SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Quantum chemical benchmark databases of gold-standard dimer interaction energies

Alexander G. Donchev[1]✉, Andrew G. Taube[1], Elizabeth Decolvenaere[1], Cory Hargus[1], Robert T. McGibbon[1], Ka-Hei Law[1], Brent A. Gregersen[1], Je-Luen Li[1], Kim Palmo[1], Karthik Siva[1], Michael Bergdorf[1], John L. Klepeis[1] & David E. Shaw[1,2]✉

Advances in computational chemistry create an ongoing need for larger and higher-quality datasets that characterize noncovalent molecular interactions. We present three benchmark collections of quantum mechanical data, covering approximately 3,700 distinct types of interacting molecule pairs. The first collection, which we refer to as DES370K, contains interaction energies for more than 370,000 dimer geometries. These were computed using the coupled-cluster method with single, double, and perturbative triple excitations [CCSD(T)], which is widely regarded as the gold-standard method in electronic structure theory. Our second benchmark collection, a core representative subset of DES370K called DES15K, is intended for more computationally demanding applications of the data. Finally, DES5M, our third collection, comprises interaction energies for nearly 5,000,000 dimer geometries; these were calculated using SNS-MP2, a machine learning approach that provides results with accuracy comparable to that of our coupled-cluster training data. These datasets may prove useful in the development of density functionals, empirically corrected wavefunction-based approaches, semi-empirical methods, force fields, and models trained using machine learning methods.

Background & Summary

Noncovalent interactions are essential determinants of the properties of molecular liquids and crystals, solvation effects, and the structure and function of biomolecules. Experimental means of quantifying individual noncovalent interactions are limited to small systems with relatively rigid intramolecular degrees of freedom[1], and computer simulations offer a much-needed alternative; quantum mechanical (QM) calculations, for example, enable the characterization of noncovalent interactions with high accuracy. Among QM-based approaches, the use of coupled-cluster singles and doubles with perturbative triples [CCSD(T)][2–4] at the complete basis set (CBS) limit is widely recognized as the gold-standard method for noncovalent interactions[4].

High-accuracy QM methods come with an intrinsically high cost; CCSD(T), for example, scales as $O(N^7)$ with system size. Publicly available databases[5–14] offer a way to amortize this cost over a large user community, thus reducing the burden on individual researchers. Such databases serve as recognized benchmarks, and are indispensable resources for both accuracy assessment and parameterization of more affordable QM approximations such as exchange-correlation functionals[12–17] in the density functional theory framework, empirically corrected wavefunction-based approaches[18–23], and semi-empirical methods[24–27] (for a comprehensive review, see summary works[28,29]). Benchmark-quality QM data, often in combination with experimental data, also feature prominently in the development of many empirical molecular mechanics–based models (so-called "force fields")[30–34]. Diverse, extensive, and consistent collections of high-quality data, moreover, can enable powerful machine learning approaches to be leveraged for molecular modeling[35–41].

Here we present three benchmark databases of quantum chemical data, including the full Cartesian coordinates of the associated geometries[42]. The first is DES370K, a database of dimer interaction energies computed at the CCSD(T)/CBS level of theory. This database features 370,959 unique geometries for 3,691 distinct dimers, which represent 392 closed-shell chemical species (both neutral molecules and ions) including, but not limited to, water and the functional groups found in proteins. An important subset of the data in the DES370K collection

[1]D. E. Shaw Research, New York, NY, 10036, USA. [2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, 10032, USA. ✉e-mail: Alexander.Donchev@DEShawResearch.com; David.Shaw@DEShawResearch.com

Database	Protocol	Monomers	Dimers	Groups	Dimer geometries
DES370K	Dimer scans based on QM optimization[a]	166	3,436	3,476	97,368
	Dimer scans based on MD configurations[b]	382	466	6,133	166,914
	Homodimer single points based on MD configurations[c]	91	91	910	42,201
	Heterodimer single points based on MD configurations[d]	261	261	2,150	64,476
	Total	392	3,691	12,669	370,959
DES15K (subset of DES370K)	Dimer scans based on QM optimization[a]	159	3,052	3,052	12,183
	Dimer scans based on MD configurations[b]	137	206	1,929	2,468
	Total	159	3,052	4,981	14,651
DES5M	Dimer scans based on QM optimization[a]	153	2,826	71,847	2,404,926
	Dimer scans based on MD configurations[b]	159	328	47,648	1,646,832
	Homodimer single points based on MD configurations[c]	138	138	12,983	464,951
	Heterodimer single points based on MD configurations[d]	163	163	14,641	439,229
	Total	206	2,967	147,119	4,955,938

Table 1. Summary information for the DES370K, DES15K, and DES5M databases. For each database, we list the protocols employed to generate particular subsets of the data, counts associated with those subsets, and the total count across subsets. The counts shown are the number of chemically distinct monomer types ("Monomers"); the number of chemically distinct dimer types ("Dimers"); the number of groups ("Groups"), where a group is a set of connected calculations, such as those from a radial profile under a dimer-scan protocol or those from a single MD frame under a single-point protocol; and the total number of dimer calculations (i.e., entries in the database) ("Dimer geometries"). [a]Reference dimer geometries were identified using QM optimization and used to construct a group of radial scan–based geometries. [b]Reference dimer geometries were extracted from MD simulations of neat liquids and solvated monomers and used to construct a group of radial scan–based geometries. [c]Reference multimer geometries were extracted from MD simulations of neat liquids and decomposed into a group of single-point dimer geometries. [d]Reference multimer geometries were extracted from MD simulations of solvated monomers and decomposed into a group of single-point dimer geometries.

consists of QM-optimized dimer structures, which were used as starting points to generate additional structures along one-dimensional radial profiles. To enhance orientational diversity and ensure adequate sampling of the internal degrees of freedom in the larger chemical species, the dataset also includes a large ensemble of structures (and corresponding radial profiles) obtained from molecular dynamics (MD) simulations (Table 1). Because many potential applications of the presented data, such as parameterizing a new exchange-correlation functional, are computationally demanding, we additionally compiled DES15K, a core subset of the most representative structures from DES370K that largely retains the chemical and orientational diversity of DES370K, but with reduced resolution of scan points in the radial profiles (Table 1).

