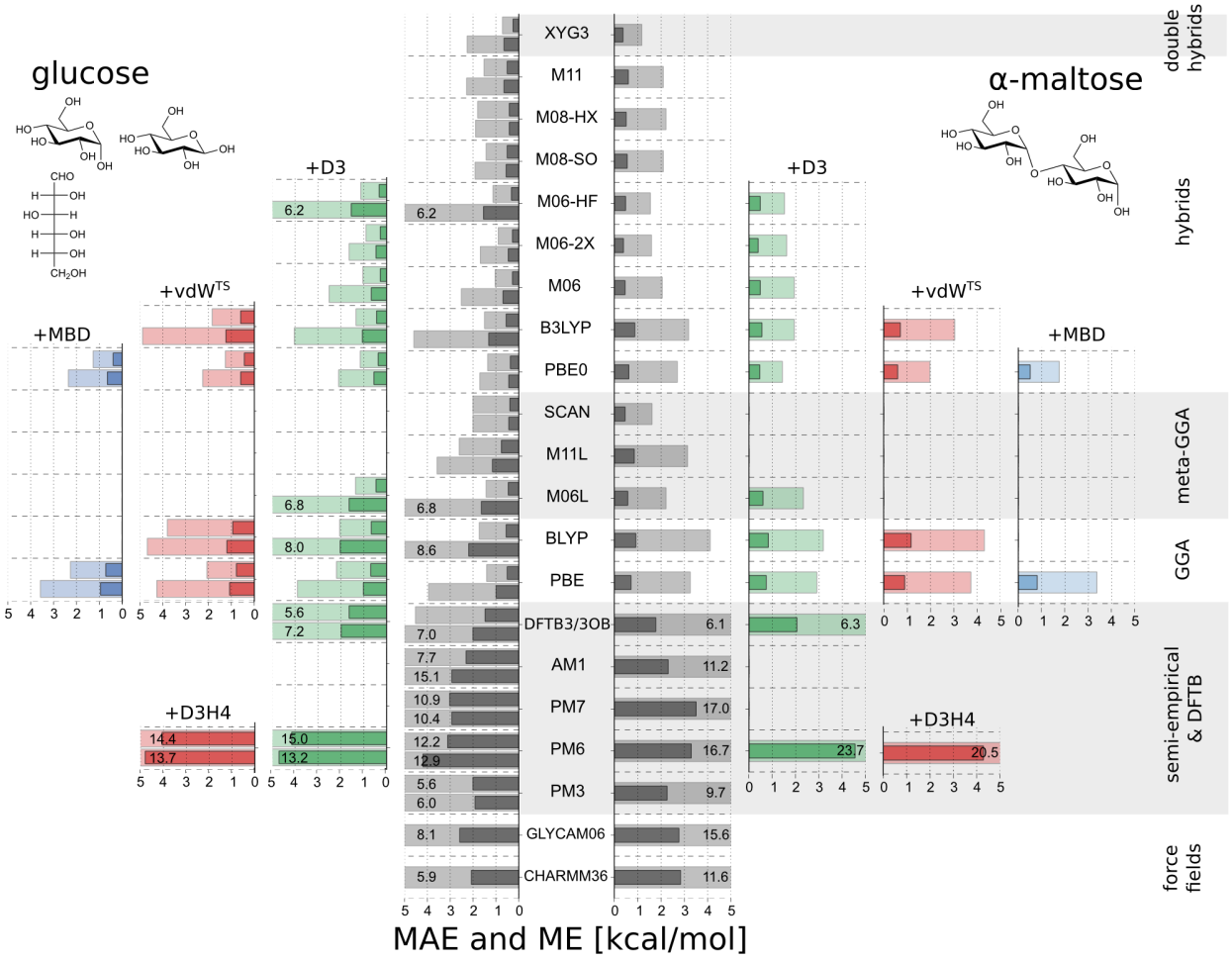

Figure 6: Calculated mean absolute errors (MAE, darker bars) and maximal errors (ME, lighter bars) for the complete benchmark set of glucose (left panel) and maltose (right panel). For glucose, we present MAE with (lower bar) or without (upper bar) open-chain glucose. If the ME exceeded maximum on the x-axis scale, the numerical value is given next to the bar.

they severely underestimate the stability of higher-energy non-chair conformations. The error originates from the parametrization derived from *ab initio* calculations of sugars in $^4C_1$ ring pucker conformations[41,42] and the inability of the model to respond to altered oxygen positions and reorganized hydrogen bonds in alternative ring puckers. Despite the apparent accuracy for the low-energy conformers, such underestimation of stability of non-chair conformations will affect simulations of long crowded carbohydrates or of carbohydrates in environments that pose sterical hindrance, *e.g.* enzymatic reaction centers, or when long-range effects can drive non-terminal sugars to non-chair ring puckers. In such cases the non-chair ring puckers might become energetically favorable, which would not be reflected in FF simulations.

