Fast and accurate quantum mechanical modeling of large molecular systems using small basis set Hartree–Fock methods corrected with atom-centered potentials

Viki Kumar Prasad [a], Alberto Otero-de-la-Roza [b,†,*] and Gino A. DiLabio [a,†,*]

[a.] Department of Chemistry, University of British Columbia, Okanagan, 3247 University Way, Kelowna, British Columbia, Canada V1V 1V7.

[b.] Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, MALTA Consolider Team, E-33006 Oviedo, Spain.

[†] These authors jointly supervised this work.

[*] E-mail correspondence: aoterodelaroza@gmail.com and gino.dilabio@ubc.ca

Abstract

There has been significant interest in developing fast and accurate quantum mechanical methods for modeling large molecular systems. In this work, by utilizing a machine-learning regression technique, we have developed new low-cost quantum mechanical approaches to model large molecular systems. The developed approaches rely on using one-electron Gaussian-type functions called atom-centered potentials (ACPs) to correct for the basis set incompleteness and the lack of correlation effects in the underlying minimal or small basis set Hartree-Fock (HF) methods. In particular, ACPs are proposed for ten elements common in organic and bio-organic chemistry (H, B, C, N, O, F, Si, P, S, and Cl) and four different base methods: two minimal basis sets (MINIs and MINIX) plus a double-ζ basis set (6-31G*) in combination with dispersion-corrected HF (HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*), and the HF-3c method. The new ACPs are trained on a very large set (73832 data points) of non-covalent properties (interaction and conformational energies) and validated additionally on a set of 32048 data points. All reference data is of complete basis set coupled-cluster quality, mostly CCSD(T)/CBS. The proposed ACP-corrected methods are shown to give errors in the tenths of a kcal/mol range for non-covalent interaction energies and up to 2 kcal/mol for molecular conformational energies. More importantly, the average errors are similar in the training and validation sets, confirming the robustness and applicability of these methods outside the boundaries of the training set. In addition, the performance of the new ACP-corrected methods is similar to complete basis set DFT but at a cost that is orders of magnitude lower, and the proposed ACPs can be used in any computational chemistry program that supports effective-core potentials without modification. It is also shown that ACPs improve the description of covalent and non-covalent bond geometries of the underlying methods and that the improvement brought about by the application of the

ACPs is directly related to the number of atoms to which they are applied, allowing the treatment of systems containing some atoms for which ACPs are not available. Overall, the ACP-corrected methods proposed in this work constitute an alternative accurate, economical, and reliable quantum mechanical approach to describe the geometries, interaction energies, and conformational energies of systems with hundreds to thousands of atoms.

1. Introduction

Quantum mechanical (QM) methods are an indispensable tool for understanding chemical phenomena. When combined with a nearly complete basis set, high-level wavefunction theory methods can predict various thermochemical and structural properties with an accuracy comparable to, or even better than, experiments. However, such approaches have limited applicability because their computational cost increases steeply with the size of the system.[1–5] This precludes high-level wavefunction methods from being applied to study chemical and biological processes involving large molecular systems, such as enzymatic catalysis, protein folding, supra-molecular host-guest complexation, and many others.[6–14]

In the past few decades, a significant amount of effort has been devoted to developing efficient and accurate QM methodologies that can be applied to large molecular systems.[15–28] The application of QM modeling begins by selecting a set of approximations to solve the Schrödinger equation. One of the simplest QM approaches with a low computational expense is the Hartree–Fock (HF) method. However, HF has a major shortcoming in that, by definition, it does not calculate any correlation energy, which results in overly repulsive dispersion interactions, bond lengths that are too short, and the poor prediction of various other molecular properties. In addition, the cost and accuracy of any QM method is strongly dependent on the choice of basis set, the set of functions used to describe the system's molecular orbitals. In HF, the computational cost scales roughly as the fourth power of the number of basis functions, with more sophisticated methods presenting an even steeper scaling. Calculations using either minimal or double-ζ basis sets are relatively inexpensive, but the use of these small basis sets introduces an additional error due to the insufficient number of basis functions. This basis set incompleteness error is severely detrimental to the method's accuracy. Therefore, even though minimal and double-ζ basis set HF offers a computationally inexpensive approach for modeling large molecular systems, a way needs to be devised to effectively mitigate the deleterious effect of missing electron correlation and basis set incompleteness error.[29,30]

Many existing semi-empirical QM methods are based on approximations to minimal basis set HF.[31–34] By construction, semi-empirical QM methods circumvent the calculation of certain two-electron integrals from the underlying minimal basis set HF approach while incorporating empirical parameters obtained by fitting to experimental or high-level theoretical reference data. These approximations substantially limit the accuracy of semi-empirical QM methods but in exchange reduce the computational cost below that of minimal basis set HF. Due to their reduced computational cost, semi-empirical QM methods have found extensive use in modeling large molecular systems. An example of a popular and more recent semi-empirical QM approach is the PM7 method of Stewart, which was also modified by Throssel and Frisch.[35,36]

Another approach that is similar in spirit to conventional semi-empirical QM methods is the HF-3c[37] method proposed by Sure and Grimme. The HF-3c method uses three separate geometry-dependent formulas[38,39] to add energy corrections ("3c") for the various deficiencies of minimal basis set HF: one to account for some of the missing dispersion interactions, and two to mitigate the effects of basis set incompleteness errors. Several other techniques have been proposed in the literature[40–54], reflecting the interest in developing computationally inexpensive methods for large systems.

Finding a good compromise between cost and accuracy is critical when modeling large molecular systems. HF-3c is an example of a QM method that strikes a good balance between these two desirable characteristics. Even though the cost of HF-3c is higher than most semi-empirical QM methods, it is still orders of magnitude cheaper than nearly complete basis set wavefunction theory or density-functional theory (DFT) based methods. On the other hand, the accuracy of HF-3c in describing molecular structures and non-covalent interaction strengths is similar to large basis set DFT.[55] These features allow HF-3c to be applied for fast geometry optimizations, conformer exploration, and prediction of non-covalent interaction energies in fairly large systems, with sizes between many hundreds and a few thousand atoms. This allows the QM description of (small) biological systems (proteins, nucleic acids, carbohydrates, lipids) as well as supramolecular host-guest complexes. The downside of HF-3c is that it is unable to accurately describe thermochemical quantities such as bond breaking and formation energies.[56] Grimme and co-workers have applied the 3c correction to a few density functional approximations to address this problem.[57–62]

Our previous works have shown that atom-centered potentials[63] (ACPs) offer a convenient means of improving the accuracy of HF and DFT based methods.[64–76] ACPs are one-electron potentials that share the same mathematical form as effective-core potentials[77,78] (ECPs) but do not replace any electrons. This allows ACPs to be used in most computational chemistry software packages without modifying the code. Additionally, ACPs are an economical way of mitigating the errors in the underlying methodology, since using them incurs only a small additional cost. In a previous proof-of-concept work, we developed a single set of ACPs for the H, C, N, and O elements to mitigate the shortcomings of dispersion-corrected minimal basis set HF.[64] The parameters for the ACPs were obtained by fitting to a set of 9814 data points of non-covalent properties (interaction, conformational, and molecular deformation energies). In that work, we demonstrated the feasibility of the ACP correction approach by showing that ACPs developed for dispersion-corrected minimal basis set HF were able to accurately predict the mentioned non-covalent properties.

In this work, we build upon our previous study[64] and develop four ACP-corrected small basis set HF based methods. In all cases, the target applications are similar to those of HF-3c and our previous work,[64] namely structures and non-covalent interaction strengths. However, ACPs are developed for a larger set of atoms (H, B, C, N, O, F, Si, P, S, and Cl) than in our previous work, greatly increasing the applicability of the proposed methods. In addition, the use of the LASSO (Least Absolute Shrinkage and Selection Operator) regression[79–81] for fitting of ACP parameters greatly simplifies ACP development and allows using a training set about eight times as large and much more diverse than in earlier works,[64,68–70,73] resulting in more robust and more widely applicable ACPs. Three of the four new ACP-corrected methods are based on HF with minimal (MINIs[82] and MINIX[37]) or small double-ζ basis sets (6-31G*[83,84]) and use Grimme's D3[38,85,86] correction to account for the missing dispersion in HF. In addition, we also present a set of ACPs designed to improve the performance of the HF-3c method. In addition, our intention with the HF-3c-ACP method is to overcome the limitation imposed by the fact that ACPs are available only for the ten elements mentioned above. Since HF-3c parameters are available for most elements in the periodic table, we expect HF-3c-ACP to reduce to HF-3c performance for the atoms for which ACPs have not been developed, which in general should be in the minority. The newly developed ACP-corrected methods are assessed using an extensive validation set, demonstrating their performance and robustness.

2. Computational Methodology

2.1 Theoretical background

The procedure employed to develop the ACPs proposed in this work is similar to our earlier proof-of-concept study[64]. The mathematical form of an ACP is:

$$\hat{V}_{ACP} = \sum_{\alpha} \left(V_{local}^{\alpha}(r) + \sum_{l=0}^{L-1} \sum_{\mathrm{m}=-l}^{l} \delta V_l^{\alpha}(r) \, |Y_{lm}\rangle\langle Y_{lm}| \right) \quad (1)$$

where $\delta V_l^{\alpha}(r) = V_l^{\alpha}(r) - V_{local}^{\alpha}(r)$, α represents the atoms on which the potentials are centered, and r is the distance to atom α. The $|Y_{lm}\rangle\langle Y_{lm}|$ represents projection operators using real spherical harmonics based on atom α with l angular quantum numbers and m magnetic quantum numbers. Equation 1 is the same general expression as ECPs[77,78]. The semi-local nature of the ACP arises from combining the first term (the *local* term), which only depends on the radial coordinate, with the second term (the *non-local* term), which incorporates the anisotropy via angular projections. The individual *local* and *non-local* terms in Equation 1 are represented by Gaussian-type functions:

$$V_{local}^{\alpha}(r) = \sum_{n=1}^{N} c_{local}^{\alpha} \exp(-\xi_{local}^{\alpha} r^2) \tag{2}$$

$$\delta V_l^{\alpha}(r) = \sum_{n=1}^{N} c_l^{\alpha} \exp(-\xi_l^{\alpha} r^2) \tag{3}$$

where the coefficients (c) and exponents (ξ) are adjustable parameters that are determined via a regularized least-squares fitting to reference data during ACP development (Section 2.2). The sum in Equations 2 and 3 runs over the total number (N) of Gaussian-type functions defined for atom α for the *local* and *non-local* potential terms. For ease of notation, we will represent the $V_{local}^{\alpha}(r)$ and $\delta V_l^{\alpha}(r)$ together as:

$$V_l^{\alpha}(r) = \sum_{n=1}^{N} c_{ln}^{\alpha} \exp(-\xi_{ln}^{\alpha} r^2) \quad for \;\; l = 0, 1, 2, \dots, L \tag{4}$$

If the ACP operator (Equation 1) with the functional form of Equation 4 is treated as a perturbative correction to any Hamiltonian then the first-order perturbation energy correction induced by the ACPs is:

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_i \langle \psi_i | \hat{V}_{ACP} | \psi_i \rangle \tag{5}$$

where the sum in Equations 5 runs over the occupied molecular orbitals. Substituting the expressions from Equations 1 and 4 into Equation 5 gives:

$$E_{ACP}(\{c_{ln}^{\alpha}\}, \{\xi_{ln}^{\alpha}\}) = \sum_{\alpha l n} c_{ln}^{\alpha} \sum_i \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle \tag{6}$$

The $\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \langle \psi_i | (|Y_{lm}\rangle \exp(-\xi_{ln}^{\alpha} r^2) \langle Y_{lm}|) | \psi_i \rangle$ integral, known as an ACP energy term, is the energy difference between the energy when an ACP with exponent ξ_{ln}^{α} is applied and the energy in absence of any ACP, divided by the ACP coefficient. Packing the coefficients c_{ln}^{α} and exponents ξ_{ln}^{α} into vectors and combining the terms in the inner sum of Equation 6 leads to:

$$E_{ACP}(\boldsymbol{c},\boldsymbol{\xi}) = \sum_{\alpha ln} c_{ln}^{\alpha}\,\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha}) = \boldsymbol{c}\cdot\boldsymbol{\Delta}\mathbf{E}(\boldsymbol{\xi})^{T} \qquad (7)$$

where $\boldsymbol{\Delta}\mathbf{E}(\boldsymbol{\xi})^{T}$ is the vector of ACP energy terms. It should be noted that Equation 7 is only correct to first order in the ACP perturbation as the coefficients c_{ln}^{α} have an influence on the underlying wavefunction. Equation 7 becomes exact only in the limit of $\boldsymbol{c} \to \boldsymbol{0}$, and it is approximately correct if the ACP coefficients are small in magnitude. The deviation between the E_{ACP} obtained using a self-consistent calculation with the corresponding ACP and the linear estimate in Equation 7, which assumes the coefficients have no influence on the underlying wavefunction, is called the *non-linearity error*.

2.2 ACP development process

ACPs have features that make them useful to develop energy corrections for a QM method (see Reference 69 for more details): (i) ACPs generate energy correction terms based on the molecular orbitals (Equation 5) and are wavefunction-dependent, which means they include information from the chemical environment and the electronic wavefunction, (ii) the angular projection operators in the potential (Equation 1) produces energy correction terms that are dependent upon the local anisotropic environment of a given atom, (iii) the exponential form of the ACPs (Equation 4) ensures that a given ACP produces a correction that decays exponentially with interatomic distances, (iv) ACPs can be used with any software that uses Gaussian-type basis sets and ECPs for geometry optimizations as well as calculation of energies and energy derived properties, and (v) the use of ACPs incurs only in a small computational cost (see Table S6 in SI for comparison of percentage change in the single-point calculation time between that of uncorrected and ACP corrected approaches for twenty selected molecules).

The first stage of the ACP development process involves assembling a comprehensive and diverse training set of target molecular properties. The training set should ideally consist of model systems composed of atoms for which ACPs are being developed. The training set should also contain molecules representing various chemical environments to ensure that the developed ACPs can be applied to diverse chemical systems. Since the focus of this work is to correct the deficiencies of small basis set HF based methods regarding molecular structures and non-covalent interactions, our training set contains data points of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies.

Next, the exponents (ξ_{ln}^{α}), atoms (α), and angular momentum channels (l) are chosen, and the corresponding ACP energy terms ($\Delta E_{ln}^{\alpha}(\xi_{ln}^{\alpha})$) are calculated. The ACP energy term evaluation process is carried out by first obtaining SCF energy for every training set entry and then evaluating the ACP terms post-SCF. This approach speeds up the ACP energy term evaluation process which is important given the size of the training set and the number of ACP terms. Once the ACP energy terms have been computed for each target method/basis-set, exponents, angular momenta, and systems in the training set, the optimal ACP coefficients c_{ln}^{α} are determined using a regularized least-squares fit subject to a constraint on the sum of the absolute values of the coefficients. This constraint limits the magnitude of the ACP contributions and ensures that the correction arising from the ACP does not lead to significant non-linearity error, which would lead to disagreements between the predictions of our linear model (Equation 7) and the results obtained from the application of the ACP in an actual self-consistent calculation.

The training set is organized into subsets, corresponding to different molecular properties and data sources from the literature. Each subset of the training set is assigned a weight in the fitting procedure. The weight of subset i is calculated in the same way as in our previous work[64]:

$$w_i = \frac{1}{M_i N_i} \tag{8}$$

where M_i is the average of the absolute value of the reference energies and N_i is the number of data points in subset i. These weights account for differences in reference data magnitude and number of points in the subsets. The error function minimized in the fit is the weighted root-mean-square-error ($wRMSE$):

$$wRMSE = \sqrt{\frac{\sum_i (w_i \sum_j^{N_i} (y_{ref,j}^{i} - y_{method,j}^{i})^2)}{\sum_i N_i}} \tag{9}$$

where j are the data points in the i^{th} subset, $y_{ref,j}^{i}$ are the high-level reference energies for system j in the i^{th} subset, and $y_{method,j}^{i}$ are the energies of the underlying method for which the ACPs are being developed.

The LASSO (Least Absolute Shrinkage and Selection Operator) regression[79–81], commonly employed in statistics and machine learning, is used to carry out the regularized least-squares fit. In LASSO, the $wRMSE$ in Equation 9 is minimized subject to the condition that l_1-norm of the ACP coefficients does not exceed a certain bound chosen beforehand:

$$\|c\|_1 = \sum_i |c_i| \tag{10}$$

The LASSO method is used to limit the magnitude of the ACP coefficients. In addition, for a given constraint, LASSO automatically selects the best subset of ACP terms and discards the others, resulting in ACPs with fewer terms, which is beneficial because it curbs the computational cost of applying the ACPs.

2.3 Training and validation data sets

The training set (Table 1) comprises non-covalent interaction energies, molecular conformational energies, and molecular deformation energies. This choice of training set properties is justified by the potential target applications of small basis set HF based methods, namely fast geometry optimizations and non-covalent interaction strengths in large systems as well as high-throughput[87] screening of conformers in combination with conformer search techniques[88–90]. These applications are useful, for instance, when performing exhaustive conformational searches of macrocyclic drugs[91–96] and other pharmaceutical candidates[97], and studying biochemical processes like protein folding[98–100] and puckering of nucleotides[101,102].

A successful method for non-covalent interactions must be able to accurately describe diverse non-covalent interaction motifs, which means that the training set must contain some of this diversity. For instance, the importance of π-π interactions is well-known in medicinal chemistry[103], structural biology[104,105], and organic electronics[106]. Such interactions also contribute to the stabilization of DNA[107,108] and proteins[109], control the strength and specificity of drug-protein interactions[110], and help in the rational design of supramolecules[111–113]. Other types of non-covalent interactions are also important in practice. For example, Berka *et al.* reported that aliphatic-aliphatic (or hydrophobic) interactions between amino acid backbone chains are the most abundant in proteins, particularly in the hydrophobic active site.[114] Hydrophobic interactions also control the structure and properties of lipid bilayers[115,116] and self-assembled supramolecules[117]. On the other hand, hydrogen bonding is probably the most studied non-covalent interaction.[118,119] Several other stabilizing non-covalent interactions with potential chemical and biological applications have also been studied, including halogen bonding[216–219], pnicogen bonding[220–222], and anionic[120–123] interactions. Therefore, we designed our training set to contain representative candidates from the interaction types mentioned above. This also allows us to assess the strengths and weaknesses of the developed ACPs regarding each interaction type.

In most cases, the subsets of the training set, and the molecular geometries and reference data in them, were adopted from the literature. Occasionally, the reference energies were re-calculated at a higher-level to improve their quality. In each subset of the training set, data points involving molecules containing atoms other than H, B, C, N, O, F, Si, P, S, and Cl were excluded. A detailed list of all the subsets used for our ACP training set is given in Table S1 of the SI. The number of data points in each subset varies depending on the availability of benchmark data sets.

