SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Quantum chemistry
» Density functional theory
» Computational chemistry

Received: 02 April 2014
Accepted: 07 July 2014
Published: 05 August 2014

Quantum chemistry structures and properties of 134 kilo molecules

Raghunathan Ramakrishnan[1], Pavlo O. Dral[2,3], Matthias Rupp[1] & O. Anatole von Lilienfeld[1,4]

Computational *de novo* design of new drugs and materials requires rigorous and unbiased exploration of chemical compound space. However, large uncharted territories persist due to its size scaling combinatorially with molecular size. We report computed geometric, energetic, electronic, and thermodynamic properties for 134k stable small organic molecules made up of CHONF. These molecules correspond to the subset of all 133,885 species with up to nine heavy atoms (CONF) out of the GDB-17 chemical universe of 166 billion organic molecules. We report geometries minimal in energy, corresponding harmonic frequencies, dipole moments, polarizabilities, along with energies, enthalpies, and free energies of atomization. All properties were calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry. Furthermore, for the predominant stoichiometry, $C_7H_{10}O_2$, there are 6,095 constitutional isomers among the 134k molecules. We report energies, enthalpies, and free energies of atomization at the more accurate G4MP2 level of theory for all of them. As such, this data set provides quantum chemical properties for a relevant, consistent, and comprehensive chemical space of small organic molecules. This database may serve the benchmarking of existing methods, development of new methods, such as hybrid quantum mechanics/machine learning, and systematic identification of structure-property relationships.

Design Type(s)	in silico design • data integration
Measurement Type(s)	Computational Chemistry
Technology Type(s)	quantum chemistry computational method
Factor Type(s)	level of theory
Sample Characteristic(s)	

[1]Department of Chemistry, Institute of Physical Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland. [2]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany. [3]Computer-Chemie-Centrum, University of Erlangen-Nuremberg, Nägelsbachstr. 25, 91052 Erlangen, Germany. [4]Argonne Leadership Computing Facility, Argonne National Laboratory, 9700S. Cass Avenue, Lemont, Illinois, 60439, USA.
Correspondence and requests for materials should be addressed to O.A.v.L. (email: anatole.vonlilienfeld@unibas.ch)

Background & summary

The goal of computationally designing novel materials and molecules with desired physicochemical properties is yet to be achieved. High-throughput screening represents the most straightforward approach towards materials design[1]. However, it presupposes that all assumptions and approximations inherent to the employed modeling techniques are applicable to the entire chemical compound space, which is the space populated by all stable molecules or materials[2]. Furthermore, due to the combinatorial scaling of chemical space with molecular size, it is difficult to explore or even navigate. Conclusive insights about the domain of applicability (transferability) are lacking even for the most popular first principle quantum chemistry methods. For example, the reliability and accuracy of density functional theory is known to dramatically depend on chemical composition and atomistic configurations[3], highlighting the importance of reliable experimental[4] or high-level quantum chemistry state-of-the-art results[5,6]. Unfortunately, the systems reported are typically small, which implies the existence of severe selection bias. One can therefore question how representative they are. The problem of representative diversity has triggered the design of special purpose chemical space libraries for method validation or molecular design[7–10].

Here, we report molecular structures and properties obtained from quantum chemistry calculations for the first 134k molecules of the chemical universe GDB-17 data base[11], covering a molecular property set of unprecedented size and consistency. The data-set corresponds to the GDB-9 subset of all neutral molecules with up to nine atoms (CONF), not counting hydrogen. The molecular size distribution of all 134k molecules is shown in Fig. 1. This data set contains small amino acids, such as GLY, ALA, as well as nucleobases cytosine, uracil, and thymine. Also pharmaceutically relevant organic building blocks, such as pyruvic acid, piperazine, or hydroxy urea are included. Among the 134k molecules, there are 621 stoichiometries, among which $C_7H_{10}O_2$ dominates with 6,095 constitutional isomers for which atomization energies and radii of gyration also are on display in Fig. 1.