Finally, it is interesting to investigate the HF energies, since this level of theory is still commonly employed in the parametrization of force fields.[41,42,127] The calculated MAE for the entire data set is equal 1.74 kcal/mol and increases to 1.85 kcal/mol for the subset of closed ring structures (Figure 1 in SI). The calculated ME are 6.60 kcal/mol and 5.99 kcal/mol, respectively. These errors are only somewhat smaller than the error of the force fields which should question the applicability of HF method to derive the reference energetics.

Semi-empirical quantum mechanics

We observe that SQM methods perform very poorly. Among the analyzed methods (AM1, PM3, PM6, PM6+D3, PM6+D3H4 and PM7), only PM3 performs with a MAE smaller than 2 kcal/mol. The method is followed by AM1 and PM7 both of which output MAEs equal to 2.93 kcal/mol. PM6 generates large MAE of 4.20 kcal/mol (Figure 5), whereas its two dispersion-corrected counterparts, PM6+D3 and PM6+D3H4, perform even worse, with MAE of 4.70 kcal/mol and 4.80 kcal/mol respectively. Moreover, among the examined methods only PM3 yields ME below 10 kcal/mol (6.01 kcal/mol). Restricting the the comparison to the subset of closed-ring conformers does in some cases even increase the MAE. In general, the investigated SQMs are incapable to reproduce even an approximate hierarchy of

carbohydrate ring puckers: most of them (except AM1) incorrectly predict $^1C_4$ ring pucker as the lowest energy ring pucker.

Density-functional tight binding

Comparable in construction to semi-empirical methods, density-functional tight-binding (here: DFTB3 with the 3OB parameters and D3 dispersion correction) method shows somewhat better performance. Likewise FFs, DFTB3 models the low-energy spectrum of the conformational energy hierarchy accurately, but, contrary to FFs, it overestimates stability of higher-energy non-chair conformers. In addition, it overemphasizes stability of the open-chain glucose. When one excludes open-chain isomer from the benchmark, the DFTB yields a MAE equal 1.61 kcal/mol and a ME of 4.04 kcal/mol. However, including the open chain form increases the MAE to 2.61 kcal/mol and ME to 8.35 kcal/mol.

Density-functional approximations

The GGA functionals PBE and BLYP surpass the accuracy of the empirical and semi-empirical methods. The computed MAEs are 1.00 kcal/mol for PBE (see Figure 5) and 2.20 kcal/mol for BLYP, whereas respective MEs reach 3.94 kcal/mol and 8.58 kcal/mol. The MAE errors computed for a subset of closed ring conformers decrease to 0.51 kcal/mol for PBE and 0.55 kcal/mol for BLYP and the respective MEs decrease to 1.41 kcal/mol and 1.73 kcal/mol. This improvement in accuracy after removing the open-chain isomer from the benchmark set was anticipated due to the well-known self-interaction error.[24,128–130] Next, we test both functionals augmented with dispersion-correction schemes designed to improve the treatment of long-range correlation, *i.e.* van der Waals interactions: vdW^{TS}, MBD, and D3 (where parametrization is available). We observe dispersion-corrections to have negligible net effect on the performance of the PBE functional. However, when tested only on the closed ring subset, they increase the PBE's MAE to 0.71 (vdW^{TS}), 0.78 (MBD) and 0.70 kcal/mol (D3). Moreover, the vdW^{TS} significantly improves the relative energetics of the open- and

closed-chain forms of glucose with BLYP functional. The β-glucose in $^1C_4$ ring conformation is especially problematic for functionals studied here with predicted MEs varying between 2.08 - 2.26 kcal/mol.