The training set used here is almost eight times larger than in our previous work[64]. In total, the training set comprises 73832 data points (167275 molecular geometries) calculated mainly with complete basis set wavefunction theory methods (Table S1 in SI). The total number of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies are 19439, 44105, and 10288, respectively. The most abundant type of non-covalent interaction in the training set is hydrogen bonding. The mixed non-covalent interactions subset is second in abundance and features a mix of all common interactions found in large molecular systems. The large size of the training set ensures that no overfitting occurs when the least-squares fit is carried out.

In order to evaluate the performance and robustness of the new ACPs, we also assembled a validation set (Table 2), different from the training set, by compiling additional data sets from the literature. A detailed list of all the data sets included in the validation set is provided in Table S2 of the SI. In total, the validation set consists of 32047 high-level data points (92161 molecular geometries) calculated mainly with complete basis set wavefunction theory. The validation set contains 27811 non-covalent interaction energies and 4237 molecular deformation energies.

The structures and reference energies of all data points in the training and validation sets are given in the SI. In addition, the subsets that comprise the training and validation sets, grouped into categories to facilitate the analysis of the results, are listed in Tables 1 and 2. It should be noted that this subset categorization is in no way an exhaustive representation of the various types of systems in the training or validation set.

Table 1. List of data sets, grouped by category, in the ACP training set.

Category	**Data set(s)**	**Data points**	**Reference energy range (kcal/mol)**	**Description**
Non-covalent interaction energies of molecular complexes[a]:				
π-stacking	Pisub[b,124,125], Pi29n[126], BzDC215[127], C2H4NT[128]	379	-18.30 to +10.33	non-stacked and stacked π-π interactions
Hydrophobic	ADIM6[38,129,130], HC12[131]	18	-5.60 to -1.30	aliphatic-aliphatic interactions
Pnicogen-bonding	PNICO23[129,132]	23	-10.97 to -0.64	pnicogen bonding interactions
Halogen-bonding	Hill18[133], X40x10[134]	238	-14.14 to +11.95	halogen bonding interactions

Hydrogen-bonding	HBC6[135,136], MiriyalaHB104[137,138], IonicHB[139], HB375x10[140], IHB100x10[140], HB300SPXx10[141], CARBH12[129]	6409	-37.01 to +16.30	hydrogen bonding interactions
Mixed NCIs	S22x5[136,142,143], S66x8[144–146], S66a8[145], A21x12[3,147,148], NBC10ext[128,136,149–151], 3B-69-DIM[152], 3B-69-TRIM[152], HW30[153]	1895	-35.76 to +9.34	mixed-character non-covalent interactions
Anionic[c]	SSI-anionic[154], WatAA-anionic[b,155], HSG-anionic[136,156], PLF547-anionic[157], IonicHB-anionic[139], IHB100x10-anionic[140], Ionic43-anionic[158]	1509	-135.11 to +88.94	anionic interactions
Biomolecule-Biomolecule	BBI[154], SSI[154], NucTAA[b,c,159–162], CarbhydBz[163], CarbhydNaph[164], CarbhydAroAA[b,165], CarbhydAro[b,166], WatAA[b, 155], HSG[136,156], PLF547[157], JSCH[142], DNAstack[167], DNA2body[167], ACHC[168], BDNA[169], NucBTrimer[b,170]	4756	-100.86 to +64.19	interactions present in various biomolecules
Gas-Ligand	CH4PAH[171,172], CO2MOF[173], CO2PAH[174], CO2NPHAC[175], BzGas[176]	876	-6.02 to +12.17	interactions between gas molecules and substrate
Water-Water	Water38[177], Water1888[128,178–180], Water-2body[d,68]	2336	-92.89 to +5.10	hydrogen-bonded water dimers and $(H_2O)_n$ clusters where n=3–10
BFSiPSCl	B-set[b,65], F-set[b,65], Si-set[b,65], P-set[b,65], S-set[b,65], Cl-set[b,65], Sulfurx8[181]	1000	-68.05 to +21.57	monomers containing B, F, Si, P, S, and Cl atoms
Molecular conformational energies[e]**:**				
Small molecule	37Conf8[182], DCONF[183], ICONF[129], MCONF[184], Torsion21[185], MolCONF[186], ANI1ccxCONF[f,187]	41224	+0.01 to +50.00	various molecules representing pharmaceuticals, catalysts, synthetic precursors, industrial chemicals, and organic compounds
Negatively charged[g]	PEPCONF-Dipeptide-anionic[b,188], MolCONF-anionic[186]	254	-0.47 to +10.96	negatively charged molecules
Biomolecule	PEPCONF-Dipeptide[b,188], TPCONF[189], P76[190], YMPJ[191], SPS[192], rSPS[193], UpU46[194], SCONF[129,195], DSCONF[196], SacchCONF[197], CCONF[198]	2082	-4.09 to +19.74	molecules representative of proteins, DNA, RNA, and carbohydrates
Hydrocarbon	ACONF[199], BCONF[200], PentCONF[201]	421	+0.14 to +16.66	hydrocarbon-like molecules
$(H_2O)_{11}$	Undecamer125[202]	124	+0.06 to +1.87	$(H_2O)_{11}$ clusters
Molecular deformation energies[h]**:**				
Deformation	MOLdef[a,65], MOLdef-H2O[d,203,204]	10288	-3.43 to +49.38	various molecules deformed along their normal modes

[a] defined as the difference between the energy of the complex and the sum of the monomer energies. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b] the reference data was recalculated in this work at the DLPNO-CCSD(T)/CBS level of theory (see SI for more details), at geometries reported in the literature.

[c] comprises non-covalently bound dimers where at least one of the monomers is negatively charged.

[d] the reference data was calculated in this work at CCSD(T)/CBS level using the same extrapolation method as in Reference 177.

[e] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[f] contains mostly conformational energies but also some molecular deformation energies.

[g] comprises negatively charged conformers.

[h)] defined as the difference between the energy of a molecule deformed along a particular normal mode and the energy of the same molecule at equilibrium.

Table 2. List of data sets, grouped by category, in the ACP validation set.

Category	Data set(s)	Data points	Reference energy range (kcal/mol)	Description
Non-covalent interaction energies of molecular complexes[a]:				
Mixed NCIs	BlindNCI[205], DES15K[206], NENCI-2021[207]	17413	-33.78 to +186.83	mixed character non-covalent interactions
Hydrogen-bonding	CE20[208,209], WaterOrg[210]	2396	-46.58 to -10.76	hydrogen bonding interactions
Halogen-bonding	XB45[211]	33	-13.11 to -0.89	halogen bonding interactions
Chalcogen-bonding	CHAL336[212]	48	-30.85 to -1.57	chalcogen bonding interactions
Repulsive contacts	R160x6[213], R739x5[214]	5290	-12.02 to +6.79	close contact interactions
Anionic[b]	HW6Cl-anionic[215,216], HW6F-anionic[215,216], FmH2O10-anionic[215,216], SW49Bind345-anionic[217], SW49Bind6-anionic[217], Anionpi-anionic[218], IL236-anionic[219], DES15K-anionic[206], NENCI-2021-anionic[207], CHAL336-anionic[212], XB45-anionic[211], S30L-anionic[220]	2525	-171.42 to +66.15	anionic interactions
$(H_2O)_{20}$ cluster	H2O20Bind10[216]	10	-200.54 to -196.59	$(H_2O)_{20}$ clusters
C_{60} dimer	C60dimer[221]	14	-6.88 to +12.07	C_{60} dimers
Large molecule	L7[13,222,223], S12L[9,11,223], S30L[220], Ni2021[224]	54	-416.08 to -1.68	large molecules relevant in supramolecular chemistry and biochemistry
Molecular conformational energies[c]:				
Small molecule	SafroleCONF[225], AlcoholCONF[226], BeranCONF[227], Torsion30[d,228]	2193	+0.001 to +12.50	Safrole or 5-(2-propenyl)-1,3-benzodioxol) molecule, small alcohol molecules, small organic molecules, and biaryl drug-like molecules
Proteinogenic	MPCONF196[e,229], PEPCONF-Tripeptide[f,188], PEPCONF-Disulfide[g,188], PEPCONF-Cyclic[g,188], PEPCONF-Bioactive[g,188]	1874	-0.47 to +81.00	peptide-like molecules
Negatively charged[h]	PEPCONF-Disulfide-anionic[g,188], PEPCONF-Bioactive-anionic[g,188]	170	+0.17 to +33.79	negatively charged molecules

[a)] defined as the difference between the energy of the complex and the sum of energy of the monomers. A negative interaction energy indicates the complex is more stable than the separated monomers.

[b)] comprises non-covalently bound complexes with at least one negatively charged monomer.

[c)] defined as the difference between the energy of a particular conformer and a lower-energy conformer of the same molecule.

[d)] only 30 systems used; we could not find the rest systems mentioned in Reference 228 in the supporting information

[e)] only macrocyclic peptides considered.

[f)] only a subset from the PEPCONF[188] database for which reference data was recalculated at the DLPNO-CCSD(T)/CBS level of theory (see SI for more details).

[g)] available reference data was calculated at LC-ωPBE-XDM/aug-cc-pVTZ level of theory.

[h)] comprises negatively charged conformers.

2.4 Technical details

Three sets of ACPs were developed for HF-D3 in combination with the minimal basis set MINIs, MINIX, and the double-ζ basis set 6-31G*. An additional set of ACPs was developed for HF-3c, which uses the MINIX basis set. The MINIX basis set was proposed at the same time as HF-3c[37], and is equivalent to MINIs for the first row atoms (H, B, C, N, O, and F) but employs an extra *d* basis function for Si, P, S, and Cl. Angular momentum channels up to the maximum angular momentum of the valence orbital basis functions for each atom present in the chosen basis set were used for ACP development. The maximum angular momentum values were: *s* for H (MINIs, MINIX, 6-31G*), *p* for B, C, N, O, F, Si, P, S, Cl (MINIs), *p* for B, C, N, O, F (MINIX), *d* for Si, P, S, Cl (MINIX), and *d* for B, C, N, O, F, Si, P, S, Cl (6-31G*).

Twenty-nine ACP exponents (ξ_{ln}^{α}) were chosen, with values: 0.12 to 0.30 in 0.02 steps, 0.40 to 2.00 in 0.10 steps, and 2.50 to 3.00 in 0.50 steps. It should be noted that the choice of exponents is different than in our previous work[64]. A careful evaluation of the computational cost associated with the calculation of ACP energy term integrals (Equation 7) with low exponent functions ($0.01 < \xi_{ln}^{\alpha} < 0.11$) suggested that ACPs containing exponents lower than 0.12 can lead to a significant increase in calculation time (especially for large molecules) compared to methods where ACPs are not applied. Therefore, choosing exponents higher than or equal to 0.12 ensures that the computational overhead is limited to a 10–30% increase relative to the uncorrected method.

The combination of ten atoms and twenty-nine exponents along with the various angular momentum channels results in 841 ACP terms for MINIs, 957 ACP terms for MINIX, and 1102 ACP terms for 6-31G*. For our training set of 73832 data points (167275 molecular geometries), ACP development required a total of 140678275 (HF- D3/MINIs), 160082175 (HF-D3/MINIX), and 184337050 (HF-D3/6-31G*) single-point energies. Combined with the self-consistent calculations used to evaluate the impact of non-linearity error (1338200) and the calculations on the validation set to evaluate the performance of the ACPs (737288), the total number of calculations for this project is 487172988. This complexity required the development of specialized software which we briefly describe next.

The parameters for the D3 dispersion correction used in this work correspond to those for the HF/aug-cc-pVTZ method with Becke-Johnson damping: $s_6 = 1.0$, $s_8 = 0.9171$, $a_1(BJ) = 0.3385$, and $a_2(BJ) = 2.8830$ Å. It should be noted that these D3 parameters are very close to those used in HF-3c[37]. The ACP energy term evaluation and fitting processes were carried out using the *dcp*[230] and *acpfit*[231] packages available in our GitHub repository[232]. These programs automatize and collate all the data required for ACP development. For the LASSO regression, we used the local linearization plus active set method proposed by Osborne *et al.*[233] and implemented in *octave/MATLAB* by Schmidt[234,235]. All single-point energy calculations with minimal or double-ζ basis set HF-D3 methods were performed with the *Gaussian-16*[236] software package. The HF-3c single-point energy calculations were performed with the *ORCA*[237] software package. All the SCF single-point energy calculations on the training and validation sets were carried out with the default settings. The post-SCF calculations for the ACP energy term evaluations (Equations 6 and 7) were executed using non-SCF multistep *Gaussian-16* jobs.

Once all the ACP energy terms for a particular target method were successfully computed, they were passed to the LASSO fit, resulting in an optimal set of ACPs for that method with minimum $wRMSE$ for a constraint of 25.0 au on the l_1-norm of coefficients. The ACPs proposed in this work contain approximately 6–19 terms per atom, and they are designed to be paired with the specific method for which they were developed (i.e., HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*, or HF-3c), and are not transferable to other methods. The ACP coefficients and exponents for each method are provided in the SI. An example of the usage of ACPs in the Gaussian software is also given in the SI.

3. Results and Discussion

3.1 Performance of ACPs for the training set

The optimal ACPs paired with their respective methods (HF-D3/MINIs, HF-D3/MINIX, HF-D3/6-31G*, and HF-3c) were applied self-consistently on the entire training set. The resulting non-covalent interaction energies, molecular conformational energies, and molecular deformation energies are compared to the corresponding reference data in Figures 1 and 2. The strip charts represent the error distribution as vertical lines for each method. The mean signed errors (MSEs) (open circle) and the standard deviations (SDs) of the errors (horizontal black lines) are also represented. The mean absolute errors (MAEs) and percentage change in the MAEs upon the application of ACPs (%ΔMAE) for each method are listed on the right. A detailed breakdown of the errors for each method and subset can be found in Table S3 of the SI.

Table S3 of the SI also lists the deviation between the prediction of the ACP performance from our linear model (the LASSO fitting procedure) against the actual results from using the ACP in self-consistent calculations. This comparison, which measures the extent of non-linearity error, shows that the deviation between the MAEs predicted by the linear model and the self-consistent calculations is under 10% for most of the subsets of the training set, with only a few exceptions. This indicates that the $\boldsymbol{l_1}$-norm constraint imposed in the LASSO fit was effective in preventing excessive non-linearity error and that the linear model used to develop ACPs is a faithful representation of their eventual performance as a correction method. In the following, the results obtained from the self-consistent application of ACPs are discussed for the different molecular properties in the training set.

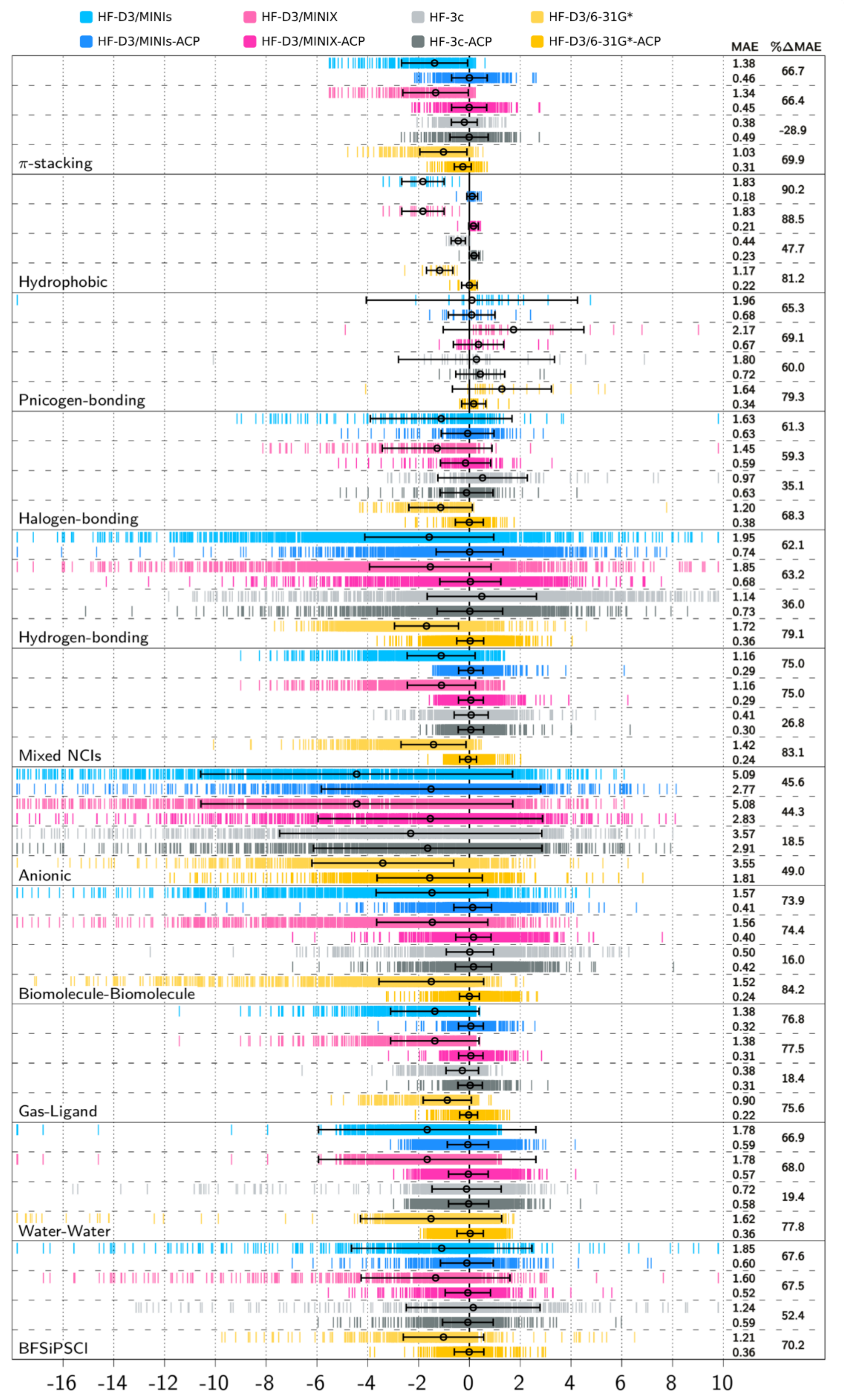


Figure 1. Error distribution (in kcal/mol) associated with non-covalent interaction energy subsets of the training set (see Table 1). Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP (pink), HF-3c (light grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -18 (left) and +10 kcal/mol (right) for clarity.

