Improved Basis-Set Incompleteness Potentials for Accurate DFT Calculations in Large Systems

A. Otero-de-la-Roza[1,*] and Gino A. DiLabio[2,3,†]

[1] *Departamento de Química Física y Analítica and MALTA-Consolider Team, Facultad de Química, Universidad de Oviedo, 33006 Oviedo, Spain*
[2] *Department of Chemistry, University of British Columbia, Okanagan, 3247 University Way, Kelowna, British Columbia, Canada V1V 1V7.*
[3] *Faculty of Management, University of British Columbia, Okanagan, 1137 Alumni Avenue, Kelowna, British Columbia, V1V 1V7, Canada.*

(Dated: May 6, 2020)

The accurate calculation of chemical properties using density-functional theory (DFT) requires the use of a nearly-complete basis set. In chemical systems involving hundreds to thousands of atoms, the cost of the calculations place practical limitations on the number of basis functions that can be used. Therefore, in most practical applications of DFT to large systems, there exists basis-set incompleteness error (BSIE). In this article, we present the next iteration of the basis-set incompleteness potentials (BSIPs), one-electron potentials designed to correct for basis-set incompleteness error. The ultimate goal associated with the development of BSIPs is to allow the calculation of molecular properties using DFT with near-complete-basis-set results at a computational cost that is similar to a small basis set calculation. In this work, we develop BSIPs for 10 atoms in the first and second rows (H, B–F, Si–Cl) and 15 common basis sets of the Pople, Dunning, Karlsruhe, and Huzinaga types. Our new BSIPs are constructed to minimize BSIE in the calculation of reaction energies, barrier heights, non-covalent binding energies, and intermolecular distances. The BSIPs were obtained using a training set of 15,944 data points. The fitting approach employed a regularized linear least-squares method with variable selection (the LASSO method), which results in a much better fit to the training data than our previous BSIPs while, at the same time, reducing the computational cost of BSIP development. The proposed BSIPs are tested on various benchmark sets and demonstrate excellent performance in practice. Our new BSIPs are also transferable, i.e., they can be used to correct BSIE in calculations that employ density functionals other than the one used in the BSIP development (B3LYP). Finally, BSIPs can be used in any quantum chemistry program that have implemented effective-core potentials without changes to the software.

I. INTRODUCTION

One of the principal limitations of quantum-mechanical methods in the calculation of molecular properties of thermochemical interest (molecular geometries, reaction energies, barrier heights, vibrational frequencies) is the increase in computational cost with the number of atoms in the system. Modeling biological systems, in particular, is a challenge because of the mixture of covalent and non-covalent interactions (NCI) that need to be treated accurately.[1,2] Due to its good accuracy, simplicity, and relatively low cost, density-functional theory (DFT) has been the leading method for thermochemical calculations since the advent of hybrid functionals, as well as the dominant technique in modeling periodic solid-state systems. With the development of dispersion-corrected methods,[3–5] DFT is now able to treat covalent and non-covalent interactions on the same footing, paving the way for the accurate description of large macromolecules.[6–10] Although common density functionals do not offer enough accuracy for the calculation of rate constants ($\sim$ 0.1–1 kcal/mol), they are, in general, good enough to provide key insights regarding chemical reaction energetics and, as such, DFT has become an essential tool in the elucidation of organic reaction mechanisms.

In the most common approach, thermochemical calculations with DFT use a set of Gaussian basis functions to describe the system's Kohn-Sham orbitals. In the limit of infinite number of basis functions (the complete-basis set limit, CBS), the Kohn-Sham orbitals are perfectly described and the calculated molecular properties are precisely those predicted by the chosen density functional. These properties may still be in error due to the approximate nature of the selected density functional, but a great deal is known about the performance of various functionals in the calculation of many properties of interest.[4,11–13] If the Gaussian basis set employed is not nearly complete, then an additional error arises from the incorrect description of the Kohn-Sham orbitals: This error is called basis-set incompleteness error (BSIE). BSIE affects all calculated properties, but it is particularly detrimental for non-covalent interactions due to their weak nature.[14–16] Because the computational cost increases with the third power of the number of basis functions employed,[17,18] BSIE is unavoidable in practice for systems with hundreds to thousands of atoms (e.g. biomolecules).

In the calculation of non-covalent interaction energies, a common method to reduce BSIE is the counterpoise correction[15,19–24] (CP), in which both interacting monomers are calculated using the basis functions of the dimer. The CP method is particularly useful in combination with wavefunction theory methods, which are in general more computationally expensive and more sensitive

to BSIE than DFT. The energy correction term arising from the application of the CP method is called basis-set superposition error (BSSE). BSSE is typically used as a proxy for the size of the BSIE, since it is often impossible to come sufficiently close to the complete-basis-set limit using wavefunction methods. The CP approach can also be applied to DFT calculations but it does not capture all the BSIE,[16] has a tendency to overcorrect,[23,25,26] and it is unclear how to apply it to systems that do not partition naturally into fragments.[27,28] However, building on the success of the CP method, alternative approaches to alleviate BSIE have been proposed, including atom-based counterpoise corrections,[27,28] parametrized *ad hoc* formulas[29–31] and using a Hubbard-U energy term,[32] among others.[33,34]

In a previous article,[16] we proposed basis-set incompleteness potentials (BSIPs) as a way to mitigate BSIE in DFT calculations on large systems using small or minimal basis sets. BSIPs are one-electron potentials with the same functional form as effective-core potentials (ECP). Instead of replacing core electrons, BSIPs are designed to yield an energy correction that matches the BSIE for a particular basis set, thus obtaining nearly complete-basis-set-quality molecular properties at a computational cost that is close to that associated with using a small basis set. In addition, BSIPs are transferable, i.e. they are not tied to a specific density functional, and can be used in any program that has implemented ECPs without modifying the software. In this article, we expand the previous work by developing BSIPs for an increased number of atoms, basis sets, and by using a much larger training set in their development. More importantly, we introduce the use of regularized linear least-squares fits with variable selection (the LASSO method), which simplifies the development of BSIPs enormously and significantly improves their performance. The developed BSIPs are subsequently tested, and shown to offer excellent performance and transferability across density functionals.

II. COMPUTATIONAL METHODS

All calculations in this work were carried out using the Gaussian package (versions 09[35] and 16[36]). Based on the functional transferability tests in our previous article,[16] we chose the B3LYP functional[37,38] for the BSIP development. A number of other functionals, sometimes including dispersion corrections (in the D3 approach[39]), were employed for testing the final BSIPs (Sec. IV). A calculation of all properties in the training set with a nearly complete basis set is required to provide reference data points for the BSIP fitting procedure. In all cases, we used the very large aug-cc-pVQZ basis set for this purpose,[40,41] which has been shown to be close enough to the CBS limit. All calculations used Gaussian's "ultrafine" grids (590×99 pruned grids).

The BSIP fitting procedure used a 1-norm regularized linear least-squares fit employing the least-absolute-shrinkage-and-selection-operator (LASSO) method.[42] The fits were carried out using the procedure proposed by Osborne et al.[43] and implemented in octave by Schmidt.[44,45]

III. BSIP DEVELOPMENT

A. Basis Set Incompleteness Error

A detailed description of the formalism and the notation used in BSIP development is given in our previous work.[16] We summarize the most salient points here for convenience. For a Gaussian basis set x, we define the basis-set incompleteness error in the energy (BSIE or BSIE(E)) as:

$$\mathrm{BSIE}_x = \mathrm{BSIE(E)}_x = E_{\mathrm{CBS}} - E_x \tag{1}$$

where E_x is the energy using basis set x and E_{CBS} is the energy in the complete-basis-set limit (approximated by aug-cc-pVQZ). We define the BSIE of any property P that maps linearly with the energy (e.g. binding energy, bond dissociation energy, multipolar moments, atomic forces,...) in an analogous way:

$$\mathrm{BSIE(P)}_x = \mathrm{P}_{\mathrm{CBS}} - \mathrm{P}_x \tag{2}$$

Because P maps linearly with the energy, BSIE(P) can be written as a linear combination of the BSIE(E) of the species involved in its calculation. The variational principle ensures that BSIE(E) is always positive whereas BSIE(P) can be positive or negative. It is important to note that BSIE is different from BSSE in that it is defined with reference to the complete basis set limit, and it applies to any property, not just intermolecular binding energies.

The impact of basis-set incompleteness depends on the calculated property and on the size of the basis set.[17,46] For instance, due to the small energies involved in their calculation and the contribution from the molecular density tails, binding energies of non-covalent dimers (BE) are particularly problematic.[14–17] The appearance of BSIE(BE) tends to result in overestimated interaction energies and at least one set of diffuse functions is required to have reasonably low BSIE(BE) (see Table II for a simple illustration). This is particularly true in strong non-covalent interactions with significant intermolecular delocalization, like hydrogen bonds[26] or halogen bonds.[47] However, a basis set like 6-31+G* is still too small for accurate non-covalent interaction (NCI) calculations and often much larger basis sets are required.[14,15] This, combined with the fact that NCIs are ever-present in supramolecular and biological systems, is one of the main motivations for the development of BSIPs. The impact of BSIE on chemical reaction energies (e.g. bond dissociation energies, BDE[48]) and barrier heights is usually smaller than for NCIs[16], although large basis sets are still required. In particular, polarization functions

are essential in minimizing BSIE in this case, especially on the atoms undergoing bond breaking and formation.

One characteristic of BSIE in DFT calculations that is relevant to this work is that it is mostly the same across density functional approximations regardless of the calculated property.[16,29] The BSIE associated with noncovalent binding energies and thermochemical properties, the reduction of which are the main objectives of associated with BSIP development, are essentially identical for any semilocal functional. (Exceptions are those functionals that show slow convergence to the CBS due to extensive parametrization, like the Minnesota functionals.[4,16]) This is convenient because BSIPs designed to minimize BSIE using a given functional are expected to be transferable to other functionals with minimal performance penalties. The BSIE also depend only mildly on the fraction of exact exchange, however we do not expect our DFT-based BSIPs to perform as well in combination with the Hartree-Fock method.