The meta-GGA functionals M06L and M11L perform well. For the conformational energy hierarchy of only the closed-ring pyranoses, M06L yields a MAE of 0.46 kcal/mol (ME of 1.43 kcal/mol) while the more recent M11L performs slightly worse with a MAE of 0.76 kcal/mol (ME of 2.61 kcal/mol). The M06L functional combined with D3 dispersion correction performs marginally better than the non-corrected one. However, similarly to the GGA's, it is the relative stability of the closed and open-chain forms of glucose that pose a major challenge for accurate energy description. Evaluation of the functionals on the complete data set increases the MAE for M06L to 1.63 kcal/mol and to 1.15 kcal/mol for M11L (MEs increase to 6.85 kcal/mol and 3.57 kcal/mol respectively). The novel non-empirical meta-GGA functional SCAN[99] behaves superior to the two Minnesota meta-GGAs, yielding a MAE of 0.44 kcal/mol and ME of 2.01 kcal/mol on the complete data set. We observe that SCAN performs accurately not only for ranking different hydrogen-bonding patterns in case of different ring puckers, but also correctly predicts the relative stability of the open-chain form of glucose (Figure 5).

The hybrid functionals PBE0 (sometimes referred to as PBE1PBE) and B3LYP predict the conformational energy hierarchies of the benchmark data set for glucose with MAEs of 0.44 kcal/mol and 1.31 kcal/mol, respectively (see Figure 5 for the PBE0 correlation plot), and the consideration of the pyranose rings subset decreases the MAE of PBE0 to 0.37 kcal/mol and the MAE of B3LYP to 0.55 kcal/mol. The difference between the accuracy of these two methods originates from an overestimation of the open-chain form by the underlying BLYP GGA functional, which is only partially corrected for by the exact-exchange energy term in the hybrid functional B3LYP form. On the other hand, PBE0 does slightly underestimate the stability of the open-chain form. Because PBE tends to overestimate the open-chain isomer, whereas PBE0 slightly overestimates it, we decided to test

intermediates value of the mixing parameter α, which selects the portion of the exact exchange that should be included in the density-functional, to minimize the MAE. The MAEs were computed for α's between 0.00 (which corresponds to plain PBE GGA) and 0.25 (as in PBE0 hybrid functional) in 0.05 steps. We observed that different fractions of the exact exchange change the relative energetics of the two forms and α equal 0.20 yields a minimal values of both MAE (0.37 kcal/mol) and ME (1.66 kcal/mol) among tested values (remaining errors and correlation plots are included in the SI). Finally, combining PBE0 with any of the dispersion-correction schemes causes the MAE to increase by 0.1 - 0.2 kcal/mol, largely due to additional over-stabilization of the open-chain conformers.

We investigate further hybrid functionals from the Minnesota family: M06, M06-2X, M06-HF, M08-SO, M08-HX, and M11. Since they are trained on a generous set of benchmark data that especially emphasizes non-bonded interactions,[102,104] their excellent performance is somehow anticipated. Two of these functionals yield MAE smaller than 0.5 kcal/mol: M06-2X with an MAE of 0.46 kcal/mol (ME 1.68 kcal/mol) and M08-HX with an MAE of 0.42 kcal/mol (ME 1.89 kcal/mol). The predicted MAEs decrease to 0.28 kcal/mol for M06-2X and 0.32 kcal/mol for M08-HX when evaluated only for the closed-ring structures. Only the M06-HF functional, that was designed to treat systems with significant charge transfers by including 100% exact exchange,[103] predicts with an error of 1.55 kcal/mol that is above chemical accuracy. Particularly, it suffers from an underestimation of the stability of open-chain glucose; limiting the benchmark set to the pyranose rings decreases the MAE to 0.32 kcal/mol. Since the Minnesota functionals have been parametrized to implicitly account for non-covalent interactions in the functional form, addition of the dispersion interactions via the D3 scheme has negligible effect on the calculated MAEs.[131,132]

Finally, we examine the performance of the double-hybrid functional XYG3. The predicted error of 0.65 kcal/mol arises mostly from overestimation of the open-chain form and decreases to 0.25 kcal/mol for the closed ring subset. Whereas XYG3 is the best performer in the restricted set, five functionals (meta-GGA SCAN and four hybdrids) produce smaller

MAEs when the open-chain form is included in the benchmark set.