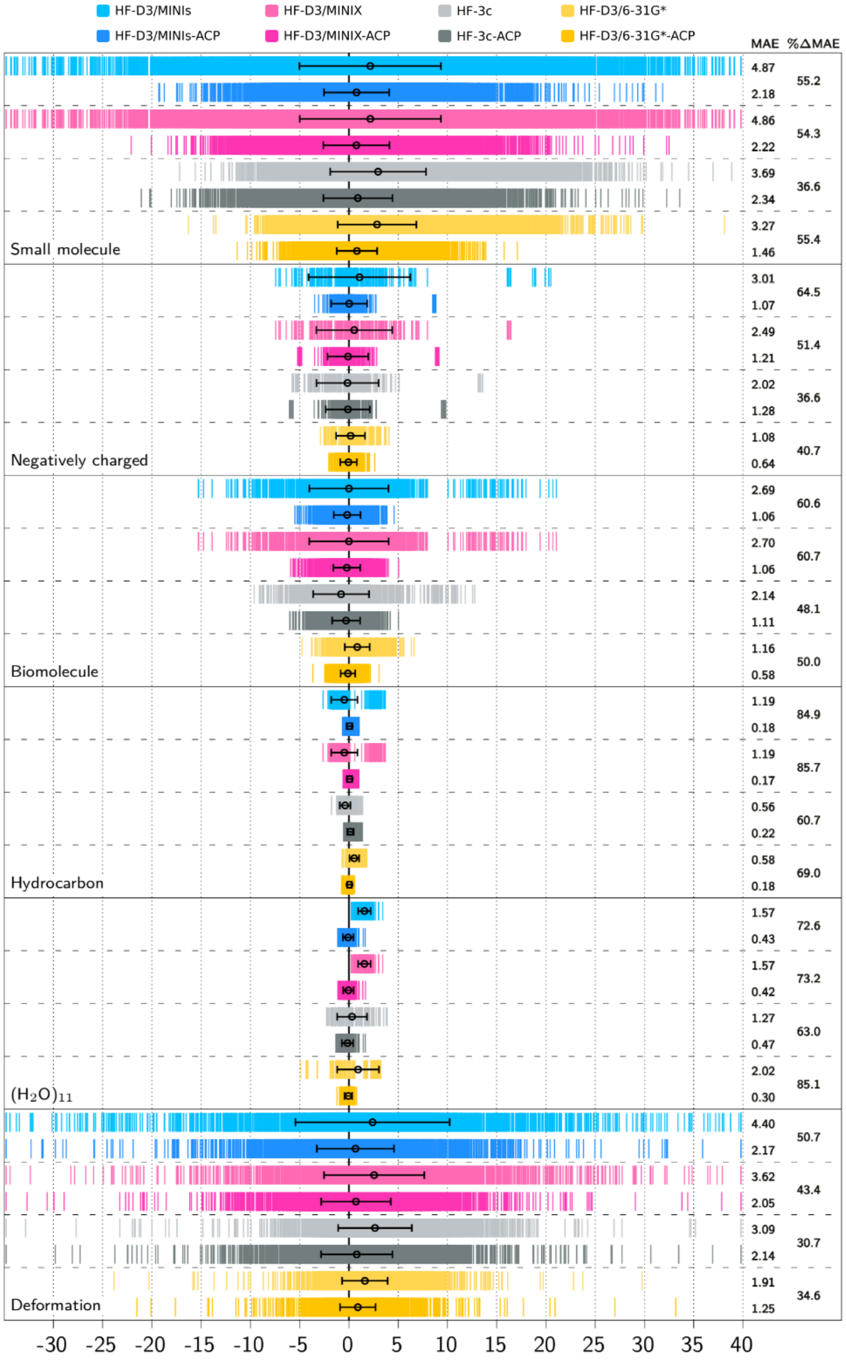


Figure 2. Error distribution (in kcal/mol) associated with molecular conformational and deformation energy subsets of the training set (see Table 1). Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP (pink), HF-3c (light grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -35 (left) and +40 kcal/mol (right) for clarity.

(i) Non-covalent interaction energies

The ACPs developed in this work have been trained on a wide range of non-covalent interaction types, including stacked and non-stacked π-π interactions, hydrophobic interactions, pnicogen bonding, halogen bonding, hydrogen bonding, and interactions of mixed and anionic nature. The proper description of each of these interactions is important for modeling large molecular systems, like proteins, where they operate co-operatively.[238] Figure 1 shows that the minimal or double-ζ basis set HF-D3 and HF-3c methods without ACPs have MAEs below 2 kcal/mol for different types of interactions except those of anionic nature (*Anionic* subset), where the MAEs are above 3.5 kcal/mol. Figure 1 also shows that HF-3c yields MAEs below 0.50 kcal/mol for *π-stacking*, *Hydrophobic*, and *Mixed NCIs* subsets, indicating that HF-3c is well suited to model systems that contain interactions of π-π, aliphatic-aliphatic, and mixed nature. The application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods mostly brings down the MAEs by about 44–90% (minimal basis set HF-D3), 49–84% (double-ζ basis set HF-D3), and 16–60% (HF-3c) for the range of interaction types covered in the subsets *π-stacking*, *Hydrophobic*, *Pnicogen-bonding*, *Halogen-bonding*, *Hydrogen-bonding*, *Mixed NCIs*, and *Anionic*.

Figure 1 shows that the application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods leads to an improved description of interactions of mixed, hydrogen bonding, and hydrophobic interaction types. π-π stacking interactions are also well described by our ACPs, except in the case of ACPs developed for HF-3c. Even though ACPs lead to a better description of anionic interactions than minimal or double-ζ basis set HF-D3 and HF-3c, the error spread is still large, evidencing the shortcomings of the underlying methods regarding anionic interactions.

Figure 1 also shows that the minimal or double-ζ basis set HF-D3 methods tend to over-estimate the interaction energies of almost all interaction types (except pnicogen bonding), with negative MSEs. On the other hand, the MSEs for HF-3c indicate that over-estimation or under-estimation of interaction energies depends on the nature of interaction type. The error spread in HF-3c is generally lower than minimal or double-ζ basis set HF-D3 methods, resulting in lower MAEs, MSEs, and SDs for this method. When applied to minimal or double-ζ basis set HF-D3, ACPs correct the over-estimation in the interaction energies, resulting in a lower spread of errors and SDs, and a corresponding decrease in MSEs. Similarly, depending on the nature of interaction type, ACPs also improve the over-estimation or under-estimation tendencies of HF-3c for certain interaction types, causing a reduction in the corresponding MSEs.

We now examine the ACP performance for the more common interaction types in the training set, namely, hydrogen bonding and mixed interactions (*Hydrogen-bonding* and *Mixed NCIs* subsets). ACPs lower the MAEs of all four methods for the *Hydrogen-bonding* subset by about 63% (minimal basis set HF-D3), 79% (double-ζ basis set HF-D3), and 36% (HF-3c). For the *Mixed NCIs* subset, ACPs lower the MAEs by 75% (minimal basis set HF-D3), 83% (double-ζ basis set HF-D3), and 27% (HF-3c). It is also evident from Figure 1 that ACPs not only reduce the MAEs of the HF-D3 and HF-3c methods but also reduce the spread of errors and SDs and the bias. This is particularly true in the case of the HF-D3 methods. Some outliers with high error exist, which is natural given the very large size of the training set, but these errors are still lower than those predicted without ACPs. The individual errors for the *Mixed NCIs* subset are mostly within ±2 kcal/mol. Some of the systems with errors beyond ±2 kcal/mol are trimers with roughly twice the reference energies than the dimers in the training set. For the *Hydrogen-bonding* subset, an inspection of the errors beyond ±5 kcal/mol reveals that the ACPs over-stabilize the hydrogen bonding interactions of some complexes with polar bonds involving electronegative S, P, F, and Cl atoms.

The *π-stacking* subset with HF-3c-ACP is the only case where ACPs slightly increase the MAE of the base method (from 0.38 kcal/mol for HF-3c to 0.49 kcal/mol for HF-3c-ACP). However, it should be noted that for *π-stacking*, the ACPs developed for minimal or double-ζ HF-D3 methods, which initially have almost three times higher MAEs than HF-3c, do lead to a reduction in the MAEs by approximately 66–70%. A similar result occurs for the *Mixed NCIs* subset where MAEs of minimal or double-ζ HF-D3 methods are almost three times higher than HF-3c, and the application of ACPs reduce the MAEs for minimal or double-ζ HF-D3 methods by about 75–80% and by only about 27% for the HF-3c method. These two examples suggest that ACPs reduce the MAEs of the underlying methods when they are high and have a lesser impact on those subsets where the MAEs of the underlying method are already low. Consequently, in a few rare instances the performance of an underlying method with low initial MAE can be negatively, but only slightly, impacted by the application of ACPs. This is the case for HF-3c-ACP applied to π-π interactions.

A particular limitation of all methods in this work is the performance for the anionic systems in the *Anionic* subset. This subset is challenging for basis sets like MINIs, MINIX, and 6-31G* because of the lack of diffuse functions required to properly describe negatively charged species. Figure 1 shows that the error spread and the SDs of the *Anionic* subset are larger than other interaction types. Because the *Anionic* subset was used in the training set, ACPs improve the performance of all four methods for anionic interaction energies, with MAE reductions of 19–49%. However, there is obvious room for improvement, and it is likely that it can only be achieved by the inclusion of diffuse basis functions, which would incur in an additional computational cost.

An interesting observation from Figure 1 (also Table S3 of SI) is that when ACPs are developed for minimal basis set HF-D3 and HF-3c, the MAEs of the resulting ACP-corrected methods are very similar irrespective of whether the ACPs are applied to minimal basis set HF-D3 or HF-3c. For example, hydrogen bonding interactions (*Hydrogen-bonding* subset) with HF-D3/MINIs, HF-D3/MINIX, and HF-3c have MAEs of 1.95, 1.85, and 1.14 kcal/mol. The application of ACPs brings these MAEs down to very similar values (0.74, 0.68, and 0.73 kcal/mol) even though the MAEs for the uncorrected methods were quite different. Such consistency in the ACP-corrected MAEs is observed for most of the other types of interactions, and they indicate that the ACPs developed for minimal basis set HF-D3 are, to some extent, able to mitigate basis set incompleteness errors just like the gCP[37,39] and SRB[37] corrections of HF-3c. Also, since the ACPs developed for HF-3c in most cases improve on HF-3c, ACPs provide additional error mitigation beyond that offered by gCP[37,39] and SRB[37]. On the other hand, ACPs developed for double-ζ basis set HF-D3 result in lower MAEs than those used in combination with minimal basis set HF-D3 for each interaction type, indicating that systematic improvement can be obtained by using ACPs with larger basis sets. However, going beyond double-ζ would lead to a significant increase in computational cost, and would result in methods with limited applicability for large molecular systems.[239] In this regard, a better alternative would be the development of ACPs for use with double-ζ DFT methods, an idea that is currently being explored in our group.[240]

The performance of the proposed ACP based methods for the different interaction types and especially for the *Mixed NCIs* subset makes them promising for various applications. Keeping in mind our goal of designing low-cost approaches for modeling supramolecular and biological systems, we also assembled subsets and generated reference interaction energy data for prototypical non-covalently bound complexes relevant in biochemistry. Non-covalent interactions present in such systems are covered by the *Biomolecule-Biomolecule* subset. The *Biomolecule-Biomolecule* subset contains model systems representative of nucleotide-nucleotide interactions as well as protein fragments interacting with carbohydrates, nucleotides, drugs, water, and with other proteins. Such interactions are relevant in applications like protein folding[241,242], protein structure refinement[243,244], protein-ligand binding[245–247], intercalation[248], nucleobase stacking[249], and protein hydration[250], to name a few. Uncorrected minimal or double-ζ basis set HF-D3 in general overestimate the interaction energies in this subset, and application of the ACPs reduces this overestimation and decreases the error spread and SDs. Specifically, ACPs reduce the MAEs by about 74% (minimal basis set HF-D3) and 84% (double-ζ basis set HF-D3). HF-3c errors are centered around the zero-error average line with a relatively small spread, indicating that HF-3c is well suited for the complexes present in the *Biomolecule-Biomolecule* subset. Application of ACPs to HF-3c further reduces the MAE (by about 16%) except for a few systems: some nucleotide trimers and some nucleotide-amino acid complexes.

The *Gas-Ligand* subset comprises small molecules like CO_2, CH_4, and N_2 interacting with benzene, coronene, polycyclic aromatic hydrocarbons, polyheterocyclic aromatic compounds, and other functionalized organic molecules. The complexes present in the *Gas-Ligand* subset are representative of potential applications in the areas of chemical sensing, gas storage, and gas separation.[251–253] By training our ACPs to this subset we expect to extend their applicability to the modeling of gas adsorption on various porous materials.[254] The application of ACPs decreases the MAEs of all considered methods for the *Gas-Ligand* by about 77% (minimal basis set HF-D3), 76% (double-ζ basis set HF-D3), and 18% (HF-3c). The bias of minimal or double-ζ basis set HF-D3 and HF-3c towards over-estimating the interaction energies are also reduced with the application of ACPs, resulting in lower error spread and SDs.

The *Water-Water* subset contains interaction energies of water dimers at various intermolecular separations as well as small water clusters $(H_2O)_n$ with n=3–10. Potential target applications of ACPs trained against this subset are the modeling of aqueous environments, the study of surfaces of astrochemical interest[255–260], as well as performing *ab initio* molecular dynamics simulations of water[261–264]. The ACPs improve the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods for *Water-Water* by about 68% (minimal basis set HF-D3), 78% (double-ζ basis set HF-D3), and 19% (HF-3c). Furthermore, Figure 1 shows that the large error spreads obtained with all underlying methods are significantly reduced by the ACPs, which bring the MSEs close to zero.

The last non-covalent interaction energy subset in the training set is *BFSiPSCl*, which contains complexes of monomers containing B, F, Si, P, S, and Cl. This subset extends the applicability of ACPs to systems like disulfide-linked proteins, covalent organic frameworks, functionalized silicon surfaces, and others. The application of ACPs results in a decrease in the MAEs of the *BFSiPSCl* subset by about 67%

(minimal basis set HF-D3), 70% (double-ζ basis set HF-D3), and 52% (HF-3c), with a decrease in the error spread and SDs in all cases. The drop in MAE observed for the HF-3c-ACP is more significant for this subset than all other interaction energy subsets, and the reduction is also close to that observed in the *Pnicogen-bonding* subset, indicating that perhaps the HF-3c parametrization is not as good for these systems as for the more "usual" non-covalent interactions in the previous sets.

Finally, we consider a few illustrative examples for which we compare the performance of our ACP-corrected methods with some commonly used DFT methods in combination with large basis sets. For this purpose, we use representative data sets from Mardirossian and Head-Gordon's benchmarking work[265], for which nearly complete basis set DFT results have been reported in the literature. Specifically, we use the following sets: BzDC215[127] for π-π stacking interactions, HC12[131] for aliphatic-aliphatic interactions, S66x8[144–146] and 3B-69-DIM[152] for interactions of mixed nature, SSI[154] and HSG[136,156] for biomolecule-biomolecule interactions, Water38[177] for water-water interactions, and Sulfurx8[181] for interactions involving S atoms. The reported MAEs (in kcal/mol) of the DFT methods with the very large def2-QZVPDD basis set as well as the MAEs of the ACP-corrected methods are shown in Table 3. The table shows that the ACPs reduce the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods in all the selected data sets and brings their MAE to a value close to or even lower than the almost complete basis set DFT methods. Therefore, Table 3 demonstrates that ACP-corrected methods have a performance similar to almost complete basis set DFT, but naturally at a cost that is reduced by orders of magnitude.

Table 3. Comparison of the mean absolute errors (MAEs) of various methods for selected data sets in the training set. The MAEs lower than those calculated with various DFT methods using the def2-QZVPDD basis set are highlighted in bold.

Data set[a]	**DFT functionals with def2-QZVPDD[b]**	**HF-D3/MINIs**	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
BzDC215[127]	0.41 [LC-ωPBE08-D3(BJ)]	0.92	**0.40**	0.85	**0.39**	**0.23**	0.44	1.27	**0.34**
HC12[131]	0.25 [M06-2X]	1.80	**0.20**	1.80	**0.22**	0.42	**0.25**	1.19	0.27
S66x8[144–146]	0.29 [CAM-B3LYP-D3(BJ)]	1.24	**0.25**	1.24	**0.26**	0.37	**0.27**	1.49	**0.23**
3B-69-DIM[152]	0.43 [M06-2X]	1.08	**0.42**	1.08	**0.41**	0.50	**0.42**	1.44	**0.25**
SSI[154]	0.17 [B3LYP-D3(BJ)]	0.87[c]	0.21[c]	0.87[c]	0.20[c]	0.28[c]	0.22[c]	0.76[c]	**0.15**[c]
HSG[136,156]	0.14 [B3LYP-D3(BJ)]	0.94[c]	0.18[c]	0.94[c]	0.19[c]	0.33[c]	0.19[c]	0.89[c]	**0.12**[c]

Water38[177]	2.85 [B3LYP-D3(BJ)]	30.62	**1.56**	30.62	**1.24**	7.67	**1.32**	19.51	**0.77**
Sulfurx8[181]	0.33 [BP86-D3(BJ)]	0.50	0.41	0.36	**0.29**	0.71	0.35	0.75	**0.19**
YMPJ[191]	0.99 [B97-D]	1.73	**0.97**	1.74	1.00	2.32	1.04	**0.80**	**0.57**
SCONF[129,195]	0.57 [LC-ωPBE08-D3(BJ)]	5.20	0.59	5.20	**0.54**	1.47	**0.57**	1.57	0.64
BCONF[200]	0.34 [CAM-B3LYP-D3(BJ)]	2.40	**0.29**	2.40	**0.27**	0.58	**0.27**	1.25	**0.34**
PentCONF[201]	0.15 [B3LYP-D3(BJ)]	0.96	0.16	0.96	**0.15**	0.55	0.21	0.47	**0.15**

[a)] details about the data sets can be found in Table S1 of the Supporting Information, [b)] from Reference 265, [c)] only non-negatively charged complexes

(ii) Molecular conformational energies

The purpose of the molecular conformational energy subsets of our training set is to inform the ACPs regarding how the potential energy surfaces of various molecules depend on the changes in rotatable bonds and torsional angles due to effects like π-conjugation, steric interactions, intramolecular hydrogen-bonding, and electron repulsion. Our *Small molecule* conformational energy subset is a good representative of such interactions that can be used to assess the performance of ACPs for conformational energies. The application of ACPs to the *Small molecule* subset leads to a reduction in the MAEs of about 55% (minimal and double-ζ HF-D3) and 37% (HF-3c), yielding MAEs ranging between 1.46–2.18 kcal/mol for ACPs with minimal or double-ζ basis set HF-D3 and 2.34 kcal/mol for ACPs with HF-3c. As seen in Figure 2, the spread of errors and SDs of the uncorrected methods is quite large: HF-D3/MINIs, for example, yields errors spanning -35 to +40 kcal/mol and an SD of 7.18 kcal/mol. The ACPs reduce the error spread of HF-D3/MINIs to about -20 to +30 kcal/mol and the SD to 3.31 kcal/mol. Similar observations can also be made for the HF-D3/MINIX, HF-D3/6-31G*, and HF-3c methods.