The DES370K collection was the source of both training and test data for a machine learning method, SNS-MP2, which we have described in full detail elsewhere[39]. Briefly, the SNS-MP2 approach combines the spin-component-scaled second-order Møller-Plesset perturbation theory (MP2) method[43] with a neural network to predict per-conformer same-spin and opposite-spin energy scaling coefficients. We found[39] that for dimer interaction energies, the SNS-MP2 method offers—at a greatly reduced cost—accuracy comparable to that of the CCSD(T)/CBS approach used to obtain the benchmark data in DES370K. The SNS-MP2 neural network also provides per-conformer confidence intervals for the predicted interaction energies[39].

Using the SNS-MP2 approach, we generated DES5M, a database of predicted gold-standard dimer interaction energies and their associated confidence intervals (Table 1). The DES5M collection contains 4,955,938 additional unique geometries originating from the same two sources as were used for DES370K: radial profiles starting from a set of QM-optimized conformers and dimer geometries extracted from MD simulations. Both the DES5M and DES370K databases also include the full set of MP2-based QM observables that serve as inputs to the SNS-MP2 procedure[39], thereby allowing for the parameterization and evaluation of other SNS-MP2-like models.

We expect that these three databases will serve as valuable benchmarks for a variety of approximate methods in computational chemistry.

Methods

Generation of monomer geometries. Input monomers were specified in the simplified molecular-input line-entry system (SMILES) string format[44]. Hydrogen atoms were added and initial three-dimensional (3D) conformations were generated using the Open Babel[45] software package. The geometry was then optimized using the OPLS_2005 force field[46], starting from a large number of perturbed initial structures (with dihedral angles sampled randomly from a uniform distribution over the range ±180° and out-of-plane angles sampled randomly from a uniform distribution over the range ±30°), to identify a set of unique stable conformers for each monomer.

Our intent was to use the QM data to fit force fields for MD simulation, and so we followed the common practice of constraining bonds to hydrogen atoms and valent angles involving two hydrogens to predefined target values. These constraints lead to more stable MD simulations, thus enabling the use of larger time steps. For a bond length, the target value was derived as a sum of three contributions: the equilibrium distance (R_e); a vibrational correction, which accounts for the anharmonicity of the stretch potential; and a correction to account for condensed-phase effects in water. The vibrational correction was estimated by approximating the monomer energy with a Morse potential[47] as a function of the bond length, $U(R) = D_e(1 - e^{-\alpha(R-R_e)})^2$, where D_e and α are fitted parameters that control the well depth and width of the potential, respectively. The equation for the Morse potential leads to the following relationship between the equilibrium (R_e) and vibrationally averaged (R_g) bond lengths: $R_g - R_e = 3\hbar/4\sqrt{2D_e\mu}$, where ħ denotes the Planck constant and μ the reduced mass of the two bonded atoms. The condensed-phase effects were estimated from the energy derivative along the stretch coordinate in a system containing the molecule of interest surrounded by a 4-Å-thick shell of solvent molecules (typically consisting of 16–26 waters). Constraint targets for valent angles were derived from their equilibrium values corrected for the condensed phase effect. We omitted angle vibrational corrections, which we expected to have only a small impact on intermolecular interactions.

We subjected our set of force field–derived monomer conformations to QM-geometry optimization, applying constraints to hydrogen-containing bond lengths and angles, at the MP2 level of theory using the density-fitting, local, and frozen-core approximations (DF-LMP2)[48–58] in the MOLPRO 2012 quantum software package (http://www.molpro.net)[59] with a triple-zeta, correlation-consistent basis set (aVTZ). (A detailed description of this basis set, and all other basis sets used in this study—including the double-zeta (aVDZ) and quadruple-zeta (aVQZ) variants of aVTZ—is provided in the Supplementary Information.)[60–76] The resulting set of unique monomer conformations were the starting point for the generation of QM-based dimer geometries.

Generation of QM-based dimer geometries. Dimer geometries were initially optimized with the OPLS_2005[46] force field starting from randomly generated relative monomer positions and orientations; the monomer conformations themselves were randomly selected from the corresponding set of QM-optimized structures (described above). The monomers were kept rigid during both this step and all subsequent QM dimer optimization steps. We identified unique dimer minima from this set and then optimized them using a two-step QM procedure: first at the relatively inexpensive DF-LMP2/aVDZ level of theory, then at the DF-MP2/aVTZ level (the convergence threshold for rigid-body optimization was 10^{-4} a.u. in both the center-of-mass gradient and torque). We note that because these minima are seeded from an empirical force field, we do not expect to necessarily recapitulate the global minimum as captured by a higher-level QM method. The set of unique QM-optimized dimer geometries served as starting points for one-dimensional radial scans along an intermolecular axis in 0.1-Å steps, probing separations that were either more compact (i.e., with the shortest intermolecular contact reaching ~1 Å) or more distant (i.e., up to 5 Å more distant than the reference). The internal monomer geometries were preserved when constructing these scans. The intermolecular axis was defined as the line connecting weighted atomic centers of the two molecules, with the weight for each atom defined as C/R^6, where R is the distance to the nearest atom from the other molecule (coefficient C is 1.0 for heavy atoms and 0.1 for hydrogens). Such a definition successfully reproduces, in an automated way, intuitively expected dissociation directions for both nonpolar complexes and hydrogen-bonded dimers; for example, in the latter case the two monomer centers reside in the vicinity of the donor and acceptor atoms.