Benchmarks for α maltose

The two carbohydrate units in maltose vastly increase the complexity of the molecular conformational space. Therefore, we neglect two possible anomeric states and focus on diversified ring puckering and glycosidic bond orientations of α-maltose. While we expect that the restriction to only one anomer should decrease the errors, α maltose is twice the size of glucose hence the errors already observed in the mono-saccharide will accumulate. Selected correlation plots are shown in Figure 5 and the MAE and ME values are summarized in the plots in Figure 6.

Force fields

The two investigated force fields, CHARMM36 and GLYCAM06, perform worse than for glucose, producing MAEs of 2.84 kcal/mol and 2.77 kcal/mol and MEs over 10 kcal/mol (11.55 kcal/mol and 15.61 kcal/mol). The HF method also experiences larger MAE than for monosaccharides, 1.99 kcal/mol, and very large ME error of 9.09 kcal/mol. While CHARMM36 maintains some accuracy in the low-energy regime (dominated by $^4C_1$ ring puckers combinations) and drastically underestimates the stability of higher-energy non-chair puckers, GLYCAM06 shows unsatisfying performance in the entire energy window.

SQM and DFTB methods

The semi-empirical methods show similarly dismal performance as already observed for glucose. Only older-generation methods, AM1 and PM3, yield MAEs below 2.5 kcal/mol, while the newer methods PM6 and PM7 as well as their dispersion-corrected counterparts yield MAEs above 3.00 kcal/mol, with the largest MAE of 4.54 kcal/mol for the PM6-D3 method. Similarly for the ME values: the lowest ME value of 9.66 kcal/mol was achieved by PM3, whereas PM6-D3 performs worst with an ME of 23.7 kcal/mol, which is even larger than

the approximate 15 kcal/mol energy window assumed for relative energies in the benchmark set. Only the DFTB3 calculations perform with almost acceptable accuracy with an MAE of 2.05 kcal/mol and ME of 6.30 kcal/mol. Again the method overestimates stability of non-chair puckers and incorrectly predicts a $^1C_4$-$B_{O,3}$ ring puckers combination to be the second lowest-energy structure.

Density-functional approximations

As we observed already for the glucose part of our benchmark set, the DFAs again perform with much better accuracy than the empirical and semi-empirical methods. With MAE values of 0.71 kcal/mol for PBE and 0.92 kcal/mol for BLYP, the two GGA functionals perform with better than chemical accuracy. The MEs are twice as large as those observed for the ring conformers of the glucose test set, 3.25 kcal/mol for PBE and 4.10 kcal/mol for BLYP. Interestingly, the MBD and vdW^{TS} dispersion corrections increase MAE and ME values. Only in case of BLYP-D3, the MAE decreases to 0.83 kcal/mol relative to the uncorrected BLYP functional.

Among the tested meta-GGA functionals, especially M06-L and SCAN should be highlighted due to their low MAEs of 0.58 kcal/mol and 0.46 kcal/mol that even outperform the more advanced hybrid functionals PBE0 (MAE = 0.62 kcal/mol) and B3LYP (MAE = 0.88 kcal/mol). Only M06 (MAE=0.46 kcal/mol), M06-2X (0.39 kcal/mol), and the double-hybrid XYG3 (0.37 kcal/mol) perform better than SCAN.

Dispersion corrections have the expected negligible impact for the Minnesota functionals, but lower the MAE of PBE0 and B3LYP. In particular the D3 correction reduces the MAE of PBE0-D3 to 0.46 kcal/mol and of B3LYP-D3 to 0.55 kcal/mol. The dispersion correction schemes vdW^{TS} and MBD show a smaller improvement of the MAE.

Conclusions

In this study we presented a benchmark data set of glucose and maltose conformations at the DLPNO-CCSD(T) level of theory as well as the assessment of the accuracy of several across-the-scale energy functions for reproducing the high-level energy data. Here we want to summarize our conclusions:

- By comparing DLPNO-CCSD(T) to canonical CCSD(T) and to focal-point extrapolated MP2+$\delta_{\texttt{CCSD(T)-MP2}}$ calculations we have demonstrated the applicability of this method for reference calculations on carbohydrate systems. Given the favorable, almost linear, scaling behavior that is described by the developers, even larger systems seem reachable.[56]

- When it comes to production methods, we have clearly seen the shortcomings of empirical FFs. Errors stem in particular from the inability of FFs to correctly rank ring puckers other than $^4C_1$. Relatively few of them have been designed to model carbohydrates, especially when compared to the plethora available for proteins. The presented new structure/energy data set can help to improve the parametrization of such empirical potentials.