B. Basis Set Incompleteness Potentials

Another important characteristic of BSIE is that it grows linearly with the number of interatomic contacts.[15] This suggests the idea of using one-electron potentials centered on the atoms (atom-centered potentials, ACP) that decay exponentially with distance for correcting BSIE. BSIPs have the same mathematical form as effective core potentials (ECP)[16]:

$$V_{\mathrm{BSIP}}(\boldsymbol{r}) = \sum_A U^A_{L_A}(r_A) + \sum_{l=0}^{L_A-1} \sum_{m=-l}^{l} |Alm\rangle U^A_l(r_A) \langle Alm| \tag{3}$$

$$U^A_l(r) = \sum_k c^A_{lk} e^{-\zeta^A_{lk} r^2} \tag{4}$$

$$E_{\mathrm{BSIP}}(\{c^A_{lk}\}, \{\zeta^A_{lk}\}) = \sum_i^{\mathrm{occ.}} \langle \psi_i | V_{\mathrm{BSIP}} | \psi_i \rangle \tag{5}$$

where A are atoms in the system, r_A is the distance to the atom, $U^A_l(r)$ is the radial potential for atom A and angular momentum l, and c^A_{lk} and ζ^A_{lk} are adjustable parameters. The U_{L_A} is called the "local" angular momentum channel. Unlike ECPs, BSIPs do not replace any electrons. Instead, BSIPs are used as a wavefunction-dependent energy correction imposed on the system (Eq. 5), with the added advantage that the corresponding term in the one-electron Hamiltonian, and therefore their effect on the self-consistent wavefunction, is already included. This has been shown to be important in the development of atom-centered potentials (ACPs) for water.[49] These ACPs, which were specifically developed to reproduce binding energies of water clusters in combination with a DFT approach, improved the calculated dipole for the water molecule to almost perfect agreement with experiment,[49] demonstrating the importance of incorporating the effect of the correction to the self-consistent wavefunction and electron density. ACPs have been extensively used in the past by DiLabio et al., particularly for the inclusion of dispersion effects in the description of intermolecular interactions,[50–55] as well as by other authors.[56,57]

The development of effective BSIPs hinges on the determination of the adjustable parameters, the exponents ζ^A_{lk} and the coefficients c^A_{lk}, for the selected atoms. We chose a total of 10 atoms, common in organic molecules, as target of our BSIP development (H, B–F, Si–Cl). The maximum angular momentum for the BSIP terms (L_A) is equal to the maximum l of the corresponding basis set and the atom for which the BSIP is being developed. With these choices, all that remains is to find a way to determine how many exponent/coefficient pairs are necessary for each atom to have an efficient BSIP, and how to determine their value. This is done by using the fact that the BSIP energy correction E_{BSIP} is approximately linear in the coefficients, provided these coefficients are relatively small:

$$E_{\mathrm{BSIP}}(\{c^A_{lk}\}, \{\zeta^A_{lk}\}) = \sum_{Alk} c^A_{lk} \Delta E^A_{lk}(\zeta^A_{lk}) \tag{6}$$

$$\Delta E^A_{lk}(\zeta^A_{lk}) = \sum_i \langle \psi_i | \left(\sum_{m=-l}^{l} |Alm\rangle e^{-\zeta^A_{lk} r^2_A} \langle Alm| \right) |\psi_i\rangle \tag{7}$$

The BSIP energy terms ($\Delta E^A_{lk}(\zeta^A_{lk})$) depend on the atom, l, and exponent, and also indirectly on the coefficients via the one-electron states ($|\psi_i\rangle$). If the coefficient and exponents are packed into vectors, we have:

$$E_{\mathrm{BSIP}}(\boldsymbol{c}, \boldsymbol{\zeta}) = \boldsymbol{c} \cdot \boldsymbol{\Delta E}(\boldsymbol{\zeta})^{\mathsf{T}} \tag{8}$$

which is linear in the coefficients vector:

$$E_{\mathrm{BSIP}}(\alpha \boldsymbol{c} + \beta \boldsymbol{d}, \boldsymbol{\zeta}) \approx \alpha E_{\mathrm{BSIP}}(\boldsymbol{c}, \boldsymbol{\zeta}) + \beta E_{\mathrm{BSIP}}(\boldsymbol{d}, \boldsymbol{\zeta}) \tag{9}$$

where α and β are scalars. This last equation is approximate because the orbitals $|\psi_i\rangle$ depend on the coefficients. However, the equation becomes exact in the limit of $\boldsymbol{c} \to \boldsymbol{0}$ and the difference between both terms, termed the "non-linearity error", depends at least on the square of the coefficients. (See Ref. 16 for the formal definition of non-linearity error.) In general, we are not interested in reproducing total energies themselves but molecular properties. For any of these properties (P), provided that they are explicitly linear functions of the energy of the involved species, we can write:

$$P_{\mathrm{BSIP}}(\boldsymbol{c}, \boldsymbol{\zeta}) = \sum_{Alk} c^A_{lk} \Delta P^A_{lk}(\zeta^A_{lk}) \tag{10}$$

and

$$P_{\mathrm{BSIP}}(\alpha \boldsymbol{c} + \beta \boldsymbol{d}, \boldsymbol{\zeta}) = \alpha P_{\mathrm{BSIP}}(\boldsymbol{c}, \boldsymbol{\zeta}) + \beta P_{\mathrm{BSIP}}(\boldsymbol{d}, \boldsymbol{\zeta}) + P_{\text{non-lin}} \tag{11}$$

where $P_{\text{non-lin}} \to 0$ at least quadratically when the coefficients vanish.

Our strategy for the development of BSIPs is to exploit the linear nature of the BSIP energy correction with respect to the coefficients. First, we choose a relatively large training set of molecular properties of interest: non-covalent binding energies, reaction energies and energy barriers, conformational energies, and molecular deformation energies (see Section III C for a detailed list). We also select a set of exponents ζ_{lk}^A that span a range relevant to the calculation of those properties (in atomic units): 0.08 to 0.30 in 0.02 steps, 0.4 to 1.0 in 0.1 steps, 1.2 to 2.0 in 0.2 steps, 3.0, and 4.0, for a total of 26 exponents per atom and angular momentum channel. This allows us to pre-calculate the effect each of the BSIP terms have on the calculated properties ($\Delta P_{lk}^A(\zeta_{lk}^A)$) for every atom ($A$), angular momentum ($l$), and exponent ($k$). Once the BSIP terms are known, the BSIP coefficients are determined using Eq. 10 via a linear least-squares fit. Specifically, we minimize the weighted root-mean-square (wRMS) of the deviation between the BSIP contribution to property P and the calculated BSIE(P) for the basis set we want to correct (x):

$$\boldsymbol{c}_{\text{opt}}(\boldsymbol{\zeta}) = \min_{\boldsymbol{c}} \left[\text{wRMS}(\boldsymbol{c}, \boldsymbol{\zeta})\right] \tag{12}$$

$$\text{wRMS}(\boldsymbol{c}, \boldsymbol{\zeta}) = \sqrt{\sum_i w_i \left[\text{BSIE}(P)_x^i - P_{\text{BSIP}}^i(\boldsymbol{c}, \boldsymbol{\zeta})\right]^2} \tag{13}$$

where i runs over all entries in the training set, $\text{BSIE}(P)_x^i$ is given by Eq. 2 for entry i, and w_i are the weights associated to each entry. These weights are chosen to both balance the contributions from each component of the training set and to make each term in the wRMS sum adimensional. In practice, they are also used to eliminate some of the subsets of the training set during the fit. The calculation of $\text{BSIE}(P)_x^i$ requires the computation of the (near) CBS property for all systems in the training set. In order to have BSIPs that are transferable across functionals, it is essential that the contribution from the uncorrected basis set (P_x^i), the BSIP terms ($\Delta P_{lk}^A(\zeta_{lk}^A)$) and the CBS reference values (P_{CBS}^i) are all calculated using the same functional.

This least-squares fitting procedure to find the BSIP coefficients is effective only as long as the coefficients resulting from the fit are small enough that non-linearity error (Eq. 11) is not a dominant contribution to the calculated properties. In our previous work,[16] we noted that BSIPs with low non-linearity, and therefore usable in practice, involve only a handful of terms, so we devised a strategy to pre-select a number of BSIP terms per atom to give both low wRMS and low non-linearity error. We chose a constraint on the 2-norm of the coefficients (1.0 Ha or 2.0 Ha, depending on the basis set) and a given number of terms per atom (5). Then, iterated over atoms to determine the combination of BSIP terms that gave the minimum wRMS while also respecting the 2-norm constraint on the coefficients.

Although relatively efficient, the BSIPs that results from this procedure are not optimal. Due to the combinatorial nature of the method, this fitting strategy is both computationally expensive and limited to at most 5 BSIP terms per atom. The alternative approach for BSIP development we use here involves a regularized linear least-squares method, in which the constraint on the coefficients is already incorporated into the fit. In particular, we use the least-absolute-shrinkage-and-selection-operator (LASSO) method of Tibshirani[42] in which we impose an *a priori* constraint (λ) on the 1-norm of the BSIP coefficients:

$$\|\boldsymbol{c}\|_1 = \sum_{lkA} |c_{lk}^A| < \lambda \tag{14}$$

The advantage of using LASSO compared to other regularization methods is that it performs variable selection, i.e., it discards unneeded BSIP terms, resulting in simpler BSIPs with fewer exponent/coefficient pairs, thus minimizing both non-linearity error and the computational cost associated with the use of the BSIP. We applied the LASSO method in a recent article,[58] where we developed a proof-of-concept set of ACPs designed to correct Hartree-Fock (HF) calculations in combination with a minimal basis set. This work is the first instance of LASSO being applied for BSIP development.

The last remaining decision to make is about the magnitude of the constraint parameter (λ). The value of this parameter must reflect a balance between non-linearity error and minimization of the wRMS in the least-squares fit. Since the LASSO fit is so fast, we decided to evaluate the performance of several BSIPs resulting from using various values of λ for each basis set. For each of these BSIPs we calculated the deviation between the BSIP-corrected small basis-set calculations and the CBS results. Then, we chose the BSIP that gave the lowest wRMS in this test. The values of λ for each basis set are given in Table II.

We note that this procedure is different from what is normally used to choose the constraint parameter λ in the development of linear models using the LASSO method, which normally involves a cross-validation approach, i.e., splitting the data set into a training set, used for fitting, and a validation set, used for validating the linear model. This is unnecessary in our case for two reasons. First, the linear model is not interesting in itself—it is merely used as a means to develop the BSIPs. It is the good performance of the latter in actual self-consistent calculations that is the true objective of this work. Second, the existence of non-linearity error provides a natural way to validate the linear model: we calculate the performance of the BSIP associated with a certain linear model, and then calculate the non-linearity error by comparing the self-consistent BSIP results on the training set with the linear model predictions. We choose the λ by ensuring the non-linearity error is kept below a certain threshold. A linear model with an inadequately high value of λ would yield a BSIP with high non-linearity error, and

therefore unusable in self-consistent calculations. Given the size of our training set, this criterion for choosing λ places a stronger constraint on its value than the requirement that the linear model does not overfit.

C. Training Set

The training set used for the BSIP development in this work is shown in Table I. The training set contains four supersets, aiming at different molecular properties of interest: non-covalent binding energies ("NCI"), conformational energies ("Conf"), chemical reactions ("Cov"), and molecular geometries ("Geom"). The objective of the NCI superset is to capture intermolecular interaction strengths and geometries. It contains binding energies of small dimers at and away from equilibrium and also relatively uncommon NCI motifs, such as charged hydrogen bonds (IonicHB) and halogen bonds (Bauza).

Two additional NCI sets were designed for this work comprising molecules containing boron (B-set) and silicon (Si-set). For consistency with the previous article,[16] these sets were designed by constructing B- and Si-containing dimers, then relaxing their geometries using the ω-B97XD functional[98,99] and the pc-2 basis set,[100–103] which has been shown to be optimal for non-covalent interactions.[14] Eight data points were generated for each dimer by increasing and decreasing the distance between the molecular centers of mass to 90, 95, 100, 105, 110, 125, 150, and 200% of the equilibrium value, in the spirit of the S66×8 set.[61–63] The geometries and reference B3LYP/aug-cc-pVQZ energies for these sets, as well as for the rest of the training set, can be found in the Supporting Information (SI).

The SSI set, proposed along with the BBI set by Burns et al.,[74] and featuring protein sidechain-sidechain interactions, is a large fraction of the NCI part of the training set. Its objective is to ensure the good performance of the BSIPs in the description of proteins and other biological systems. In addition to non-covalent binding energies and geometries, our training set also contains the "Conf" superset, which focuses on molecular conformational energies. The Conf superset contains conformational energies of peptides (P26), hydrocarbons (ACONF), carbohydrates (SCONF) and specific molecules (cysteine and 1,4-butanediol). Since we are interested in modeling chemical reactions as well as intermolecular interactions, we included the Cov superset, featuring reaction energies and barrier heights for chemical reactions involving covalent bond breaking and formation. In the previous BSIP work,[16] this was handled by the BDEx set, which contained stretched and compressed geometries for only a handful of molecules. In this work, this superset comprises several benchmark sets from the literature, including reaction energies (BDE03, BSR36, DARC), isomerization energies (ISO34), proton affinities (PA26), atomization energies (W4-11), and barrier heights (BHPeri, BH76).