Generation of MD-based dimer geometries. To more closely mimic biologically relevant physical conditions, we derived a large set of dimer geometries from condensed-phase MD simulations. For a given molecule, two types of simulations were run (both with the OPLS_2005 force field and MD sampling under the NVT ensemble using the Desmond software package)[77,78]: First, a neat liquid was simulated at the temperature closest to 298 K under which the system remains a liquid at atmospheric pressure, with the density set to the experimentally determined value for that liquid; second, a single solute molecule solvated in a cubic water box (30 Å × 30 Å × 30 Å) was simulated at 298 K and a pre-solute density of 0.997 g cm^{-3}. Dimer configurations were extracted from the MD simulation frames and clustered as follows: (i) randomly select an MD dimer configuration as a center of the first cluster and remove from the ensemble M/N structures closest to the center, where M is the number of MD configurations and N is the desired number of clusters; (ii) select as a center of the second cluster the configuration most distant from the first center and remove from the remaining unassigned ensemble M/N structures closest to the second center; (iii) repeat step (ii) until N centers are selected based on the largest distance to the closest previously selected center. The distance between two conformers is defined according to the bag-of-bonds[79] approach. Such a procedure achieves the twin objective of obtaining samples that are both representative and diverse. These dimer configurations were then used to generate radial scans following the same protocol as for QM-optimized conformers.

Multimer configurations were typically extracted from the same MD simulation frames. The multimer configurations extracted from neat liquid simulations were decomposed into the set of all possible homodimer geometries, and those extracted from the water-solvated monomer simulations were decomposed into the set of all possible heterodimer geometries (unless water was both the solute and solvent, in which case water dimer geometries were generated). These multimer-derived dimer configurations were used in single-point QM calculations (not used to seed radial scans).

QM calculation of dimer interaction energies. For all dimer geometries (including at every point along each radial scan), the interaction energy was computed at the DF-MP2/aVQZ level of theory and counterpoise-corrected for basis-set-superposition error (BSSE)[80]. The resulting MP2 interaction energies form the basis of all datasets presented herein.

The DES370K dataset, which includes CCSD(T) interaction energies, was constructed using the QM- and MD-based protocols for generating dimer geometries described above, but with a more limited set of conformers. QM-derived dimer configurations were restricted to the scans containing the most stable dimer structure for each chemically distinct dimer type. In the case of MD-derived dimer configurations, the number of scans and

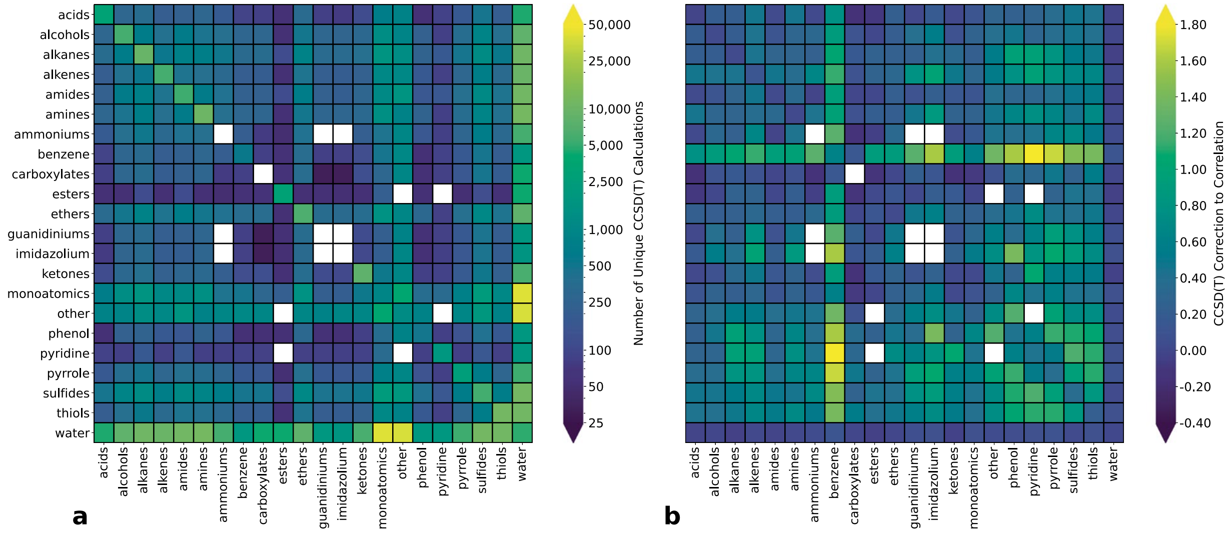

Fig. 1 Heatmaps of (**a**) dimer counts and (**b**) ΔCCSD(T) in kcal mol^{-1} for DES370K. The rows and columns of the matrices correspond to the molecule classes of the two monomers. A full list of the monomer SMILES strings assigned to each molecule class is provided in the Supplementary Information. Figure made with Matplotlib[93] and Seaborn[94].

multimers was limited to ~10 for each chemically distinct dimer type included in the dataset. We excluded the most compact, and thus very repulsive, conformers from all scans.

For each dimer in the DES370K dataset, we calculated a benchmark CCSD(T) interaction energy by using the "gold-standard" method of combining canonical MP2 energies extrapolated to the CBS limit with the difference between the CCSD(T) and MP2 energy estimated in a smaller basis set[5]. For MP2/CBS extrapolation, we used a two-point extrapolation[81] of DF-MP2/aVTZ and DF-MP2/aVQZ counterpoise-corrected interaction energies. The post-MP2 interaction energy correction (denoted ΔCCSD(T)) was estimated by the difference between counterpoise-corrected CCSD(T) and MP2 interaction energies in the largest basis set that we could afford; this basis set varied from aVQZ for the smallest systems (e.g., a water dimer) to aVDZ for the largest (e.g., a phenol dimer).

Figure 1 shows heatmaps of dimer counts and ΔCCSD(T) for DES370K, grouped according to the molecule class of the two monomers. (A full list of SMILES strings assigned to each molecule class is provided in the Supplementary Information).