- Semi-empirical quantum mechanics promises first-principles accuracy at only a fraction of its costs. We however see serious limitations, especially of recent SQM parametrizations.

- The evaluated DFTB3 parametrization 3OB augmented by Grimme's D3 dispersion correction performs comparably well, however the error remains higher than chemical accuracy.

- DFT methods offer a substantial improvement of accuracy. If only the energetics of closed-ring conformers are considered, the MAEs are below 1 kcal/mol and thus reach "chemical accuracy". Getting the difference between open-chain and closed-chain

conformers of glucose poses a challenge that is particularly well resolved by the GGA functional PBE, the meta-GGA SCAN, the hybrid functionals PBE0, M06, M06-2X, M08-SO, M08-HX, and M11, and the double-hybrid XYG3. The often used functionals BLYP and B3LYP do not match the accuracy of the aforementioned alternatives.

- We cannot observe a clear trend for the impact of dispersion corrections on the accuracy of SQM and DFTB since these approaches predict too large errors by themselves that cover the minuscule impact of the dispersion corrections. In case of DFAs, the errors are typically small already for the uncorrected functionals. This hints on a limited importance of long-range dispersion for carbohydrates of monomeric or dimeric size.

We compare conformational energy hierarchies derived from the PES of different methods that inherently differ in construction. The (semi)empirical energy functions are designed to yield thermodynamic quantities in specific environment. Nevertheless, we argue that observable quantities are derived from underlying physical model, PES ultimately, and any robust model should be in good agreement with accurate benchmark methods.

The choice of the right production method for large-scale sampling, molecular dynamics, or QM/MM simulation does of course depend on an efficient use of the available computing resources. While we refrain from comparing timings of very different approximations across different software implementations, convergence criteria, and computational setups, we want to comment on the timings of energy evaluations for α-glucose conformers using different DFA classes computed with FHI-aims. Assuming that a computation with the GGA PBE consumes one time unit, the calculations using meta-GGA required from 2.4 (SCAN) to 2.5 (M06-L) time units whereas hybrid functionals required from 15 (B3LYP) to 18 (Minnesota functionals), with the exception of M11 for which the self consistency cycle took 35 time units to converge. The timings, complemented with presented MAEs, argue again in favor of SCAN functional. For a sixth of the computational cost of and more efficient scaling than a hybrid functional, SCAN provides better than chemical accuracy when describing carbohydrates.

Overall, the DFT methods promise the desired 'chemical accuracy' of 1 kcal/mol. Hybrid DFT is suitable for the parametrization of empirical force fields, in stark contrast to HF, which shows a rather poor performance for carbohydrates. The computationally less demanding (meta)-GGA methods should be considered for production simulations like structure sampling, molecular dynamics, or as the quantum-mechanical part in QM/MM studies. DFT calculations remain computationally expensive, but the speed-up that we get with more approximate/empirical methods results in a limited accuracy that can only lead to false conclusions. Potential energy is the primary property we compute, errors that are made at this level can only pile up when deriving secondary properties like vibrational spectra, relative free energies, *etc.*

Supporting Information available

- Details of the GA structure searches for α, β, and open-chain glucose as well as for α maltose.
- All correlation plots of energy functions discussed here against the relative DLPNO-CCSD(T) energies.
- Cartesian coordinates of all conformers available in `xyz` format.
- Tables with the relative energies (or heats of formation) computed with the energy functions discussed here in `csv` format.
- Examples of input control files used with the simulation packages Gromacs, DFTB+, Mopac, ORCA and FHI-aims.

Acknowledgments

The authors are grateful to Matthias Scheffler (FHI Berlin) for continuous support of their work. CB acknowledges discussions with Jan Řezáč and Pavel Hobza (IOCB Prague) about semi-empirical calculations and accuracy. MM and CB are grateful to Pan Chen (KTH Stockholm) for discussions about carbohydrates in general and for force field parameters in particular.

References

(1) Carpita, N.; Gibeaut, D. *Plant J.* **1993**, *3*, 1–30.

(2) Dwek, R. A. *Chem. Rev.* **1996**, *96*, 683–720.

(3) Cosgrove, D. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 850–861.

