



OPEN

DATA DESCRIPTOR

Accurate Chemistry Collection: Coupled cluster atomization energies for broad chemical space

Sebastian Ehlert^{1,6}✉, Jan Hermann^{2,6}✉, Thijs Vogels¹, Victor Garcia Satorras¹, Stephanie Lanius², Marwin Segler^{1,3}, Klaas J. H. Giesbertz¹, Derk P. Kooi¹, Kenji Takeda⁴, Chin-Wei Huang¹, Giulia Luise³, Rianne van den Berg¹, Paola Gori-Giorgi¹ & Amir Karton^{1,5}✉

Accurate thermochemical data with sub-chemical accuracy (within 1 kcal mol⁻¹ of the empirical ground truth) are essential for advancing computational chemistry methods. However, existing datasets that reach this level of accuracy remain limited in size or scope. This hinders the development of data-driven methods with predictive accuracy across the broad chemical space of closed-shell, neutral molecules. Here we present Microsoft Research Accurate Chemistry Collection (MSR-ACC) and its first release, MSR-ACC/TAE25, comprising 73,040 total atomization energies at the CCSD(T)/CBS level obtained with the W1-F12 thermochemical protocol. The dataset is constructed to exhaustively cover the chemical space of closed-shell, charge-neutral, covalently bound equilibrium molecular structures containing up to 5 non-hydrogen atoms drawn from elements up to argon and lacking significant multireference character. The dataset and its canonical train and validation splits are openly available on Zenodo in the QCSchema format under the CDLA Permissive 2.0 license. This first release of MSR-ACC enables data-driven approaches for developing predictive computational chemistry methods with unprecedented accuracy and scope.

Background & Summary

The enthalpy of formation is the most fundamental thermodynamic property of a molecule, from which many other thermodynamic properties are derived, for example, reaction and combustion enthalpies, bond dissociation energies, Gibbs-free energies of formation, and equilibrium constants. The enthalpy of formation—i.e., the energy change associated with forming a molecule from its constituent elements in their standard states—is cognate to the total atomization energy (TAE), which is the energy required to break down a molecule into its constituent atoms in the gas phase. These two thermochemical properties are readily interconverted via the atomic enthalpies of formation at 0 K. From a theoretical perspective, the TAE provides a fundamental reference for determining rigorous error bars for electronic structure theories^{1–6}. The reason for this is not only because the TAE is the most fundamental thermodynamic quantity, but also because it is one of the most challenging thermodynamic properties for electronic structure methods. It is well established that thermochemical transformations become increasingly more challenging for approximate electronic structure methods as they conserve less of the chemical environments between reactants and products^{7–19}. For example, the performance of any approximate electronic structure method is expected to improve along the reaction hierarchy:

atomization → isogyric →isodesmic → hypohomodesmotic → homodesmotic → hyperhomodesmotic

This improvement is due to progressively more systematic error cancellation between reactants and products occurring along this series^{7,11,15}. Thus, TAEs, which involve the complete dissociation of a molecule into its constituent atoms, represent the most extreme scenario where there is no conservation of the chemical

¹Microsoft Research, AI for Science, Amsterdam, Netherlands. ²Microsoft Research, AI for Science, Berlin, Germany.

³Microsoft Research, AI for Science, Cambridge, United Kingdom. ⁴Microsoft Research, Cambridge, United Kingdom.

⁵School of Science and Technology, University of New England, Armidale, Australia. ⁶These authors contributed equally: Sebastian Ehlert, Jan Hermann. ✉e-mail: sehlert@microsoft.com; jan.hermann@microsoft.com; amir.karton@une.edu.au

environments between reactants and products. It follows that the performance of electronic structure methods against TAEs provides upper-bound errors, and the performance for less challenging chemical systems and properties should improve along this series.

Over the past two decades, a number of high-level theoretical datasets for TAEs and heats of formation labeled with *ab initio* wavefunction methods have been developed. They include the W4 series of datasets of TAEs of small compounds containing H, B, C, N, O, F, Al, Si, P, S, and Cl^{9,20,21} calculated at the full configuration interaction (FCI) complete-basis-set (CBS) limit with the W4 composite wavefunction method (or variants thereof)^{22–25}. The prohibitive computational cost of these calculations limits the size and number of the systems, but offers the possibility to include also molecules with strong multireference character, since the W4 protocol provides confident sub-chemical accuracy even for such systems^{9,22–26}. Moving to computationally more economical coupled-cluster methods, namely with single, double, and quasiperturbative triple excitations (CCSD(T))²⁷, allows for the construction of datasets with many more molecules. In a tour de force study Narayanan *et al.*²⁸ calculated G4(MP2) energies for 133k organic species (composed of H, C, N, O, and F) with up to 9 non-hydrogen atoms in the GDB-9 dataset (GDB9-G4MP2)²⁹. G4(MP2) is a computationally efficient composite wavefunction method^{30,31}, which includes an empirical ‘higher-level correction’ (HLC) term to account for efficiency-driven compromises in the theoretical model. As a result, G4(MP2) is applicable even to large organic systems, such as large polycyclic aromatic hydrocarbons and fullerenes^{17,32–35}, but it may not achieve chemical accuracy for nonstandard bonding situations and for systems that are not well-represented in the experimental G3/05² training set^{36–41}. Composite wavefunction methods such as Wn ⁴², Wn -F12⁴³, and WnX ⁴⁴ ($n = 1, 2$) overcome these limitations by extrapolating the Hartree–Fock, CCSD, and (T) components separately to the CBS limit and also including a core-valence (CV) correction. They were recently used to construct a number of datasets of heats of formation and TAEs^{45–47}, but these are mostly limited to organic systems such as those already covered by the GDB9-G4MP2 dataset²⁸. Departing from the coupled cluster family of methods Khan *et al.*⁴⁸ developed the VQM24 dataset of chemically diverse molecules from the first three periods of the periodic table, and labeled the subset of the 10k smallest molecules with up to 4 non-hydrogen atoms with single-determinant diffusion quantum Monte Carlo (DMC). However, DMC with a single-determinant Slater–Jastrow ansatz is insufficiently accurate for atomization energies⁴⁹, with mean absolute error on the G1 dataset⁵⁰ of ~ 3 kcal mol^{−1}, and can only be made more accurate through much more expensive multideterminant ansatzes.

A significant gap in the literature is thus the absence of a chemical dataset covering a broad and diverse set of closed-shell, neutral TAEs for organic and inorganic compounds at sub-chemical accuracy. Here, we partially fill that gap by introducing Microsoft Research Accurate Chemistry Collection (MSR-ACC) and its first dataset, designated MSR-ACC/TAE25⁵¹, comprising 73,040 total atomization energies at the CCSD(T)/CBS level obtained with the W1-F12 thermochemical protocol. The dataset is constructed to exhaustively cover the chemical space of closed-shell, charge-neutral, covalently bound equilibrium molecular structures containing up to 5 non-hydrogen atoms drawn from elements up to argon and lacking significant multireference character. Rare-gas elements were *a priori* excluded as they do not form covalent bonds, while all other elements up to argon (i.e., H, Li, Be, B, C, N, O, F, Na, Mg, Al, Si, P, S, and Cl) are featured prominently. The molecular structures were obtained by exhaustive sampling of all possible molecular graphs subject in the most general case only to the constraint on a maximum valency of each element. The resulting graphs were then converted to equilibrium 3D structures by a cascade of structure optimization steps with increasingly accurate methods. All molecular structures that dissociated into covalently disconnected fragments, for which the triplet state was found lower in energy than a singlet state, and for which after the W1-F12 labeling the multireference character was found too strong, were discarded. Next to the W1-F12 reference values, the dataset also includes a number of auxiliary values such as atomization energies with DFT or quantities used for filtering that serve for technical validation.