SNS-MP2 predictions. For every dimer included in DES5M, the interaction energy was computed using our SNS-MP2 approach, described in full detail elsewhere[39]. In addition to predicting an energy value, SNS-MP2 quantifies the uncertainty of that prediction: Each SNS-MP2 energy is accompanied by the upper and lower bounds of a 90% confidence interval associated with that prediction.

Calculation of other QM quantities. Beyond MP2 energies, SNS-MP2 requires additional features that encode the interaction in a geometry-independent manner, leveraging commonalities between chemically disparate dimers. An account of all inputs to the neural network can be found elsewhere[39]; here we describe only these additional quantities, which are included as entries in all three datasets. All of the below quantities are calculated automatically when using the SNS-MP2 plugin[39] (https://github.com/DEShawResearch/sns-mp2) which relies on the Psi4 quantum chemistry software package[82].

For each monomer in our set of dimers, we calculated the Hartree-Fock (HF) wavefunction (and thus density matrix) and the MP2 density matrix in the monomer basis. From these quantities, we calculated three properties of the dimer interaction: the classical electrostatic interaction energy, the Heitler-London energy, and the density-matrix overlap.

For each dimer, we calculated the following SAPT0/aVTZ energy components[82–86]: the second-order dispersion, induction, exchange-dispersion, and exchange-induction energies; the same-spin component of the second-order dispersion and exchange-dispersion energies; the first-order electrostatic and exchange energies; and the first-order exchange energy computed in the S^2 approximation.

Figure 2 portrays a SAPT interaction energy analysis of the DES370K and DES15K datasets.

Core subset DES15K. DES15K is a core subset of the most representative structures from DES370K, and was assembled with a focus on retaining the chemical and orientational diversity of DES370K. DES15K consists of dimer configurations from both QM- and MD-derived scans, though with reduced resolution of scan points in the radial profiles. In the case of QM-optimized dimers, up to four conformers were selected, all from the radial scan corresponding to the most stable QM-optimized structure. These conformers correspond to the minimum, a point less compact than the minimum, the zero-crossing point at a more compact structure (if the minimum

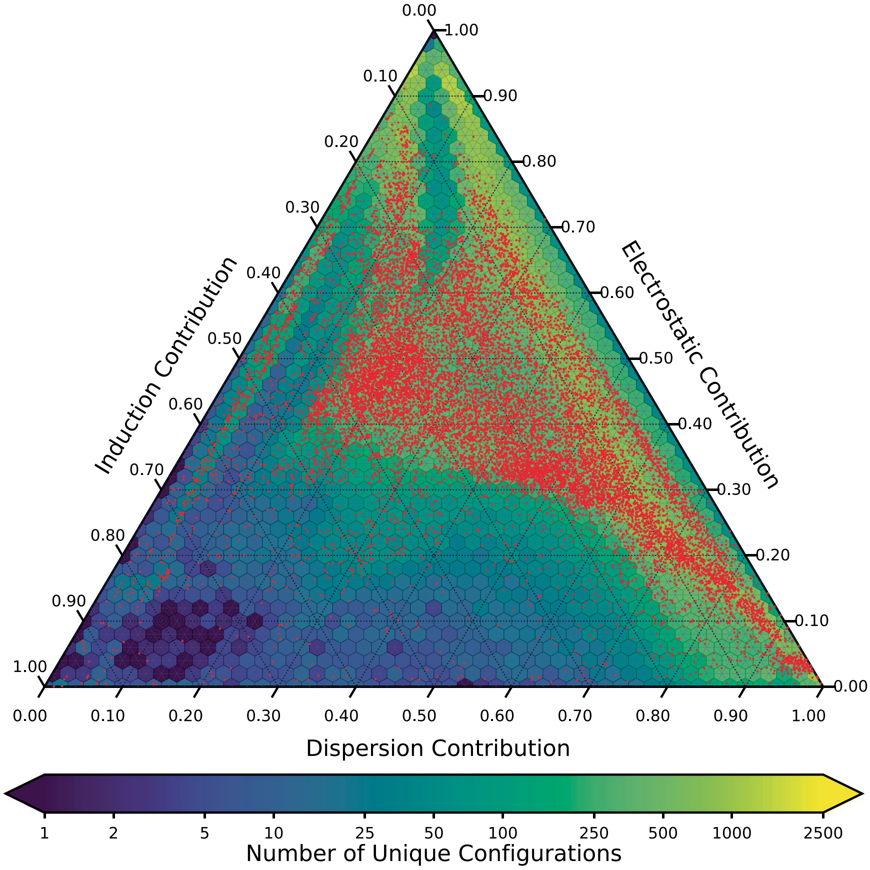

Fig. 2 Ternary plot showing the relationship between the electrostatic, dispersion, and induction energy components, as calculated using SAPT, for the dimers in DES370K and DES15K. Counts for DES370K are colored according to the color bar, and DES15K dimers are indicated by red points. Figure made with Matplotlib[93] and Python-Ternary[95].

is not too deep), and a point representative of the repulsive wall at short distances. The exact definition of these points is specified in Fig. 3, which shows a typical dimer interaction energy profile, highlighting points along the scan that are included in DES15K. The DES15K dimers extracted from MD simulations include up to 10 conformers, and in addition to the MD-observed dimer, the minima along the corresponding radial scans at least half as deep as the most stable QM-optimized configuration. Based on these selection criteria, the MD-based component of DES15K includes only dimers for which we have the corresponding QM-optimized structure and sufficiently attractive scans. DES15K thus features a smaller set of monomers than DES370K, though most of the removed monomers are alkylated forms of the monomers still included in DES15K (and so the chemical diversity of the dimers—that is, at the level of functional-group interactions—is largely maintained).