Conformers in the *Negatively charged* subset have an overall negative charge, which, as mentioned previously, is problematic for the minimal and double-ζ basis sets used in this work. Similar to the *Anionic* subset of non-covalent interaction energies, all uncorrected methods are inadequate for conformational energies of negatively charged species, which results in MAEs for the *Negatively charged* subset higher than for the other molecular conformational energy subsets. However, the application of ACPs yields relatively low MAEs (0.64–1.28 kcal/mol) compared to the uncorrected methods (1.08–3.01 kcal/mol), indicating that ACP-corrected methods are better suited to model molecular conformational energies of anionic systems.

Some other molecular conformational energy subsets used in the training set include the *Biomolecule*, *Hydrocarbon*, and *$(H_2O)_{11}$* subsets. The *Biomolecule* subset contains conformers of molecules that are biologically relevant, like proteins, DNA, RNA, and carbohydrates. The *Hydrocarbon* subset incorporates model systems of aliphatic nature relevant in lipids, polymers, fossil fuels, and organic chemistry. The *$(H_2O)_{11}$* contains systems relevant in the description of aqueous media.[266–268] The application of ACPs to the *Biomolecule*, *Hydrocarbon*, and *$(H_2O)_{11}$* subsets results in a significant drop in MAEs relative to the underlying methods, by about 61–86% (minimal basis set HF-D3), 50–85% (double-ζ basis set HF-D3), and 48–63% (HF-3c). Figure 2 shows that the error spread, SDs, and MSEs of minimal or double-ζ basis set HF-D3 and HF-3c methods are all reduced upon application of ACPs.

Same as for non-covalent interaction energies, Figure 2 shows that the application of ACPs brings down the MAEs of various molecular conformational energy subsets to similar values irrespective of whether the ACPs are applied to minimal basis set HF-D3 or HF-3c. For example, the MAEs of HF-D3/MINIs and HF-3c for the *Biomolecule* subset are 2.69 kcal/mol and 2.14 kcal/mol, respectively. Application of the corresponding ACPs results in a reduction of the MAEs to the very similar values of 1.06 kcal/mol and 1.11 kcal/mol. Like non-covalent interaction energies, the ACP for HF-D3/6-31G* yields lower MAEs compared to minimal basis set HF-D3 or HF-3c. For molecular conformational energies, the MAEs of the HF-3c-ACP method are notably lower (by about 31–61%) than that of HF-3c. Therefore, the ACPs developed for HF-3c offer a significant improvement beyond gCP[37,39] and SRB[37] for molecular conformation energies.

Finally, we compare the performance of our ACP-corrected methods relative to nearly complete basis set DFT results from the literature. For this, we consider a few representative data sets such as YMPJ[191] for amino acid conformers, SCONF[129,195] for carbohydrate-like conformers, BCONF[200] for butane-1,2-diol conformers, and PentCONF[201] for pentane conformers. The MAEs (in kcal/mol) of the DFT/def2-QZVPDD and the ACP-corrected methods for these data sets are shown in Table 3. Similar to non-covalent interaction energies, the application of ACPs reduces the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods in all the selected data sets and are close to or even lower than the MAEs reported for the various functionals. Table 3 demonstrates that the proposed ACP-corrected methods are able to predict the conformational energies with an accuracy similar to large basis set DFT methods at a significantly lower computational cost.

(iii) Molecular deformation energies

The *Deformation* subset of the ACP training set contains energy differences between a molecule at its equilibrium geometry and the same molecule deformed along its various normal modes. Our intention with this subset is to improve the description of the molecular potential energy surfaces around the equilibrium geometries, and consequently improve the prediction of bond lengths and molecular geometries in general.

The fact that small basis set HF methods predict erroneous geometries is important for the study of large molecules like proteins, as discussed by Kulik *et al.*[269] and Schmitz *et al.*[270] Their findings suggest that small basis set HF methods without any correction give, in general, quite inaccurate protein structures. This is likely the reason why the HF-3c[37] method employs the semi-empirical SRB correction. In fact, the SRB correction itself was parametrized by fitting to the geometries of 107 small organic molecules computed at a higher level of theory.

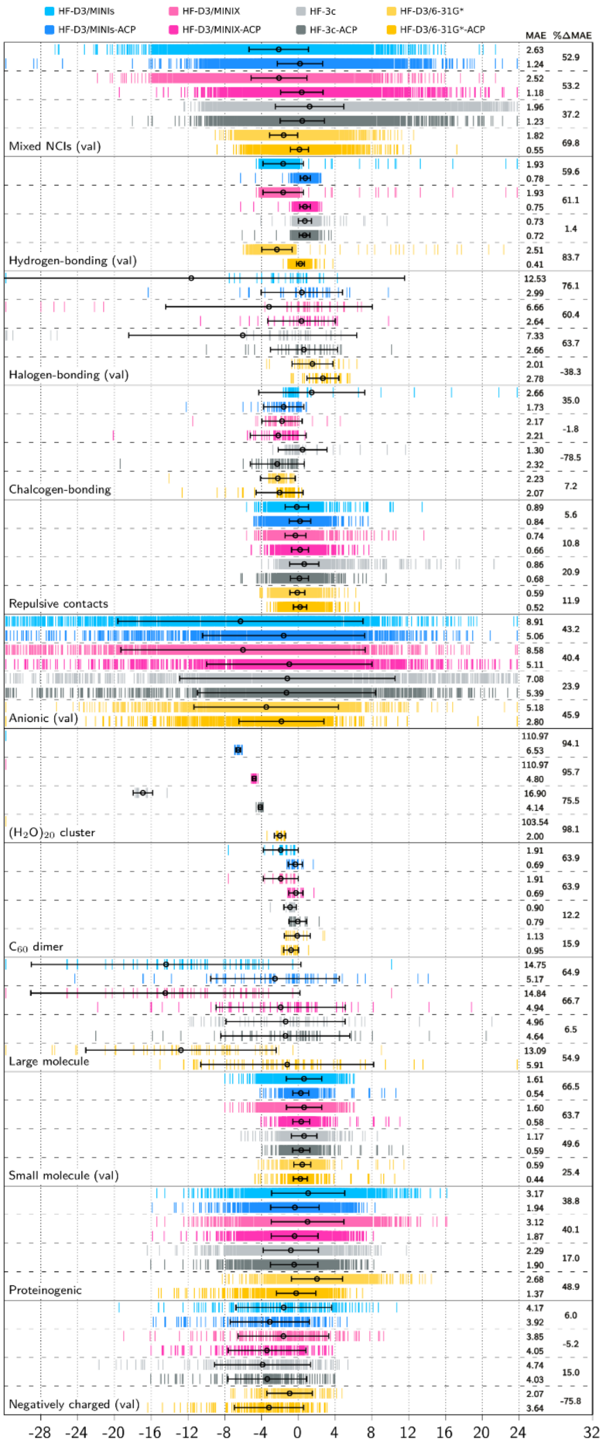
The performance of our ACP-corrected methods for actual geometry optimizations is discussed in Section 3.4. The results for the *Deformation* subset in Figure 2 already suggest that ACPs improve the prediction of molecular geometries substantially. On application of ACPs, the MAEs of all four methods for this subset are reduced by about 35–51% (minimal and double-ζ HF-D3) and 31% (HF-3c). Figure 2 shows that even though the decrease in the spread of errors using ACPs is modest, the under-estimation in the prediction of molecular deformation energies of the underlying methods is greatly corrected by the ACPs, and the MSEs as well as the SDs decrease. Molecular deformations that are farthest from equilibrium have relatively high reference energies and result in errors higher than ±5 kcal/mol. Nevertheless, the application of ACPs predict individual errors that are lower than ±5 kcal/mol for 85% (or more) of the data points out of a total of 10,288.

3.2 Performance of ACPs for the validation set

The results regarding the application of ACPs to the systems in the validation set (Table 2) are presented in Figure 3. The figure includes the signed error distribution, MSEs, MAEs, and SDs of minimal or double-ζ basis set HF-D3 and HF-3c methods with and without ACPs, as well as the percentage change in MAEs upon application of ACPs (%ΔMAE) for each method. A detailed breakdown of the errors by method and subset can be found in Table S4 of the SI.

Figure 3. Error distribution (relative to the reference data, kcal/mol) associated with the validation set (see Table 2). The top nine panels represent non-covalent interaction energy subsets while the bottom three panels represent molecular conformational energy subsets. Methods shown include HF-D3/MINIs (light blue), HF-D3/MINIs-ACP (blue), HF-D3/MINIX (light pink), HF-D3/MINIX-ACP (pink), HF-3c (light

grey), HF-3c-ACP (grey), HF-D3/6-31G* (light yellow), and HF-D3/6-31G*-ACP (yellow). The black circles represent the mean signed errors (MSEs, kcal/mol) and the black error bars are the standard deviations of the error (SDs, kcal/mol). The numbers on the right hand side of each panel are the mean absolute errors (MAEs, kcal/mol) and the percentage change in MAEs upon the application of ACPs (%ΔMAE) for each method. %ΔMAE is defined as [MAE(base method) – MAE(ACP-corrected method)] / MAE(base method) x 100%. The X-axis has been capped at -32 (left) and +24 kcal/mol (right) for clarity. The black circles and error bars of HF-D3/MINIs, HF-D3/MINIX, and HF-D3/6-31G* methods for *$(H_2O)_{20}$ cluster* subset are absent from the figure due to MAEs being higher than 100 kcal/mol.



The results show that the when the ACPs are applied to the *Mixed NCIs* validation subset ("*Mixed NCIs (val)*"), the MAEs of all methods decrease, by 53% (minimal basis set HF-D3), 70% (double-ζ basis set HF-D3), and 37% (HF-3c). This reduction in MAEs is similar to what was observed for the mixed character interactions in the training set, confirming the robustness of the ACPs for non-covalent interactions when applied to systems outside the training set. One particular data set present in the *Mixed NCIs (val)* subset is BlindNCI[205]. Taylor *et al.*[205] reported an MAE of 0.34 kcal/mol with the M11/aug-cc-pVTZ method for the BlindNCI data set, which is almost equivalent to the MAE obtained with HF-D3/MINIs-ACP, HF-D3/MINIX-ACP, and HF-3c-ACP, and almost 28% higher than HF-D3/6-31G*-ACP (see Table 4). As in the training set, the performance of ACPs in the description of non-covalent interaction energies in the validation set is similar in quality to large basis set DFT methods.

Table 4. Comparison of the mean absolute errors (MAEs) of various methods for selected data sets in the validation set. The MAEs that are lower than the DFT methods are highlighted in bold.

Data set[a]	**DFT functional with a large basis set**	**HF-D3/MINIs**	**HF-D3/MINIs-ACP**	**HF-D3/MINIX**	**HF-D3/MINIX-ACP**	**HF-3c**	**HF-3c-ACP**	**HF-D3/6-31G***	**HF-D3/6-31G*-ACP**
BlindNCI[205]	0.34[b] [M11/aug-cc-pVTZ]	1.00	**0.34**	1.00	0.36	0.38	0.35	1.10	**0.25**
CE20[208,209]	1.72[c] [M06-2X/6-311+G(3df,2p)]	16.64	**1.63**	16.64	**1.55**	3.32	1.84	11.10	**1.31**
WaterOrg[210]	0.44[d] [B3LYP-D3(BJ)/6-31+G**-BSIP]	1.81	0.77	1.81	0.75	0.71	0.71	2.44	**0.40**
CHAL336[212]	1.18[e] [BLYP-D3(BJ)/ma-def2-QZVPP]	2.66[l]	1.73[l]	2.17[l]	2.21[l]	1.30[l]	2.32[l]	2.23[l]	2.07[l]
H2O20Bind10[216]	8.84[f] [B3LYP-D3(BJ)/def2-QZVPPD]	110.97	**6.53**	110.97	**4.80**	16.90	**4.14**	103.54	**2.00**
C60dimer[221]	2.85[g] [BP86-D3(BJ)/def2-TZVP]	**1.91**	**0.69**	**1.91**	**0.69**	**0.90**	**0.79**	**1.13**	**0.95**

L7[222,223]	1.62[h] [B3LYP-NL/def2-TZVP]	3.64	**1.42**	3.64	**1.39**	**1.37**	**1.48**	3.61	**0.82**
S12L[9,11,223]	6.44[h] [BLYP-NL/def2-TZVP]	14.63	**6.41**	14.65	**5.96**	**6.05**	**5.58**	13.51	**6.22**
S30L[220]	6.60[i] [PBE-D3/CBS]	13.07	**5.65**	13.25	**5.23**	**4.80**	**4.74**	11.71	**5.34**
Ni2021[224]	3.20[j,k] [B3LYP-D3(BJ)/triple-ζ]	25.53	5.60	25.53	5.89	6.74	5.98	22.01	10.05

[a] details about the data sets can be found in Table S2 of the Supporting Information, [b] from Reference 205, [c] from Reference 208, [d] from Reference 210, [e] from Reference 212, [f] from Reference 265, [g] from Reference 221, [h] from Reference 195, [i] from Reference 55, [j] from Reference 224, [k] aug-cc-pVTZ basis set for six systems and cc-pVTZ basis set for other seven systems, [l] only non-negatively charged complexes

Two data sets (CE20[208,209], WaterOrg[210]) were used to validate the ACPs for hydrogen bonding interactions. The MAEs of minimal or double-ζ basis set HF-D3 methods for the *Hydrogen-bonding* validation subset ("*Hydrogen-bonding (val)*") are improved on applying the ACPs by 61% (minimal basis set HF-D3) and 84% (double-ζ basis set HF-D3). As in the case of the mixed NCIs, this improvement is close to the one observed in the training set. On the other hand, application of the ACPs to HF-3c neither improves nor deteriorates the MAE for *Hydrogen-bonding (val),* and the MAE is almost the same as the MAE for hydrogen bonding interactions (0.73 kcal/mol) in the training set. As observed for the training set, ACPs improve methods whose errors are higher, and barely affect methods that already have low MAEs. Comparing to the results obtained with DFT and a large basis set (Table 4), the MAEs of the CE20 data set with ACPs are close to or lower than most of the benchmarked DFT methods with a 6-311+G(3df,2p) basis set in the work of Chan *et al.*[208] Also, the B3LYP-D3(BJ)/6-31+G**-BSIP method that yields results that are close to B3LYP-D3(BJ)/aug-cc-pVQZ has an MAE of 0.44 kcal/mol for WaterOrg data set, which is close that predicted via HF-D3/6-31G*-ACP approach.[210]

The assessment of the ACPs on the validation set also helps us understand what types of interactions are poorly represented in the training set. Based on the analysis performed with the validation set, these interactions are halogen bonding, chalcogen bonding, and close contact repulsions, as discussed below. It should be noted that an assessment of ACPs on interaction types such as π-π stacking, pnicogen bonding, and hydrophobic interactions, discussed earlier for the training set, was not possible in the validation stage because of the scarcity of high-level reference data in the literature.

For the halogen bonding interactions in the validation set ("*Halogen-bonding (val)*" subset), all methods in absence of ACPs show a large over-estimation of the interaction energies. The application of ACPs correct for this over-estimation and lead to a decrease in the MAEs by about 60–76%. Figure 3 shows that the MAEs of the minimal basis set HF-D3 and HF-3c methods without ACPs are almost three times higher than HF-D3/6-31G* (2.01 kcal/mol). Figure 3 also shows that HF-D3/6-31G* has a positive MSE, suggesting an under-estimation in the interaction energies for halogen bonding interactions. This observation for HF-D3/6-31G* is opposite to what was found in the training set. Nonetheless, the spread of errors is decreased when the ACP corrections are used, including HF-D3/6-31G*-ACP, leading to lower SDs than without ACPs.

Model systems representative of chalcogen bonding interactions ("*Chalcogen-bonding*" subset) were absent from the training set. As expected, the improvements in the MAEs when ACPs are applied are not significant for the minimal or double-ζ basis set HF-D3 methods. At the same time, ACPs applied to HF-3c over-estimate the interaction energies and lead to an increase in the MSE and MAE. A slight improvement in the description of chalcogen bonding interactions is observed with ACPs for minimal or double-ζ basis set HF-D3 methods, probably due to the presence of O and S containing complexes in the training set that are not purely chalcogen-bonded. This suggests that increasing the representation of such interactions in the training set could improve the performance and applicability of ACPs. Chalcogen bonding interactions are a difficult test not only for the methods considered in this work but also for many other electronic structure methods. For example, several dispersion-corrected DFT methods tested with ma-def2-QZVPP basis set have MAEs above 1 kcal/mol for the entire CHAL336[212] data set.[212] In this context, Figure 3 suggests that the HF-3c method is the best suited among the minimal basis set HF methods for modeling chalcogen bonding interactions.

Steric repulsive interactions, even though found in some molecules that are forced to be in close contact due to the presence of other attractive interactions or external pressure, seldom occur naturally.[271] Repulsive interactions ("*Repulsive contacts*" subset) are captured well by the minimal or double-ζ basis set HF-D3 and HF-3c methods (MAEs of 0.59–0.89 kcal/mol) and only small improvements are seen with the application of ACPs. Specific subsets for repulsive interactions were missing from our training set. Still, the slight reduction in the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods observed with ACPs probably comes from some of the data sets in our training set that contain some data points with repulsive character (e.g. S22x5[136,142,143], S66x8[144–146], S66a8[145], A21x12[3,147,148], and NBC10ext[128,136,149–151]). It should be noted that the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods with and without ACPs are lower than the newly reparametrized PM6-D3H4R, DFTB3-D3H4R, PM6-D3H4X, and DFTB3-D3H4X methods (MAEs of 0.94–1.48 kcal/mol). These methods attempt to capture the repulsive interactions via the use of a repulsive energy correction term (parametrized against R160x6[213] and R739x5[214] data sets that constitute the validation *Repulsive contacts* subset) specifically designed for PM6 and DFTB3 methods with the D3H4 correction for dispersion and hydrogen-bonding.[214,272]

Next, we turn our attention to the *$(H_2O)_{20}$ cluster* (H2O20Bind10[216] data set), *C_{60} dimer* (C60dimer[221] data set), and *Large molecule* (L7[13,222,223], S12L[9,11,223], S30L[220], Ni2021[224] data sets) subsets. These subsets are a good test for ACPs as they contain non-covalently bound complexes that are relatively large and at the same time feature multiple co-operative interactions including one or more of hydrogen bonds, halogen bonds, π-π stacking, H-π, ion-dipole, dispersion, etc. The systems present in *$(H_2O)_{20}$ cluster*, *C_{60} dimer*, and *Large molecule* subsets are known to be challenging not only for minimal or double-ζ basis set HF-D3 and HF-3c methods but also for many other electronic structure methods. The absolute reference interaction energies of many complexes in these subsets range from 25 kcal/mol to 416 kcal/mol. It should also be noted that due to the large size of the systems, the most feasible way to generate the reference data of such systems is via the use of methods like CCSD(T)-F12a (H2O20Bind10), DLPNO-CEPA/1 (C60dimer), and CIM-DLPNO-CCSD(T) (Ni2021) or back-correction of experimental data (*S12L* and *S30L*). The reference data for these data sets are expected to be of lower quality than the others, which are typically calculated at CCSD(T)/CBS. The application of ACPs to minimal or double-ζ basis set HF-D3 and HF-3c methods on *$(H_2O)_{20}$ cluster*, *C_{60} dimer*, and *Large molecule* show a general improvement in the MAEs of the underlying methods. ACPs for minimal basis set HF-D3 reduces the MAEs by about 64–94%, while that for double-ζ basis set HF-D3 reduces the MAEs by about 16–98%. ACPs for HF-3c also reduce the MAEs by about 7–76%. The systems with large errors, higher than ±8 kcal/mol even after the use of ACPs, are, as expected, those that have very large reference energies. Table 4 shows a comparison of MAEs of various HF based approaches with large basis set DFT methods for the H2OBind10, C60dimer, L7, S12L, S30L, and Ni2021 data sets. It can be seen that, with the exception of Ni2021 data set, the MAEs in Table 4 for all other data sets shows that ACPs have a performance similar to large basis set DFT, making the approach particularly promising for modeling interaction energies in large molecular systems.