MSR-ACC/TAE25 opens the way for developing and evaluating generally applicable machine-learning, density-functional theory, and semi-empirical methods, and has already been used for training the first exchange–correlation functional to reach chemical accuracy for atomization energies⁵². In contrast to smaller or more specialized TAE databases, MSR-ACC/TAE25 provides a large and chemically diverse test set for identifying systematic errors and validating approximate electronic structure methods, for example, in describing bond energies across a wide range of challenging s- and p-block compounds. The size, diversity, and accuracy of the dataset make it useful not only for developing deep-learning DFT methods⁵², but also for training and validating models such as graph neural networks, critically testing their ability to generalize beyond the typical organic chemistry focus of datasets like GDB-9. Finally, the dataset can be filtered to create highly specific benchmarks to answer practical questions, such as selecting the most accurate functional for a particular type of chemical system (e.g., phosphorus sulfide compounds)³⁹. We envision continuing to expand MSR-ACC in the future with datasets of quality equal to or exceeding that of the dataset presented in this work.

Methods

Structure generation. The molecular geometries were generated to cover the chemical space for elements of the first three periods of the periodic table (H–Ar) excluding the rare-gas atoms. The structure generation has two phases: molecular graph generation and conversion of molecular graphs to molecular 3D structures.

Molecular graphs. We constrain the number of electrons to be even (closed-shell molecules) and the total molecular charge to be zero. Compared to what is common in organic chemistry, we use much looser criteria on which molecular graphs are allowed and put the only constraint on the *maximum* valency of each element: 1 for Li and Na; 2 for Be, O, and Mg; 4 for B, C, N, and Cl; and 6 for every other element. The molecular graphs were generated using different approaches to maximize the diversity of all possible bonding situations (Fig. 1a). First, we use a brute-force enumeration of all possible molecular graphs based on bare template graphs for a given number of non-hydrogen atoms (Version 1). While this approach has combinatorial complexity, it allows us to

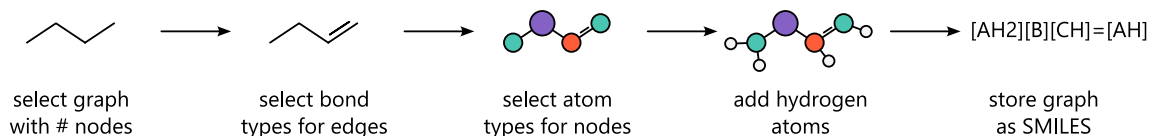
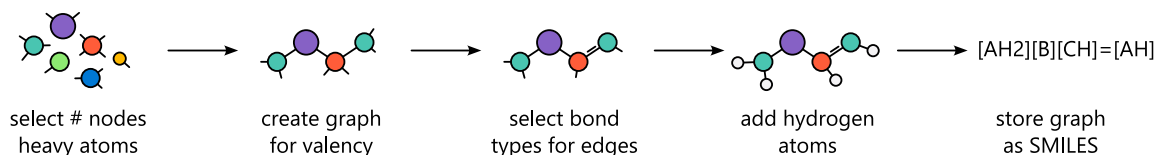
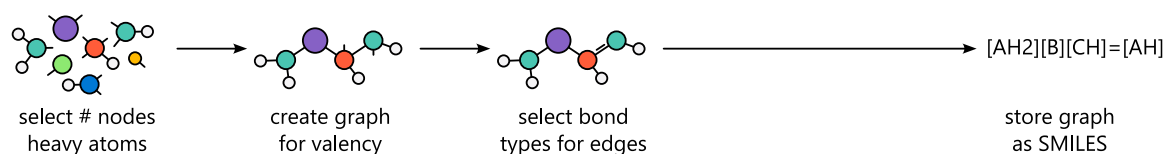
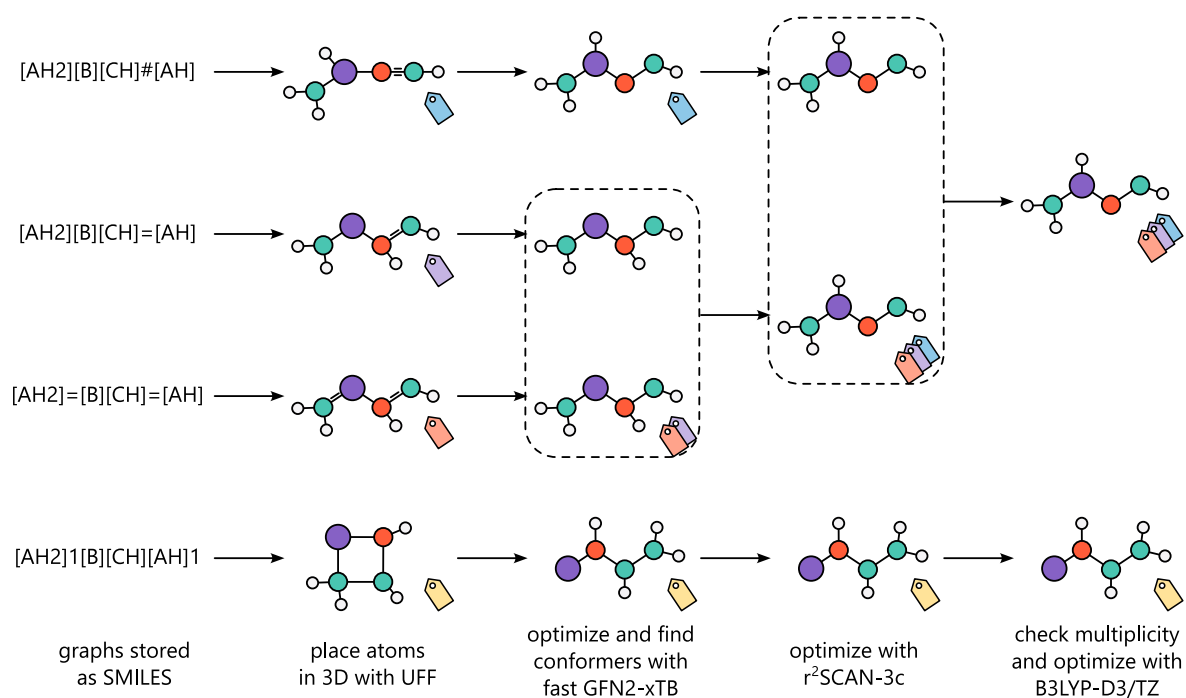
a Version 1: combinatorial enumeration with implicit hydrogen**Version 2: degree sequence sampling with implicit hydrogen****Version 3: degree sequence sampling with explicit hydrogen****b**

Fig. 1 End-to-end molecular structure generation. **(a)** Three methods used for generating new molecular graphs (as SMILES strings). (1) Enumerate all possible graphs, bond types, and atom types before saturating with hydrogen atoms. (2) Sample nonhydrogen atoms, a graph and bond types to satisfy maximum allowed valencies, and saturate with hydrogen atoms. (3) Sample nonhydrogen atoms with explicit hydrogens and then a graph and bond types to satisfy maximum allowed valencies. **(b)** From molecular graphs to final B3LYP-D3(BJ)/def2-TZVPP optimized structures. Atoms are placed using UFF initially and then optimized using GFN2-xTB which also samples the conformational space. Finally, the geometry is optimized using a meta-GGA (r^2 SCAN-3c) and a hybrid functional (B3LYP). Throughout, duplicates based on reconstructed molecular graph and matching TAE are merged (indicated by label symbols). Spin states are checked with B3LYP.