Data Records

Datasets are provided as CSV files (one file each for DES370K, DES15K, DES5M, DESS66, and DESS66x8) in a Figshare data repository[42]. A table providing column names and a description of the contents of each column can be found in the Supplementary Information. The Pandas[87] and Scipy and Numpy[88] packages were used in data processing and packaging for the CSV files.

Technical Validation

Validation of monomer geometries. Employing an established strategy[89], we used molecular graph connectivity to confirm that the final set of dimer conformers had not undergone extreme or unintended changes in geometry during any stage of the protocols used to generate those geometries. Connectivity for each dimer, based on the original SMILES string, was compared against the graph assigned by Open Babel[45] from the final atom positions of each conformer. Bond order, formal charge, and stereochemistry were ignored in order to avoid ambiguities in molecular graph construction. That is, we verified that the molecular graphs are isomorphic (i.e., they have identical edges between nodes labeled by element). We also calculated monomer energies with the OPLS_2005 force field[46] and rejected dimers that included any monomer with excitation energy $>30\,\mathrm{kcal\,mol^{-1}}$.

Comparison to the S66 and S66x8 datasets. We applied the present protocol for estimating CCSD(T)/CBS dimer interaction energies to all 66 conformers from the S66 dataset[9] and to all 528 conformers from the S66x8 dataset[8], using the reference geometries in both cases. The results of these calculations are provided in

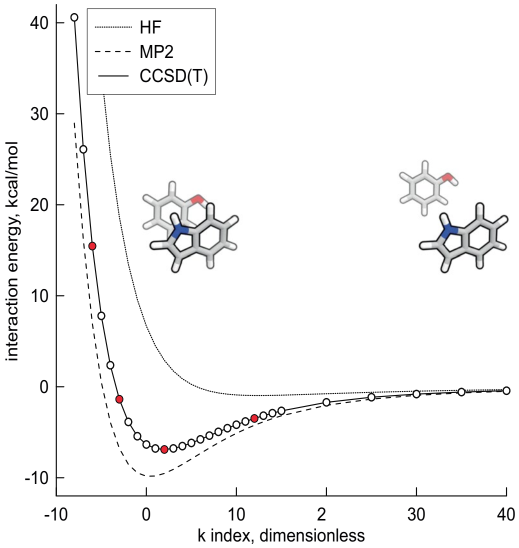

Fig. 3 Typical dimer interaction energy profile, showing HF (at aVQZ), MP2 (at the CBS limit), and CCSD(T) (computed using the hybrid "gold-standard[5]" method with ΔCCSD(T)/aVDZ) interaction energies. The plot corresponds to the most stable phenol-indole dimer (obtained using QM optimization), but is representative of other dimer scans. The x-axis shows the k index, with k = 0 corresponding to the reference geometry (in this case, the most stable QM-optimized geometry, based on MP2, for the phenol-indole dimer) used to construct the radial scan. Each k unit corresponds to a 0.1 Å step along the intermolecular axis (defined in the "Generation of QM-based dimer geometries" section of the manuscript). These steps are, in general, taken in both the negative (more compact) and positive (more separated) directions with respect to the reference geometry (k = 0). All circles shown on the CCSD(T) curve correspond to data points included in the DES370K dataset. Red circles on the CCSD(T) curve correspond to data points included in the DES15K dataset. In the case of QM-derived dimer scans (as is the case here), we selected four conformers to include in DES15K: the conformer with the lowest energy, designated E_{min}; the conformer that was less compact than the lowest-energy conformer and had an energy nearest to $E_{min} + 0.5E_{exc}$, where the positive excitation energy is defined as $E_{exc} = \min(|E_{min}|, 10\,\mathrm{kcal\,mol^{-1}})$; the conformer representing the zero of the interaction energy curve (when $|E_{min}| < 10\,\mathrm{kcal\,mol^{-1}}$), such that it was more compact than the lowest-energy conformer with an energy nearest to $E_{min} + E_{exc}$; and the conformer with an energy nearest to $E_{min} + 3E_{exc}$, which is representative of the repulsive region of the radial scan. Figure made with Grapher™ (Golden Software, LLC; http://www.goldensoftware.com).

the DESS66 and DESS66x8 datasets, respectively. We found good correspondence between our estimates and the values reported in Tables 6 and 7 of Kesharwani *et al.*[90]: mean unsigned errors of 0.07 kcal mol^{-1} (S66) and 0.05 kcal mol^{-1} (S66x8) and mean signed errors of −0.02 kcal mol^{-1} (S66) and −0.01 kcal mol^{-1} (S66x8).

Comparison of bond-length constraints to published experimental data. To validate the present protocol for determining the values for bond constraints involving hydrogen atoms, we compared the values to published experimental data[91]. For the methyl (CH_3) and methylene (CH_2) functional groups, our values for the CH bond length—1.101 Å and 1.107 Å, respectively—compare favorably with the experimental distribution, which features a median value of 1.107 Å. We observed similarly good agreement between computed (1.098 Å) and experimental (median of 1.094 Å) bond lengths between hydrogen and aromatic carbon. The present approach yields values for NH bond lengths in primary and secondary amines of 1.026 Å and 1.028 Å, respectively, which are close to the experimentally measured length of 1.021 (±0.006) Å[91]. For bonds between hydrogen and oxygen, experimental uncertainties are even larger. For alcohols, our protocol yielded a bond length of 0.976 Å before the condensed-phase correction and 0.980 Å after the correction; the former value compares very well with the experimentally determined bond length of 0.975 (±0.010) Å in methanol[91]. In carboxylic acids, our approach yields a bond length of 0.983 Å before the condensed-phase correction and 0.996 Å after the correction; the former is comparable to the experimentally measured value of 0.981 (±0.003) Å for formic acid.