The last superset (Geom) comprises a single set of molecular deformation energies (Mol-def). A molecular deformation energy is the energy difference between a molecule at its equilibrium geometry and a small deformation of the same molecule. The aim of the Mol-def set is to improve the description of intramolecular geometries, which are known to be affected by BSIE when small and minimal basis sets are used.[16] In addition, we use the complete set of normal modes of the chosen molecules for the deformations, which ensures that every possible deformation is a combination of those included in the set. By doing this, we also seek to improve the description of the energy surfaces around the equilibrium geometries and, consequently, minimize BSIE in the calculation of the interatomic force constant matrix, vibrational frequencies, and the vibrational contribution to thermochemical quantities. However, it is important to note that, because molecular deformation energies are relative to the equilibrium geometry, this set does not address the BSIE inherent to bond breaking and formation processes. In fact, this is useful because we want to avoid correcting for this BSIE in very small and minimal basis sets (see Table II), but we still need to correct the BSIE on the intramolecular geometries. In addition, we expect the inclusion of the Mol-def set to indirectly improve the description of strong non-covalent interactions like hydrogen bonds by improving the accuracy of the calculated monomer deformation energies.[104,105]

The Mol-def set developed in this work is much larger than the set we developed previously (termed BDExrel) and presented in Ref. 16. A total of 107 molecules featuring all 10 atoms for which BSIPs are developed were chosen. We relaxed the molecular geometries using LC-ωPBE[106,107] combined with the XDM dispersion correction[108,109] and the aug-cc-pVTZ basis set.[40,41] Normal modes calculated at the equilibrium geometry were then used to construct the molecular deformations, with 7 deformations around equilibrium per normal mode and with a maximum deformation designed to yield a maximum change in energy of approximately 20 kcal/mol.

The whole training set comprises 15,944 data points, much larger than the 3,397 used for previous BSIP development.[16] The large size of the training set ensures that no overfitting will be present when the least-squares fit is carried out.

D. Target Basis Sets

The basis sets used as target for BSIP development are listed in Table II. A total of 15 basis sets were used. The list includes the smallest basis sets in Pople's 3-21G and 6-31G series, STO-3G, the Karlsruhe basis sets Def2-SV(P) and def2-SVP, Huzinaga's MINI, scaled MINI (MINIs), and MIDIh, and the smallest Jensen's polarization consistent basis sets (pc-0, and pc-1). (Segmented versions of these basis sets, called "pcs-n", have been proposed recently.[126]) The BSIPs developed in our

TABLE I. Data sets used in the training set, grouped into supersets by type of molecular property. The number of points in each set is given in the "Num" column.

Set	Description	Num.	Ref.
	Non-covalent interactions (NCI)		
S22×5	Interaction energies of 22 small dimers with only C, H, N, and O; 5 intermolecular distances per dimer.	110	59,60
S66×8	Interaction energies of 66 small dimers with only C, H, N, and O; 8 intermolecular distances per dimer.	528	61–63
F-set	Interaction energies of 20 F-containing small dimers; 8 intermolecular distances per dimer.	160	16
P-set	Interaction energies of 15 P-containing small dimers; 8 intermolecular distances per dimer.	120	16
S-set	Interaction energies of 18 S-containing small dimers; 8 intermolecular distances per dimer.	144	16
Cl-set	Interaction energies of 20 Cl-containing small dimers; 8 intermolecular distances per dimer.	160	16
B-set	Interaction energies of 20 B-containing small dimers; 8 intermolecular distances per dimer.	160	This work
Si-set	Interaction energies of 19 Si-containing small dimers; 8 intermolecular distances per dimer.	152	This work
KB49	Binding energies of 49 small dimers.	49	64,65
Water	Binding energies of water clusters with up to 10 molecules.	38	66
A24	Binding energies of small dimers, some not at equilibrium. The two Ar-containing dimers were left out.	22	67
HSG	Interaction energies of small dimers from the active site of the HIV-II protease/indinavir complex. Some dimers are charged.	21	68,69
IonicHB	Binding energies of small hydrogen-bonded dimers, at least one of which is charged.	120	70
ADIM6	Binding energies of $n-$alkane dimers, with monomers up to C_7H_{16}.	6	11,71
X40×10	Interaction energies of 40 halogenated dimers; 10 intermolecular distances per dimer. The Br- and I-containing dimers were left out.	220	72
S×8	Interaction energies of 14 S-containing dimers; 8 intermolecular distances per dimer.	112	73
BBI	Binding energies of small dimers from protein backbones.	100	74
SSI	Binding energies of small dimers from protein sidechains. Some dimers are charged.	3380	74
ACHC	Binding energies of adenine/cytosine nucleobase dimers.	54	75,76
HBC6	Interaction energies of 6 double-hydrogen-bonded dimers; ≈20 intermolecular distances per dimer.	118	76–78
NBC10ext	Interaction energies of 10 dispersion-bound dimers; ≈20 intermolecular distances per dimer.	195	76,78,79
Bauza	Halogen-bonded dimers. The Br-, Se-, and As-containing dimers were left out. Some dimers are charged.	10	47,80
	Molecular conformations (Conf)		
P26	Conformational energies of dipeptides and tripeptides with aromatic side-chains.	69	81
ACONF	Conformational energies of alkenes.	15	11,82
CYCONF	Conformational energies of cysteine.	10	11,83
SCONF	Conformational energies of carbohydrates.	17	11,84
BCONF	Conformational energies in 1,4-butanediol	64	85
	Covalent bond breaking and formation (Cov)		
BDE03	Bond dissociation energies. The Br-containing dimers were left out.	73	86
BHPeri	Barrier heights of pericyclic reactions.	26	11,87–90
BSR36	Separation energies of alkanes.	36	11,91
DARC	Diels-Alder reaction energies.	14	11,92
BH76	Barrier heights of reactions between small molecules.	76	11,12,93,94
ISO34	Isomerization energies of organic molecules.	34	11,95
PA26	Proton affinities.	26	12,96
W4-11	Atomization energies. The Be- and Al-containing molecules were left out.	133	12,97
	Intra-molecular geometries (Geom)		
Mol-def	Molecular deformation energies.	9372	This work

TABLE II. List of basis functions for which BSIPs are designed in this work. The columns indicate the type of basis set (dif = one set of diffuse functions, pol = one set of polarization functions on non-hydrogen atoms, polH = one set of polarization functions on all atoms), the number of basis functions on the H, O, and S atoms, the mean absolute error (MAE) for the S22×5 and W4-11 sets (in kcal/mol), the weight scheme employed in the fit, the value of the 1-norm constraint parameter (in Hartree), and the total number of terms in the set of proposed BSIP for all atoms (N_{terms}).

Basis set	Type	H/O/S	S22×5	W4-11	Weights	λ	N_{terms}	Ref.
MINI	minimal	1s/2s1p/3s2p	3.56	116.80	Small	10	86	110,111
MINIs	minimal	1s/2s1p/3s2p	1.85	55.55	Small	15	102	110,111
STO-3G	minimal	1s/1s1sp/1s2sp	2.47	42.93	Small	5	71	112
3-21G	2ζ	2s/1s2sp/1s3sp	3.82	29.07	Small	5	79	113–117
6-31G	2ζ	2s/1s2sp/1s3sp	1.85	35.44	Small	5	83	118–123
MIDIh	2ζ	2s/3s2p/4s3p	3.61	37.33	Small	5	75	110,111
pc-0	2ζ	2s/3s2p/4s3p	4.35	24.51	Small	5	84	100–103
6-31+G	2ζ+dif	2s/1s3sp/1s4sp	0.94	38.32	Small	10	104	118–123
3-21G*	2ζ+pol	2s/1s2sp/1s3sp1d	3.82	19.34	Small	5	89	113–117
6-31G*	2ζ+pol	2s/1s2sp1d/1s3sp1d	1.43	5.53	Large	15	130	118–123
Def2-SV(P)[a]	2ζ+pol	2s/3s2p1d/4s3p1d	1.60	6.49	Large	5	108	124,125
Def2-SVP[b]	2ζ+polH	2s1p/3s2p1d/4s3p1d	1.58	3.97	Large	5	108	124,125
pc-1	2ζ+polH	2s1p/3s2p1d/4s3p1d	1.16	8.11	Large	10	137	100–103
6-31+G*	2ζ+dif+pol	2s/1s3sp1d/1s4sp1d	0.42	7.38	Large	10	147	118–123
6-31+G**	2ζ+dif+polH	2s1p/1s3sp1d/1s4sp1d	0.38	5.44	Large	10	160	118–123

[a] In Gaussian, this basis set is accessed via the "Def2SVPP" keyword. [b] In Gaussian, this basis set is accessed via the "Def2SVP" keyword.

previous work[16] targeted only five basis sets: scaled MINI (MINIs), MINIs(d) (MINIs with a set of polarization functions on the second-row atoms), 6-31G*, pc-1, and 6-31+G**. All of them are included in the present work, except for MINIs(d), a single-ζ basis set plus polarization, which we found not to improve substantially upon MINIs. In all cases, we use the default choice in Gaussian regarding Cartesian or spherical primitives. Namely, all Pople basis sets use Cartesian functions and the Karlsruhe and Jensen's basis sets use spherical functions.

The basis sets considered are either minimal (MINI, MINIs, STO-3G) or double-ζ. Among the latter, we have included a variety of function types and basis-set sizes, with sets of polarization functions on the non-hydrogen atoms (pol), on all atoms (polH), and/or diffuse functions (dif). Table II also gives the mean absolute error (MAE) of the uncorrected basis sets using the B3LYP functional against the near-CBS reference values on two representative sets of the NCI and Cov supersets: the S22×5 set of non-covalent interactions energies and the W4-11 set of atomization energies (see Table I). The MAE in both sets roughly decreases with basis set size, but we can see that diffuse functions significantly reduce errors for NCIs (e.g. 6-31+G) whereas polarization functions are important for the atomization energies (e.g. 6-31G*). The table also gives the number of basis functions for some representative atoms that are targets in the BSIP development (H, O, and S). This serves as an indication of the characteristic size of the basis sets as well as their computational cost in practice.

The weights associated to each of the entries in the training set (Table I) depend on the basis set employed. Table II shows that there is a large gap in BSIE for atomization energies using different basis sets. While basis sets with polarization functions (other than 3-21G*) show relatively small errors, the MAEs for the rest are so high that we decided not to include the treatment of covalent bond breaking and formation as a target of the BSIP fitting procedure for those basis sets. Consequently, we used two different schemes for assigning weights during the fitting procedure. We call these the "small" and "large" weighting schemes in Table II. In the large weighting scheme, we assign a default weight of 1 to all systems in the training set, with the following exceptions. First, in sets composed of compressed and stretched geometries around equilibrium, the weight of the equilibrium geometry is equal to the number of geometries in each series. Thus, for instance, the equilibrium dimer geometries in the S66×8 have $w_i = 8$, and the non-equilibrium dimers have $w_i = 1$. Second, each element in the Mol-def set has $w_i = 0.1$ to account for the large number of elements in that set.