(4) Daniels, R.; Kurowski, B.; Johnson, A. E.; Hebert, D. N. *Mol. Cell* **2003**, *11*, 79–90.

(5) Molinari, M. *Nat. Chem. Biol.* **2007**, *3*, 313–320.

(6) Varki, A. *Trends Mol. Med.* **2008**, *14*, 351–360.

(7) Hofmann, J.; Hahm, H. S.; Seeberger, P. H.; Pagel, K. *Nature* **2015**, *526*, 241–244.

(8) Struwe, W. B.; Baldauf, C.; Hofmann, J.; Rudd, P. M.; Pagel, K. *Chem. Commun.* **2016**, *52*, 12353–12356.

(9) Himmel, M. E.; Ding, S.-Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.

(10) Mascal, M.; Nikitin, E. B. *Angew. Chem. Int. Ed.* **2008**, *47*, 7924–7926.

(11) Bornscheuer, U.; Buchholz, K.; Seibel, J. *Angew. Chem. Int. Ed.* **2014**, *53*, 10876–10893.

(12) Chen, P.; Marianski, M.; Baldauf, C. *ACS Macro Lett.* **2016**, *5*, 50–54.

(13) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. *Sci. Data* **2016**, *3*, 160009.

(14) Ropo, M.; Blum, V.; Baldauf, C. *Sci. Rep.* **2016**, *6*, 35772.

(15) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(16) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

(17) Řezáč, J.; Hobza, P. *Chem. Rev.* **2016**, *116*, 5038–5071.

(18) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(19) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

(20) Hemmingsen, L.; Madsen, D. E.; Esbensen, A. L.; Olsen, L.; Engelsen, S. B. *Carbohydr. Res.* **2004**, *339*, 937–948.

(21) Stortz, C. A.; Johnson, G. P.; French, A. D.; Csonka, G. I. *Carbohydr. Res.* **2009**, *344*, 2217–2228.

(22) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. *J. Chem. Theory Comput.* **2009**, *5*, 679–692.

(23) Csonka, G. I.; Kaminsky, J. *J. Chem. Theory Comput.* **2011**, *7*, 988–997.

(24) Sameera, W. M. C.; Pantazis, D. A. *J. Chem. Theory Comput.* **2012**, *8*, 2630–2645.

(25) Govender, K. K.; Naidoo, K. J. *J. Chem. Theory Comput.* **2014**, *10*, 4708–4717.

(26) Neese, F.; Hansen, A.; Wennmohs, F.; Grimme, S. *Acc. Chem. Res.* **2009**, *42*, 641–648.

(27) Appell, M.; Strati, G.; Willett, J. L.; Momany, F. A. *Carbohydr. Res.* **2004**, *339*, 537–551.

(28) Momany, F. A.; Appell, M.; Willett, J. L.; Schnupf, U.; Bosma, W. B. *Carbohydr. Res.* **2006**, *341*, 525–537.

(29) Schnupf, U.; Willett, J. L.; Bosma, W. B.; Momany, F. A. *Carbohydr. Res.* **2007**, *342*, 196–216.

(30) Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. *J. Am. Chem. Soc.* **2014**, *136*, 1008–1022.

(31) Szczepaniak, M.; Moc, J. *J. Phys. Chem. A* **2014**, *118*, 7925–38.

(32) Szczepaniak, M.; Moc, J. *J. Phys. Chem. A* **2015**, *119*, 10946–10958.

(33) Çarçabal, P.; Jockusch, R. A.; Hünig, I.; Snoek, L. C.; Kroemer, R. T.; Davis, B. G.; Gamblin, D. P.; Compagnon, I.; Oomens, J.; Simons, J. P. *J. Am. Chem. Soc.* **2005**, *127*, 11414–11425.

(34) Brauer, B.; Pincu, M.; Buch, V.; Bar, I.; Simons, J. P.; Gerber, R. B. *J. Phys. Chem. A* **2011**, *115*, 5859–5872.

(35) Alonso, J. L.; Lozoya, M. A.; Peña, I.; López, J. C.; Cabezas, C.; Mata, S.; Blanco, S. *Chem. Sci.* **2014**, *5*, 515.

(36) Yuriev, E.; Farrugia, W.; Scott, A. M.; Ramsland, P. A. *Immunol Cell Biol* **2005**, *83*, 709–717.