For all 134k molecules, we have calculated equilibrium geometries, frontier orbital eigenvalues, dipole moments, harmonic frequencies, polarizabilities, and thermochemical energetics corresponding to atomization energies, enthalpies, and entropies at ambient temperature. These properties have been obtained at the B3LYP/6-31G(2df,p) level of theory which forms the basis for the more accurate state-of-the art G*n* methods which are on par with experimental accuracy[12]. For the 6,095 constitutional isomers of the predominant stoichiometry, $C_7H_{10}O_2$, we report the energetics at the significantly more accurate G4MP2[12] level of theory.

This report is structured as follows. We first describe the genesis of the results. Thereafter, we discuss the validation of our DFT results by comparison to (i) G4MP2, (ii) G4, and (iii) CBS-QB3 results for

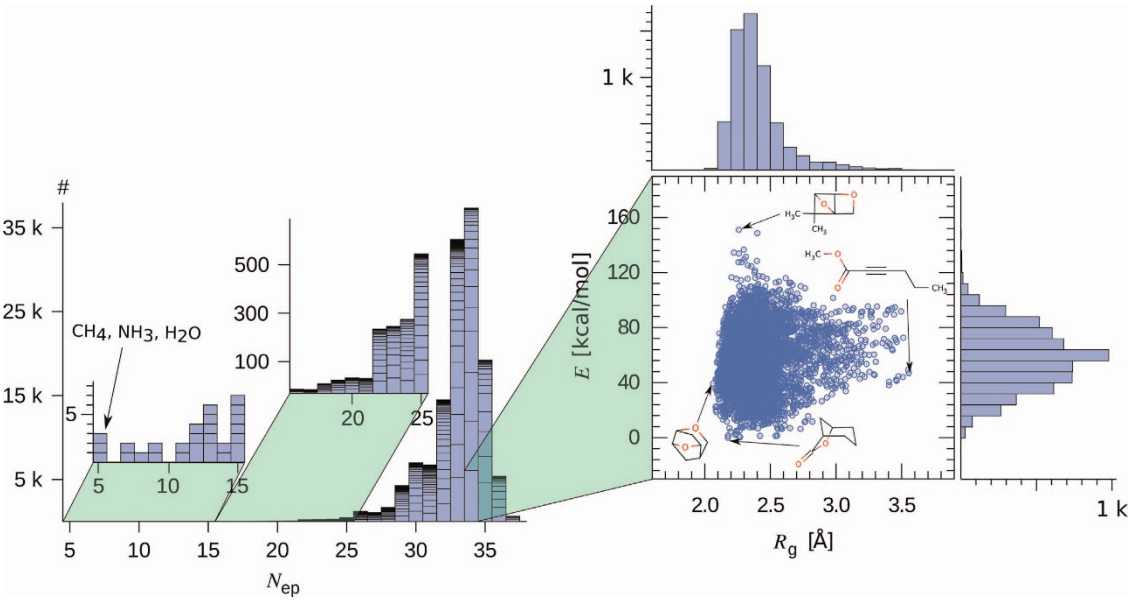

Figure 1. Illustration of the scaling of chemical space with system size. For the smallest 134k molecules, with up to 9 heavy atoms CONF (not counting hydrogens) taken from the chemical universe GDB-17[11], the distribution of molecular size is shown as a function of number of occupied electron orbitals, i.e. number of electron pairs, $N_{ep} = N_e/2$. Each black box denotes the number of constitutional isomers for one out of the 621 stoichiometries present in the 134k molecules. The two left-hand side insets correspond to zoom-ins for smaller compounds. The right-hand side inset zooms in on the predominant stoichiometry, $C_7H_{10}O_2$, and features a scatter plot of G4MP2 relative (w.r.t. global minimum) potential energies of atomization E versus molecular radius of gyration, R_g. Joined projected distributions are shown as well.

100 molecules, randomly chosen out of the 134k set. This data can serve the development, training and evaluation of inductive statistical data analysis-based machine learning (ML) models[13]. It might also assist the search and discovery of hitherto unknown trends, structure-property relationships, and molecular materials design[1,14,15].