exhaustively enumerate all possible graphs up to four non-hydrogen atoms. Second, a degree sequence sampling approach is used where candidate graphs are considered from a list of atoms with a degree up to the respective

open valency. The resulting bare graphs were dressed with different bond types (single, double, triple), respecting the maximum valency for each atom. The degree sequence sampling was performed once with implicitly added hydrogen atoms (Version 2) and once with explicit hydrogen atoms in the degree sampling step (Version 3). A pseudocode for this molecular graph generation algorithm is released alongside the dataset⁵¹. Third, an autoregressive model based on a GPT-2 transformer architecture was used for predicting new molecular graphs. The model was trained on all SMILES strings with 5 non-hydrogen atoms generated by the graph sampling described above that could be successfully converted and optimized at B3LYP-D3(BJ)/def2-TZVPP level of theory (~6M). The SMILES generative model was then sampled for ~1.5M novel molecular graphs which were not part of the training data with a 85% novelty success rate. After 3D structure generation and optimization (see below), this led to about ~20% of the resulting molecular structures with 5 non-hydrogen atoms originating from the GPT-2 SMILES model. All three approaches taken together, this exhaustive generation of molecules can be considered a bottom-up counterpart to the top-down fragment extraction of Huang and von Lilienfeld⁵³.

3D structures. The initial 3D structures were generated with UFF⁵⁴ using OpenBabel⁵⁵ to convert SMILES to 3D geometries (Fig. 1a). The resulting structure guess was optimized using GFN2-xTB⁵⁶ with the *xtb*⁵⁷ program package and reoptimized using r²SCAN-3c⁵⁸ and refined using B3LYP-D3(BJ)^{59,60}/def2-TZVPP (as prescribed by the W1 protocol) with the Orca⁶¹ program package. In each optimization step, structures were compared with a global index to find duplicates based on the molecular graph⁶² and the total atomization energy at the respective level of theory. Importantly, all the species in the dataset were verified to be equilibrium structures by confirming they have all-real harmonic frequencies at the same level of theory.

To ensure that the dataset contains only molecules in the electronic ground state, we filter out structures for which the singlet–triplet gap $S_0 - T_1$ is negative according to B3LYP/def2-TZVP—this filters out ~5% of all molecules we find. To select the structures which can be safely treated with CCSD(T)-based composite wavefunction protocols, the fraction of TAE accounted for by the parenthetical connected triple excitations (%TAE[(T)]) is computed at the CCSD(T)/6-31G(d) level of theory, all structures with more than 6% %TAE[(T)] were not considered for labeling. This led to exclusion of additional ~5% of the generated structures. We later verified that the %TAE[(T)] diagnostic taken from W1-F12 theory also does not exceed 6%. Finally, we filter out all structures that have dissociated to more than one covalently bound fragment. We determine the molecular graph through a combination of the bond model of GFN-FF⁶² and a simple heuristic based on a sum of covalent radii.

For all structures we evaluate a range of electronic structure methods, including semi-empirical methods GFN1-xTB⁶³ and GFN2-xTB⁵⁷, composite electronic structure methods PBEh-3c⁶⁴, B97-3c⁶⁵, r²SCAN-3c⁵⁸, (meta-)GGAs B97M-V⁶⁶ and r²SCAN-D3(BJ)⁶⁷, (range-separated) hybrid functionals B3LYP-D3(BJ)^{59,60}, M06-2X⁶⁸, ω B97X-V⁶⁹, and ω B97M-V⁷⁰, and double hybrid functionals revDSD-PBEP86-D3(BJ)⁷¹ and ω B97X-2⁷².

Subsampling. After generating ~1M unique molecular 3D structures with up to 5 non-hydrogen atoms, we subsample the structures for labeling with the high-accuracy W1-F12 protocol (see below) to maximize the chemical diversity while minimizing the total computational cost of the labeling. We subsample along three axes: (i) number of non-hydrogen atoms, (ii) presence of s-block elements, and (iii) presence of elements from the 3rd period of the periodic table. For each bracket, we have estimated the computational cost of W1-F12 labeling based on preliminary calculations on a small subset of molecules and then allocated a computational budget for each bracket, attempting to maximize the eventual training signal of the whole dataset at a fixed total computational cost.

Labeling with the composite W1-F12 protocol. Using the fully optimized B3LYP-D3(BJ)/def2-TZVPP structures, we obtain the nonrelativistic, all-electron CCSD(T)/CBS TAE using the high-level W1-F12 composite wavefunction protocol⁴³. W1-F12 theory is an explicitly correlated^{73–75} version of the original W1 theory⁴². The computational details of W1-F12 theory have been specified and rationalized in great detail elsewhere⁴³. In brief, the Hartree–Fock (HF) energy is extrapolated to the complete basis-set limit from the cc-pVDZ-F12 and cc-pVTZ-F12 basis sets⁷⁶, using the $E(L) = E_\infty + A/L^\alpha$ two-point extrapolation formula, with $\alpha = 5$. The complementary auxiliary basis set (CABS) singles correction is included in the Hartree–Fock (HF) energy^{77–79}. The valence CCSD-F12 correlation energy (denoted by Δ CCSD hereinafter) is extrapolated from the same basis sets, using an optimal extrapolation exponent of $\alpha = 3.67$ for elements up to argon. Optimal values for the geminal Slater exponents were taken from Peterson *et al.*⁷⁶ and Hill *et al.*⁸⁰, and the resolution of the identity (RI) approximation was applied using the OptRI auxiliary basis sets⁸¹. The diagonal, fixed-amplitude 3C(FIX) ansatz and the CCSD-F12b approximation were used in these calculations^{78,79}. The valence parenthetical connected triple excitations contribution (denoted by $\Delta(T)$) is obtained from standard CCSD(T) calculations, as is the case in W1w theory^{22,42}. Specifically, the $\Delta(T)$ component is extrapolated from the jul-cc-pV(D+d)/Z and jul-cc-pV(T+d) Z basis sets using the above two-point extrapolation formula with $\alpha = 3.22$ ^{82–85}. The CCSD inner-shell contribution is calculated with the core-valence weighted correlation-consistent cc-pwCVTZ basis set of Peterson and Dunning⁸⁶, while the (T) inner-shell contribution is calculated with the cc-pwCVTZ basis set without the f functions. Since we are primarily interested in assessing the performance of nonrelativistic computational chemistry methods on the electronic potential energy surface, the scalar relativistic, spin-orbit, and diagonal Born–Oppenheimer corrections are not considered in the present work. All the calculations involved in W1-F12 theory were carried out with the Molpro 2024.1 program suite^{87–89}.

We have labeled all found molecules with up to 4 non-hydrogen atoms, and a smaller sample of all found molecules with 5 non-hydrogen atoms.

Key	Description	Unit
graph:all	Connectivity of all atoms	list of indices (0-based)
graph:non-h	Connectivity of all non-hydrogen atoms	list of indices (0-based)
tae@method	Total atomization energy	hartree
tae:frac[(T)]@{method}	%TAE[(T)] as a fraction	1
singlet-triplet-gap-s0-t1@{method}	Singlet-triplet gap S_0-T_1	hartree
tae[component]@w1-f12	Component of the W1-F12 TAE	hartree

Table 1. Labels included in the `extras` field of the data records.