(37) Werz, D. B.; Ranzinger, R.; Herget, S.; Adibekian, A.; Von der Lieth, C. W.; Seeberger, P. H. *ACS Chem. Biol.* **2007**, *2*, 685–691.

(38) Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.

(39) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R. *Org. Lett.* **2007**, *9*, 1851–1854.

(40) Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C. *Theor. Chem. Acc.* **2010**, *127*, 429–442.

(41) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655.

(42) Raman, E. P.; Guvench, O.; MacKerell, A. D. *J. Phys. Chem. B* **2010**, *114*, 12981–12994.

(43) Spiwok, V.; Králová, B.; Tvaroška, I. *Carbohydr. Res.* **2010**, *345*, 530–537.

(44) Biarnés, X.; Ardèvol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2007**, *129*, 10686–10693.

(45) Vitalini, F.; Mey, a. S. J. S.; Noé, F.; Keller, B. G. *J. Chem. Phys.* **2015**, *142*, 084101.

(46) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2014**, *10*, 3066–3073.

(47) Bohac, E. J.; Marshall, M. D.; Miller, R. E. *J. Chem. Phys.* **1992**, *96*, 6681–6695.

(48) Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2011**, *7*, 3162–3180.

(49) Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

(50) Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.

(51) East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638–4650.

(52) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

(53) Liakos, D. G.; Neese, F. *J. Phys. Chem. A* **2012**, *116*, 4801–4816.

(54) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *135*.

(55) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M. L.; Neese, F. *J. Chem. Theory Comput.* **2015**, *11*, 1525–1539.

(56) Liakos, D. G.; Neese, F. *J. Chem. Theory Comput.* **2015**, *11*, 4054–4063.

(57) Riplinger, C.; Neese, F. *J. Chem. Phys.* **2013**, *138*, 034106.

(58) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. *J. Chem. Phys.* **2013**, *139*, 134101.

(59) Edmiston, C. *J. Chem. Phys.* **1966**, *45*, 1833.

(60) Ahlrichs, R.; Lischka, H.; Staemmler, V.; Kutzelnigg, W. *J. Chem. Phys.* **1975**, *62*, 1235.

(61) Friedrich, J.; Hänchen, J. *J. Chem. Theory Comput.* **2013**, *9*, 5381–5394.

(62) Jesus, A. J. L.; Rosado, M. T. S.; Reva, I.; Fausto, R.; Eusébio, M. E. S.; Redinha, J. S. *J. Phys. Chem. A* **2008**, *112*, 4669–4678.

(63) Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. L. *J. Phys. Chem. A* **2013**, *117*, 2269–2277.

(64) Supady, A.; Blum, V.; Baldauf, C. *J. Chem. Inf. Model.* **2015**, *55*, 2338–2348.

(65) `https://github.com/adrianasupady/fafoom`.

(66) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

(67) Hill, A. D.; Reilly, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 1031–1035.

(68) Neese, F. *WIREs Comput Mol Sci* **2012**, *2*, 73–78.

(69) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(70) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–305.

(71) Kossmann, S.; Neese, F. *Chem. Phys. Lett.* **2009**, *481*, 240–243.

(72) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(73) Truhlar, D. *Chem. Phys. Lett.* **1998**, *294*, 45–48.

(74) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(75) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330–333.

(76) Neese, F.; Valeev, E. F. *J. Chem. Theory Comput.* **2011**, *7*, 33–43.

(77) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1-2*, 19–25.

(78) Thiel, W. *WIREs Comput Mol Sci* **2014**, *4*, 145–157.

(79) J. J. P. Stewart, MOPAC2012. `http://openmopac.net`.

(80) Maia, J. D. C.; Carvalho, G. A. U.; Carlos Peixoto Mangueira, J.; Santana, S. R.; Cabral, L. A. F.; Rocha, G. B. *J. Chem. Theory Comput.* **2012**, *8*, 3072–3081.

(81) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(82) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209–220.

(83) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(84) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

(85) Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1–32.

(86) http://openmopac.net/manual/SCF_calc_hof.html.

(87) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(88) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(89) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

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

(91) http://www.thch.uni-bonn.de/tc/index.php?section=downloads\&subsection=DFT-D3.

(92) Perdew, J. P.; Schmidt, K. *AIP Conf. Proc.* **2001**, *577*, 1–20.