Methods

Generation of atomic coordinates

Starting with ref. 11, we use all SMILES[16] strings for molecules with up to nine heavy atoms. Cations, anions, and molecules containing S, Br, Cl, or I, have been excluded, resulting in 133,885 molecules. 1,705 zwitterions have been kept in the data due to their occurrence in small biomolecules, such as amino acids. Initial Cartesian coordinates for all molecules were generated by parsing the corresponding SMILES strings using Corina (Version 3.491 2013)[17]. We subsequently carried out geometry relaxations at the PM7 semi-empirical level of theory using MOPAC (Version 13.136L 2012)[18]. In the PM7 calculations, we invoked tight electronic and geometric convergence thresholds, using `precise` keyword. PM7 equilibrium geometries have subsequently been used as input for B3LYP geometry relaxations using Gaussian 09[19]. We iteratively refined the electronic and geometry thresholds. For the first iteration, Gaussian 09's default electronic and geometry thresholds have been used for all molecules. For those molecules which failed to reach SCF convergence `ultrafine` grids have been invoked within a second iteration for evaluating the XC energy contributions. Within a third iteration on the remaining unconverged molecules, we identified those which had relaxed to saddle points, and further tightened the SCF criteria using the keyword `scf(maxcycle=200,verytight)`. All those molecules which still featured imaginary frequencies entered the fourth iteration using keywords, `opt(calcfc, maxstep=5, maxcycles=1000)`. `calcfc` constructs a Hessian in the first step of the geometry relaxation for eigenvector following. Within the fifth and final iteration, all molecules which still failed to reach convergence, have subsequently been converged using `opt(calcall, maxstep=1, maxcycles=1000)`. `calcall` constructs a Hessian for all steps through the the geometry relaxation. After all these measures taken, eleven problematic molecules still failed to converge to a minimal geometry. Out of these eleven molecules, six can be converged with low threshold using the `opt(loose)`-keyword. In the remaining five there were two near-linear molecules which converged to saddle points with very low imaginary frequencies ($\omega_0 < i10\ \mathrm{cm}^{-1}$ for the lowest mode). In the `readme.txt` file of this report, all these 11 molecules are specified using their indices in the database.

In the case of the 6,095 constitutional isomers of $C_7H_{10}O_2$, all molecules converged to local minima during the B3LYP geometry relaxation. To compute atomization energies, we have also performed spin-unrestricted calculations for all atoms with spin-multiplicities 2,3,4,3,2 for the atoms H, C, N, O, F, respectively.

Data Records

Molecular structures and properties are publicly available at Figshare (Data Citation 1) in a plain text XYZ-like format described below. Deposited files include the 133, 885 GDB-1 to GDB-9 molecules (`dsgdb9nsd.xyz.tar.bz2`), the 6,095 constitutional isomers of $C_7H_{10}O_2$ (`dsC7O2H10nsd.xyz.tar.bz2`), the 100 validation molecules (see Table 1) enthalpies of atomization (`validation.txt`), and atomic reference data (`atomref.txt`).

File format

For each molecule, atomic coordinates and calculated properties are stored in a file named `dataset_index.xyz`. The XYZ format (originally developed for the XMol program by the Minnesota Supercomputer Center) is a widespread plain text format for encoding Cartesian coordinates of molecules, with no formal specification. It contains a header line specifying the number of atoms n_a, a comment line, and n_a lines containing element type and atomic coordinates, one atom per line. We have extended this format as indicated in Table 2. Now, the comment line is used to store all scalar properties,

Reference	MAE	RMSE	maxAE
G4MP2	5.0	6.1	16.0
G4	4.9	5.9	14.4
CBS-QB3	4.5	5.5	13.4

Table 1. Validation of atomization enthalpies at B3LYP/6-31G(2df,p)-level. For 100 molecules randomly drawn out of the pool of 134k molecules, mean absolute error (MAE), root mean square error (RMSE), and maximal absolute error (maxAE) with respect to more accurate reference methods are reported. All values are in kcal/mol.

Line	Content
1	Number of atoms n_a
2	Scalar properties (see Table 3)
3,…,n_a+2	Element type, coordinate (x, y, z, in Å), Mulliken partial charges (in e) on atoms
n_a+3	Harmonic vibrational frequencies ($3n_a$–5 or $3n_a$-6, in cm^{-1})
n_a+4	SMILES strings from GDB-17 and from B3LYP relaxation
n_a+5	InChI strings for Corina and B3LYP geometries

Table 2. XYZ-like file format for molecular structure and properties. n_a = number of atoms.