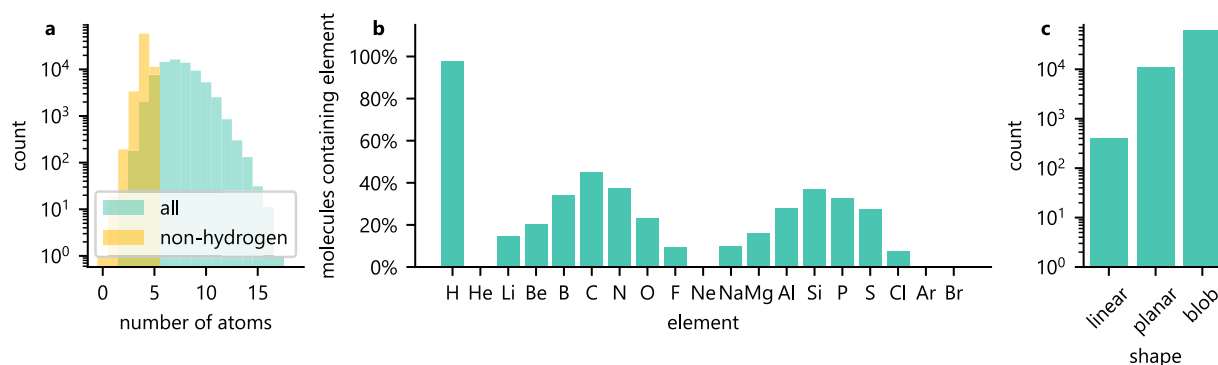


Fig. 2 Elemental and structural distributions. **(a)** Molecular size as atom count. **(b)** Elemental occurrence. Hydrogen is present in 97.6% of molecules. **(c)** General 3D shape of the molecules.

Data Records

The data records for 73,040 molecules of MSR-ACC/TAE25 are available in Zenodo under the CDLA Permissive 2.0 license⁵¹. To aid machine-learning applications, we also release training and validation subsets of the full dataset, which constitute canonical 99% and 1% splits, respectively, of MSR-ACC/TAE25 after removing the overlap with the popular W4-17 and GMTKN55 benchmark sets. The overlap was constructed such that it contains any atomization reaction where the reconstructed molecular graphs of the molecules match.

Molecules are formatted using the QCSchema standard, including the Cartesian coordinates in bohr (geometry field), element symbols (symbols field), total charge (molecular_charge field) and spin state (molecular_multiplicity field). The `extras` field of each molecule contains the W1-F12 atomization energy labels (tae@w1-f12) as well as additional information including molecular graphs, DFT atomization energies, singlet-triplet gaps, and W1-F12 energy components described in Table 1.

Technical Validation

Figure 2 gives an overview of basic structural distributions over MSR-ACC/TAE25. The molecular sizes range from zero (H_2) to five non-hydrogen atoms and from 2 to 17 (isopentane) atoms including hydrogen. Defining organic systems as molecules containing at least one carbon atom, there are 45.1% organic systems and 54.9% inorganic systems. Thus, similarly to the much smaller and chemically less diverse W4-17 dataset, there is a similar representation of organic and inorganic systems. The most represented elements in the dataset are:

$$C > N \text{ and } Si > B \text{ and } P > S \text{ and } Al > O > Be > Mg > Li > Na > F > Cl$$

Importantly, about 75% of the systems are mixed second/third-period species, 17% are pure second-period systems, and about 8% are pure third-period systems. The geometry of the molecules spans linear (0.6%), planar (15.2%) and general 3D structures (84.3%).

The elemental composition of the molecules in MSR-ACC gives rise to a diverse bond types, with many molecules involving nontraditional bonding situations. Overall, there are 287k bonds between non-hydrogen atoms in MSR-ACC/TAE25 according to our bond heuristic, covering all possible combinations of elements in the first two rows except for F-F. The occurrence of element pairs involved in bonding follows the pattern that could be expected from typical valencies of individual elements (Fig. 3a). Notable outliers in occurrence are pairs of typically monovalent elements: H-H (only once in H_2), H-halogen (20), Cl-halogen (7), and F-F (none, as F_2 is associated with %TAE[(T)] > 6%). This can be contrasted to two other large atomization energy datasets labeled with ab initio wavefunction methods. VQM24/DMC follows the same bond distributions with non-hydrogen elements from groups 1, 2, and 13 (Li, Na, Be, Mg, B, Al) omitted and Br added (following the same pattern as other halogens). GDB-9 is severely restricted in diversity by its focus on organic drug-like molecules, resulting in only C-C, C-N, C-O, C-F, N-N, and N-O non-hydrogen bonds, with the latter three occurring with low relative frequency.

A different metric of chemical diversity is obtained by counting unique 1st-neighbor environments in the molecular graphs for each element (Fig. 3b). In the full dataset, all elements except for H and halogens occur