(93) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(94) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

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

(96) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(97) Peverati, R.; Truhlar, D. G. *J. Phys. Chem. Lett.* **2012**, *3*, 117–124.

(98) Sun, J.; Ruzsinszky, A.; Perdew, J. P. *Phys. Rev. Lett.* **2015**, *115*, 036402.

(99) Sun, J.; Remsing, R. C.; Zhang, Y.; Sun, Z.; Ruzsinszky, A.; Peng, H.; Yang, Z.; Paul, A.; Waghmare, U.; Wu, X.; Klein, M. L.; Perdew, J. P. *Nat. Chem.* **2016**, 1–6.

(100) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(101) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982.

(102) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(103) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.

(104) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.

(105) Peverati, R.; Truhlar, D. G. *J. Phys. Chem. Lett.* **2011**, *2*, 2810–2817.

(106) Ren, X.; Rinke, P.; Blum, V.; Wieferink, J.; Tkatchenko, A.; Sanfilippo, A.; Reuter, K.; Scheffler, M. *New J. of Phys.* **2012**, *14*.

(107) Zhang, Y.; Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4963–4968.

(108) Ying, I.; Xu, X.; Zhang, I. Y. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12554–70.

(109) Zhang, I. Y.; Ren, X.; Rinke, P.; Blum, V.; Scheffler, M. *New J. of Phys.* **2013**, *15*, 123033.

(110) Zhang, I. Y.; Luo, Y.; Xu, X. *J. Phys. Chem.* **2010**, *133*.

(111) Dunning Jr, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(112) Grimme, S. *WIREs Comput Mol Sci* **2011**, *1*, 211–228.

(113) Vydrov, O. A.; Van Voorhis, T. *J. Phys. Chem.* **2010**, *133*, 244103.

(114) DiStasio Jr., R. A.; Gobre, V. V.; Tkatchenko, A. *J. Phys. Condens. Matter* **2014**, *26*.

(115) Tkatchenko, A.; Rossi, M.; Blum, V.; Ireta, J.; Scheffler, M. *Phys. Rev. Lett.* **2011**, *106*, 118102.

(116) Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M. *Chem. Eur. J.* **2013**, *19*, 11224–11234.

(117) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. *J. Phys. Chem. A* **2014**, *118*, 7349–7359.

(118) Schubert, F.; Rossi, M.; Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Filsinger, F.; Kupser, P.; Meijer, G.; Salwiczek, M.; Koksch, B.; Scheffler, M.; Blum, V. *Phys. Chem. Chem. Phys.* **2015**, *17*, 7373–7385.

(119) Baldauf, C.; Rossi, M. *J. Phys. Condens. Matter* **2015**, *27*, 493002.

(120) Marianski, M.; Asensio, A.; Dannenberg, J. J. *J. Chem. Phys.* **2012**, *137*, 044109.

(121) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 73005.

(122) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. *J. Chem. Phys.* **2014**, *140*, 18A508.

(123) Reilly, A. M.; Tkatchenko, A. *Phys. Rev. Lett.* **2014**, *113*, 055701.

(124) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

(125) Vorlová, B.; Nachtigallová, D.; Jirásková-Vaníčková, J.; Ajani, H.; Jansa, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P.; Konvalinka, J.; Lepšík, M. *Eur. J. Med. Chem.* **2015**, *89*, 189 – 197.

(126) Pecina, A.; Meier, R.; Fanfrlík, J.; Lepsik, M.; Řezáč, J.; Hobza, P.; Baldauf, C. *Chem. Commun.* **2016**, *52*, 3312–3315.

(127) Pol-Fachin, L.; Rusu, V. H.; Verli, H.; Lins, R. D. *J. Chem. Theory Comput.* **2012**, *8*, 4681–4690.

(128) Schreiner, P. R. *Angew. Chem. Int. Ed.* **2007**, *46*, 4217–4219.

(129) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. T. *Science* **2008**, *321*, 792.

(130) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Chem. Rev.* **2012**, *112*, 289–320.

(131) Mardirossian, N.; Head-Gordon, M. *J. Chem. Theory Comput.* **2016**, *12*, 4303–4325.

(132) Marom, N.; Tkatchenko, A.; Rossi, M.; Gobre, V. V.; Hod, O.; Scheffler, M.; Kronik, L. *J. Chem. Theory Comput.* **2011**, *7*, 3944–3951.

Table of contents figure