No.	Property	Unit	Description
1	tag	—	'gdb9' string to facilitate extraction
2	i	—	Consecutive, 1-based integer identifier
3	A	GHz	Rotational constant
4	B	GHz	Rotational constant
5	C	GHz	Rotational constant
6	μ	D	Dipole moment
7	α	a_0^3	Isotropic polarizability
8	ϵ_{HOMO}	Ha	Energy of HOMO
9	ϵ_{LUMO}	Ha	Energy of LUMO
10	ϵ_{gap}	Ha	Gap ($\epsilon_{LUMO} - \epsilon_{HOMO}$)
11	$\langle R^2 \rangle$	a_0^2	Electronic spatial extent
12	zpve	Ha	Zero point vibrational energy
13	U_o	Ha	Internal energy at 0 K
14	U	Ha	Internal energy at 298.15 K
15	H	Ha	Enthalpy at 298.15 K
16	G	Ha	Free energy at 298.15 K
17	C_v	$\frac{cal}{molK}$	Heat capacity at 298.15 K

Table 3. Calculated properties. Properties are stored in the order given by the first column.

Mulliken charges are added as a fifth column. Harmonic vibrational frequencies, SMILES and InChI are appended as respective additional lines.

Properties

All molecular geometries were relaxed, and properties calculated, at the DFT/B3LYP/6-31G(2df,p) level of theory. The list of properties of the 134k molecules is summarized in Table 3. For a subset of 6,095 isomers of $C_7H_{10}O_2$, energetics (properties 12–16) were additionally calculated at the G4MP2 level of theory. For a validation set of 100 randomly drawn molecules from the 133,885 GDB-9 set, enthalpies of atomization were calculated at the DFT/B3LYP/6-31G(2df,p), G4MP2, G4 and CBS-QB3 levels of theory.

Technical Validation

Validation of geometry consistency

To validate the consistency of the relaxed B3LYP geometries, we have used them to generate the corresponding InChI[20] strings with Corina and Open Babel (Version 2.3.0 2011)[21]. InChI corresponds to

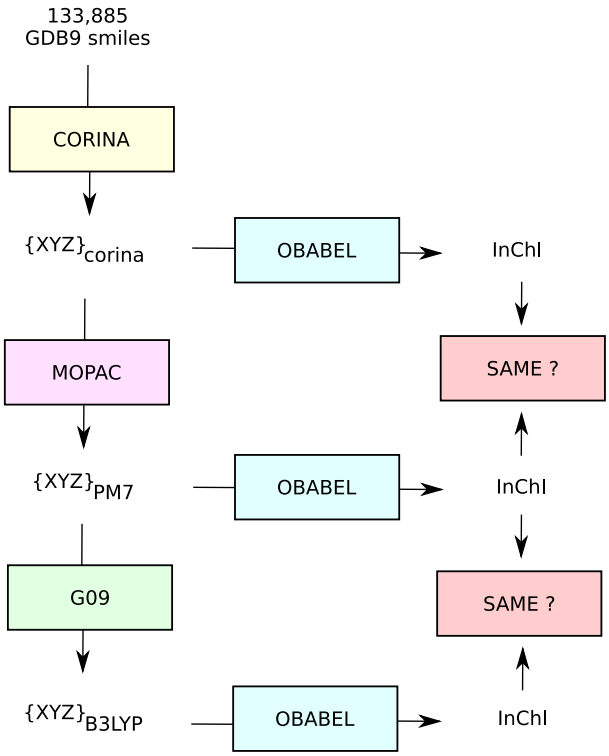

Figure 2. Schematic flow chart used for geometry consistency check.

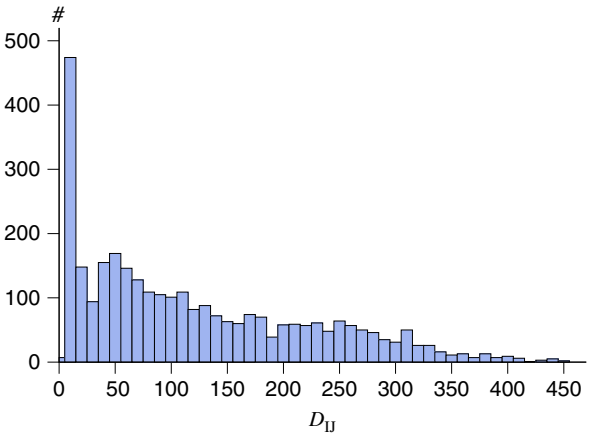

Figure 3. Histogram of Coulomb-matrix distances. For 3,054 molecules which failed the consistency test shown in Fig. 2 Coulomb-matrix distances, D_{IJ} in Ha, between B3LYP and Corina geometries are shown.