For the basis sets using the small weighting scheme we focus on non-covalent binding energies and intra- and inter-molecular geometries. This weighting scheme uses the same weights as the large scheme but assigns zero weight to all sets in the Cov superset (except BSR36) and also to the Bauza set, which involves non-covalent dimers with somewhat covalent character. The reason for including BSR36 in the small weighting scheme is that BSR36 comprises isodesmic reactions,[127] i.e. reactions where the number and types of bonds are preserved. BSIE is known to have minimal impact on the calculated reaction energies of isodesmic reactions (hence their popularity in the early days of computational chemistry), and therefore it is not detrimental to include it in the fit even for very small and minimal basis sets. A consequence of the weight assignment in Table II is that the BSIP-corrected basis sets using the small weighting scheme should not be used for modeling bond breaking and formation.

Lastly, Table II also indicates the 1-norm constraint used in the LASSO method (λ) for each basis set. As mentioned previously, λ was determined by running a series of BSIP-corrected self-consistent calculations with the BSIPs generated by different λ values, then choosing the λ that gave the lowest wRMS.

E. BSIP development

To recapitulate our choices, we developed BSIPs for ten atoms (H, B–F, Si–Cl) with angular momentum channels in the BSIP up to the maximum l in the basis set for a particular atom and 26 exponents per channel, from 0.08 up to 4.0. Fifteen basis sets of different sizes were used (Table II), grouped into large (BSIPs fitted using the whole training set) and small (the subsets with covalent bond breaking and formation left out). The training set (Table I) contains entries for non-covalent interactions, covalent bond energies, conformational energies, and molecular deformations. The computation of the uncorrected ("bare") basis set result plus the BSIP terms on the training set involved a total of 498,048,831 calculations plus an additional 38,191 B3LYP/aug-cc-pVQZ calculations for the reference data. The calculation of the BSIP terms was efficiently carried out by first running the self-consistent calculation with the indicated basis-set without BSIPs, then computing all the BSIP terms post-SCF.

The LASSO fit was carried out with the 1-norm constraints given in Table II, and the resulting BSIPs were evaluated by running BSIP-corrected self-consistent calculations on the training set. The number of terms in the proposed BSIPs is given in Table II (there are approximately $N_{\text{terms}}/10$ BSIP terms per atom). The performance of these BSIPs is shown in Table III. A detailed list of results for each individual subset in the training set, as well as the BSIPs themselves, are given in the SI.

Table III shows that the MAE for all sets and all basis sets decreases with the application of the BSIP, except in some of the cases where the set in question was not part of the fit (the Cov superset from MINI to 3-21G* in Table III). The MAE is reduced by a factor of 3–5 in most cases. For NCIs, the improvement is general and the average error is under 1 kcal/mol for all basis sets with a simultaneous decrease in the bias measured by the mean

TABLE III. Mean absolute errors (MAE) and mean errors (ME) in the supersets of the training set (Table I) using B3LYP compared to the B3LYP/aug-cc-pVQZ reference for the 15 basis sets in this work, in kcal/mol. The three results columns are: Bare = the uncorrected basis set, LS = the least-squares fit prediction of the BSIP-corrected result, Eval = the actual BSIP result when used in a self-consistent calculation.

		NCI			Conf			Cov			Geom		
		Bare	LS	Eval	Bare	LS	Eval	Bare	LS	Eval	Bare	LS	Eval
MINI	MAE	4.03	0.66	0.83	6.94	1.62	1.83	53.02	49.28	47.35	9.92	1.71	1.80
	ME	−3.91	0.00	−0.48	5.50	−0.26	−0.00	−43.72	−35.98	−34.40	−3.45	0.10	−0.05
MINIs	MAE	2.11	0.54	0.56	4.93	1.00	0.99	27.28	34.98	34.15	4.53	0.98	0.99
	ME	−1.86	0.03	−0.13	3.75	−0.30	−0.10	−12.72	−24.82	−22.69	0.33	0.06	−0.09
STO-3G	MAE	2.81	0.66	0.87	5.33	1.19	1.21	27.27	42.17	28.77	3.56	1.16	1.46
	ME	−2.51	0.07	−0.32	4.21	−0.37	−0.04	−0.34	−31.47	−18.32	1.05	0.26	0.63
3-21G	MAE	4.01	0.47	0.80	4.22	0.95	0.96	14.52	17.38	18.26	1.70	0.79	0.90
	ME	−3.97	−0.01	−0.54	3.58	−0.40	0.08	−11.21	−10.48	−7.73	0.12	0.15	0.27
6-31G	MAE	2.06	0.32	0.53	2.04	0.60	0.57	14.40	14.84	15.64	1.80	0.53	0.57
	ME	−1.99	−0.03	−0.37	1.80	−0.16	0.21	−11.63	−4.44	−6.03	0.09	0.03	0.07
MIDIh	MAE	3.98	0.51	0.78	4.07	0.78	0.95	16.92	16.17	16.44	2.10	0.80	0.97
	ME	−3.93	0.00	−0.43	3.44	−0.12	0.28	−13.22	−8.57	−7.33	−0.04	0.15	0.26
pc-0	MAE	4.26	0.43	0.85	5.65	0.87	1.03	14.13	20.17	20.25	2.24	0.72	0.82
	ME	−4.20	−0.01	−0.61	4.53	−0.18	0.32	−9.03	−13.41	−10.66	0.07	0.13	0.26
6-31+G	MAE	1.01	0.21	0.29	1.45	0.44	0.37	14.82	12.92	13.00	1.81	0.52	0.53
	ME	−0.89	−0.00	−0.16	1.31	−0.18	0.04	−12.00	−6.07	−5.23	0.01	−0.06	−0.01
3-21G*	MAE	3.99	0.43	0.80	4.20	0.81	1.02	11.14	19.20	18.61	1.04	0.70	0.81
	ME	−3.94	−0.01	−0.56	3.56	−0.24	0.21	−8.25	−12.18	−7.78	0.23	0.21	0.36
6-31G*	MAE	1.59	0.37	0.41	1.36	0.43	0.45	3.87	1.57	1.69	0.64	0.49	0.44
	ME	−1.56	−0.05	−0.24	1.11	−0.01	0.15	−2.45	−0.04	0.11	0.08	0.34	0.28
Def2-SV(P)[a]	MAE	1.91	0.39	0.48	1.77	0.49	0.50	4.57	1.51	1.72	1.02	0.51	0.59
	ME	−1.87	−0.04	−0.18	1.48	0.00	0.16	−2.97	−0.04	0.21	−0.15	0.36	0.44
Def2-SVP[b]	MAE	1.84	0.34	0.44	1.63	0.41	0.43	3.00	1.40	1.61	0.91	0.41	0.45
	ME	−1.81	−0.03	−0.20	1.37	0.07	0.19	−1.43	−0.04	0.17	−0.13	0.23	0.28
pc-1	MAE	1.21	0.27	0.29	1.16	0.41	0.43	3.86	1.02	1.22	0.72	0.30	0.30
	ME	−1.19	−0.04	−0.14	0.86	−0.03	−0.01	−3.30	−0.10	0.17	−0.09	0.14	0.16
6-31+G*	MAE	0.39	0.16	0.17	0.66	0.30	0.33	3.45	0.62	0.68	0.67	0.21	0.21
	ME	−0.33	−0.00	−0.05	0.52	−0.03	0.01	−2.83	−0.10	−0.02	0.00	0.04	0.04
6-31+G**	MAE	0.36	0.12	0.14	0.50	0.24	0.26	2.48	0.45	0.50	0.61	0.14	0.14
	ME	−0.32	−0.00	−0.06	0.38	−0.02	0.00	−2.00	−0.07	0.00	−0.03	0.02	0.03

error (ME). Thanks to its diffuse functions, the performance of BSIP-corrected 6-31+G is particularly good for NCIs, with an MAE of only 0.29 kcal/mol. In general, the MAE for the NCI subsets differ little from the MAE predicted by the least-squares fitting procedure, indicating that our choice of 1-norm constraint was appropriate. However, there are two noteworthy exceptions: the ionic-HB and the water set show a significant deviation between the self-consistent and the least-squares MAEs, probably related to the strong induction effects present in these systems (many-body induction in the case of the water clusters). It is also important to note that the water MAE is dominated by the large water clusters in the set, whose reference binding energies are in the hundreds of kcal/mol. The tendency to large non-linearity error during ACP development in water clusters had already been noted by Holmes et al.[49] in the development of water-specific ACPs, where it was found that a very tight constraint was necessary to prevent large non-linearity error effects.

The excellent performance of our BSIPs extends to the conformational energy set. In this case, the MAE from the self-consistent calculations is above 1 kcal/mol for MINI and STO-3G but not for MINIs (0.99 kcal/mol), which is the best-performing minimal-basis-set method overall. The conformational energy MAEs decrease substantially for the larger basis sets, with 6-31+G (0.37 kcal/mol) and 6-31G* (0.42 kcal/mol) offering particularly low MAEs for their cost. The MAEs for the molecular deformation superset (Geom) have a very similar behavior to the conformational energies, following essentially the same trends and with similar error values. Lastly, there are clear differences between basis sets for the chemical reaction superset ("Cov"). For the basis sets in which Cov was part of the training set (6-31G* and below in Table III), the MAEs are greatly reduced. For the others, the performance is erratic, showing improvement for some basis sets but not for others.

It is instructive to compare the performance of the BSIPs developed in this work with the previous version presented in Ref. 16. Table IV shows the MAEs in the subsets of the training set using B3LYP against the CBS reference for four basis sets for which the previous BSIPs are available (MINIs(d) was left out). Compared to the

TABLE IV. Mean absolute errors (MAE) of B3LYP combined with six basis sets compared to the B3LYP/aug-cc-pVQZ reference in the subsets of the training set (Table I), in kcal/mol. Bare = uncorrected basis set, BSIP1 = corrected with the BSIPs proposed in Ref. 16, gCP = geometric counterpoise correction of Kruse et al.[29], BSIP2 = BSIPs proposed in this work. The subsets (rows) labeled with an asterisk were in the training set of the BSIPs proposed in Ref. 16.