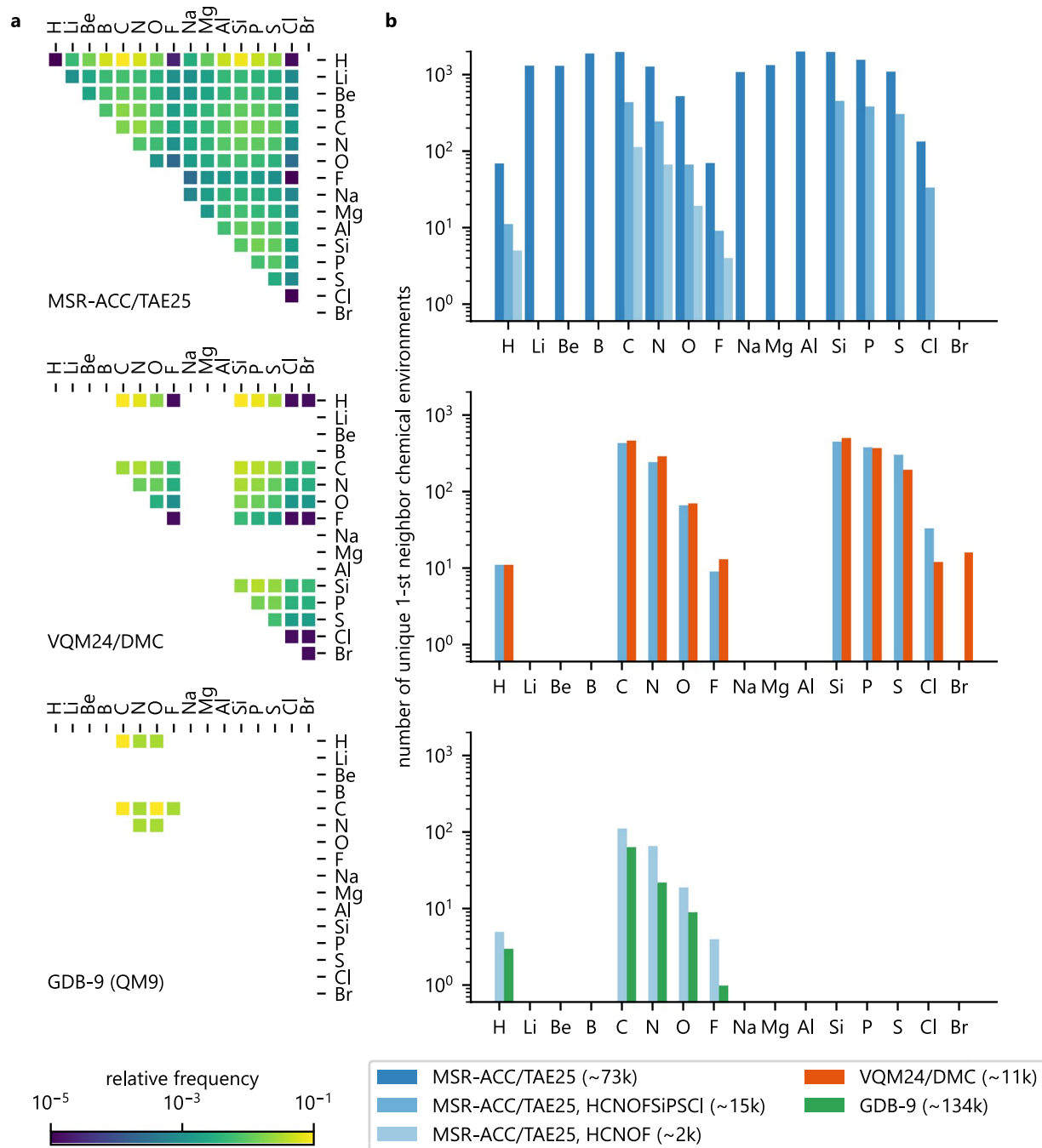


Fig. 3 Reconstructed molecular graphs compared across three datasets: MSR-ACC/TAE25, VQM24/DMC, and GDB-9 (also known as QM9). Graphs are reconstructed from 3D structures using a bond model from GFN-FF and a simple heuristic based on a sum of covalent radii. **(a)** Distribution of element pairs in bonds. **(b)** Unique 1-st neighbor environments per element. Subsets of MSR-ACC/TAE25 constrained to elemental subspaces are also shown.

in ~1000 unique chemical environments, while H and halogens occur in ~100 unique chemical environments. Hydrogen gains this diversity through various two-electron three-centric bonds with one atom from groups 1, 2, and 13, and some other non-hydrogen element. When restricting to molecules with no non-hydrogen atom from groups 1, 2, and 13, the number of unique chemical environments drops by half an order of magnitude, and its distribution over elements is remarkably similar to VQM24/DMC. This indirectly verifies that the structure generation for both datasets exhaustively covers all possible local chemical environments in the respective elemental spaces. When further restricting to the HCNOF subset, we find that the number of unique environments in MSR-ACC is still 2–3 times larger than in GDB-9, owing to covering more than just ‘drug-like’ organic chemistry.

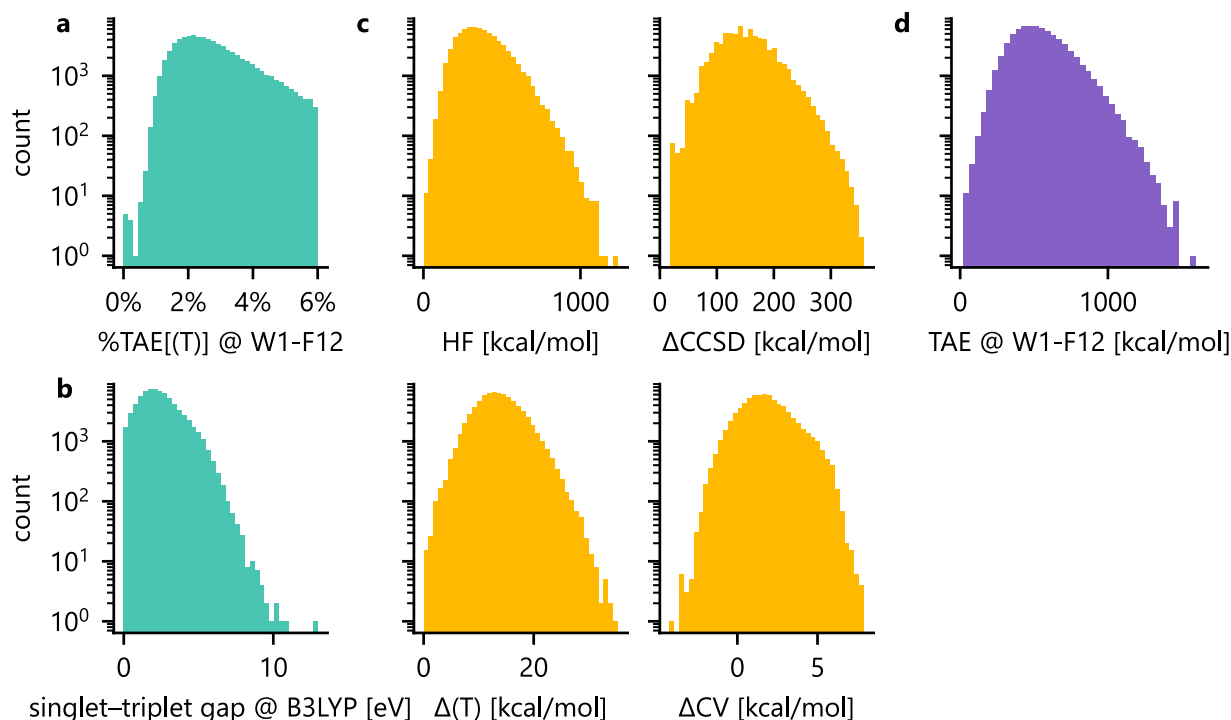


Fig. 4 Energetic distributions. (a) Multireference diagnostic %TAE[(T)]. (b) Vertical singlet-triplet gaps S_0-T_1 from B3LYP. (c) All four components of the W1-F12 total atomization energy. (d) Reference W1-F12 total atomization energy.

Yet another validation of the coverage of the chemical space is comparison of the overlap between MSR-ACC/TAE25 and the closed-shell single-referential subset of W4-17 up to 4 non-hydrogen atoms, which we expect to be covered exhaustively. We find that only two molecules from this W4-17 subset are missing in MSR-ACC/TAE25 for fundamental chemical reasons: diborane, B_2H_6 , which due to its uniqueness cannot be constructed by our molecular graph generation, and cyclobutadiene, C_4H_4 , likely due to its instability, since we did find two dozen of its other isomers. For three other missing molecules (hydrogen peroxide, H_2O_2 , acetaldehyde, CH_3CHO , and hydrazine, N_2H_4), our structure generation yielded the corresponding molecular graph, but the subsequent 3D structure optimization yielded a saddle point rather than an equilibrium structure, so they were filtered out.

It is well established that the CCSD(T) method cannot generally achieve chemical accuracy for systems dominated by moderate-to-severe nondynamical correlation effects^{9,21–26,90–94}. It is therefore essential to estimate the contributions to the TAEs from post-CCSD(T) excitations. A highly successful a priori predictor for such contributions is the so-called %TAE[(T)] diagnostic^{9,22,23}, which is defined as the percentage of the TAE accounted for by parenthetical connected triple excitations:

$$\%TAE[(T)] = \frac{TAE[CCSD(T)] - TAE[CCSD]}{TAE[CCSD(T)]}$$

where TAE[CCSD] and TAE[CCSD(T)] are the TAEs calculated at the CCSD and CCSD(T) levels. This simple multireference diagnostic is particularly well suited for eliminating systems with significant post-CCSD(T) contributions to the TAEs. It has been suggested that %TAE[(T)] values up to ~5% indicate that post-CCSD(T) contributions to the TAEs should not exceed ~0.5 kcal mol⁻¹²³. In the present work, we have chosen to exclude all systems with %TAE[(T)] > 6%. The distribution of %TAE[(T)] (Fig. 4a) therefore ends sharply at 6%, while peaking at ~2% and tapering off towards 0%. The second major filtering criterion is positive vertical singlet-triplet S_0-T_1 gap, in order to have only molecules in their electronic ground states. The distribution of the S–T gap (Fig. 4b) therefore ends sharply at 0 eV, peaks at ~2 eV, and tapers off to ~10 eV.