'IUPAC International Chemical Identifier'. The resulting strings have been compared to the InChI strings obtained from the initial Cartesian coordinates (generated by Corina using the original GDB-17 SMILES strings). See Fig. 2 for a flow-chart of this consistency check. Out of the 134k molecules, 3,054 molecules did not pass this test. This is due to the fact that SMILES and InChI representations are not unique because transformation of Cartesian coordinates to string based chemical identifiers is prone to implementation specific artifacts. For molecules with same topology, small differences in interatomic distances, bond and dihedral angles can lead to different molecular graphs encoded by the string. To quantify this artifact, the distribution of Coulomb-matrix distances[13], D_{IJ} [Hartree], using the Manhattan or L1 metric, between Corina generated Cartesian coordinates (see Fig. 2) and B3LYP coordinates is on display in Fig. 3 for all the 3,054 molecules.

Consider, for example, molecule indexed 58 in the 134k set, which is among the 3,054 molecules for which the consistency check failed. Its original GDB-17 SMILES corresponds to NC(=N)C#N, and

Cartesian coordinates can be generated using Corina. When feeding back these coordinates to Open Babel to perform the inverse task of reproducing the initial SMILES string, [NH]C(=[NH2])C#N is obtained instead. By contrast, performing first a geometry relaxation of the Corina generated Cartesian coordinates using PM7 followed by B3LYP, and only then parsing through Open Babel, recovers the original SMILES string. In this case, D_{IJ} using the Coulomb-matrices of Corina and B3LYP geometries is rather small (~27 Ha) when compared to distances between geometries of other failed molecules, see distance distribution plot in Fig. 3. By contrast, failed molecules with large D_{IJ} between Corina and B3LYP geometries, see Fig. 3, correspond to molecules for which the B3LYP relaxation induces dramatic distortion with significant alteration and rearrangement of covalent bonding patterns. As a result, parsing back these geometries with Open Babel yields different SMILES strings. Note that all the 6,095 constitutional isomers of $C_7H_{10}O_2$ for which G4MP2 calculations have been performed, passed this geometry consistency check, shown as a flow-chart in Fig. 2.

Validation of quantum chemistry results

All 134k molecules have been modeled using B3LYP/6-31G(2df,p) based DFT. Previously, B3LYP has been validated for several subsets, containing up to a few hundred small molecules. These benchmarks are of limited use since they are not necessarily sufficiently representative for gauging B3LYP's performance in general. In the case of DFT's systematic errors this issue is particularly pertinent[22]. Experimental data assembled in the NIST database is very sparse by comparison to our 134k organic molecules made up of CHONF atoms. Consequently, we have performed additional benchmark calculations for a subset of 100 randomly selected molecules using high level theories G4MP2[12], G4[23], and CBS-QB3[24,25].

The predictive power of the G4MP2 method is widely considered to be on par with experimental uncertainties. For example, comparison to the G3/05 test set[26,27] with 454 experimental energies (including enthalpies of formation, ionization potentials, electron affinities, proton affinities, and hydrogen bond energies) of small molecules yields MAE, and RMSE of 1.0, and 1.5 kcal/mol, respectively[12]. For the same properties and molecules, the slightly more accurate, and considerably more expensive method G4[12,23] yields errors of MAE = 0.8 kcal/mol, RMSE = 1.2 kcal/mol. G4MP2 has been shown to deviate only by 1.4 kcal/mol from 261 bond dissociation enthalpies computed with the highly accurate W1w composite procedure[28,29] for the BDE261 data set[28]. Consequently, we believe these calculations to be sufficiently suitable to validate the quality of the B3LYP energetics predictions. Various resulting deviations are summarized in Table 1. For the 100 molecules, the mean absolute error of B3LYP heats of atomization amounts to no more than 5 kcal/mol.