		MINIs				6-31G*				Def2-SV(P)			Def2-SVP			pc-1			6-31+G**		
	Set	Bare	BSIP1	gCP	BSIP2	Bare	BSIP1	gCP	BSIP2	Bare	gCP	BSIP2	Bare	gCP	BSIP2	Bare	BSIP1	BSIP2	Bare	BSIP1	BSIP2
*	S22×5	1.85	0.52	0.70	0.35	1.43	0.47	0.51	0.39	1.60	0.63	0.41	1.58	0.65	0.37	1.16	0.35	0.31	0.38	0.22	0.15
*	S66×8	1.36	0.46	0.59	0.32	1.39	0.48	0.47	0.34	1.63	0.51	0.43	1.57	0.49	0.38	1.10	0.39	0.30	0.38	0.23	0.16
*	F-set	2.02	0.62	0.99	0.69	1.29	0.46	3.99	0.29	1.67	1.33	0.39	1.50	1.96	0.35	1.18	0.41	0.27	0.41	0.21	0.15
*	P-set	1.80	0.52	0.81	0.43	0.85	0.44	0.49	0.35	1.21	0.58	0.49	1.17	0.56	0.48	0.83	0.38	0.24	0.39	0.17	0.14
*	S-set	0.67	0.46	0.53	0.39	0.86	0.36	0.32	0.29	1.50	0.37	0.37	1.49	0.34	0.36	0.94	0.33	0.25	0.45	0.18	0.15
*	Cl-set	2.18	0.43	0.97	0.44	0.73	0.46	0.52	0.39	1.51	0.58	0.50	1.45	0.63	0.48	0.81	0.45	0.31	0.39	0.19	0.16
	B-set	4.50	2.83	1.05	1.06	1.67	0.65	1.60	0.50	2.23	0.94	0.59	2.14	1.12	0.60	1.11	0.55	0.40	0.45	0.28	0.19
	Si-set	1.85	1.40	0.95	0.73	1.24	0.63	0.63	0.49	1.69	0.76	0.55	1.64	0.72	0.54	0.91	0.51	0.35	0.42	0.30	0.17
	KB49	1.84	0.57	0.72	0.55	1.59	0.48	1.05	0.33	1.95	0.72	0.43	1.91	0.85	0.44	1.32	0.40	0.26	0.42	0.23	0.16
	Water	56.39	2.81	12.11	4.38	28.35	7.87	10.13	2.85	39.06	5.44	5.28	34.35	7.54	4.64	22.94	3.42	1.01	10.56	3.59	1.28
	A24	0.93	0.45	0.58	0.52	1.12	0.49	0.60	0.36	1.27	0.54	0.35	1.25	0.56	0.38	0.78	0.29	0.21	0.32	0.22	0.15
	HSG	1.73	0.82	0.76	0.56	1.77	0.82	0.85	0.46	2.05	0.62	0.48	1.99	0.62	0.44	1.25	0.61	0.26	0.23	0.19	0.09
	IonicHB	6.73	1.74	1.69	1.65	3.49	1.85	1.80	1.19	4.48	1.46	1.48	4.11	1.57	1.23	3.07	1.37	0.82	1.17	0.78	0.44
	ADIM6	0.51	0.05	0.42	0.35	0.70	1.38	1.52	0.37	0.77	1.16	0.56	0.57	0.92	0.23	0.34	0.76	0.07	0.02	0.45	0.06
	X40×10	2.14	1.01	0.82	0.80	1.31	0.62	0.77	0.51	1.94	0.85	0.56	1.82	0.84	0.56	1.40	0.59	0.38	0.53	0.34	0.20
	S×8	0.75	0.33	0.48	0.33	0.96	0.38	0.37	0.28	1.65	0.37	0.41	1.64	0.36	0.40	1.02	0.33	0.25	0.49	0.20	0.18
	BBI	1.85	0.93	0.34	0.49	1.80	0.48	0.36	0.17	2.34	0.66	0.21	2.33	0.50	0.17	1.42	0.38	0.15	0.27	0.18	0.09
	SSI	1.40	0.66	0.68	0.47	1.41	0.78	0.75	0.37	1.50	0.64	0.38	1.49	0.61	0.38	0.96	0.52	0.25	0.18	0.19	0.10
	ACHC	0.84	1.21	0.76	0.51	1.79	0.30	0.81	0.33	2.40	0.21	0.45	2.45	1.05	0.59	1.82	0.17	0.38	0.56	0.47	0.22
	HBC6	6.94	0.76	1.77	0.65	2.74	0.66	0.96	0.65	3.47	1.38	0.72	3.60	1.18	0.87	2.43	0.71	0.36	0.37	0.23	0.14
	NBC10ext	0.89	0.48	0.47	0.39	0.83	0.43	0.26	0.28	0.75	0.35	0.35	0.75	0.25	0.31	0.65	0.31	0.21	0.26	0.15	0.11
	Bauza	5.80	13.78	1.84	15.29	1.94	3.38	1.16	2.37	7.39	1.17	4.02	7.25	1.10	4.16	2.92	4.60	2.47	1.19	1.45	0.81
	P26	5.12	1.71	1.07	1.41	0.82	0.84	0.52	0.43	0.99	0.86	0.56	0.97	0.63	0.51	0.69	1.21	0.53	0.30	0.45	0.38
	ACONF	1.30	0.93	0.15	0.62	0.12	0.54	0.35	0.47	0.19	0.28	0.36	0.17	0.24	0.20	0.13	0.21	0.28	0.04	0.22	0.16
	CYCONF	3.19	0.81	0.29	1.38	0.61	0.26	0.35	0.28	1.01	0.28	0.31	1.15	0.32	0.19	0.54	0.56	0.27	0.21	0.14	0.38
	SCONF	9.52	1.17	0.94	1.51	3.12	0.27	1.00	0.48	3.12	0.34	0.63	2.88	0.37	0.46	2.42	0.45	0.38	0.60	0.19	0.14
	BCONF	4.62	0.30	0.46	0.43	1.88	0.30	0.71	0.49	2.74	0.31	0.52	2.42	0.35	0.64	1.67	0.25	0.39	0.84	0.35	0.17
	BDE03	15.42	10.83	6.94	14.74	2.16	1.22	2.37	1.46	3.19	3.70	1.39	2.12	3.13	1.41	1.16	0.87	1.20	1.00	0.66	0.52
	BHPeri	13.38	10.19	2.01	5.15	3.63	0.97	1.64	1.28	3.14	1.00	1.36	2.74	1.02	1.03	2.01	1.06	0.76	1.74	0.59	0.37
	BSR36	2.37	16.20	0.69	1.95	0.72	2.18	1.86	0.62	3.67	1.02	0.70	0.83	1.55	1.01	0.64	3.00	0.60	0.46	0.32	0.13
	DARC	29.36	30.78	3.52	19.38	8.03	1.69	3.26	0.65	8.20	1.64	0.89	7.17	1.84	0.96	4.08	2.54	0.71	4.03	0.58	0.35
	BH76	19.09	13.03	6.64	11.82	4.26	3.67	4.16	2.33	4.28	4.71	2.17	3.61	4.38	1.97	3.14	2.90	1.44	0.97	0.83	0.52
	ISO34	8.00	7.93	3.32	4.75	2.26	1.27	2.16	1.45	2.63	2.47	1.36	1.23	2.24	1.27	0.91	0.92	0.90	0.60	0.78	0.45
	PA26	14.47	31.30	8.52	33.89	3.66	4.70	3.28	2.86	2.91	4.06	2.49	2.15	3.42	1.84	2.23	1.37	1.56	1.35	0.94	0.82
	W4-11	55.55	46.39	25.33	81.79	5.53	2.85	9.68	1.77	6.49	13.58	2.01	3.97	12.76	2.31	8.11	2.56	1.43	5.44	2.02	0.58
	BDExrel	4.53	1.27	1.82	0.99	0.64	0.67	0.80	0.44	1.02	0.97	0.50	0.91	0.90	0.42	0.72	0.48	0.30	0.61	0.22	0.14

previous version, our new BSIPs show decreased MAE for most of the subsets in the training set; all of them in the case of the 6-31+G** basis set. It is surprising that the reduction in MAE is also quite substantial for the subsets that were part of the original training set in Ref. 16, since the current training set contains almost five times as many data points as the previous one. This can be justified by the improved performance in the least-squares fit afforded by the LASSO method and also by the change in density functional used to develop the BSIPs (B3LYP in this work, BLYP in Ref. 16). The wRMS, which is the target of our fit, almost halves when comparing the old BSIPs (MINIs = 1.72, 6-31G* = 1.11, pc-1 = 0.81, and 6-31+G** = 0.43 kcal/mol) to the new ones (MINIs = 0.84, 6-31G* = 0.65, pc-1 = 0.47, and 6-31+G** = 0.21 kcal/mol). Our new BSIPs also solve the spurious behavior of the old potentials in the description of isodesmic reaction energies with minimal basis sets (represented by BSR36 in Table IV), which showed a higher MAE than the uncorrected basis set. Table IV also compares the performance of the proposed BSIPs with the results obtained using the geometric counterpoise correction (gCP) of Kruse and Grimme using the MINIs, 6-31G∗, Def2-SV(P), and Def2-SVP basis sets. In general, our new BSIPs improve upon gCP across the board, in some cases by 100% or more. The only exceptions are the treatment of covalent bond breaking with MINIs, in which gCP yields lower MAEs than the uncorrected basis set or the same basis set corrected with any of the BSIPs, and the Bauza set, which was specifically excluded from our fit.

IV. BSIP TESTING

Table V contains the MAE for uncorrected and BSIP-corrected B3LYP-D3 for a few benchmark sets. Unlike in previous sections, these average errors are computed against the reference data for the corresponding sets, rather than the near-CBS result. Our objective is to evaluate whether B3LYP-D3 in combination with a BSIP-corrected small basis set is useful in practice. Some of the sets included in the table are part of our training set (Table I) but others (the S12L[8,9], S30L[10], and PEPCONF,[128]) are not. Given the scarcity of data not in the training set, we have included an additional validation set based on the X23 molecular crystal lattice energy benchmark.[129] The geometry of the 23 crystals in the X23 set were relaxed using the B86bPBE-XDM functional, and then all unique close-contact dimers were extracted. The reference binding energy for these 106 dimers was calculated at the B3LYP-D3/aug-cc-pVQZ level.

For the NCI sets that are part of the training set (S22, S22×5, S66, S66×8, and KB49), the MAE drops significantly after applying the BSIPs, indicating an excellent performance of B3LYP-D3 with BSIP-corrected small basis sets for non-covalent interactions. In general, larger basis sets with diffuse functions give lower MAEs, even when BSIPs are used. However, calculations using MINIs with BSIPs gives a remarkably low MAEs for these sets, around or under 0.5 kcal/mol. The improvement with the application of BSIPs extends to the S12L and S30L, which comprise NCI binding energies in large molecular dimers, and were not part of the training set. In this case. the MAE reduces from 20–30 kcal/mol to 6–8 kcal/mol upon application of the BSIPs. For reference, the MAE of B3LYP-D3 with a quadruple-ζ basis set in the S30L set is 4.1 kcal/mol,[10] and the MAE of the similar B3LYP-XDM method in the S12L set is 4.0 kcal/mol.[6] Interestingly, and unlike the other NCI sets considered, the MAEs for S12L and S30L do not have a clear decreasing trend as the size of the basis set increases. For instance, B3LYP-D3/STO-3G-BSIP is the best-performing method in the S12L set. This is probably a result of error cancellation between BSIE, errors due to the approximate nature of the density functional, and errors from the applied experimental back-correction in the reference data. For the NCI binding energies in the X23, which were also not in the training set, there is a substantial reduction in MAE upon application of the BSIPs, particularly for the small basis sets.

We consider now the results for the conformational energies, measured by the PEPCONF set. The PEPCONF set[128] is a reference benchmark set for the development and testing of force fields and similar inexpensive methods in the description of peptide-peptide interactions. It contains a number of subsets designed to target specific interactions, as indicated in Table V. The dipeptide and tripeptide conformation energy sets comprise the majority of the PEPCONF set, but specific systems of interest are also included, like peptides connected by disulfide bridges ("disulfide" subset), residues associated with biofunctionality ("bioactive"), and cyclic oligopeptides ("cyclic"). The calculation level of the reference data for the PEPCONF set is LC-ωPBE-XDM/aug-cc-pVTZ.[128]

Table V shows that the MAEs in the PEPCONF set, as well as in all its subsets individually, decrease for all basis sets with the application of BSIPs. The decrease is most noticeable for the dipeptide and tripeptide subsets, with MAEs in the 1–2 kcal/mol range for the basis sets with the small weighting scheme and in the 0.3–1 kcal/mol range for the others. This gap between groups of basis sets is mostly absent for the uncorrected MAE results, suggesting that including the Cov superset in the fit may indirectly improve the description of conformational energies. The improvement caused by BSIPs extends to the other subsets of PEPCONF, with a reduction in MAE of up to a third of the uncorrected value. Reassuringly, this improvement occurs also for the disulfide subset, even though disulfide bonds were not part of the training set.

We turn our attention to the covalent bond breaking and formation sets in Table V. The BDE03 set[86] comprises homolytic bond dissociation energies of small molecules. The reference BDEs in the set were obtained from experimental formation enthalpies with a back-

TABLE V. Mean absolute errors (MAE) of uncorrected ("Bare") and BSIP-corrected B3LYP-D3 in combination with 15 basis sets for a few selected benchmark sets, in kcal/mol. The MAE is calculated relative to the corresponding reference data reported in the literature. The sets preceded by an asterisk were not part of the training set.