Ionic system tests. The DES370K and DES5M collections include two types of ionic systems: dimers with only one of the monomers carrying a charge (−1, +1, or +2) and the other monomer neutral, and salts composed of a monovalent cation (+1) and a monovalent anion (−1). At large separations and in the absence of solvent, the desired biologically relevant monomer charges frequently do not represent the ground state. As a precautionary step, we clipped radial scans containing salts or divalent cations to separations between −0.5 and 0.5 Å from the reference structure used to seed the scan. We performed a natural population analysis (NPA)[92], implemented in Molpro 2015.1 (http://www.molpro.net)[59], to confirm that the interaction energies corresponded to the desired charges. We required that the NPA charges of each monomer differed from the target by <0.3 electrons, were smooth (as described in the next section), and approached the correct asymptotic limit.

Smoothness and curvature tests for radial scans. For each radial scan, we imposed two additional conditions, requiring both that the interaction energy and all components were smooth functions of the separation and that they asymptotically converged to zero at large distances. The smoothness was validated by fitting each radial profile with a cubic spline and assessing the impact of individually removing each data point from the fit. In addition, we ensured that along a given scan, the total interaction energy featured no more than one local minimum. Scans without a local minimum were considered valid only if the interaction energy was strictly positive. Scans with a negative local minimum were allowed to exhibit at most one local maximum with a positive interaction energy.

Received: 23 July 2020; Accepted: 14 December 2020;
Published online: 10 February 2021

References

1. Hobza, P., Zahradník, R. & Müller-Dethlefs, K. The world of non-covalent interactions: 2006. *Collect. Czech. Chem. Commun.* **71**, 443–531 (2006).
2. Raghavachari, K., Trucks, G. W., Pople, J. A. & Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **157**, 479–483 (1989).
3. Urban, M., Noga, J., Cole, S. J. & Bartlett, R. J. Towards a full CCSDT model for electron correlation. *J. Chem. Phys.* **83**, 4041–4046 (1985).
4. Bartlett, R. J. & Musial, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79**, 291–352 (2007).
5. 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.* **9**, 2151–2155 (2013).
6. Jurečka, P., Šponer, J., Cerný, 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.* **8**, 1985–1993 (2006).
7. Marshall, M. S., Burns, L. A. & Sherrill, C. D. Basis set convergence of the coupled-cluster correction, $\delta_{\mathrm{MP2}}^{\mathrm{CCSD(T)}}$: best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. *J. Chem. Phys.* **135**, 194102 (2011).
8. Brauer, B., Kesharwani, M. K., Kozuch, S. & Martin, J. M. L. The S66x8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **18**, 20905–20925 (2016).
9. Řezáč, J., Riley, K. E. & Hobza, P. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.* **7**, 2427–2438 (2011).
10. Řezáč, J., Riley, K. E. & Hobza, P. Benchmark calculations of noncovalent interactions of halogenated molecules. *J. Chem. Theory Comput.* **8**, 4285–4292 (2012).
11. Burns, L. A. *et al.* The Bio-Fragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.* **147**, 161727 (2017).
12. Schneebeli, S. T., Bochevarov, A. D. & Friesner, R. A. Parameterization of a B3LYP specific correction for noncovalent interactions and basis set superposition error on a gigantic dataset of CCSD(T) quality noncovalent interaction energies. *J. Chem. Theory Comput.* **7**, 658–668 (2011).
13. Mardirossian, N. & Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110 (2016).
14. Smith, D. G. A., Burns, L. A., Patkowski, K. & Sherrill, C. D. Revised damping parameters for the D3 dispersion correction to density functional theory. *J. Phys. Chem. Lett.* **7**, 2197–2203 (2016).
15. Yu, H. S., He, X. & Truhlar, D. G. MN15-L: A new local exchange-correlation functional for Kohn-Sham density functional theory with broad accuracy for atoms, molecules, and solids. *J. Chem. Theory Comput.* **12**, 1280–1293 (2016).
16. Tkatchenko, A., DiStasio, R. A. Jr, Car, R. & Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).
17. Goerigk, L., Kruse, H. & Grimme, S. Benchmarking density functional methods against the S66 and S66x8 datasets for non-covalent interactions. *ChemPhysChem* **12**, 3421–3433 (2011).
18. DiStasio, R. A. Jr & Head-Gordon, M. Optimized spin-component scaled second-order Møller-Plesset perturbation theory for intermolecular interaction energies. *Mol. Phys.* **105**, 1073–1083 (2007).
19. Marchetti, O. & Werner, H. J. Accurate calculations of intermolecular interaction energies using explicitly correlated coupled cluster wave functions and a dispersion-weighted MP2 method. *J. Phys. Chem. A* **113**, 11580–11585 (2009).
20. Takatani, T., Hohenstein, E. G. & Sherrill, C. D. Improvement of the coupled-cluster singles and doubles method via scaling same- and opposite-spin components of the double excitation correlation energy. *J. Chem. Phys.* **128**, 124111 (2008).
21. Pitoňák, M., Neogrady, P., Cerný, J., Grimme, S. & Hobza, P. Scaled MP3 non-covalent interaction energies agree closely with accurate CCSD(T) benchmark data. *ChemPhysChem* **10**, 282–289 (2009).
22. Hesselmann, A. Improved supermolecular second order Møller-Plesset intermolecular interaction energies using time-dependent density functional response theory. *J. Chem. Phys.* **128**, 144112 (2008).
23. Burns, L. A., Marshall, M. S. & Sherrill, C. D. Appointing silver and bronze standards for noncovalent interactions: a comparison of spin-component-scaled (SCS), explicitly correlated (F12), and specialized wavefunction approaches. *J. Chem. Phys.* **141**, 234111 (2014).
24. McNamara, J. P. & Hillier, I. H. Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules. *Phys. Chem. Chem. Phys.* **9**, 2362–2370 (2007).
25. Řezáč, J. & Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **8**, 141–151 (2011).
26. Christensen, A. S., Elstner, M. & Cui, Q. Improving intermolecular interactions in DFTB3 using extended polarization from chemical-potential equalization. *J. Chem. Phys.* **143**, 084123 (2015).
27. Christensen, A. S., Kubař, T., Cui, Q. & Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **116**, 5301–5337 (2016).
28. Patkowski, K. Benchmark databases of intermolecular interaction energies: design, construction, and significance. *Annu. Rep. Comput. Chem.* **13**, 3–91 (2017).
29. Řezáč, J. & Hobza, P. Benchmark calculations of interaction energies in noncovalent complexes and their applications. *Chem. Rev.* **116**, 5038–5071 (2016).
30. Wang, L.-P. *et al.* Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15. *J. Phys. Chem. B* **121**, 4023–4039 (2017).
31. Lopes, P. E. M. *et al.* Polarizable force field for peptides and proteins based on the classical drude oscillator. *J. Chem. Theory Comput.* **9**, 5430–5449 (2013).