References

1. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature Mater* **12,** 191–201 (2013).
2. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432,** 823 (2004).
3. Koch, W. & Holthausen, M. C. *A Chemist's Guide to Density Functional Theory* (Wiley, 2002).
4. National institute of standards and technology. http://srdata.nist.gov (accessed 31 March 2014).
5. Jurečka, P., Šponer, J., Černý, J. & Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **8,** 1985–1993 (2006).
6. 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).
7. Lynch, B. J. & Truhlar, D. G. Small representative benchmarks for thermochemical calculations. *J. Phys. Chem. A* **107,** 8996–8999 (2003).
8. Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew. Chem. Int. Ed.* **44,** 1504–1508 (2005).
9. Martin, K. & Grimme, S. Mindless DFT benchmarking. *J. Chem. Theory Comput.* **5,** 993–1003 (2009).
10. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **19,** 7296–7303 (2013).
11. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52,** 2864–2875 (2012).
12. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **127,** 124105 (2007).
13. 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).
14. Hachmann, J. *et al.* The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2,** 2241–2251 (2011).
15. Norskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chem.* **1,** 37–46 (2009).
16. Weininger, D. SMILES, a chemical language and information system. 1.Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* **28,** 31–36 (1988).
17. Sadowski, J. & Gasteiger, J. From atoms and bonds to 3-dimensional atomic coordinates - automatic model builders. *Chem. Rev.* **93,** 2567–2581 (1993).
18. Stewart, J. J. P. *MOPAC2012, Version 13.136L, Stewart Computational Chemistry* (Colorado Springs, 2012).
19. Frisch, M. J. *et al. Gaussian 09, Revision d.01* (Gaussian, Inc., 2009).
20. Heller, S. R. & McNaught, A. D. The IUPAC international chemical identifier (InChI). *Chemistry International* **31,** 7–9 (2009).
21. O'Boyle, N. M. *et al.* Open Babel: an open chemical toolbox. *J. Chem. Inf.* **3,** 33 (2011).
22. Wodrich, M. D., Corminboeuf, C., Schreiner, P. R., Fokin, A. A. & Schleyer, P. v. R. How accurate are DFT treatments of organic energies? *Org. Lett.* **9,** 1851–1854 (2007).
23. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **126,** 084108 (2007).

24. Montgomery, J. A. Jr, Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VI. use of density functional geometries and frequencies. *J. Chem. Phys.* **110,** 2282–2827 (1999).
25. Montgomery, J. A. Jr, Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VII. use of the minimum population localization method. *J. Chem. Phys.* **112,** 6532–6542 (2000).
26. Curtiss, L. A., Redfern, P. C., Raghavachari, K. & Pople, J. A. Gaussian-3X (G3X) theory: use of improved geometries, zero-point energies, and Hartree-Fock basis sets. *J. Chem. Phys.* **114,** 108–117 (2001).
27. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Assessment of Gaussian-3 and density-functional theories on the G3/05 test set of experimental energies. *J. Chem. Phys.* **123,** 124107 (2005).
28. Chan, B. & Radom, L. BDE261: a comprehensive set of high-level theoretical bond dissociation enthalpies. *J. Phys. Chem. A* **116,** 4975–4986 (2012).
29. Boese, A. D. *et al.* W3 theory: robust computational thermochemistry in the kJ/mol accuracy range. *J. Chem. Phys.* **120,** 4129–4141 (2004).

Data Citation

1. Ramakrishnan, R., Dral, P. O., Rupp, M. & von lilienfeld, O. A. Figshare. http://dx.doi.org/10.6084/m9.figshare.978904 (2014).

Acknowledgements

The authors thank James Stewart and Christof H. Schwab for providing trial licenses for the packages MOPAC and Corina, respectively. The authors are thankful for CPU time at the Universitätsrechenzentrum, University of Basel. We greatly acknowledge J.-L. Reymond and his group for extensive discussions, ideas, and access to GDB-17 data. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under contract DE-AC02-06CH11357. Funding from the Swiss National Science foundation is acknowledged (No. PPOOP2 138932).

Author Contributions

All authors designed and performed research, and wrote the paper.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ramakrishnan, R. *et al.* Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1:140022 doi: 10.1038/sdata.2014.22 (2014).