We now consider the results for the conformational energies evaluated with the subsets "*Small molecule (val)*" and "*Proteinogenic*". The *Small molecule (val)* subset contains conformational energies of various small organic and biaryl drug-like molecules. On the other hand, the *Proteinogenic* subset contains a collection of polypeptide conformers like tripeptides, peptides with disulfide linkages, macrocyclic peptides, and peptide sequences with associated bio-functionality. For the *Small molecule (val)* subset, application of ACPs brings down the MAEs of minimal or double-ζ basis set HF-D3 and HF-3c methods by about 66% (minimal basis set HF-D3), 25% (double-ζ basis set HF-D3), and 50% (HF-3c). For the *Proteinogenic* subset, the improvement in MAEs seen on the application of ACPs is about 40% (minimal basis set HF-D3), 49% (double-ζ basis set HF-D3), and 17% (HF-3c). Figure 3 shows that the spread of errors is more or less symmetric about the zero-error average line, except for some systems with errors higher than ±8 kcal/mol, corresponding to the disulfide linkages which were not present in the training set.

For non-covalent interaction energies of anionic interactions ("*Anionic (val)*" subset), although the application of ACPs leads to a reduction in MAEs of minimal or double-ζ HF-D3 and HF-3c methods by about 24–43%, these MAEs still range between 2.65–5.21 kcal/mol making the overall approach not usable for modeling anionic interactions. The good performance of ACPs for most of the conformational energies in the training and validation sets does not translate to the *Negatively charged (val)* subset. Nevertheless, upon application of ACPs, the MAEs (in kcal/mol) of HF-D3/MINIs is slightly reduced from 4.17 to 3.92 and from 4.74 to 4.03 for HF-3c. The respective MAEs of HF-D3/MINIX-ACP and HF-D3/6-31G*-ACP increase by about 5% and 76%. Similar to the results in the training set, the poor results in the *Anionic (val)* and *Negatively charged (val)* subsets are another indication of the serious problems associated with using minimal and double-ζ basis sets without diffuse functions for negatively charged systems. A possible solution to deal with anionic systems would be to develop ACPs for minimally augmented basis sets[273].

3.3 Performance of ACPs for molecular geometries

One attractive possible use of ACPs is for fast geometry optimizations. To gauge the performance of the various HF based methods for this task, we compared and analyzed the structures obtained after energy relaxation with those obtained using dispersion-corrected DFT methods with large basis sets. The 296 structures used for the test contain both non-covalently bound complexes and single molecules, ranging in size between 2 and 205 atoms. For single molecule structures, we used the equilibrium geometries of small organic molecules taken from our *MOLdef* data set, a variety of organic molecules taken from Reference 274, selected structures from LB12[57] and CLB18[275] data sets, and polypeptide structures from Reference 270. For non-covalently bound structures, we selected the equilibrium complex structures from the A21[148], S66[144], L7[222], and S30L[220] data sets. Wherever high-level geometries were not available, we obtained reference geometries using dispersion-corrected DFT and a reasonably large basis set (CAM-B3LYP-D3(BJ)/6-311++G**) with the "tight" convergence criteria in *Gaussian-16*. All the optimized and reference geometries used for testing are provided in the SI. We used Kabsch's algorithm[276] to compare the optimized structures with the reference.

The results are summarized in Table 5. The table shows that the root-mean-square-deviation (RMSD) of the atomic coordinates for the small basis set HF based methods with ACPs are generally lower than those without ACPs, and this decrease happens for single molecules and non-covalently bound complexes, including charged systems. These results indicate that ACPs are generally able to yield better geometries than the uncorrected methods. The RMSD values for the individual methods and geometries can be found in Table S5 of SI.

To examine the source of the improvement in the molecular geometries upon application of ACPs, we calculated the average error in the intermolecular separation distances for the dimer complexes. We also compared the average error in a few selected bond lengths and angles for the single molecules. Table 5 shows that, on average, the intermolecular separation distances are improved, and the under-estimation in the separation distances yielded by the small basis set HF based methods is corrected by the ACPs, leading to better geometries for non-covalently bound complexes. Furthermore, the average deviations in bond lengths presented in Table 5 indicate that the average deviation of the small basis set HF based methods is between 0.002 Å to 0.092 Å for the selected bonds. It can be seen that the inclusion of ACPs with small basis set HF based methods leads to a better description of bond lengths as the polarity of a bond increases, leading to lower errors in bond lengths (except for HF-3c-ACP). A general improvement in the prediction of bond angles is also observed on application of ACPs. The combination of low average errors in bond lengths and angles leads to better overall geometries of single molecule structures. Despite overall good geometries for both single molecules and complexes, upon application of ACPs some tested geometries tend to have slightly higher RMSDs (greater than 0.7 Å) than the uncorrected methods due to slight deterioration in the bond lengths of C-H and C-C bonds. Such deviations relative to the reference geometry are visible in Table S5 of SI for some purely planar systems and peptides with highly flexible backbones.

Table 5. Results of various methods for equilibrium structures. RMSD is the root-mean-square deviation in the atomic coordinates, MAE is the mean absolute error, and MSE is the mean signed error.

	HF-D3/MINIs	HF-D3/MINIs-ACP	HF-D3/MINIX	HF-D3/MINIX-ACP	HF-3c	HF-3c-ACP	HF-D3/6-31G*	HF-D3/6-31G*-ACP
Overall geometry:								
Mean RMSD (Å) *(complexes)*	0.326	0.289	0.326	0.284	0.223	0.211	0.320	0.229
Mean RMSD (Å) *(single molecules)*	0.205	0.167	0.182	0.163	0.158	0.187	0.084	0.049
Mean RMSD (Å) *(charged single molecules)*	0.759	0.533	0.750	0.513	0.483	0.684	0.516	0.208
Overall mean RMSD (Å)	0.254	0.217	0.240	0.212	0.184	0.196	0.180	0.122
Inter-molecular separation distance[a,b]:								
MAE (Å)	0.222	0.124	0.221	0.126	0.112	0.111	0.157	0.084
MSE (Å)	-0.171	0.015	-0.172	0.017	-0.022	0.002	-0.087	0.005
Selected bond lengths:								
C-H bond (MAE / MSE) (Å)	0.005 / -0.004	0.014 / 0.014	0.005 / -0.004	0.017 / 0.017	0.007 / -0.006	0.018 / 0.018	0.010 / -0.010	0.002 / 0.001
C-C bond (MAE / MSE) (Å)	0.019 / 0.018	0.034 / -0.033	0.019 / 0.018	0.028 / -0.024	0.016 / 0.013	0.028 / -0.023	0.011 / -0.011	0.005 / 0.003
C-N bond (MAE / MSE) (Å)	0.036 / 0.034	0.030 / -0.029	0.036 / 0.034	0.031 / -0.029	0.018 / 0.011	0.033 / -0.030	0.013 / -0.012	0.010 / -0.009
C-O bond (MAE / MSE) (Å)	0.057 / 0.057	0.012 / -0.001	0.057 / 0.057	0.011 / -0.003	0.011 / 0.001	0.014 / -0.008	0.017 / -0.017	0.005 / -0.002
C-F bond (MAE / MSE) (Å)	0.068 / 0.068	0.020 / -0.014	0.068 / 0.068	0.022 / -0.019	0.010 / -0.004	0.030 / -0.030	0.017 / 0.017	0.006 / -0.001
C-Cl bond (MAE / MSE) (Å)	0.092 / 0.092	0.042 / -0.042	0.026 / 0.026	0.039 / -0.039	0.015 / 0.015	0.093 / -0.078	0.017 / -0.017	0.016 / 0.016
Selected bond angles:								
C-C-H angle *(MAE / MSE)*	0.546 / -0.039	0.365 / 0.070	0.554 / -0.043	0.378 / 0.049	0.421 / -0.063	0.381 / 0.059	0.228 / -0.001	0.197 / 0.007
C-C-C angle *(MAE / MSE)*	0.792 / -0.434	0.404 / -0.132	0.781 / -0.445	0.443 / -0.216	0.597 / -0.340	0.460 / -0.220	0.336 / -0.193	0.274 / -0.128
C-C-N angle *(MAE / MSE)*	1.385 / -0.444	1.026 / -0.092	1.381 / -0.456	1.114 / -0.350	1.280 / -0.392	1.108 / -0.404	0.583 / -0.171	0.366 / -0.071
C-C-O angle *(MAE / MSE)*	1.402 / 0.503	0.891 / 0.094	1.325 / 0.491	0.975 / 0.171	1.125 / 0.568	1.108 / 0.248	0.455 / -0.025	0.328 / 0.145
C-C-F angle *(MAE / MSE)*	0.265 / 0.086	0.218 / 0.045	0.233 / 0.078	0.183 / 0.017	0.267 / 0.154	0.253 / -0.023	0.106 / 0.045	0.224 / 0.133
C-C-Cl angle *(MAE / MSE)*	0.493 / -0.374	0.377 / 0.078	0.327 / 0.204	0.422 / 0.228	0.435 / 0.323	0.407 / 0.264	0.124 / 0.025	0.305 / -0.121

[a)] calculated as the distance between the centers of mass of each monomer.

[b)] excluding the geometries of the non-dimer complexes from the L7 data set for simplicity.

3.4 Applications of ACPs developed for HF-3c

This section explores the use of HF-3c-ACP for modeling systems where most but not all atoms in the system have an associated ACP. For the atoms for which ACPs are not available, our intention is that

HF-3c, which is the overall best of the underlying methods in this work will still give a reasonable description of the system. A particular example of an application where HF-3c-ACP could be used is in modeling metalloproteins where the atoms for which ACPs are unavailable are the metal ion(s) in the active site.

We calculated the interaction energies of two systems from Reference 224 using HF-3c and HF-3c-ACP to demonstrate the above idea. These two systems represent the adsorption of ethanol and benzene with different-sized cluster models of zeolite ZSM-5.[277,278] The ZSM-5 zeolite complexes are mainly composed of H, C, O, and Si atoms for which ACPs are available. However, they also contain an additional aluminum atom. For the ZSM-5 zeolite complexes, the high-level (CIM-DLPNO-CCSD(T)) interaction energy reported in Reference 224 is -12.35 kcal/mol (benzene and zeolite or Benzene-ZSM5) and -36.55 kcal/mol (ethanol and zeolite or Ethanol-ZSM5). The HF-3c approach overestimates the interaction energies and yields -21.38 kcal/mol for Benzene-ZSM5 and -43.86 kcal/mol for Ethanol-ZSM5. The ACPs help reduce the interaction energies over-estimated by HF-3c and brings them closer to the reference: The corrected interaction energies predicted by HF-3c-ACP are -19.41 kcal/mol and -39.75 kcal/mol, respectively.

We now explore the same idea by taking some subsets of the validation set and purposefully applying only part of the available ACPs so that not all atoms in the system have an associated correction. Table 6 presents a summary of the MAEs using various methods for two data sets from the validation set: the DES15K[206] set of non-covalent interaction energies (11474 data points) and the Torsion30[228] set of molecular conformational energies (2107 data points). The table shows that the MAEs of the HF/MINIX method are 4.83 and 0.92 kcal/mol for the DES15K and Torsion30 data sets, respectively. Using the HF-3c method, the MAE decreases for DES15K (2.14 kcal/mol) and increases to 1.18 kcal/mol for Torsion30. Table 6 shows that using HF-3c-ACP but applying ACPs only to hydrogen and one of the non-hydrogen atoms indicates that ACPs improve the results progressively as the correction is applied to more atoms in the system. If only hydrogen and one of the non-hydrogen atoms are corrected, the performance of HF-3c-ACP is similar to HF-3c. If the correction is applied to hydrogen and two non-hydrogen atoms, most atoms are corrected and the MAEs decrease substantially and resemble HF-3c-ACP, in which all atoms receive an ACP. Therefore, we conclude that the application of ACPs is, in general, beneficial, and greater performance is obtained as more atoms receive an ACP, so the use of ACPs is recommended even in systems containing atoms for which ACPs are not available.

Table 6. Mean absolute error (MAE) for the DES15K data set of non-covalent interaction energies and Torsion30 of conformational energies for HF/MINIX, HF-3c, and HF-3c with application of ACPs to various atoms.

Subset	Mean absolute error (in kcal/mol)									
	HF/MINIX	**HF-3c**	**HF-3c with H & O ACPs**	**HF-3c with H & N ACPs**	**HF-3c with H & C ACPs**	**HF-3c with H, N, & O ACPs**	**HF-3c with H, C, & O ACPs**	**HF-3c with H, C, & N ACPs**	**HF-3c with H, C, N, & O ACPs**	**HF-3c-ACP**
DES15K[a]	4.83	2.14	1.95	2.24	2.17	1.83	1.78	2.07	1.68	1.32
Torsion30[b]	0.92	1.18	1.20	1.09	0.70	1.06	0.70	0.58	0.57	0.59

[a] atom frequency: H = 138132, C = 62376, O = 11604, N = 9250, other (F, P, S, Cl) = 7936.

[b] atom frequency: C = 23744, H = 18473, N = 4143, O = 985, other (S) = 70.

4. Summary and Outlook

An important field of research in modern computational chemistry is the development of new quantum mechanical methods that are accurate and can be applied to model large molecular systems. Small basis set Hartree–Fock (HF) methods are orders of magnitude less expensive than more accurate nearly complete basis set wavefunction theory or DFT methods, but suffer from basis set incompleteness error and lack of electronic correlation. Provided these shortcomings can be addressed, such methods could be applied for modeling large molecular systems as well as for routine applications of fast geometry optimizations, conformational exploration, and prediction of non-covalent interaction strengths.

In this work, we show that HF with small and minimal basis sets can be effectively corrected by applying atom-centered potentials (ACPs, one-electron potentials similar to effective-core potentials) that are designed to correct for the inaccuracies in the underlying method. Four new sets of ACPs were developed for use with HF-D3 and small basis sets (MINIs, MINIX, 6-31G*) and HF-3c. The advantages of ACPs include that they can be used in most computational chemistry software packages without changes to the code and that they incur only a modest computational cost. The ACPs developed in this work apply to ten elements (H, B, C, N, O, F, Si, P, S, Cl), and our purpose is that the presented ACPs serve to address problems in organic chemistry and biochemistry. For the occasional system containing atoms for which no ACPs are available, we have shown that the improvement of the performance of the underlying method is progressive with the number of atoms where ACPs have been applied. Therefore, the use of ACPs is beneficial even if some atoms are not corrected, and in this case, we recommend the use of HF-3c-ACP. We anticipate that the ACP based approaches developed in this work will allow efficient and accurate modeling of biomolecules such as proteins, nucleic acids, carbohydrates, lipids, and other molecules containing B, Si, and halogen atoms such as covalent organic frameworks, functionalized polyaromatic hydrocarbons, functionalized silicon surfaces, and more.

The ACPs were developed by using a large training set of 73832 data points calculated at a very high level of theory (CCSD(T)/CBS, in general). The training set contains a mixture of non-covalent interaction energies, molecular conformational energies, and molecular deformation energies. We expected that the size of the training set ensures the robustness and applicability of the ACPs. To test this, we validated the new ACPs on a validation set with 32047 data points. The assessment of minimal and double-ζ basis set HF-D3 and HF-3c methods, before and after the application of their corresponding ACPs showed that ACPs lower the MAEs of most subsets in the training set and that this good performance is carried over to the validation set with approximately the same performance in terms of average error. Relative to the uncorrected methods, ACP-corrected approaches improve the prediction of non-covalent interaction energies and molecular conformational energies. Furthermore, the addition of molecular deformation energies to the training set results in an improvement of the equilibrium molecular structures upon application of the ACPs. However, ACP-corrected methods showed relatively poor performance for some interaction types that were not part of the training set, such as chalcogen bonding or repulsive contacts, indicating that more diverse systems need to be included in the training set for greater robustness of the resulting methods.

Our analysis of representative data sets indicates that our ACP-corrected methods yield results similar to almost complete basis set DFT methods, naturally at a much lower computational cost. Nonetheless, there remains a limitation regarding the description of negatively charged systems probably caused by the lack of diffuse functions in the basis sets employed. In spite of this, our ACPs offer a modest improvement even in this case. ACPs for small basis sets that include some diffuse functions are currently under development. We are also currently working on expanding the set of ACPs to DFT-D3 methods with small basis sets for prediction of accurate thermochemical properties along with non-covalent properties. Despite the limitation, we have shown that ACPs provide a way of developing methods that combine low-cost with robustness and wide applicability. We anticipate that ACPs will be useful to practitioners interested in modeling large systems or other time-intensive applications.

Acknowledgements

We thank the Natural Sciences and Engineering Research Council of Canada, Canadian Foundation for Innovation, B.C. Knowledge Development Fund, and the University of British Columbia for financial support. This research was enabled by the computational resource support provided by WestGrid, Compute Canada, and the University of British Columbia's advanced research computing services. AOR thanks the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (AEI) for financial support (projects PGC2018-097520-A-100 and RED2018-102612-T), and the Spanish MINECO for a Ramón y Cajal fellowship (RyC-2016-20301).

Supporting information

The supporting Information (SI) is available free of charge on the ACS publications website at DOI: http://dx.doi.org/xx.xxxx/acs.jctc.xxxxxxxx.