32. Laury, M. L., Wang, L.-P., Pande, V. S., Head-Gordon, T. & Ponder, J. W. Revised parameters for the AMOEBA polarizable atomic multipole water model. *J. Phys. Chem. B* **119**, 9423–9437 (2015).
33. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **119**, 5113–5123 (2015).
34. Harder, E. *et al*. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2015).
35. Bereau, T., Andrienko, D. & von Lilienfeld, O. A. Transferable atomic multipole machine learning models for small organic molecules. *J. Chem. Theory Comput.* **11**, 3225–3233 (2015).
36. Gao, T. *et al*. A machine learning correction for DFT non-covalent interactions based on the S22, S66 and X40 benchmark databases. *J. Cheminformatics* **8**, 24 (2016).
37. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
38. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
39. McGibbon, R. T. *et al*. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *J. Chem. Phys.* **147**, 161725 (2017).
40. Faber, F. A. *et al*. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
41. Hegde, G. & Bowen, R. C. Machine-learned approximations to density functional theory Hamiltonians. *Sci. Rep.* **7**, 42669 (2017).
42. Donchev, A. G. *et al*. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *figshare* https://doi.org/10.6084/m9.figshare.c.5070644 (2021).
43. Grimme, S. Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* **118**, 9095–9102 (2003).
44. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
45. O'Boyle, N. M. *et al*. Open Babel: an open chemical toolbox. *J. Cheminformatics* **3**, 33–47 (2011).
46. Banks, J. L. *et al*. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **26**, 1752–1780 (2005).
47. Morse, P. M. Diatomic molecules according to the wave mechanics. II. *Vibrational levels. Phys. Rev.* **34**, 57–64 (1929).
48. Polly, R., Werner, H.-J., F. Manby, R. & Knowles, P. J. Fast Hartree-Fock theory using local density fitting approximations. *Mol. Phys.* **102**, 2311–2321 (2004).
49. Köppl, C. & Werner, H.-J. Parallel and low-order scaling implementation of Hartree-Fock exchange using local density fitting. *J. Chem. Theory Comput.* **12**, 3122–3134 (2016).
50. Pipek, J. & Mezey, P. G. A fast intrinsic localization procedure applicable for ab initio and semiempirical linear combination of atomic orbital wave functions. *J. Chem. Phys.* **90**, 4916–4926 (1989).
51. El Azhary, A., Rauhut, G., Pulay, P. & Werner, H.-J. Analytical energy gradients for local second-order Møller-Plesset perturbation theory. *J. Chem. Phys.* **108**, 5185–5193 (1998).
52. Schütz, M., Werner, H.-J., Lindh, R. & Manby, F. R. Analytical energy gradients for local second-order Møller-Plesset perturbation theory using density fitting approximations. *J. Chem. Phys.* **121**, 737–750 (2004).
53. Hetzer, G., Pulay, P. & Werner, H.-J. Multipole approximation of distant pair energies in local MP2 calculations. *Chem. Phys. Lett.* **290**, 143–149 (1998).
54. Schütz, M., Hetzer, G. & Werner, H.-J. Low-order scaling local electron correlation methods. I: Linear scaling local MP2. *J. Chem. Phys.* **111**, 5691–5705 (1999).
55. Hetzer, G., Schütz, M., Stoll, H. & Werner, H.-J. Low-order scaling local correlation methods. II: Splitting the Coulomb operator in linear scaling local second-order Møller-Plesset perturbation theory. *J. Chem. Phys.* **113**, 9443–9455 (2000).
56. Werner, H.-J., Manby, F. R. & Knowles, P. J. Fast linear scaling second-order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations. *J. Chem. Phys.* **118**, 8149–8160 (2003).
57. Lindh, R., Bernhardsson, A., Karlström, G. & Malmqvist, P.-A. On the use of a Hessian model function in molecular geometry optimizations. *Chem. Phys. Letters* **241**, 423–428 (1995).
58. Lindh, R., Bernhardsson, A. & Schütz, M. Force-constant weighted redundant coordinates in molecular geometry optimizations. *Chem. Phys. Letters* **303**, 567–575 (1999).
59. Werner, H.-J., Knowles, P. J., Knizia, G., Manby, F. R. & Schütz, M. A general-purpose quantum chemistry program package. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 242–253 (2012).
60. Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).
61. Woon, D. E. & Dunning Jr., T.H. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **100**, 2975–2988 (1994).
62. Kendall, R. A., Dunning, T. H. Jr. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96**, 6796–6806 (1992).
63. Woon, D. E. & Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through hydrogen. *J. Chem. Phys.* **98**, 1358–1371 (1993).
64. Dunning, T. H. Jr., Peterson, K. A. & Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations: X. The atoms aluminum through argon revisited. *J. Chem. Phys.* **114**, 9244–9253 (2001).
65. Peterson, K. A. & Dunning, T. H. Jr. Accurate correlation consistent basis sets for molecular core–valence correlation effects: The second row atoms Al–Ar, and the first row atoms B–Ne revisited. *J. Chem. Phys.* **117**, 10548–10560 (2002).
66. Prascher, B., Woon, D. E., Peterson, K. A., Dunning, T. H. Jr. & Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. VII. Valence, core-valence, and scalar relativistic basis sets for Li, Be, Na, and Mg. *Theor. Chem. Acc.* **128**, 69–82 (2011).
67. Koput, J. & Peterson, K. A. Ab initio potential energy surface and vibrational-rotational energy levels of $X^2\Sigma^+$ CaOH. *J. Phys. Chem. A* **106**, 9595–9599 (2002).
68. Lim, I. S., Schwerdtfeger, P., Metz, B. & Stoll, H. All-electron and relativistic pseudopotential studies for the group 1 element polarizabilities from K to element 119. *J. Chem. Phys.* **122**, 104103 (2005).
69. Lim, I. S., Stoll, H. & Schwerdtfeger, P. Relativistic small-core energy-consistent pseudopotentials for the alkaline-earth elements from Ca to Ra. *J. Chem. Phys.* **124**, 034107 (2006).
70. Peterson, K. A. & Yousaf, K. E. Molecular core-valence correlation effects involving the post-d elements Ga-Rn: benchmarks and new pseudopotential-based correlation consistent basis sets. *J. Chem. Phys.* **133**, 174116 (2010).
71. Peterson, K. A., Shepler, B. C., Figgen, D. & Stoll, H. On the spectroscopic and thermochemical properties of ClO, BrO, IO, and their anions. *J. Phys. Chem. A* **110**, 13877–13883 (2006).
72. Peterson, K. A., Figgen, D., Goll, E., Stoll, H. & Dolg, M. Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-d group 16–18 elements. *J. Chem. Phys.* **119**, 11113–11123 (2003).
73. Wilson, A. K., Woon, D. E., Peterson, K. A. & Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. IX. *The atoms gallium through krypton. J. Chem. Phys.* **110**, 7667–7676 (1999).