	Set		MINI	MINIs	STO-3G	3-21G	6-31G	MIDIh	pc-0	6-31+G
	S22	Bare	4.79	4.34	5.16	4.93	3.23	3.99	6.06	2.87
		BSIP	0.88	0.48	0.84	1.17	0.85	1.11	1.43	0.59
	S22×5	Bare	2.89	2.81	3.53	3.24	2.22	2.76	3.91	1.83
		BSIP	0.63	0.45	0.59	0.91	0.63	0.88	1.07	0.38
	S66	Bare	3.49	3.32	3.59	3.81	2.75	3.31	4.80	2.91
		BSIP	0.69	0.36	0.60	0.86	0.58	0.80	1.20	0.47
	S66×8	Bare	2.41	2.39	2.72	2.81	2.09	2.50	3.47	2.13
		BSIP	0.56	0.31	0.46	0.74	0.49	0.73	0.96	0.32
	KB49	Bare	3.63	3.17	3.72	3.76	2.54	3.29	4.83	2.40
		BSIP	1.03	0.67	1.12	1.10	0.73	1.16	1.25	0.51
*	S12L	Bare	13.39	20.77	23.02	16.85	19.57	13.51	17.82	21.67
		BSIP	8.76	8.46	6.20	8.02	8.54	8.25	12.45	9.92
*	S30L	Bare	25.62	29.37	31.58	23.93	25.48	19.70	25.29	28.11
		BSIP	7.70	8.11	8.03	7.54	7.43	9.38	11.17	9.04
*	X23	Bare	12.92	6.97	4.47	4.65	0.90	5.29	6.02	3.57
		BSIP	2.39	1.82	2.12	1.60	0.49	1.89	1.15	0.88
*	PEPCONF...all	Bare	5.02	4.08	5.10	3.61	2.76	3.13	4.80	3.04
		BSIP	2.59	1.78	2.35	1.62	1.30	1.95	1.85	1.09
*	..... bioactive	Bare	5.96	5.10	6.37	4.25	2.83	3.50	5.30	3.07
		BSIP	2.78	1.84	2.65	1.68	1.13	2.16	2.22	0.93
*	.....cyclic	Bare	6.17	4.47	6.78	6.43	3.77	5.27	9.48	3.24
		BSIP	6.07	3.40	4.74	2.29	2.07	2.20	3.56	1.48
*	.....dipeptide	Bare	3.67	2.86	3.83	2.18	1.46	1.91	2.68	1.57
		BSIP	1.74	1.16	1.54	1.10	0.77	1.40	1.07	0.65
*	.....disulfide	Bare	6.69	5.91	7.19	5.01	4.89	4.48	6.31	5.56
		BSIP	3.53	2.70	3.28	2.86	2.46	3.22	3.02	2.26
*	..... tripeptide	Bare	4.72	3.77	4.36	3.19	2.34	2.76	4.42	2.71
		BSIP	1.92	1.37	1.88	1.17	0.92	1.59	1.37	0.72
	BDE03	Bare	17.21	12.99	16.28	6.32	9.75	8.82	6.35	11.18
		BSIP	21.35	18.47	13.12	12.13	11.13	12.38	14.49	9.58
	BHPeri	Bare	14.76	9.09	12.02	5.09	7.64	3.97	2.95	7.65
		BSIP	5.75	4.50	4.53	4.53	3.60	3.90	4.00	3.16

	Set		3-21G*	6-31G*	Def2-SV(P)[a]	Def2-SVP[b]	pc-1	6-31+G*	6-31+G**	HF-3c
	S22	Bare	4.93	2.63	2.92	2.84	2.57	3.40	3.36	0.51
		BSIP	1.20	0.58	0.44	0.46	0.42	0.48	0.45	
	S22×5	Bare	3.24	1.83	2.02	1.98	1.75	2.15	2.12	0.52
		BSIP	0.95	0.54	0.48	0.50	0.42	0.36	0.33	
	S66	Bare	3.81	2.19	2.54	2.52	2.45	2.90	2.88	0.36
		BSIP	0.90	0.46	0.44	0.45	0.39	0.37	0.36	
	S66×8	Bare	2.81	1.68	1.93	1.92	1.83	2.14	2.12	0.37
		BSIP	0.76	0.44	0.42	0.46	0.36	0.28	0.25	
	KB49	Bare	3.65	2.11	2.43	2.44	2.12	2.63	2.56	0.89
		BSIP	1.08	0.54	0.59	0.60	0.40	0.40	0.39	
*	S12L	Bare	16.85	21.12	20.63	20.75	19.64	26.46	26.28	5.39
		BSIP	8.84	7.32	6.73	6.43	8.86	6.92	7.44	
*	S30L	Bare	23.78	26.37	26.10	26.15	25.32	32.49	32.29	5.57
		BSIP	8.07	6.90	7.64	7.55	8.54	6.42	6.78	
*	X23	Bare	4.65	3.74	1.64	1.44	1.19	0.87	1.09	4.26
		BSIP	1.24	1.08	0.82	0.81	0.60	0.38	0.67	
*	PEPCONF...all	Bare	3.64	2.60	2.53	2.41	2.65	3.26	3.29	2.05
		BSIP	1.52	1.04	1.32	1.29	1.02	0.88	0.83	
*	..... bioactive	Bare	4.26	2.61	2.53	2.45	2.63	3.26	3.28	2.64
		BSIP	1.62	0.82	1.12	1.07	0.82	0.59	0.54	
*	.....cyclic	Bare	6.43	2.52	2.62	2.54	2.53	2.79	2.79	3.22
		BSIP	2.04	1.14	2.05	1.70	1.08	1.45	1.28	
*	.....dipeptide	Bare	2.20	1.44	1.38	1.25	1.34	1.64	1.67	1.41
		BSIP	0.94	0.51	0.66	0.65	0.53	0.39	0.36	
*	.....disulfide	Bare	5.07	4.79	4.62	4.47	5.12	6.09	6.15	3.37
		BSIP	2.78	2.53	2.77	2.80	2.41	2.15	2.13	
*	..... tripeptide	Bare	3.22	2.29	2.24	2.12	2.30	3.04	3.05	1.49
		BSIP	1.14	0.65	0.90	0.89	0.66	0.46	0.41	
	BDE03	Bare	3.75	5.19	4.66	5.41	6.21	6.53	6.21	23.97
		BSIP	12.49	3.63	3.93	4.08	3.76	4.00	4.01	
	BHPeri	Bare	5.03	8.01	8.37	9.85	7.82	8.37	7.99	6.63
		BSIP	4.24	3.00	3.12	3.28	2.74	2.95	2.97	

correction calculated using B3LYP/6-31G(2df,p). The BHPeri set[11,87–90] was part of the BSIP training set (Table I) and benchmarks barrier heights of pericyclic reactions. In this case, Table I shows that there is a clear difference between the basis sets whose BSIPs were developed with the Cov superset in the training set (6-31G* to 6-31+G** in Table I) and the ones where the small weighting scheme was used (MINI to 3-21G*). For reference, the MAE using B3LYP-D3 and an almost complete basis set (aug-cc-pVQZ) is 4.1 kcal/mol for BDE03 and 1.8 kcal/mol for BHPeri.

For the minimal and small basis sets, without Cov in their training set, the performance of the BSIP is erratic, with occasional improvement of the BDE03 MAE and general improvement for BHPeri. In the case of the large basis sets, the MAE is decreased in all cases, although to a lesser extent than in the case of the NCIs, evidencing the smaller impact of BSIE in the calculation of these properties. For the BDE03 set, the reduction in MAE is such that almost the same MAE is obtained with the BSIP-corrected basis sets as with the near-CBS basis set (4.1 kcal/mol). This is not the case for BHPeri (the B3LYP-D3/aug-cc-pVQZ MAE is 1.8 kcal/mol) but the MAE from the BSIP-corrected basis sets is reasonably close and there is a very significant improvement relative to the uncorrected MAE, indicating a very important influence of basis-set incompleteness in barrier height calculations. Curiously, once BSIP-corrected, the small and minimal basis sets that used the small weighting scheme in the BSIP fit produce a similar, but slightly higher MAE.

For comparison, Table V also shows the MAEs for the HF-3c method,[130,131] based on HF/MINIs plus three additional corrections, one of which is gCP. The B3LYP-D3/MINIs-BSIP method, which is comparable to HF-3c in cost, achieves lower MAEs in almost all sets, except S12 and S30L. Larger basis sets have even lower MAEs. The basis sets for which the Cov set was included in the training set can be used for thermochemical calculations.

It is also interesting to consider whether the proposed BSIPs are transferable, that is, if they can be used with functionals other than the one used in their development (B3LYP). We have shown in previous work[16] that BSIE is almost functional-independent, and is mostly determined by the fraction of exact exchange in the functional definition. Table VI shows the results of our transferability tests for the proposed BSIPs. We chose a few functionals (BLYP, PBE, B3LYP, PBE0, and LC-ωPBE, all with the D3 dispersion correction) and a few test sets (S22, S66, BDE03, and BHPeri) and calculated the MAE using seven BSIP-corrected basis sets of various sizes (MINIs, 6-31G, 6-31+G, 6-31G*, Def2-SVP, 6-31+G*, and 6-31+G**). The almost-complete-basis-set results, at the aug-cc-pVQZ level, are also given in the table for reference.

Except for BDE03 combined with the basis sets for which the Cov superset was not a target of the BSIP fit (MINIs, 6-31G, and 6-31+G), the results are excellent and close to the near-CBS value regardless of functional and basis set. In the NCI sets the MAEs are always lower than 1 kcal/mol (S22) and 0.6 kcal/mol (S66). For the basis sets that include at least one diffuse function, the MAE is only a few tenths of a kcal/mol above the near-CBS reference value. In the case of the bond dissociation energies (BDE03 set), there is very little difference between the near-CBS and the BSIP-corrected MAEs. The fact that many of these MAEs are smaller than the one obtained using a near-complete-basis-set indicates that there is error cancellation present and that the BSIE no longer dominates the overall error in the BDE calculation. The same can be said about the barrier height set (BHPeri), in which even the BSIP-corrected small and minimal basis sets achieve an MAE similar to the near-CBS reference value. Therefore, it is clear from Table VI that our BSIPs can be utilized in combination with any of the usual functionals, other than the Minnesota functionals.

As a final test, we consider the energy difference (ΔE) between the folded and unfolded conformation of crambin, a protein with 642 atoms.[29] In the original paper by Kruse et al.[29] and in our previous work,[16] only BLYP-D3 was considered. Table VII shows the ΔE between both forms of crambin with several basis sets, functionals, and with and without BSIPs. Calculations using the pc-2 triple-ζ basis set were used as reference. As the pc-2 result likely contains significant BSIE, we also report in the table the standard deviation of the ΔE calculated with the six basis sets.

The table shows that the new BSIPs greatly improve the agreement with the pc-2 reference for every basis set and functional compared to the uncorrected calculation and, at the same time, they also reduce the spread of the ΔE calculated with different basis sets. The best performance is obtained with B3LYP, which is the functional used in the development of the BSIPs, with an 8-fold decrease in MAE. However, BLYP-D3 and LC-ωPBE-D3 are also significantly improved. The proposed BSIPs in combination with BLYP also improve slightly upon the BSIPs proposed in our previous article,[16] which is particularly encouraging since BLYP was the functional used to develop the BSIPs in that work.