Sample Gaussian16 input file demonstrating the use of ACPs, formulas for the error measures, extrapolation scheme used for generation of new reference data, tables listing the details of data sets in the ACP training and validation set, tables associated with the detailed error analysis of training and validation sets, comparison of root-mean-square deviation in geometries, and comparison of relative single-point energy calculation time. (PDF)

Basis set and ACP files, detailed results, database files, and Cartesian coordinates used in the study. (ZIP)

Conflicts of interest

Authors declare no conflicts of interest.

References

(1) Hohenstein, E. G.; Sherrill, C. D. Wavefunction Methods for Noncovalent Interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 304–326.

(2) Bachrach, S. M. Challenges in Computational Organic Chemistry. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 482–487.

(3) 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.

(4) Al-Hamdani, Y. S.; Tkatchenko, A. Understanding Non-Covalent Interactions in Larger Molecular Complexes from First Principles. *J. Chem. Phys.* **2019**, *150*, 10901.

(5) Helgaker, T.; Jørgensen, P.; Olsen, J. Calibration of the Electronic-Structure Models. In *Molecular Electronic-Structure Theory*; John Wiley & Sons, Inc.: Chichester, UK, 2014; pp 817–883.

(6) Cui, Q. Perspective: Quantum Mechanical Methods in Biochemistry and Biophysics. *J. Chem. Phys.* **2016**, *145*, 140901.

(7) Ratcliff, L. E.; Mohr, S.; Huhs, G.; Deutsch, T.; Masella, M.; Genovese, L. Challenges in Large Scale Quantum Mechanical Calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1290.

(8) Riley, K. E.; Hobza, P. Noncovalent Interactions in Biochemistry. *Wiley Interdiscip. Rev. Comput Mol Sci* **2011**, *1*, 3–17.

(9) Ambrosetti, A.; Alfè, D.; DiStasio Jr., R. A.; Tkatchenko, A. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. *J. Phys. Chem. Lett.* **2014**, *5*, 849–855.

(10) Otero-de-la-Roza, A.; Johnson, E. R. Predicting Energetics of Supramolecular Systems Using the XDM Dispersion Model. *J. Chem. Theory Comput.* **2015**, *11*, 4033–4040.

(11) Risthaus, T.; Grimme, S. Benchmarking of London Dispersion-Accounting Density Functional Theory Methods on

Very Large Molecular Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 1580–1591.

(12) Brandenburg, J. G.; Burke, K.; Civalleri, B.; Cole, D. J.; Csányi, G.; David, G.; Gidopoulos, N. I.; Gowland, D.; Helgaker, T.; Herbst, M. F.; Hourahine, B.; Irons, T. J. P.; Jacob, C. R.; Loos, P. F.; Mehta, N.; Mulay, M. R.; Neugebauer, J.; Pernal, K.; Pribram-Jones, A.; Romaniello, P.; Ryder, M. R.; Savin, A.; Sirbu, D.; Skylaris, C. K.; Truhlar, D. G.; Wetherell, J.; Yang, W. Challenges for Large Scale Simulation: General Discussion. *Faraday Discuss.* **2020**, *224*, 309–332.

(13) Al-Hamdani, Y. S.; Nagy, P. R.; Barton, D.; Kállay, M.; Brandenburg, J. G.; Tkatchenko, A. Interactions between Large Molecules: Puzzle for Reference Quantum-Mechanical Methods. *Nat. Comm.* **2021**, *12*, 3927.

(14) Antony, J.; Sure, R.; Grimme, S. Using Dispersion-Corrected Density Functional Theory to Understand Supramolecular Binding Thermodynamics. *Chem. Commun.* **2015**, *51*, 1764–1774.

(15) Sherrill, C. D. Frontiers in Electronic Structure Theory. *J. Chem. Phys.* **2010**, *132*, 110902.

(16) Hofer, T. S. From Macromolecules to Electrons—Grand Challenges in Theoretical and Computational Chemistry. *Front. Chem.* **2013**, *1*, 6.

(17) Grimme, S.; Schreiner, P. R. Computational Chemistry: The Fate of Current Methods and Future Challenges. *Angew. Chemie – Int. Ed.* **2018**, *57*, 4170–4176.

(18) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50*, 539–543.

(19) Merz, K. M. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47*, 2804–2811.

(20) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632–672.

(21) Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115*, 5607–5642.

(22) Li, S.; Li, W.; Ma, J. Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates. *Acc. Chem. Res.* **2014**, *47*, 2712–2720.

(23) Li, W.; Dong, H.; Ma, J.; Li, S. Structures and Spectroscopic Properties of Large Molecules and Condensed-Phase Systems Predicted by Generalized Energy-Based Fragmentation Approach. *Acc. Chem. Res.* **2021**, *54*, 169–181.

(24) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643–5677.

(25) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. Fragment Quantum Mechanical Calculation of Proteins and Its Applications. *Acc. Chem. Res.* **2014**, *47*, 2748–2757.

(26) Ramabhadran, R. O.; Raghavachari, K. The Successful Merger of Theoretical Thermochemistry with Fragment-Based Methods in Quantum Chemistry. *Acc. Chem. Res.* **2014**, *47*, 3596–3604.

(27) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2776–2785.

(28) Pruitt, S. R.; Bertoni, C.; Brorsen, K. R.; Gordon, M. S. Efficient and Accurate Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2786–2794.

(29) Sure, R.; Brandenburg, J. G.; Grimme, S. Small Atomic Orbital Basis Set First-Principles Quantum Chemical Methods for Large Molecular and Periodic Systems: A Critical Analysis of Error Sources. *ChemistryOpen* **2016**, *5*, 94–109.

(30) Goerigk, L.; Collyer, C. A.; Reimers, J. R. Recommending Hartree–Fock Theory with London-Dispersion and Basis-Set-Superposition Corrections for the Optimization or Quantum Refinement of Protein Structures. *J. Phys. Chem. B* **2014**, *118*, 14612–14626.

(31) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116*, 5301–5337.

(32) Yilmazer, N. D.; Korth, M. Enhanced Semiempirical QM Methods for Biomolecular Interactions. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 169–175.

(33) Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Weber, W.; Steiger, R.; Scholten, M.; Thiel, W. Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters. *J. Chem. Theory Comput.* **2016**, *12*, 1082–1096.

(34) Thiel, W. Semiempirical Quantum-Chemical Methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 145–157.

(35) Throssell, K. T. Evaluating and Improving Approximate LCAO-MO Theory with Restored Overlap and Bond Order Bond Energy Corrections. PhD thesis, Wesleyan University: Middletown, CT, 2018.

(36) 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.

(37) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

(38) 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.

(39) Kruse, H.; Grimme, S. A Geometrical Correction for the Inter- and Intra-Molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large Systems. *J. Chem. Phys.* **2012**, *136*, 154101.

(40) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All Spd-Block Elements (*Z* =1–86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(41) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB–An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(42) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chemie* **2020**, *132*, 15795–15803.

(43) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended Tight-Binding Quantum Chemistry Methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, e1493.

(44) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for Large Molecules. *ChemRxiv* **2019**. DOI:10.26434/chemrxiv.8326202.v1.

(45) Gani, T. Z. H.; Kulik, H. J. Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+U. *J. Chem. Theory Comput.* **2016**, *12*, 5931–5945.

(46) Alizadegan, R.; Hsia, K. J.; Martinez, T. J. A Divide and Conquer Real Space Finite-Element Hartree-Fock Method. *J. Chem. Phys.* **2010**, *132*, 034101.

(47) Garcia, J.; Szalewicz, K. Ab Initio Extended Hartree-Fock plus Dispersion Method Applied to Dimers with Hundreds of Atoms. *J. Phys. Chem. A* **2020**, *124*, 1196–1203.

(48) Chen, Y.; Zhang, L.; Wang, H.; Weinan, W. Ground State Energy Functional with Hartree-Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155–7165.

(49) Altun, A.; Neese, F.; Bistoni, G. HFLD: A Nonempirical London Dispersion-Corrected Hartree-Fock Method for the Quantification and Analysis of Noncovalent Interaction Energies of Large Molecular Systems †. *J. Chem. Theory Comput.* **2019**, *15*, 5894–5907.

(50) Barca, G. M. J.; Galvez-Vallejo, J. L.; Poole, D. L.; Rendell, A. P.; Gordon, M. S. High-Performance, Graphics Processing Unit-Accelerated Fock Build Algorithm. *J. Chem. Theory Comput.* **2020**, *16*, 7232–7238.

(51) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(52) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 1–8.

(53) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(54) McKemmish, L. K.; Gilbert, A. T. B.; Gill, P. M. W. Mixed Ramp-Gaussian Basis Sets. *J. Chem. Theory Comput.* **2014**, *10*, 4369–4376.

(55) Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S. Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *J. Phys. Chem. Lett.* **2014**, *5*, 4275–4284.

(56) Iron, M. A.; Janes, T. Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange–Correlation Functionals: The MOBH35 Benchmark Database. *J. Phys. Chem. A* **2019**, *123*, 3761–3781.

(57) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent Structures and Interactions by Density Functional Theory with Small Atomic Orbital Basis Sets. *J. Chem. Phys.* **2015**, *143*, 054107.

(58) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A Revised Low-Cost Variant of the B97-D Density Functional Method. *J. Chem. Phys.* **2018**, *148*, 064104.

(59) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. R 2 SCAN-3c: A "Swiss Army Knife" Composite Electronic-Structure Method . *J. Chem. Phys.* **2021**, *154*, 064103.

(60) Pracht, P.; Grant, D. F.; Grimme, S. Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra. *J. Chem. Theory Comput.* **2020**, *16*, 7044–7060.

(61) Brandenburg, J. G.; Caldeweyher, E.; Grimme, S. Screened Exchange Hybrid Density Functional for Accurate and Efficient Structures and Interaction Energies. *Phys. Chem. Chem. Phys.* **2016**, *18*, 15519–15523.

(62) Caldeweyher, E.; Brandenburg, J. G. Simplified DFT Methods for Consistent Structures and Energies of Large Systems. *J. Phys. Condens. Matter* **2018**, *30*, 213001.

(63) DiLabio, G. A. Atom-Centered Potentials for Noncovalent Interactions and Other Applications. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., 2017; pp 221–240.

(64) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Atom-Centered Potentials with Dispersion-Corrected Minimal-Basis-Set Hartree–Fock: An Efficient and Accurate Computational Approach for Large Molecular Systems. *J. Chem. Theory Comput.* **2018**, *14*, 726–738.

(65) Otero-de-la-Roza, A.; DiLabio, G. A. Improved Basis-Set Incompleteness Potentials for Accurate Density-Functional Theory Calculations in Large Systems. *J. Chem. Theory Comput.* **2020**, *16*, 4176–4191.

(66) Otero-de-la-Roza, A.; DiLabio, G. A. Transferable Atom-Centered Potentials for the Correction of Basis Set Incompleteness Errors in Density-Functional Theory. *J. Chem. Theory Comput.* **2017**, *13*, 3505–3524.

(67) van Santen, J. A.; DiLabio, G. A. Dispersion Corrections Improve the Accuracy of Both Noncovalent and Covalent Interactions Energies Predicted by a Density-Functional Theory Approximation. *J. Phys. Chem. A* **2015**, *119*, 6703–6713.

(68) DiLabio, G. A.; Koleini, M. Dispersion-Correcting Potentials Can Significantly Improve the Bond Dissociation Enthalpies and Noncovalent Binding Energies Predicted by Density-Functional Theory. *J. Chem. Phys.* **2014**, *140*, 18A542.

(69) DiLabio, G. A.; Koleini, M.; Torres, E. Extension of the B3LYP–Dispersion-Correcting Potential Approach to the Accurate Treatment of Both Inter- and Intra-Molecular Interactions. *Theor. Chem. Acc.* **2013**, *132*, 1389.

(70) Torres, E.; DiLabio, G. A. A (Nearly) Universally Applicable Method for Modeling Noncovalent Interactions Using B3LYP. *J. Phys. Chem. Lett.* **2012**, *3*, 1738–1744.

(71) Mackie, I. D.; DiLabio, G. A. Interactions in Large, Polyaromatic Hydrocarbon Dimers: Application of Density Functional Theory with Dispersion Corrections. *J. Phys. Chem. A* **2008**, *112*, 10968–10976.

(72) DiLabio, G. A. Accurate Treatment of van Der Waals Interactions Using Standard Density Functional Theory Methods with Effective Core-Type Potentials: Application to Carbon-Containing Dimers. *Chem. Phys. Lett.* **2008**, *455*, 348–353.

(73) Mackie, I. D.; DiLabio, G. A. Accurate Dispersion Interactions from Standard Density-Functional Theory Methods with Small Basis Sets. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6092.

(74) Torres, E.; DiLabio, G. A. Density-Functional Theory with Dispersion-Correcting Potentials for Methane: Bridging the Efficiency and Accuracy Gap between High-Level Wave Function and Classical Molecular Mechanics Methods. *J. Chem. Theory Comput.* **2013**, *9*, 3342–3349.

(75) Mackie, I. D.; DiLabio, G. A. CO_2 Adsorption by Nitrogen-Doped Carbon Nanotubes Predicted by Density-Functional Theory with Dispersion-Correcting Potentials. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2780–2787.

(76) Holmes, J. D.; Otero-de-la-Roza, A.; DiLabio, G. A. Accurate Modeling of Water Clusters with Density-Functional Theory Using Atom-Centered Potentials. *J. Chem. Theory Comput.* **2017**, *13*, 4205–4215.

(77) Cao, X.; Dolg, M. Pseudopotentials and Modelpotentials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 200–210.

(78) Dolg, M.; Cao, X. Relativistic Pseudopotentials: Their Development and Scope of Applications. *Chem. Rev.* **2012**, *112*, 403–480.

(79) Tibshirani, R. Regression Shrinkage and Selection via the Lasso: A Retrospective. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)* **2011**, *73*, 273–282.

(80) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.

(81) Osborne, M. R.; Presnell, B.; Turlach, B. A. On the LASSO and Its Dual. *J. Comput. Graph. Stat.* **2000**, *9*, 319–337.

(82) Tatewaki, H.; Huzinaga, S. A Systematic Preparation of New Contracted Gaussian-Type Orbital Sets. III. Second-Row Atoms from Li through Ne. *J. Comput. Chem.* **1980**, *1*, 205–228.

(83) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent Molecular Orbital Methods. XXIII. A Polarization-type Basis Set for Second-row Elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(84) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(85) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(86) Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.* **2006**, *124*, 174104.

(87) Habgood, M.; James, T.; Heifetz, A. Conformational Searching with Quantum Mechanics. In *Methods in Molecular Biology*; Humana Press Inc., 2020; Vol. 2114, pp 207–229.

(88) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(89) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

(90) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044–2058.

(91) Rezai, T.; Bock, J. E.; Zhou, M. V.; Kalyanaraman, C.; Lokey, R. S.; Jacobson, M. P. Conformational Flexibility, Internal Hydrogen Bonding, and Passive Membrane Permeability: Successful in Silico Prediction of the Relative Permeabilities of Cyclic Peptides. *J. Am. Chem. Soc.* **2006**, *128*, 14073–14080.

(92) Poongavanam, V.; Danelius, E.; Peintner, S.; Alcaraz, L.; Caron, G.; Cummings, M. D.; Wlodek, S.; Erdelyi, M.; Hawkins, P. C. D.; Ermondi, G.; et al. Conformational Sampling of Macrocyclic Drugs in Different Environments: Can We Find the Relevant Conformations? *ACS Omega* **2018**, *3*, 11742–11757.

(93) Diaz, D. B.; Appavoo, S. D.; Bogdanchikova, A. F.; Lebedev, Y.; McTiernan, T. J.; dos Passos Gomes, G.; Yudin, A.

K. Illuminating the Dark Conformational Space of Macrocycles Using Dominant Rotors. *Nat. Chem.* **2021**, *13*, 218–225.

(94) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.

(95) Gutten, O.; Bím, D.; Řezáč, J.; Rulíšek, L. Macrocycle Conformational Sampling by DFT-D3/COSMO-RS Methodology. *J. Chem. Inf. Model.* **2018**, *58*, 48–60.

(96) Saha, I.; Dang, E. K.; Svatunek, D.; Houk, K. N.; Harran, P. G. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 24679–24690.

(97) Butler, K. T.; Luque, F. J.; Barril, X. Toward Accurate Relative Energy Predictions of the Bioactive Conformation of Drugs. *J. Comput. Chem.* **2009**, *30*, 601–610.

(98) Sorokina, I.; Mushegian, A. The Role of the Backbone Torsion in Protein Folding. *Biol. Direct* **2016**, *11*, 1–5.

(99) Culka, M.; Rulíšek, L. Factors Stabilizing β-Sheets in Protein Structures from a Quantum-Chemical Perspective. *J. Phys. Chem. B* **2019**, *123*, 6453–6461.

(100) Culka, M.; Galgonek, J.; Vymětal, J.; Vondrášek, J.; Rulíšek, L. Toward Ab Initio Protein Folding: Inherent Secondary Structure Propensity of Short Peptides from the Bioinformatics and Quantum-Chemical Perspective. *J. Phys. Chem. B* **2019**, *123*, 1215–1227.

(101) Mattelaer, C.-A.; Mattelaer, H.-P.; Rihon, J.; Froeyen, M.; Lescrinier, E. Efficient and Accurate Potential Energy Surfaces of Puckering in Sugar-Modified Nucleosides. *J. Chem. Theory Comput.* **2021**, *17*, 3814–3823.

(102) Huang, M.; Giese, T. J.; Lee, T. S.; York, D. M. Improvement of DNA and RNA Sugar Pucker Profiles from Semiempirical Quantum Methods. *J. Chem. Theory Comput.* **2014**, *10*, 1538–1545.

(103) Di Fenza, A.; Heine, A.; Koert, U.; Klebe, G. Understanding Binding Selectivity toward Trypsin and Factor Xa: The Role of Aromatic Interactions. *ChemMedChem* **2007**, *2*, 297–308.

(104) McGaughey, G. B.; Gagné, M.; Rappé, A. K. π-Stacking Interactions. Alive and Well in Proteins. *J. Biol. Chem.* **1998**, *273*, 15458–15463.

(105) Sal-Man, N.; Gerber, D.; Bloch, I.; Shai, Y. Specificity in Transmembrane Helix-Helix Interactions Mediated by Aromatic Residues. *J. Biol. Chem.* **2007**, *282*, 19753–19761.

(106) Ravva, M. K.; Risko, C.; Brédas, J. L. Noncovalent Interactions in Organic Electronic Materials. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*; Elsevier Inc., 2017; pp 277–302.

(107) Černý, J.; Kabeláč, M.; Hobza, P. Double-Helical → Ladder Structural Transition in the B-DNA Is Induced by a Loss of Dispersion Energy. *J. Am. Chem. Soc.* **2008**, *130*, 16055–16059.