74. DeYonker, N. J., Peterson, K. A. & Wilson, A. K. Systematically convergent correlation consistent basis sets for molecular core—valence correlation effects: the third-row atoms gallium through Krypton. *J. Phys. Chem. A* **111**, 11383–11393 (2007).
75. Weigend, F. A fully direct RI-HF algorithm: Implementation, optimized auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **4**, 4285–4291 (2002).
76. Weigend, F. Hartree–Fock exchange fitting basis sets for H to Rn. *J. Comput. Chem.* **29**, 167–175 (2008).
77. Bowers, K. J. *et al*. Scalable algorithms for molecular dynamics simulations on commodity clusters. Proc. ACM/IEEE Conf. Supercomput. (ACM, 2006).
78. Bergdorf, M., Baxter, S., Rendleman, C. A. & Shaw, D. E. Desmond/GPU performance as of November 2016. D. E. Shaw Research Technical Report DESRES/TR—2016-01. (2016).
79. Hansen, K. *et al*. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
80. Boys, S. F. & Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **19**, 553–566 (1970).
81. Halkier, A., Helgaker, T., Jorgensen, P., Klopper, W. & Olsen, J. Basis-set convergence of the energy in molecular Hartree–Fock calculations. *Chem. Phys. Lett.* **302**, 437–446 (1999).
82. Turney, J. M. *et al*. Psi4: An open-source ab initio electronic structure program. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 556–565 (2012).
83. Jeziorski, B., Moszynski, R. & Szalewicz, K. Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes. *Chem. Rev.* **94**, 1887–1930 (1994).
84. Hohenstein, E. G. & Sherrill, C. D. Wavefunction method for noncovalent interactions. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 304–326 (2012).
85. Hohenstein, E. G. & Sherrill, C. D. Density fitting and Cholesky decomposition approximations in symmetry-adapted perturbation theory: implementation and application to probe the nature of π-π interactions in linear acenes. *J. Chem. Phys.* **132**, 184111 (2010).
86. Hohenstein, E. G., Parrish, R. M., Sherrill, C. D., Turney, J. M. & Schaefer, H. F. Large-scale symmetry-adapted perturbation theory computations via density fitting and Laplace transformation techniques: investigating the fundamental forces of DNA-intercalator interactions. *J. Chem. Phys.* **135**, 174107 (2011).
87. McKinney, W. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 56–61 (2010).
88. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).
89. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
90. Kesharwani, M. K., Karton, A., Sylvetsky, N. & Nitai, J. M. L. The S66 non-covalent interactions benchmark reconsidered using explicitly correlated methods near the basis set limit. *Aust. J. Chem.* **71**, 238–248 (2018).
91. Ma, B., Lii, J.-H., Schaefer, H. F. & Allinger, N. L. Systematic comparison of experimental, quantum mechanical, and molecular mechanical bond lengths for organic molecules. *J. Phys. Chem.* **100**, 8763–8769 (1996).
92. Reed, A. E., Weinstock, R. B. & Weinhold, F. Natural population analysis. *J. Chem. Phys.* **83**, 735–746 (1985).
93. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
94. Waskom, M. *et al*. mwaskom/seaborn: v0.9.0 (July 2018). *Zenodo* https://doi.org/10.5281/zenodo.1313201 (2018).
95. Marc *et al*. marcharper/python-ternary: Corner label functions. *Zenodo* https://doi.org/10.5281/zenodo.1220444 (2018).

Acknowledgements

The authors thank Jessica McGillen and Berkman Frank for editorial assistance.

Author contributions

A.G.D. and D.E.S. designed the project; A.G.D., A.G.T., E.D., C.H., R.T.M., K.-H.L., B.A.G., J.-L.L., K.P., K.S., M.B. and J.L.K. developed and applied the methods for dataset generation; E.D. and K.-H.L. prepared the data for release; E.D. formatted the data for release; E.D. and A.G.D. generated the figures; A.G.D., A.G.T., E.D., J.L.K. and D.E.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-00833-x.

Correspondence and requests for materials should be addressed to A.G.D. or D.E.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.