V. CONCLUSIONS

In this article, we developed basis-set incompleteness potentials (BSIP) for a number of small and minimal basis sets and a set of atoms relevant in organic chemistry and biochemistry. The objective of BSIPs is to alleviate the detrimental effect of basis-set incompleteness on calculated molecular properties using density-functional theory (DFT). In large systems, the near-complete basis set necessary to describe non-covalent interactions and chemical reactions accurately cannot be used due to their computational cost. BSIPs are one-electron potentials, similar to effective-core potentials, that provide an en-

TABLE VI. Mean absolute error (MAE) of various D3-corrected density functionals with some BSIP-corrected basis sets in a few benchmark sets, in kcal/mol. The MAEs are calculated relative to the corresponding reference data reported in the literature. The last column gives the MAE with an almost-complete basis set (aQZ = aug-cc-pVQZ).

Set	Functional	MINIs	6-31G	6-31+G	6-31G*	Def2-SVP	6-31+G*	6-31+G**	aQZ
S22	BLYP-D3	0.68	0.91	0.41	0.64	0.58	0.33	0.30	0.22
	PBE-D3	0.90	0.73	0.51	0.61	0.58	0.45	0.44	0.57
	B3LYP-D3	0.48	0.85	0.59	0.58	0.46	0.48	0.45	0.31
	PBE0-D3	0.68	0.66	0.65	0.50	0.44	0.52	0.52	0.53
	LC-ωPBE-D3	0.53	0.47	0.59	0.37	0.46	0.54	0.50	0.31
S66	BLYP-D3	0.47	0.60	0.36	0.48	0.52	0.27	0.25	0.19
	PBE-D3	0.54	0.57	0.45	0.46	0.58	0.39	0.40	0.39
	B3LYP-D3	0.36	0.58	0.47	0.46	0.45	0.37	0.36	0.22
	PBE0-D3	0.53	0.55	0.51	0.43	0.53	0.42	0.42	0.36
	LC-ωPBE-D3	0.44	0.40	0.46	0.33	0.56	0.36	0.33	0.21
BDE03	BLYP-D3	19.21	12.25	11.04	4.91	5.39	5.68	5.48	5.08
	PBE-D3	16.20	10.75	9.30	3.99	4.04	3.79	3.75	4.43
	B3LYP-D3	18.47	11.13	9.58	3.63	4.08	4.00	4.01	4.12
	PBE0-D3	17.38	10.70	9.16	3.25	3.63	3.33	3.37	4.01
	LC-ωPBE-D3	17.23	10.77	8.71	2.96	3.27	2.99	3.12	3.57
BHPeri	BLYP-D3	4.01	3.74	3.11	2.56	2.98	2.72	2.80	4.18
	PBE-D3	4.26	4.95	4.33	4.06	4.32	4.32	4.23	6.69
	B3LYP-D3	4.50	3.60	3.16	3.00	3.28	2.95	2.97	1.18
	PBE0-D3	4.03	4.02	3.55	3.33	3.35	3.25	3.08	3.27
	LC-ωPBE-D3	4.31	2.50	2.25	3.34	3.42	2.86	2.75	2.18

TABLE VII. Total energy differences between the unfolded and folded forms of crambin using selected functionals and basis sets and with and without BSIP correction. For BLYP, the current BSIPs (BSIP2) are compared to the results using the BSIPs from our previous work.[16] The reference energies were calculated using the pc-2 basis set[100–103] (B3LYP-D3 = 245.1; BLYP-D3 = 234.2; LC-ωPBE-D3 = 245.7 kcal/mol). The mean absolute error (MAE) and standard deviation are indicated in the bottom row. All values in kcal/mol.

Basis Set	B3LYP-D3		BLYP-D3			LC-ωPBE-D3	
	Bare	BSIP2	Bare	BSIP2	BSIP1	Bare	BSIP2
MINIs	242.6	246.6	238.0	227.9	228.7	225.0	224.3
pc-0	312.9	224.5	312.2	222.8		307.3	212.2
6-31G*	305.2	248.2	305.6	227.0	219.5	299.3	215.7
Def2-SVP	258.4	250.7	249.0	242.1		267.6	261.3
pc-1	388.9	256.1	391.4	254.9		367.1	231.1
6-31+G**	292.7	244.2	282.2	233.6	225.3	306.1	250.3
MAE	55.9	7.1	62.2	9.0	9.7	56.6	20.0
Std. dev.	51.3	10.9	55.2	11.9		47.2	19.5

ergy correction that changes depending on the chemical environment and the system's Kohn-Sham orbitals. The two types of adjustable parameters in BSIPs, the coefficients and exponents of the radial Gaussian functions, are determined in a fitting procedure in which we preselect a number of exponents for each atom, then calculate the coefficients by linear least-squares minimization of the BSIP energy compared to the difference between the near-CBS energy and the uncorrected energy using the B3LYP density functional. Since basis-set incompleteness error is mostly transferable between density functionals, using the B3LYP near-CBS to generate the fitting and reference data ensures that the developed BSIPs can be applied to a wide range of density functionals and are tied only to the corresponding basis sets.

In this work, BSIPs were developed for ten atoms (H, B–F, Si–Cl) and for 15 double-ζ and minimal basis sets, some with polarization and/or diffuse functions (MINI, MINIs, STO-3G, 3-21G, 6-31G, MIDIh, pc-0, 6-31+G, 3-21G*, 6-31G*, Def2-SV(P)[a], Def2-SVP[b], pc-1, 6-31+G*, and 6-31+G**). The BSIPs involve the local plus all angular momentum channels up to the maximum l in each basis set, and 26 pre-selected exponents per atom and channel. The training set contains a mixture of molecular properties, including non-covalent binding energies, reaction energies, barrier heights, conformational energies, and molecular deformations, for a total of 15,944 data points. The linear least-squares fit used to determine the BSIP coefficients is carried out using the LASSO regularization method, which serves the double purpose of keeping the 1-norm of the coefficients below a given threshold and performing variable selection, thus resulting in smaller and therefore more efficient BSIPs.

The BSIPs were tested by running self-consistent

BSIP-corrected calculations on the training set, demonstrating excellent performance in reproducing near-complete-basis-set results, much better in fact than our previous version of the same potentials that used a different fitting technique. The BSIPs were also tested in several sets for large molecules not contained in the training set (S12L and S30L) as well as in the description of conformational equilibria in polypeptides (the PEPCONF set). In addition, we showed that the new BSIPs can be used with functionals other than B3LYP, and that the description of molecular properties in very large systems, such as the difference between the folded and unfolded conformation of crambin, is also improved. We are confident that these new BSIPs offer a simple way of carrying out quantum chemical calculations in large systems with improved accuracy at the computational cost of using a small or minimal basis set.

VI. ACKNOWLEDGEMENTS

AOR thanks the Spanish government for a Ramón y Cajal fellowship (RyC-2016-20301) and for financial support (projects PGC2018-097520-A-100 and RED2018-102612-T), and the MALTA Consolider supercomputing centre and Compute Canada for computational resources. GAD thanks the National Science and Engineering Research Council, the Canadian Foundation for Innovation, and the British Columbia Knowledge Development Fund for financial support.

SUPPORTING INFORMATION

The Supporting Information contains:

- `si.pdf`: B3LYP MAEs for all subsets of the training set and all basis sets with and without BSIPs. Exponents and coefficients for all the proposed BSIPs
- `si_files/bsip/`: BSIPs in Gaussian ECP format.
- `si_files/db/`: training set database.
- `si_files/gbs/`: specification for the basis sets that do not have a Gaussian keyword, in Gaussian format.

* aoterodelaroza@gmail.com
† gino.dilabio@ubc.ca

[1] Goerigk, L.; Reimers, J. R. Efficient methods for the quantum chemical treatment of protein structures: the effects of london-dispersion and basis-set incompleteness on peptide and water-cluster geometries. *J. Chem. Theory Comput.* **2013**, *9*, 3240–3251.
[2] Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J. How large should the QM region be in QM/MM calculations? The case of catechol O-methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381–11394.
[3] Otero-de-la-Roza, A., DiLabio, G., Eds. *Non-covalent Interactions in Quantum Chemistry and Physics*; Elsevier, in press, 2017.
[4] DiLabio, G. A.; Otero-de-la-Roza, A. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Ed.; Wiley-VCH, 2016; Vol. 29; pp 1–97.
[5] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion Interactions in Density-Functional Theory. *J. Phys. Org. Chem.* **2009**, *22*, 1127.
[6] 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.
[7] 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.
[8] Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.
[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] Sure, R.; Grimme, S. Comprehensive benchmark of association (free) energies of realistic host–guest complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.
[11] Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.
[12] 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.
[13] Koch, W.; Holthausen, M. C. *A chemist's guide to density functional theory*; Wiley-Vch: Weinheim, 2001.
[14] Johnson, E. R.; Otero-de-la-Roza, A.; Dale, S. G.; DiLabio, G. A. Efficient basis sets for non-covalent interactions in density-functional theory. *J. Chem. Phys.* **2013**, *139*, 214109.
[15] Witte, J.; Neaton, J. B.; Head-Gordon, M. Push it to the limit: Characterizing the convergence of common sequences of basis sets for intermolecular interactions as described by density functional theory. *J. Chem. Phys.* **2016**, *144*, 194306.
[16] 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.
[17] Davidson, E. R.; Feller, D. Basis set selection for molecular calculations. *Chem. Rev* **1986**, *86*, 681–696.
[18] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publications, 1996.
[19] Boys, S. F.; Bernardi, F. The calculation of small molec-

ular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553–566.

[20] Gutowski, M.; Chal/asiński, G. Critical evaluation of some computational approaches to the problem of basis set superposition error. *J. Chem. Phys.* **1993**, *98*, 5540–5554.

[21] van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. State-of-the-Art in Counterpoise Theory. *Chem. Rev.* **1994**, *94*, 1873–1885.

[22] Mentel, Ł.; Baerends, E. Can the counterpoise correction for basis set superposition effect be justified? *J. Chem. Theory Comput.* **2013**, *10*, 252–267.

[23] Burns, L. A.; Marshall, M. S.; Sherrill, C. D. Comparing counterpoise-corrected, uncorrected, and averaged binding energies for benchmarking noncovalent interactions. *J. Chem. Theory Comput.* **2014**, *10*, 49–57.

[24] Brauer, B.; Kesharwani, M. K.; Martin, J. M. Some observations on counterpoise corrections for explicitly correlated calculations on noncovalent interactions. *J. Chem. Theory Comput.* **2014**, *10*, 3791–3799.

[25] Mackie, I. D.; DiLabio, G. A. Approximations to complete basis set-extrapolated, highly correlated non-covalent interaction energies. *J. Chem. Phys.* **2011**, *135*, 134318.

[26] 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.

[27] Galano, A.; Alvarez-Idaboy, J. R. A new approach to counterpoise correction to BSSE. *J. Comput. Chem.* **2006**, *27*, 1203–1210.

[28] Jensen, F. An atomic counterpoise method for estimating inter-and intramolecular basis set superposition errors. *J. Chem. Theory Comput.* **2010**, *6*, 100–106.

[29] 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*, 04B613.

[30] Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

[31] 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* **2015**, *5*, 94.

[32] Kulik, H. J.; Seelam, N.; Mar, B. D.; Martínez, T. J. Adapting DFT+U for the chemically motivated correction of minimal basis set incompleteness. *J. Phys. Chem. A* **2016**, *120*, 5939–5949.

[33] Deng, J.; Gilbert, A.; Gill, P. Communication: efficient counterpoise corrections by a perturbative approach. *J. Chem. Phys.* **2011**, *135*, 081105.