(108) Karabiyik, H.; Sevinçek, R.; Karabiyik, H. π-Cooperativity Effect on the Base Stacking Interactions in DNA: Is There a Novel Stabilization Factor Coupled with Base Pairing H-Bonds? *Phys. Chem. Chem. Phys.* **2014**, *16*, 15527–15538.

(109) Wilson, K. A.; Wetmore, S. D. A Survey of DNA–Protein π–Interactions: A Comparison of Natural Occurrences and Structures, and Computationally Predicted Structures and Strengths. In *Challenges and Advances in Computational Chemistry and Physics*; Springer, 2015; Vol. 19, pp 501–532.

(110) Salonen, L. M.; Ellermann, M.; Diederich, F. Aromatic Rings in Chemical and Biological Recognition: Energetics and Structures. *Angew. Chemie - Int. Ed.* **2011**, *50*, 4808–4842.

(111) Schneider, H. J. Binding Mechanisms in Supramolecular Complexes. *Angew. Chemie - Int. Ed.* **2009**, *48*, 3924–3977.

(112) Deng, J. H.; Luo, J.; Mao, Y. L.; Lai, S.; Gong, Y. N.; Zhong, D. C.; Lu, T. B. Π-π Stacking Interactions: Non-Negligible Forces for Stabilizing Porous Supramolecular Frameworks. *Sci. Adv.* **2020**, *6*, eaax9976.

(113) Hwang, J. wun; Li, P.; Shimizu, K. D. Synergy between Experimental and Computational Studies of Aromatic Stacking Interactions. *Org. Biomol. Chem.* **2017**, *15*, 1554–1564.

(114) Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Energy Matrix of Structurally Important Side-Chain/ Side-Chain Interactions in Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 2191–2203.

(115) MacCallum, J. L.; Drew Bennett, W. F.; Peter Tieleman, D. Distribution of Amino Acids in a Lipid Bilayer from Computer Simulations. *Biophys. J.* **2008**, *94*, 3393–3404.

(116) Mbaye, M. N.; Hou, Q.; Basu, S.; Teheux, F.; Pucci, F.; Rooman, M. A Comprehensive Computational Study of Amino Acid Interactions in Membrane Proteins. *Sci. Rep.* **2019**, *9*, 1–14.

(117) Busseron, E.; Ruff, Y.; Moulin, E.; Giuseppone, N. Supramolecular Self-Assemblies as Functional Nanomaterials. *Nanoscale* **2013**, *5*, 7098–7140.

(118) Desiraju, G. R.; Steiner, T. *The Weak Hydrogen Bond In Structural Chemistry and Biology*; Oxford University Press: Oxford and New York, 1999.

(119) Herschlag, D.; Pinney, M. M. Hydrogen Bonds: Simple after All? *Biochemistry* **2018**, *57*, 3338–3352.

(120) Bauzá, A.; Deyà, P. M.; Frontera, A. Anion-π Interactions in Supramolecular Chemistry and Catalysis. In *Challenges and Advances in Computational Chemistry and Physics*; Springer, 2015; Vol. 19, pp 471–500.

(121) Schottel, B. L.; Chifotides, H. T.; Dunbar, K. R. Anion-Π Interactions. *Chem. Soc. Rev.* **2008**, *37*, 68–83.

(122) Lucas, X.; Bauzá, A.; Frontera, A.; Quiñonero, D. A Thorough Anion-π Interaction Study in Biomolecules: On the Importance of Cooperativity Effects. *Chem. Sci.* **2016**, *7*, 1038–1050.

(123) Borozan, S. Z.; Zlatović, M. V.; Stojanović, S. Anion–π Interactions in Complexes of Proteins and Halogen-Containing

Amino Acids. *J. Biol. Inorg. Chem.* **2016**, *21*, 357–368.

(124) Sanders, J. M. Optimal π-Stacking Interaction Energies in Parallel-Displaced Aryl/Aryl Dimers Are Predicted by the Dimer Heavy Atom Count. *J. Phys. Chem. A* **2010**, *114*, 9205–9211.

(125) Parrish, R. M.; Sherrill, C. D. Quantum-Mechanical Evaluation of π-π Versus Substituent-π Interactions in π Stacking: Direct Evidence for the Wheeler-Houk Picture. *J. Am. Chem. Soc.* **2014**, *136*, 17386–17389.

(126) Steinmann, S. N.; Corminboeuf, C. Exploring the Limits of Density Functional Approximations for Interaction Energies of Molecular Precursors to Organic Electronics. *J. Chem. Theory Comput.* **2012**, *8*, 4305–4316.

(127) Crittenden, D. L. A Systematic CCSD(T)Study of Long-Range and Noncovalent Interactions between Benzene and a Series of First- and Second-Row Hydrides and Rare Gas Atoms. *J. Phys. Chem. A* **2009**, *113*, 1663–1669.

(128) 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.

(129) 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.

(130) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. Estimated MP2 and CCSD(T) Interaction Energies of n-Alkane Dimers at the Basis Set Limit: Comparison of the Methods of Helgaker et Al. and Feller. *J. Chem. Phys.* **2006**, *124*, 114304.

(131) Granatier, J.; Pitoňák, M.; Hobza, P. Accuracy of Several Wave Function and Density Functional Theory Methods for Description of Noncovalent Interaction of Saturated and Unsaturated Hydrocarbon Dimers. *J. Chem. Theory Comput.* **2012**, *8*, 2282–2292.

(132) Setiawan, D.; Kraka, E.; Cremer, D. Strength of the Pnicogen Bond in Complexes Involving Group VA Elements N, P, and AS. *J. Phys. Chem. A* **2015**, *119*, 1642–1656.

(133) Hill, J. G.; Legon, A. C. On the Directionality and Non-Linearity of Halogen and Hydrogen Bonds. *Phys. Chem. Chem. Phys.* **2015**, *17*, 858–867.

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

(135) Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 88–96.

(136) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis Set Convergence of the Coupled-Cluster Correction, P2CCSD(T): Best Practices for Benchmarking Non-Covalent Interactions and the Attendant Revision of the S22, NBC10, HBC6, and HSG Databases. *J. Chem. Phys.* **2011**, *135*, 194102.

(137) 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.

(138) Miriyala, V. M.; Řezáč, J. Description of Non-Covalent Interactions in SCC-DFTB Methods. *J. Comput. Chem.* **2017**, *38*, 688–697.

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

(140) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J. Chem. Theory Comput.* **2020**, *16*, 141–151.

(141) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space. *J. Chem. Theory Comput.* **2020**, *16*, 6305–6316.

(142) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(143) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

(144) 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.

(145) 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.

(146) DiLabio, G. A.; Johnson, E. R.; Otero-de-la-Roza, A. Performance of Conventional and Dispersion-Corrected Density-Functional Theory Methods for Hydrogen Bonding Interaction Energies. *Phys. Chem. Chem. Phys.* **2013**, *15*, 12821–12828.

(147) 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.

(148) Witte, J.; Goldey, M.; Neaton, J. B.; Head-Gordon, M. Beyond Energies: Geometries of Nonbonded Molecular

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

(149) David Sherrill, C.; Takatani, T.; Hohenstein, E. G. An Assessment of Theoretical Methods for Nonbonded Interactions: Comparison to Complete Basis Set Limit Coupled-Cluster Potential Energy Curves for the Benzene Dimer, the Methane Dimer, Benzene-Methane, and Benzene-H2S. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.

(150) Hohenstein, E. G.; Sherrill, C. D. Effects of Heteroatoms on Aromatic π-π Interactions: Benzene-Pyridine and Pyridine Dimer. *J. Phys. Chem. A* **2009**, *113*, 878–886.

(151) Takatani, T.; David Sherrill, C. Performance of Spin-Component-Scaled Møller-Plesset Theory (SCS-MP2) for Potential Energy Curves of Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6106–6114.

(152) Řezáč, J.; Huang, Y.; Hobza, P.; Beran, G. J. O. Benchmark Calculations of Three-Body Intermolecular Interactions and the Performance of Low-Cost Electronic Structure Methods. *J. Chem. Theory Comput.* **2015**, *11*, 3065–3079.

(153) Copeland, K. L.; Tschumper, G. S. Hydrocarbon/Water Interactions: Encouraging Energetics and Structures from Dft but Disconcerting Discrepancies for Hessian Indices. *J. Chem. Theory Comput.* **2012**, *8*, 1646–1656.

(154) 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. *J. Chem. Phys.* **2017**, *147*, 161727.

(155) Černý, J.; Schneider, B.; Biedermannová, L. WatAA: Atlas of Protein Hydration. Exploring Synergies between Data Mining and: Ab Initio Calculations. *Phys. Chem. Chem. Phys.* **2017**, *19*, 17094–17102.

(156) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.

(157) Kříž, K.; Řezáč, J. Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 1453–1460.

(158) Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **2015**, *11*, 2473–2486.

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

(160) Hostaš, J.; Jakubec, D.; Laskowski, R. A.; Gnanasekaran, R.; Řezáč, J.; Vondrášek, J.; Hobza, P. Representative Amino Acid Side-Chain Interactions in Protein-DNA Complexes: A Comparison of Highly Accurate Correlated Ab Initio Quantum Mechanical Calculations and Efficient Approaches for Applications to Large Systems. *J. Chem. Theory Comput.* **2015**, *11*, 4086–4092.

(161) Jakubec, D.; Laskowski, R. A.; Vondrasek, J. Sequence-Specific Recognition of DNA by Proteins: Binding Motifs Discovered Using a Novel Statistical/Computational Analysis. *PLoS One* **2016**, *11*, e0158704.

(162) Stasyuk, O. A.; Jakubec, D.; Vondrášek, J.; Hobza, P. Noncovalent Interactions in Specific Recognition Motifs of Protein-DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13*, 877–885.

(163) Kozmon, S.; Matuška, R.; Spiwok, V.; Koča, J. Three-Dimensional Potential Energy Surface of Selected Carbohydrates' CH/π Dispersion Interactions Calculated by High-Level Quantum Mechanical Methods. *Chem. - A Eur. J.* **2011**, *17*, 5680–5690.

(164) Kozmon, S.; Matuška, R.; Spiwok, V.; Koča, J. Dispersion Interactions of Carbohydrates with Condensate Aromatic Moieties: Theoretical Study on the CH-π Interaction Additive Properties. *Phys. Chem. Chem. Phys.* **2011**, *13*, 14215–14222.

(165) Stanković, I. M.; Blagojević Filipović, J. P.; Zarić, S. D. Carbohydrate – Protein Aromatic Ring Interactions beyond CH/π Interactions: A Protein Data Bank Survey and Quantum Chemical Calculations. *Int. J. Biol. Macromol.* **2020**, *157*, 1–9.

(166) Kumari, M.; Sunoj, R. B.; Balaji, P. V. Conformational Mapping and Energetics of Saccharide-Aromatic Residue Interactions: Implications for the Discrimination of Anomers and Epimers and in Protein Engineering. *Org. Biomol. Chem.* **2012**, *10*, 4186–4200.

(167) Kruse, H.; Banáš, P.; Šponer, J. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.* **2019**, *15*, 95–115.

(168) Parker, T. M.; Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *J. Chem. Theory Comput.* **2015**, *11*, 4197–4204.

(169) Banáš, P.; Mládek, A.; Otyepka, M.; Zgarbová, M.; Jurečka, P.; Svozil, D.; Lankaš, F.; Šponer, J. Can We Accurately Describe the Structure of Adenine Tracts in B-DNA? Reference Quantum-Chemical Computations Reveal Overstabilization of Stacking by Molecular Mechanics. *J. Chem. Theory Comput.* **2012**, *8*, 2448–2460.

(170) Kabeláč, M.; Valdes, H.; Sherer, E. C.; Cramer, C. J.; Hobza, P. Benchmark RI-MP2 Database of Nucleic Acid Base Trimers: Performance of Different Density Functional Models for Prediction of Structures and Binding Energies. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5000–5008.

(171) Smith, D. G. A.; Patkowski, K. Toward an Accurate Description of Methane Physisorption on Carbon Nanotubes. *J. Phys. Chem. C* **2014**, *118*, 544–550.

(172) Smith, D. G. A.; Patkowski, K. Interactions between Methane and Polycyclic Aromatic Hydrocarbons: A High

Accuracy Benchmark Study. *J. Chem. Theory Comput.* **2013**, *9*, 370–389.

(173) Vogiatzis, K. D.; Klopper, W.; Friedrich, J. Non-Covalent Interactions of CO2 with Functional Groups of Metal-Organic Frameworks from a CCSD(T) Scheme Applicable to Large Systems. *J. Chem. Theory Comput.* **2015**, *11*, 1574–1584.

(174) Smith, D. G. A.; Patkowski, K. Benchmarking the CO2 Adsorption Energy on Carbon Nanotubes. *J. Phys. Chem. C* **2015**, *119*, 4934–4948.

(175) Li, S.; Smith, D. G. A.; Patkowski, K. An Accurate Benchmark Description of the Interactions between Carbon Dioxide and Polyheterocyclic Aromatic Compounds Containing Nitrogen. *Phys. Chem. Chem. Phys.* **2015**, *17*, 16560–16574.

(176) Li, W.; Grimme, S.; Krieg, H.; Möllmann, J.; Zhang, J. Accurate Computation of Gas Uptake in Microporous Organic Molecular Crystals. *J. Phys. Chem. C* **2012**, *116*, 8865–8871.

(177) Temelso, B.; Archer, K. A.; Shields, G. C. Benchmark Structures and Binding Energies of Small Water Clusters with Anharmonicity Corrections. *J. Phys. Chem. A* **2011**, *115*, 12034–12046.

(178) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; Van Der Avoird, A. Water Pair Potential of near Spectroscopic Accuracy. I. Analysis of Potential Surface and Virial Coefficients. *J. Chem. Phys.* **2000**, *113*, 6687–6701.

(179) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315*, 1249–1252.

(180) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Polarizable Interaction Potential for Water from Coupled Cluster Calculations. I. Analysis of Dimer Potential Energy Surface. *J. Chem. Phys.* **2008**, *128*, 094313.

(181) Mintz, B. J.; Parks, J. M. Benchmark Interaction Energies for Biologically Relevant Noncovalent Complexes Containing Divalent Sulfur. *J. Phys. Chem. A* **2012**, *116*, 1086–1092.

(182) Sharapa, D. I.; Genaev, A.; Cavallo, L.; Minenkov, Y. A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules. *ChemPhysChem* **2018**, *20*, 92–102.

(183) Sellers, B. D.; James, N. C.; Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **2017**, *57* (6), 1265–1275.

(184) Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. L. The Melatonin Conformer Space: Benchmark and Assessment of Wave Function and DFT Methods for a Paradigmatic Biological and Pharmacological Molecule. *J. Phys. Chem. A* **2013**, *117*, 2269–2277.

(185) Tahchieva, D. N.; Bakowies, D.; Ramakrishnan, R.; Von Lilienfeld, O. A. Torsional Potentials of Glyoxal, Oxalyl Halides, and Their Thiocarbonyl Derivatives: Challenges for Popular Density Functional Approximations. *J. Chem. Theory Comput.* **2018**, *14*, 4806–4817.

(186) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies Using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2021**, *121*, e26381.

(187) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7*, 1–10.

(188) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. PEPCONF, A Diverse Data Set of Peptide Conformational Energies. *Sci. Data*. **2019**, *6*, 180310.

(189) Goerigk, L.; Karton, A.; Martin, J. M. L.; Radom, L. Accurate Quantum Chemical Energies for Tetrapeptide Conformations: Why MP2 Data with an Insufficient Basis Set Should Be Handled with Caution. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7028.

(190) Valdes, H.; Pluháčková, K.; Pitonák, M.; Řezáč, J.; Hobza, P. Benchmark Database on Isolated Small Peptides Containing an Aromatic Side Chain: Comparison between Wave Function and Density Functional Theory Methods and Empirical Force Field. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747.

(191) Kesharwani, M. K.; Karton, A.; Martin, J. M. L. Benchmark *Ab Initio* Conformational Energies for the Proteinogenic Amino Acids through Explicitly Correlated Methods. Assessment of Density Functional Methods. *J. Chem. Theory Comput.* **2016**, *12*, 444–454.

(192) Mládek, A.; Krepl, M.; Svozil, D.; Čech, P.; Otyepka, M.; Banáš, P.; Zgarbová, M.; Jurečka, P.; Šponer, J. Benchmark Quantum-Chemical Calculations on a Complete Set of Rotameric Families of the DNA Sugar-Phosphate Backbone and Their Comparison with Modern Density Functional Theory. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7295–7310.

(193) Mládek, A.; Banáš, P.; Jurečka, P.; Otyepka, M.; Zgarbová, M.; Šponer, J. Energies and 2′-Hydroxyl Group Orientations of RNA Backbone Conformations. Benchmark CCSD(T)/CBS Database, Electronic Analysis, and Assessment of DFT Methods and MD Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 463–480.

(194) Kruse, H.; Mladek, A.; Gkionis, K.; Hansen, A.; Grimme, S.; Sponer, J. Quantum Chemical Benchmark Study on 46 RNA Backbone Families Using a Dinucleotide Unit. *J. Chem. Theory Comput.* **2015**, *11*, 4972–4991.

(195) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. Evaluation of Density Functionals and Basis Sets for Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5*, 679–692.

(196) Chan, B. Aqueous-Phase Conformations of Lactose, Maltose, and Sucrose and the Assessment of Low-Cost DFT Methods with the DSCONF Set of Conformers for the Three Disaccharides. *J. Phys. Chem. A* **2020**, *124*, 582–590.

(197) Sameera, W. M. C.; Pantazis, D. A. A Hierarchy of Methods for the Energetically Accurate Modeling of Isomerism in Monosaccharides. *J. Chem. Theory Comput.* **2012**, *8*, 2630–2645.

(198) Marianski, M.; Supady, A.; Ingram, T.; Schneider, M.; Baldauf, C. Assessing the Accuracy of Across-the-Scale Methods for Predicting Carbohydrate Conformational Energies for the Examples of Glucose and α-Maltose. *J. Chem. Theory Comput.* **2016**, *12*, 6157–6168.

(199) Gruzman, D.; Karton, A.; Martin, J. M. L. Performance of Ab Initio and Density Functional Methods for Conformational Equilibria of C_nH_{2n+2} Alkane Isomers ($n = 4-8$). *J. Phys. Chem. A* **2009**, *113*, 11974–11983.

(200) Kozuch, S.; Bachrach, S. M.; Martin, J. M. L. Conformational Equilibria in Butane-1,4-Diol: A Benchmark of a Prototypical System with Strong Intramolecular H-Bonds. *J. Phys. Chem. A* **2014**, *118*, 293–303.

(201) Martin, J. M. L. What Can We Learn about Dispersion from the Conformer Surface of N-Pentane? *J. Phys. Chem. A* **2013**, *117*, 3118–3132.