Figure 4c gives an overview of the HF, Δ CCSD, $\Delta(T)$, and Δ CV contributions to the W1-F12 TAEs. The HF component spreads over a wide range from 1.3 to 1238.8 kcal mol⁻¹. We note, however, that small HF TAEs below 32.5 kcal mol⁻¹ are obtained for only 12 systems, nearly all of which consist of Li bonding with Na, Be, or Mg. The overall all-electron CCSD(T)/CBS TAEs for these 12 systems are also relatively small and range between 20.1 and 62.2 kcal mol⁻¹. HF TAEs above 1000 kcal mol⁻¹ are obtained for 38 systems, nearly all of which contain 3–4 carbon. We note that there is no apparent correlation between the HF component and the number of non-hydrogen atoms, however, the squared correlation coefficient between the total number of atoms and the HF component is $R^2 = 0.757$.

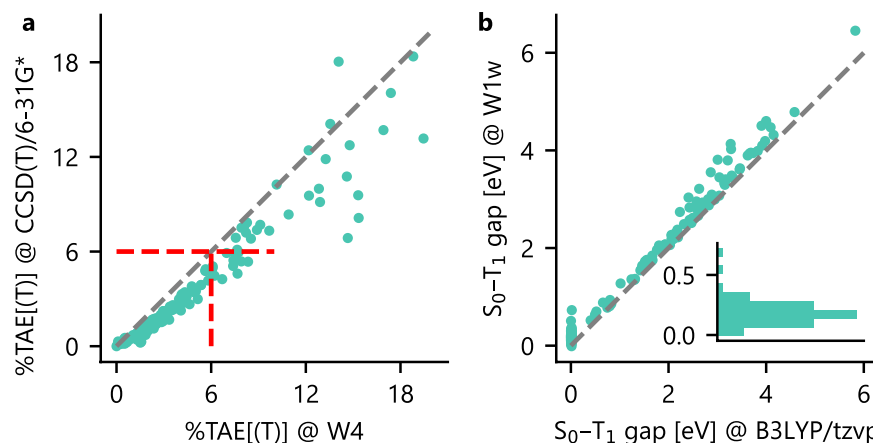


Fig. 5 Validation of filtering criteria. **(a)** Multireference descriptor %TAE[(T)] evaluated on the W4-17 dataset in the small 6-31G(d) basis almost always underestimates the true CBS value (from W4), leading to false negatives but no false positives of multireference character. **(b)** Singlet–triplet gaps are only overestimated by B3LYP, never underestimated. A random sample of molecules with the B3LYP gap close to zero (inset) has the CCSD(T) gap (from W1w) ranging from 0.0 to 0.4 eV.

The ΔCCSD component spreads over a wide range of $340 \text{ kcal mol}^{-1}$ from 18.4 to $358 \text{ kcal mol}^{-1}$. The largest 40 ΔCCSD contributions range between 327 and $358.1 \text{ kcal mol}^{-1}$. Most of these systems are nitrogen-rich species. The systems with the smallest ΔCCSD contributions are dominated by beryllium compounds. We note that overall, for the 73,040 systems in the dataset, there is a noticeable but weak statistical correlation between ΔCCSD component and the number of valence electrons, with $R^2 = 0.613$.

The $\Delta(T)$ component varies between 0.0 and $35.3 \text{ kcal mol}^{-1}$. The lowest $\Delta(T)$ contributions are obtained, as expected, for lithium and sodium systems. Excluding the systems with only two valence electrons (H_2 , HLi , HNa , LiNa) for which the $\Delta(T)$ contribution is zero by definition, the lowest 40 $\Delta(T)$ contributions range between 0.0 and $1.8 \text{ kcal mol}^{-1}$. Of these 40 systems, 23 contain either Na or Li, or both, where in most cases there are multiple Li and Na atoms in the same molecule. Similarly to the ΔCCSD component, the largest $\Delta(T)$ contributions are obtained for nitrogen-rich species.

The ΔCV contribution can be either negative or positive, meaning that ΔCV correlation is more or less stabilizing, respectively, in the constituent atoms than in the molecule. Overall, the ΔCV contributions range from -4.2 to $+7.9 \text{ kcal mol}^{-1}$, where 12% of the ΔCV values are negative and 88% are positive. In line with observations^{23–25} on the W4-17 dataset²¹, 87.5% of the systems with a negative CV correction contain Al, Si, or both atoms. The resulting TAEs range from 20.1 (LiNa) to $1560 \text{ kcal mol}^{-1}$ (isopentane).

We introduced two primary filtering criteria when constructing MSR-ACC. First, %TAE[(T)] computed in the small 6-31G(d) basis is used to identify multireference molecules to be filtered out before labeling them with the much more expensive W1-F12 calculation. We have verified on the W4-17 dataset that this almost never leads to excluding molecules that are in fact single-referential, as the small-basis version of the metric strictly underestimates it (Fig. 5a). After calculating the W1-F12 TAEs, we again removed any additional systems with a %TAE[(T)] diagnostic above 6%, where the %TAE[(T)] CBS values were taken from W1-F12 theory. It is important to note that this filtering process may result in some ‘false negatives’ (i.e., erroneously flagging well-behaved molecules as having multireference character) but is highly unlikely to produce ‘false positives’ (i.e., erroneously including multireference molecules)²³. Second, singlet–triplet gap S_0-T_1 is approximated by B3LYP/def2-TZVP to exclude molecules with a triplet ground state. We have verified on a random sample of molecules from the dataset against gaps calculated with W1w that B3LYP strictly underestimates the gap (Fig. 5b). This leads to a small fraction of molecules being unnecessarily filtered out that in fact have a singlet ground state, but we consider this worth the saved cost of having to run extra W1-F12 triplet calculations. This ensures that the molecules that pass this filter are reliably singlet ground-state systems. The conservative combination of the %TAE[(T)] diagnostic and the B3LYP singlet–triplet gap ensures the thermochemical reliability of the final MSR-ACC/TAE25 dataset.

Next to the reference W1-F12 values, the dataset contains TAEs from KS-DFT with a number of popular exchange–correlation (XC) functionals. We found that the distribution of errors of all functionals with respect to the reference values follows the normal distribution very closely, which enables us to identify outliers confidently (Fig. 6). This can then serve as an indirect validation of the reference values, where we ensure that there are no cases of different XC functionals agreeing among each other on the TAE but disagreeing with W1-F12. The accuracy of the different XC functionals follows the general trends one would expect from the Jacob’s ladder categorization.