[34] Mao, Y.; Horn, P. R.; Mardirossian, N.; Head-Gordon, T.; Skylaris, C.-K.; Head-Gordon, M. Approaching the basis set limit for DFT calculations using an environment-adapted minimal basis with perturbation theory: Formulation, proof of concept, and a pilot implementation. *J. Chem. Phys.* **2016**, *145*, 044109.

[35] Frisch, M. J. et al. Gaussian 09 Revision A.1. Gaussian Inc. Wallingford CT 2009.

[36] Frisch, M. J. et al. Gaussian˜16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

[37] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[38] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

[39] 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.

[40] Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[41] Woon, D. E.; Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

[42] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **1996**, *58*, 267–288.

[43] Osborne, M. R.; Presnell, B.; Turlach, B. A. A new approach to variable selection in least squares problems. *IMA J. Num. Anal.* **2000**, *20*, 389–403.

[44] Schmidt, M. Graphical Model Structure Learning with l1-Regularization. Ph.D. thesis, The University of British Columbia, Vancouver, 2010.

[45] Schmidt, M. *Optimization Methods for l1-Regularization*; 2009.

[46] Andzelm, J.; Kłobukowski, M.; Radzio-Andzelm, E. Compact contracted Gaussian-type basis sets for halogen atoms. Basis-set superposition effects on molecular properties. *J. Comput. Chem.* **1984**, *5*, 146–161.

[47] Otero-de-la-Roza, A.; Johnson, E. R.; DiLabio, G. A. Halogen bonding from dispersion-corrected density-functional theory: the role of delocalization error. *J. Chem. Theory Comput.* **2014**, *10*, 5436–5447.

[48] In the literature, BDE is normally used to denote bond dissociation enthalpy instead of bond dissociation energy.

[49] 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.

[50] 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.

[51] 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.

[52] 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*, 6092–6098.

[53] Torres, E.; DiLabio, G. A. A (nearly) universally applicable method for modeling noncovalent interactions using B3LYP. *J. Phys. Chem. Lett.* **2012**, *3*, 1738–1744.

[54] 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.

[55] 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.

[56] Von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. Optimization of effective atom centered potentials for London dispersion forces in density functional theory. *Phys. Rev. Lett.* **2004**, *93*, 153004.

[57] DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. Simple one-electron quantum capping potentials for use in hybrid QM/MM studies of biological molecules. *J. Chem. Phys.* **2002**, *116*, 9578–9584.

[58] 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.

[59] Jurečka, P.; Šponer, J.; Černỳ, 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.

[60] 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.

[61] 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.

[62] 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.

[63] Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. The S66x8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905–20925.

[64] Kannemann, F. O.; Becke, A. D. Van der waals interactions in density-functional theory: Rare-gas diatomics. *J. Chem. Theory Comput.* **2009**, *5*, 719–727.

[65] Kannemann, F. O.; Becke, A. D. van der Waals Interactions in Density-Functional Theory: Intermolecular Complexes. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.

[66] 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.

[67] 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.

[68] Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz Jr, K. M. Formal estimation of errors in computed absolute interaction energies of protein-ligand complexes. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.

[69] Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, delta[sub MP2][sup CCSD(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.

[70] Řezáč, J.; Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **2011**, *8*, 141–151.

[71] 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.

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

[73] 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.

[74] Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G.; Vanommeslaeghe, K.; MacKerell Jr, A. D.; Merz Jr, 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.

[75] 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.

[76] Smith, D. G.; 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.

[77] 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.* **2010**, *7*, 88–96.

[78] Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, δ MP2 CCSD (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.

[79] Burns, L. A.; Mayagoitia, Á. V.; Sumpter, B. G.; Sherrill, C. D. Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *J. Chem. Phys.* **2011**, *134*, 084107.

[80] Bauza, A.; Alkorta, I.; Frontera, A.; Elguero, J. On the reliability of pure and hybrid DFT methods for the evaluation of halogen, chalcogen, and pnicogen bonds involving anionic and neutral electron donors. *J. Chem. Theory Comput.* **2013**, *9*, 5201–5210.

[81] 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*, 2747–2757.

[82] Gruzman, D.; Karton, A.; Martin, J. M. Performance of Ab Initio and Density Functional Methods for Conformational Equilibria of CnH2n+ 2 Alkane Isomers (n= 4- 8). *J. Phys. Chem. A* **2009**, *113*, 11974–11983.

[83] Wilke, J. J.; Lind, M. C.; Schaefer III, H. F.; Császár, A. G.; Allen, W. D. Conformers of gaseous cysteine. *J. Chem. Theory Comput.* **2009**, *5*, 1511–1523.

[84] 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.
[85] Kozuch, S.; Bachrach, S. M.; Martin, J. M. Conformational equilibria in butane-1, 4-diol: a benchmark of a prototypical system with strong intramolecular H-bonds. *J. Phys. Chem. A* **2013**, *118*, 293–303.
[86] Johnson, E. R.; Clarkin, O. J.; DiLabio, G. A. Density Functional Theory Based Model Calculations for Accurate Bond Dissociation Enthalpies. 3. A Single Approach for X-H, X-X, and X-Y (X, Y = C, N, O, S, Halogen) Bonds. *J. Phys. Chem. A* **2003**, *107*, 9953–9963.
[87] Karton, A.; Tarnopolsky, A.; Lamere, J.-F.; Schatz, G. C.; Martin, J. M. Highly accurate first-principles benchmark data sets for the parametrization and validation of density functional and other approximate methods. Derivation of a robust, generally applicable, double-hybrid functional for thermochemistry and thermochemical kinetics. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.
[88] Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. A standard set of pericyclic reactions of hydrocarbons for the benchmarking of computational methods: the performance of ab initio, density functional, CASSCF, CASPT2, and CBS-QB3 methods for the prediction of activation barriers, reaction energetics, and transition state geometries. *J. Phys. Chem. A* **2003**, *107*, 11445–11459.
[89] Ess, D. H.; Houk, K. Activation energies of pericyclic reactions: Performance of DFT, MP2, and CBS-QB3 methods for the prediction of activation barriers and reaction energetics of 1, 3-dipolar cycloadditions, and revised activation enthalpies for a standard set of hydrocarbon pericyclic reactions. *J. Phys. Chem. A* **2005**, *109*, 9542–9553.
[90] Dinadayalane, T.; Vijaya, R.; Smitha, A.; Sastry, G. N. Diels- Alder Reactivity of Butadiene and Cyclic Five-Membered Dienes ((CH) 4X, X= CH2, SiH2, O, NH, PH, and S) with Ethylene: A Benchmark Study. *J. Phys. Chem. A* **2002**, *106*, 1627–1633.
[91] Krieg, H.; Grimme, S. Thermochemical benchmarking of hydrocarbon bond separation reaction energies: Jacob's ladder is not reversed! *Mol. Phys.* **2010**, *108*, 2655–2666.
[92] Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Delocalization errors in density functionals and implications for main-group thermochemistry. *J. Chem. Phys.* **2008**, *129*, 204112.
[93] Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.
[94] Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark database of barrier heights for heavy atom transfer, nucleophilic substitution, association, and unimolecular reactions and its use to test theoretical methods. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.
[95] Grimme, S.; Steinmetz, M.; Korth, M. How to compute isomerization energies of organic molecules with quantum chemical methods. *J. Org. Chem.* **2007**, *72*, 2118–2126.
[96] Parthiban, S.; Martin, J. M. Assessment of W1 and W2 theories for the computation of electron affinities, ionization potentials, heats of formation, and proton affinities. *J. Chem. Phys.* **2001**, *114*, 6014–6029.
[97] Karton, A.; Daon, S.; Martin, J. M. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **2011**, *510*, 165–178.
[98] Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.
[99] Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
[100] Jensen, F. Polarization consistent basis sets: Principles. *J. Chem. Phys.* **2001**, *115*, 9113–9125.
[101] Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn–Sham basis set limit. *J. Chem. Phys.* **2002**, *116*, 7372–7379.
[102] Jensen, F. Polarization consistent basis sets. III. The importance of diffuse functions. *J. Chem. Phys.* **2002**, *117*, 9234–9240.
[103] Jensen, F.; Helgaker, T. Polarization consistent basis sets. V. The elements Si–Cl. *J. Chem. Phys.* **2004**, *121*, 3463–3470.
[104] Kalescky, R.; Kraka, E.; Cremer, D. Accurate determination of the binding energy of the formic acid dimer: The importance of geometry relaxation. *J. Chem. Phys.* **2014**, *140*, 084315.
[105] Otero-de-la-Roza, A.; DiLabio, G. A.; Johnson, E. R. Exchange-correlation effects for non-covalent interactions in density-functional theory. *J. Chem. Theory Comput.* **2016**, *12*, 3160–3175.
[106] Vydrov, O. A.; Scuseria, G. E. Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.* **2006**, *125*, 234109.
[107] Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *J. Chem. Phys.* **2006**, *125*, 074106.
[108] Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *J. Chem. Phys.* **2007**, *127*, 154108.
[109] Otero-de-la Roza, A.; Johnson, E. R. Non-Covalent Interactions and Thermochemistry using XDM-Corrected Hybrid and Range-Separated Hybrid Density Functionals. *J. Chem. Phys.* **2013**, *138*, 204109.
[110] 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.
[111] Huzinaga, S.; Andzelm, J.; Klobukowski, M.; Radzio-Andzelm, E.; Sakai, Y.; Tatewaki, H. *Gaussian Basis Sets for Molecular Calculations*; Elsevier: Amsterdam, 1984.
[112] Hehre, W. J.; Stewart, R. F.; Pople, J. A. self-consistent molecular-orbital methods. i. use of gaussian expansions of Slater-type atomic orbitals. *J. Chem. Phys.* **1969**, *51*, 2657–2664.
[113] Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.
[114] Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements. *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.
[115] Pietro, W. J.; Francl, M. M.; Hehre, W. J.; DeFrees, D. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods. 24. Supplemented small split-valence basis sets for second-row elements. *J. Am. Chem. Soc.* **1982**, *104*, 5039–5048.

[116] Dobbs, K.; Hehre, W. Molecular orbital theory of the properties of inorganic and organometallic compounds 5. Extended basis sets for first-row transition metals. *J. Comput. Chem.* **1987**, *8*, 861–879.
[117] Dobbs, K.; Hehre, W. Molecular orbital theory of the properties of inorganic and organometallic compounds. 6. Extended basis sets for second-row transition metals. *J. Comput. Chem.* **1987**, *8*, 880–893.
[118] Dithcfield, R.; Hehre, W.; Pople, J. Self-consistent molecular-orbital methods. 9. Extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54*, 2921–2923.
[119] Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
[120] Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.
[121] Hariharan, P.; Pople, J. Accuracy of AH n equilibrium geometries by single determinant molecular orbital theory. *Mol. Phys.* **1974**, *27*, 209–214.
[122] Gordon, M. S. The isomers of silacyclopropane. *Chem. Phys. Lett.* **1980**, *76*, 163–168.
[123] 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.
[124] Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
[125] Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.
[126] Jensen, F. Unifying general and segmented contracted basis sets. Segmented polarization consistent basis sets. *J. Chem. Theory Comput.* **2014**, *10*, 1074–1085.
[127] Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular orbital theory of the electronic structure of organic compounds. V. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.
[128] 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.
[129] Otero-de-la-Roza, A.; Johnson, E. R. A benchmark for non-covalent interactions in solids. *J. Chem. Phys.* **2012**, *137*, 054103.
[130] Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.
[131] Brandenburg, J. G.; Grimme, S. Dispersion Corrected Hartree-Fock and Density Functional Theory for Organic Crystal Structure Prediction. *Top. Curr. Chem.* **2014**, *345*, 1–24.