(202) Temelso, B.; Klein, K. L.; Mabey, J. W.; Pérez, C.; Pate, B. H.; Kisiel, Z.; Shields, G. C. Exploring the Rich Potential Energy Surface of (H2O)11 and Its Physical Implications. *J. Chem. Theory Comput.* **2018**, *14*, 1141–1153.

(203) Smith, B. J.; Swanton, D. J.; Pople, J. A.; Schaefer III, H. F.; Radom, L. Transition Structures for the Interchange of Hydrogen Atoms within the Water Dimer. *J. Chem. Phys.* **1990**, *92*, 1240–1247.

(204) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer III, H. F.; Quack, M. Anchoring the Water Dimer Potential Energy Surface with Explicitly Correlated Computations and Focal Point Analyses. *J. Chem. Phys.* **2002**, *116*, 690–701.

(205) Taylor, D. E.; Ángyán, J. G.; Galli, G.; Zhang, C.; Gygi, F.; Hirao, K.; Song, J. W.; Rahul, K.; Anatole Von Lilienfeld, O.; Podeszwa, R.; et al. Blind Test of Density-Functional-Based Methods on Intermolecular Interaction Energies. *J. Chem. Phys.* **2016**, *145*, 124105.

(206) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; et al. Quantum Chemical Benchmark Databases of Gold-Standard Dimer Interaction Energies. *Sci. Data* **2021**, *8*, 1–9.

(207) Sparrow, Z. M.; Ernst, B. G.; Joo, P. T.; Lao, K. U.; DiStasio Jr., R. A. NENCI-2021 Part I: A Large Benchmark Database of Non-Equilibrium Non-Covalent Interactions Emphasizing Close Intermolecular Contacts. *arXiv preprint*, https://arxiv.org/abs/2102.02354 (accessed Nov. 8, 2021).

(208) Chan, B.; Gilbert, A. T. B.; Gill, P. M. W.; Radom, L. Performance of Density Functional Theory Procedures for the Calculation of Proton-Exchange Barriers: Unusual Behavior of M06-Type Functionals. *J. Chem. Theory Comput.* **2014**, *10*, 3777–3783.

(209) Karton, A.; O'Reilly, R. J.; Chan, B.; Radom, L. Determination of Barrier Heights for Proton Exchange in Small Water, Ammonia, and Hydrogen Fluoride Clusters with G4(MP2)-Type, MPn, and SCS-MPn Procedures-a Caveat. *J. Chem. Theory Comput.* **2012**, *8*, 3128–3136.

(210) Romero-Montalvo, E.; DiLabio, G. A. Computational Study of Hydrogen Bond Interactions in Water Cluster-Organic Molecule Complexes. *J. Phys. Chem. A* **2021**, *125*, 3369–3377.

(211) Oliveira, V.; Kraka, E.; Cremer, D. The Intrinsic Strength of the Halogen Bond: Electrostatic and Covalent Contributions Described by Coupled Cluster Theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 33031–33046.

(212) Mehta, N.; Fellowes, T.; WHITE, J.; Goerigk, L. The CHAL336 Benchmark Set: How Well Do Quantum-Chemical Methods Describe Chalcogen-Bonding Interactions? *J Chem. Theory Comput.* **2021**, *17*, 2783–2806.

(213) Miriyala, V. M.; Řezáč, J. Testing Semiempirical Quantum Mechanical Methods on a Data Set of Interaction Energies Mapping Repulsive Contacts in Organic Molecules. *J. Phys. Chem. A* **2018**, *122*, 2801–2808.

(214) Kříž, K.; Nováček, M.; Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 3: Repulsive Contacts. *J. Chem. Theory Comput.* **2021**, *17*, 1548–1561.

(215) Lao, K. U.; Herbert, J. M. An Improved Treatment of Empirical Dispersion and a Many-Body Energy Decomposition Scheme for the Explicit Polarization plus Symmetry-Adapted Perturbation Theory (XSAPT) Method. *J. Chem. Phys.* **2013**, *139*, 034107.

(216) Lao, K. U.; Herbert, J. M. Accurate and Efficient Quantum Chemistry Calculations for Noncovalent Interactions in Many-Body Systems: The XSAPT Family of Methods. *J. Phys. Chem. A.* **2015**, *119*, 235–252.

(217) Mardirossian, N.; Lambrecht, D. S.; McCaslin, L.; Xantheas, S. S.; Head-Gordon, M. The Performance of Density Functionals for Sulfate-Water Clusters. *J. Chem. Theory Comput.* **2013**, *9*, 1368–1380.

(218) Mezei, P. D.; Csonka, G. I.; Ruzsinszky, A.; Sun, J. Accurate, Precise, and Efficient Theoretical Methods to Calculate Anion-π Interaction Energies in Model Structures. *J. Chem. Theory Comput.* **2015**, *11*, 360–371.

(219) Zahn, S.; Macfarlane, D. R.; Izgorodina, E. I. Assessment of Kohn-Sham Density Functional Theory and Møller-Plesset Perturbation Theory for Ionic Liquids. *Phys. Chem. Chem. Phys.* **2013**, *15*, 13664–13675.

(220) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.

(221) Sharapa, D. I.; Margraf, J. T.; Hesselmann, A.; Clark, T. Accurate Intermolecular Potential for the C60 Dimer: The Performance of Different Levels of Quantum Theory. *J. Chem. Theory Comput.* **2017**, *13*, 274–285.

(222) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for

Large Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

(223) Calbo, J.; Ortí, E.; Sancho-García, J. C.; Aragó, J. Accurate Treatment of Large Supramolecular Complexes by Double-Hybrid Density Functionals Coupled with Nonlocal van Der Waals Corrections. *J. Chem. Theory Comput.* **2015**, *11*, 932–939.

(224) Ni, Z.; Guo, Y.; Neese, F.; Li, W.; Li, S. Cluster-in-Molecule Local Correlation Method with an Accurate Distant Pair Correction for Large Systems. *J. Chem. Theory Comput.* **2021**, *17*, 756–766.

(225) Zhang, H.; Krupa, J.; Wierzejewska, M.; Biczysko, M. The Role of Dispersion and Anharmonic Corrections in Conformational Analysis of Flexible Molecules: The Allyl Group Rotamerization of Matrix Isolated Safrole. *Phys. Chem. Chem. Phys.* **2019**, *21*, 8352–8364.

(226) Kirschner, K. N.; Heiden, W.; Reith, D. Small Alcohols Revisited: CCSD(T) Relative Potential Energies for the Minima, First- and Second-Order Saddle Points, and Torsion-Coupled Surfaces. *ACS Omega* **2018**, *3*, 419–432.

(227) Greenwell, C.; Beran, G. J. O. Inaccurate Conformational Energies Still Hinder Crystal Structure Prediction in Flexible Organic Molecules. *Cryst. Growth Des.* **2020**, *20*, 4875–4881.

(228) Lahey, S. L. J.; Thien Phuc, T. N.; Rowley, C. N. Benchmarking Force Field and the ANI Neural Network Potentials for the Torsional Potential Energy Surface of Biaryl Drug Fragments. *J. Chem. Inf. Model.* **2020**, *60*, 6258–6268.

(229) Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *J. Chem. Theory Comput.* **2018**, *14*, 1254–1266.

(230) The dcp package, https://github.com/aoterodelaroza/dcp (accessed Nov. 8, 2021).

(231) The acpfit package, https://github.com/aoterodelaroza/acpfit (accessed Nov. 8, 2021).

(232) AOR GitHub repository, https://github.com/aoterodelaroza (accessed Nov. 8, 2021).

(233) Osborne, M. R.; Presnell, B.; Turlach, B. A. A New Approach to Variable Selection in Least Squares Problems. *IMA J. Numer. Anal.* **2000**, *20*, 389–403.

(234) Schmidt, M. Graphical model structure learning using L_1-regularization, PhD Thesis, Unviversity of British Columbia, 2010.

(235) Schmidt, M.; Fung, G.; Rosales, R. Optimization Methods for L1-Regularization, https://www.cs.ubc.ca/tr/2009/tr-2009-19 (accessed Nov. 8, 2021).

(236) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; calmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 16, Revision B.01; Gaussian Inc.: Wallingford, CT, 2016.

(237) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73–78.

(238) Mahadevi, A. S.; Sastry, G. N. Cooperativity in Noncovalent Interactions. *Chem. Rev.* **2016**, *116*, 2775–2825.

(239) Hostaš, J.; Řezáč, J. Accurate DFT-D3 Calculations in a Small Basis Set. *J. Chem. Theory Comput.* **2017**, *13*, 3575–3585.

(240) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Performance of Small Basis Set Hartree–Fock Methods for Modeling Non-Covalent Interactions. *Electron. Struc.* **2021**, *3*, 034007.

(241) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Sherrill, C. D.; Merz, K. M. The Energy Computation Paradox and Ab Initio Protein Folding. *PLoS One* **2011**, *6*, 18868.

(242) Cutini, M.; Bechis, I.; Corno, M.; Ugliengo, P. Balancing Cost and Accuracy in Quantum Mechanical Simulations on Collagen Protein Models. *J. Chem. Theory Comput.* **2021**, *17*, 2566–2574.

(243) Hsiao, Y. W.; Sanchez-Garcia, E.; Doerr, M.; Thiel, W. Quantum Refinement of Protein Structures: Implementation and Application to the Red Fluorescent Protein DsRed.M1. *J. Phys. Chem. B* **2010**, *114*, 15413–15423.

(244) Antony, J.; Grimme, S. Fully Ab Initio Protein-Ligand Interaction Energies with Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2012**, *33*, 1730–1739.

(245) Lepšík, M.; Řezáč, J.; Kolář, M.; Pecina, A.; Hobza, P.; Fanfrlík, J. The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design. *ChemPlusChem* **2013**, *78*, 921–931.

(246) Cavasotto, C. N. Binding Free Energy Calculation Using Quantum Mechanics Aimed for Drug Lead Optimization. In *Methods in Molecular Biology*; Humana Press Inc., 2020; Vol. 2114, pp 257–268.

(247) Harding, D. P.; Kingsley, L. J.; Spraggon, G.; Wheeler, S. E. Importance of Model Size in Quantum Mechanical Studies of DNA Intercalation. *J. Comput. Chem.* **2020**, *41*, 1175–1184.

(248) Zhang, C.; Qin, S.; Hu, B.; Lv, J.; Yang, Z.; Yan, W.; Wang, J.; Huang, N.; Huang, Z. Disruption of Nucleobase Stacking to Restore Reactivity. *Nucleosides, Nucleotides and Nucleic Acids* **2019**, *38*, 567–577.

(249) Fogarty, A. C.; Duboué-Dijon, E.; Sterpone, F.; Hynes, J. T.; Laage, D. Biomolecular Hydration Dynamics: A Jump

Model Perspective. *Chem. Soc. Rev.* **2013**, *42*, 5672–5683.

(250) Britz, D. A.; Khlobystov, A. N. Noncovalent Interactions of Molecules with Single Walled Carbon Nanotubes. *Chem. Soc. Rev.* **2006**, *35*, 637–659.

(251) Kauffman, D. R.; Star, A. Carbon Nanotube Gas and Vapor Sensors. *Angew. Chemie – Int. Ed.* **2008**, *47*, 6550–6570.

(252) Cao, D.; Zhang, X.; Chen, J.; Wang, W.; Yun, J. Optimization of Single-Walled Carbon Nanotube Arrays for Methane Storage at Room Temperature. *J. Phys. Chem. B* **2003**, *107*, 13286–13292.

(253) Lohse, M. S.; Bein, T. Covalent Organic Frameworks: Structures, Synthesis, and Applications. *Adv. Funct. Mater.* **2018**, *28*, 1705553.

(254) Germain, A.; Ugliengo, P. Modeling Interstellar Amorphous Solid Water Grains by Tight-Binding Based Methods: Comparison Between GFN-XTB and CCSD(T) Results for Water Clusters. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH, 2020; Vol. 12253 LNCS, pp 745–753.

(255) Martínez-Bachs, B.; Ferrero, S.; Rimola, A. Binding Energies of N-Bearing Astrochemically-Relevant Molecules on Water Interstellar Ice Models. a Computational Study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH, 2020; Vol. 12251 LNCS, pp 683–692.

(256) Rimola, A.; Ferrero, S.; Germain, A.; Corno, M.; Ugliengo, P. Computational Surface Modelling of ICES and Minerals of Interstellar Interest—Insights and Perspectives. *Minerals* **2021**, *11*, 1–25.

(257) Ferrero, S.; Zamirri, L.; Ceccarelli, C.; Witzel, A.; Rimola, A.; Ugliengo, P. Binding Energies of Interstellar Molecules on Crystalline and Amorphous Models of Water Ice by Ab Initio Calculations. *Astrophys. J.* **2020**, *904*, 11.

(258) Van Dishoeck, E. F.; Herbst, E.; Neufeld, D. A. Interstellar Water Chemistry: From Laboratory to Observations. *Chem. Rev.* **2013**, *113*, 9043–9085.

(259) Steber, A. L.; Pérez, C.; Temelso, B.; Shields, G. C.; Rijs, A. M.; Pate, B. H.; Kisiel, Z.; Schnell, M. Capturing the Elusive Water Trimer from the Stepwise Growth of Water on the Surface of the Polycyclic Aromatic Hydrocarbon Acenaphthene. *J. Phys. Chem. Lett.* **2017**, *8*, 5744–5750.

(260) Ruiz Pestana, L.; Mardirossian, N.; Head-Gordon, M.; Head-Gordon, T. Ab Initio Molecular Dynamics Simulations of Liquid Water Using High Quality Meta-GGA Functionals. *Chem. Sci.* **2017**, *8*, 3554–3565.

(261) Gaiduk, A. P.; Gustafson, J.; Gygi, F.; Galli, G. First-Principles Simulations of Liquid Water Using a Dielectric-Dependent Hybrid Functional. *J. Phys. Chem. Lett.* **2018**, *9*, 3068–3073.

(262) DiStasio Jr., R. A.; Santra, B.; Li, Z.; Wu, X.; Car, R. The Individual and Collective Effects of Exact Exchange and Dispersion Interactions on the *Ab Initio* Structure of Liquid Water. *J. Chem. Phys.* **2014**, *141*, 084502.

(263) Chen, M.; Ko, H.-Y.; Remsing, R. C.; Calegari Andrade, M. F.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; et al. Ab Initio Theory and Modeling of Water. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10846–10851.

(264) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(265) Hartke, B. Efficient Global Geometry Optimization of Clusters: Method, and Application to Water Clusters. In *Eur. Phys. J. D* **2003**, *24*, 57–60.

(266) Guimaräes, F. F.; Belchior, J. C.; Johnston, R. L.; Roberts, C. Global Optimization Analysis of Water Clusters (H2O)n (11≤n≤13) through a Genetic Evolutionary Approach. *J. Chem. Phys.* **2002**, *116*, 8327–8333.

(267) Lenz, A.; Ojamäe, L. A Theoretical Study of Water Clusters: The Relation between Hydrogen-Bond Topology and Interaction Energy from Quantum-Chemical Computations for Clusters with up to 22 Molecules. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1905–1911.

(268) Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. Ab Initio Quantum Chemistry for Protein Structures. *J. Phys. Chem. B* **2012**, *116*, 12501–12509.

(269) Schmitz, S.; Seibert, J.; Ostermeir, K.; Hansen, A.; Göller, A. H.; Grimme, S. Quantum Chemical Calculation of Molecular and Periodic Peptide and Protein Structures. *J. Phys. Chem. B* **2020**, *124,* 3636–3646.

(270) Vorlová, B.; Nachtigallová, D.; Jirásková-Vaníčková, J.; Ajani, H.; Jansa, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P.; Konvalinka, J.; et al. Malonate-Based Inhibitors of Mammalian Serine Racemase: Kinetic Characterization and Structure-Based Computational Study. *Eur. J. Med. Chem.* **2015**, *89*, 189–197.

(271) Miriyala, V. M.; Řezáč, J. Testing Semiempirical Quantum Mechanical Methods on a Data Set of Interaction Energies Mapping Repulsive Contacts in Organic Molecules. *J. Phys. Chem. A* **2018**, *122*, 2801–2808.

(272) Zheng, J.; Xu, X.; Truhlar, D. G. Minimally Augmented Karlsruhe Basis Sets. *Theor. Chem. Acc.* **2011**, *128*, 295–305.

(273) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.

(274) Morgante, P.; Peverati, R. CLB18: A New Structural Database with Unusual Carbon–Carbon Long Bonds. *Chem. Phys. Lett.* **2021**, *765*, 138281.

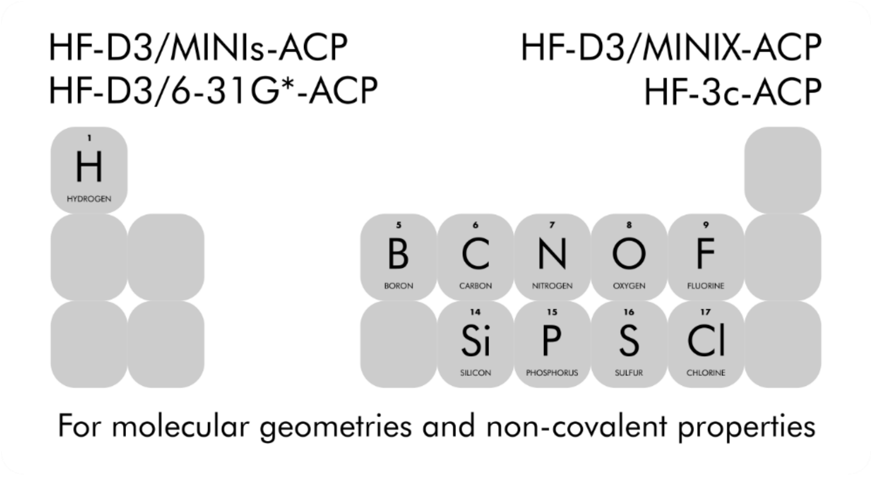
(275) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923.

(276) Boekfa, B.; Choomwattana, S.; Khongpracha, P.; Limtrakul, J. Effects of the Zeolite Framework on the Adsorptions

and Hydrogen-Exchange Reactions of Unsaturated Aliphatic, Aromatic, and Heterocyclic Compounds in ZSM-5 Zeolite: A Combination of Perturbation Theory (MP2) and a Newly Developed Density Functional Theory (M06-2X) in ONIOM Scheme. *Langmuir* **2009**, *25*, 12990–12999.

(277) Kim, S.; Robichaud, D. J.; Beckham, G. T.; Paton, R. S.; Nimlos, M. R. Ethanol Dehydration in HZSM-5 Studied by Density Functional Theory: Evidence for a Concerted Process. *J. Phys. Chem. A* **2015**, *119*, 3604–3614.

Table of Content Graphic:



For molecular geometries and non-covalent properties