The MSR-ACC/TAE25 dataset provides a large-scale, internally consistent dataset of CCSD(T)/CBS TAEs calculated with the high-level W1-F12 composite wavefunction method⁴³. Our technical validation focuses on ensuring that the molecules included in the dataset can be reliably treated at this level of theory. By applying the %TAE[(T)] diagnostic and the singlet–triplet gap, we systematically filter out molecules for which the CCSD(T) method would be less reliable. It is important to note that W1-F12 theory has been extensively benchmarked against

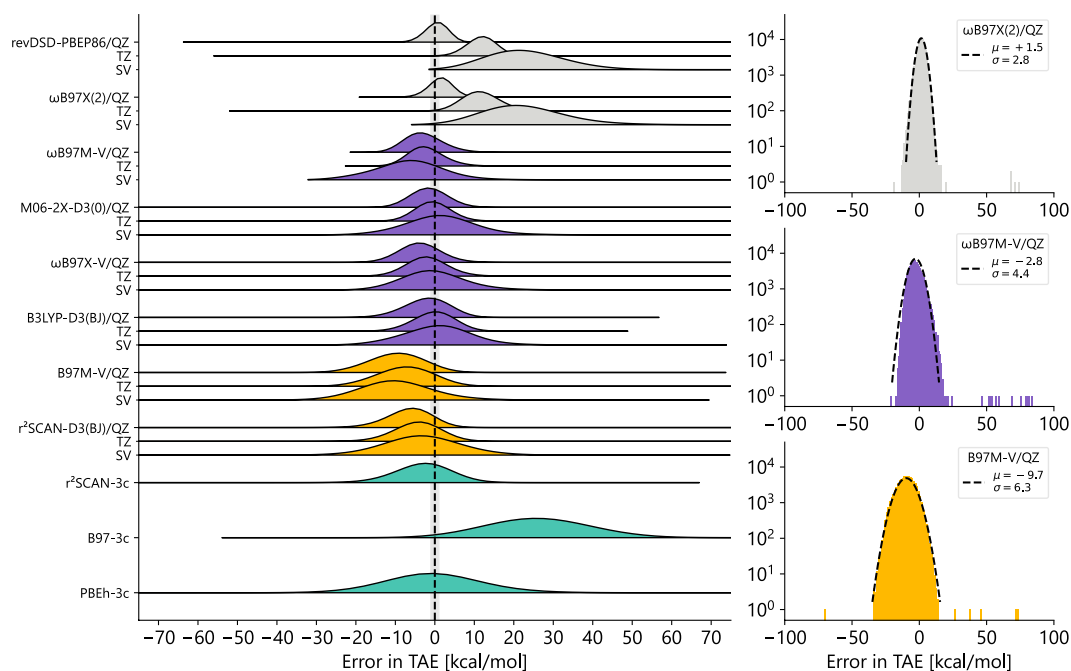


Fig. 6 Error distribution of common exchange–correlation (XC) functionals with respect to W1-F12 reference values. The error distribution of each functional is shown in def2-QZVP, def2-TZVP, and def2-SVP from top to bottom, except for composite DFT methods. The error range of ± 1 kcal mol⁻¹ is highlighted as light gray area. Jacob's ladder rungs are denoted with color. For selected functionals the distribution is replotted in log scale with a fitted normal distribution.

highly accurate experimental and theoretical TAEs. For example, for the 137 TAEs in the W4-11 database, which provides reference values approximating the relativistic, all-electron FCI/CBS level of theory via W4 theory^{9,22}, W1-F12 theory attains a mean absolute deviation (MAD) of 0.51 kcal mol⁻¹ and a root-mean-square deviation (RMSD) of 0.74 kcal mol⁻¹.⁴³ These error statistics are reduced to MAD = 0.32 and RMSD = 0.65 kcal mol⁻¹ when post-CCSD(T) contributions are removed from the W4 reference values⁴³. It should be noted that W4 theory attains a MAD and RMSD of 0.065 and 0.085 kcal mol⁻¹, respectively, relative to highly accurate experimental TAEs obtained from the active thermochemical tables (ATcTs) thermochemical network^{9,95–97}. While the primary goal of the present work is to provide all-electron CCSD(T)/CBS TAEs for the parameterization and validation of electronic structure methods, these reference TAEs can be directly compared with experimental values by adding scalar relativistic, spin-orbit, diagonal Born-Oppenheimer corrections (DBOCs), and zero-point vibrational energies (ZPVEs)^{23–25}.

Data availability

The data is available in Zenodo under the CDLA Permissive 2.0 license at <https://doi.org/10.5281/zenodo.15387279>.

Code availability

OpenBabel⁵⁵ and *xtb*⁵⁷ are open-source software. Orca⁶¹ and Molpro^{87–89} are commercial software. The SMILES generation is specified in the Methods section and no further custom code was used.

Received: 28 July 2025; Accepted: 31 March 2026;

Published online: 25 April 2026

References

1. Sprague, M. K. & Irikura, K. K. Quantitative estimation of uncertainties from wavefunction diagnostics. *Theor. Chem. Acc.* **133**, 1544 (2014).
2. 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, <https://doi.org/10.1063/1.2039080> (2005).
3. Ruscic, B. *et al.* Introduction to active thermochemical tables: Several “key” enthalpies of formation revisited. *J. Phys. Chem. A* **108**, 9979–9997, <https://doi.org/10.1021/jp047912y> (2004).
4. Ruscic, B. *et al.* Active thermochemical tables: Thermochemistry for the 21st century. *J. Phys.: Conf. Ser.* **16**, 561, <https://doi.org/10.1088/1742-6596/16/1/078> (2005).
5. Curtiss, L. A., Raghavachari, K., Redfern, P. C. & Pople, J. A. Assessment of gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **112**, 7374 (2000).
6. Curtiss, L. A., Raghavachari, K., Trucks, G. W. & Pople, J. A. Gaussian-2 theory for molecular energies of first-row and second-row compounds. *J. Chem. Phys.* **94**, 7221–7230, <https://doi.org/10.1063/1.460205> (1991).
7. Wheeler, S. E., Houk, K. N., Schleyer, P. v. R. & Allen, W. D. Hierarchy of homodesmotic reactions for thermochemistry. *J. Am. Chem. Soc.* **131**, 2547, <https://doi.org/10.1021/ja808278b> (2009).

8. Karton, A., Gruzman, D. & Martin, J. M. L. Benchmark Thermochemistry of the C_nH_{2n+2} Alkane Isomers ($n = 2-8$) and Performance of DFT and Composite Ab Initio Methods for Dispersion-Driven Isomeric Equilibria. *The Journal of Physical Chemistry A*, 113 29 8434–8447, 2009. 1520–5215. <https://doi.org/10.1021/jp904369h>.
9. Karton, A., Daon, S. & Martin, J. M. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chemical Physics Letters* 510(4–6), 165–178, <https://doi.org/10.1016/j.cplett.2011.05.007> (2011).
10. Ramabhadran, R. O. & Raghavachari, K. Theoretical thermochemistry for organic molecules: Development of the generalized connectivity-based hierarchy. *J. Chem. Theory Comput.* 7, 2094 (2011).
11. Wheeler, S. E. Homodesmotic reactions for thermochemistry. *WIREs Comput. Mol. Sci.* 2, 204, <https://doi.org/10.1002/wcms.88> (2012).
12. Ramabhadran, R. O. & Raghavachari, K. Connectivity-based hierarchy for theoretical thermochemistry: Assessment using wave function-based methods. *J. Phys. Chem. A* 116, 7531 (2012).
13. Ramabhadran, R. O. & Raghavachari, K. The successful merger of theoretical thermochemistry with fragment-based methods in quantum chemistry. *Acc. Chem. Res.* 47, 3596 (2014).
14. Yu, L.-J. & Karton, A. Assessment of theoretical procedures for a diverse set of isomerization reactions involving double-bond migration in conjugated dienes. *Chem. Phys.* 441, 166 (2014).
15. Karton, A., Schreiner, P. R. & Martin, J. M. L. Heats of formation of platonic hydrocarbon cages by means of high-level thermochemical procedures. *J. Comput. Chem.* 37, 49, <https://doi.org/10.1002/jcc.23986> (2016).
16. Chan, B., Collins, E. & Raghavachari, K. Applications of isodesmic-type reactions for computational thermochemistry. *WIREs Comput. Mol. Sci.* 11, e1501, <https://doi.org/10.1002/wcms.1501> (2021).
17. Karton, A. & Chan, B. Pah335 - a diverse database of highly accurate ccsd(t) isomerization energies of 335 polycyclic aromatic hydrocarbons. *Chem. Phys. Lett.* 824, 140544 (2023).
18. 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.* 92, 4796–4801 (1970).
19. Radom, L., Hehre, W. J. & Pople, J. A. Molecular orbital theory of the electronic structure of organic compounds. vii. systematic study of energies, conformations, and bond interactions. *J. Am. Chem. Soc.* 93, 289–300 (1971).
20. Karton, A., Tarnopolsky, A., Lamere, J.-F., Schatz, G. C. & Martin, J. M. L. Highly accurate first-principles benchmark data sets for the parameterization 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* 112, 12868–12886 (2008).
21. Karton, A., Sylvetsky, N. & Martin, J. M. L. W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *Journal of Computational Chemistry*, 38(24), 2063–2075, <https://doi.org/10.1002/jcc.24854> (2017).
22. Karton, A., E., Rabinovich, Martin, J. M. L. & Ruscic, B. W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions. *The Journal of Chemical Physics*, 125(14), 144108, <https://doi.org/10.1063/1.2348881> (2006).
23. Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *WIREs Comput. Mol. Sci.* 6, 292–310 (2016).
24. Karton, A. Quantum mechanical thermochemical predictions 100 years after the Schrödinger equation. In *Annual Reports in Computational Chemistry*, volume 18, pages 123–166. ISBN 978-0-323-99092-9, <https://doi.org/10.1016/bs.arcc.2022.09.003> (Elsevier, 2022).
25. Karton, A. Benchmark accuracy in thermochemistry, kinetics, and noncovalent interactions. In R. J. Boyd and M. Yanez, editors, *Comprehensive Computational Chemistry*, volume 1, pages 47–68. 1 edition, ISBN 9780128232569, (Elsevier, 2023).
26. Karton, A., Taylor, P. R. & Martin, J. M. L. Basis set convergence of post-CCSD contributions to molecular atomization energies. *J. Chem. Phys.* 127, 064104, <https://doi.org/10.1063/1.2755751> (2007).
27. 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).
28. Narayanan, B., Redfern, P. C., Assary, R. S. & Curtiss, L. A. Accurate quantum chemical energies for 133000 organic molecules. *Chem. Sci.* 10, 7449 (2019).
29. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 1(1), 140022, <https://doi.org/10.1038/sdata.2014.22> (2014).
30. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* 127, 124105 (2007).
31. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gn theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 1, 810 (2011).
32. Wan, W. & Karton, A. Heat of formation for c60 by means of the g4(mp2) thermochemical protocol through reactions in which c60 is broken down into corannulene and sumanene. *Chem. Phys. Lett.* 643, 34–38 (2016).
33. Karton, A., Waite, S. L. & Page, A. J. Performance of dft for c60 isomerization energies: A noticeable exception to jacob's ladder. *J. Phys. Chem. A* 123, 257–262 (2019).
34. Karton, A. Fullerenes pose a strain on hybrid density functional theory. *J. Phys. Chem. A* 126, 4709–4717 (2022).
35. Karton, A. & Chan, B. Performance of local g4(mp2) composite ab initio procedures for fullerene isomerization energies. *Comput. Theor. Chem.* 1217, 113874 (2022).
36. Karton, A., Haasler, M. & Kaupp, M. Post-CCSD(T) thermochemistry of chlorine fluorides as a challenging test case for evaluating density functional theory and composite ab initio methods. *Chem. Phys. Chem.* 26, e202400750 (2025).
37. Karton, A. High-level thermochemistry for the octasulfur ring: A converged coupled cluster perspective for a challenging second-row system. *Chem. Phys. Impact* 3, 100047 (2021).
38. Yu, L.-J., Dalem, S. G., Chan, B. & Karton, A. Benchmark study of DFT and composite methods for bond dissociation energies in argon compounds. *Chem. Phys.* 531, 110676 (2020).
39. Kroeger, A. & Karton, A. Thermochemistry of phosphorus sulfide cages: An extreme challenge for high-level ab initio methods. *Struct. Chem.* 30, 1665 (2019).
40. Sarrami, F., Yu, L.-J. & Karton, A. Thermochemistry of icosahedral closo-dicarbaboranes: A composite ab initio quantum-chemical perspective. *Can. J. Chem.* 94, 1082–1089 (2016).
41. Karton, A., Tarnopolsky, A. & Martin, J. Atomization energies of the carbon clusters C_n ($n = 2-10$) revisited by means of W4 theory as well as density functional, G_m , and CBS methods. *Mol. Phys.* 107, 977–990 (2009).
42. Martin, J. M. L. & de Oliveira, G. Towards standard methods for benchmark quality ab initio thermochemistry—w1 and w2 theory. *J. Chem. Phys.* 111, 1843–1856 (1999).
43. Karton, A. & Martin, J. M. L. Explicitly correlated wn theory: W1-f12 and w2-f12. *J. Chem. Phys.* 136, 124114 (2012).
44. Chan, B. & Radom, L. W1x-1 and w1x-2: W1-quality accuracy with an order of magnitude reduction in computational cost. *J. Chem. Theory Comput.* 8, 4259–4269 (2012).
45. Chan, B. High-level quantum chemistry reference heats of formation for a large set of c, h, n, and o species in the nist chemistry webbook and the identification and validation of reliable protocols for their rapid computation. *J. Phys. Chem. A* 126, 4981–4990 (2022).
46. Chan, B. Reliable quantum-chemistry heats of formation for an extensive set of c-, h-, n-, o-, f-, s-, cl-, br-containing molecules in the nist chemistry webbook. *J. Phys. Chem. A* 129, 3578–3586 (2025).

47. Karton, A. A highly diverse and accurate database of 3366 total atomization energies calculated at the CCSD(T)/CBS level by means of W1-F12 theory. *Chem. Phys. Lett.* **868**, 142030 (2025).
48. Khan, D., Benali, A., Kim, S. Y. H., von Rudorff, G. F. & von Lilienfeld, O. A. Quantum mechanical dataset of 836k neutral closed shell molecules with upto 5 heavy atoms from cnoofsclbr. <https://arxiv.org/abs/2405.05961> (2024).
49. Morales, M. A., McMinis, J., Clark, B. K., Kim, J. & Scuseria, G. E. Multideterminant wave functions in quantum monte carlo. *Journal of Chemical Theory and Computation*. **8**(7), 2181–2188, <https://doi.org/10.1021/ct3003404> (2012).
50. Curtiss, L. A., Jones, C., Trucks, G. W., Raghavachari, K. & Pople, J. A. Gaussian-1 theory of molecular energies for second-row compounds. *The Journal of Chemical Physics* **93**(4), 2537–2545, <https://doi.org/10.1063/1.458892> (1990).
51. Ehlert, S. *et al.* Msr-acc/tae25: 73k coupled cluster atomization energies for broad chemical space, <https://doi.org/10.5281/zenodo.15387279> (2025).
52. Luise, G. *et al.* Accurate and scalable exchange-correlation with deep learning. arXiv:2506.14665 [physics], <http://arxiv.org/abs/2506.14665> (2025).
53. Huang, B. & von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature Chemistry* **12**, 945–951, <https://doi.org/10.1038/s41557-020-0527-z> (2020).
54. Rappe, A. K. *et al.* UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **114**(25), 10024–10035, <https://doi.org/10.1021/ja00051a040> (1992).
55. O’Boyle, N. M. *et al.* Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, **3**(1), 33, <https://doi.org/10.1186/1758-2946-3-33> (2011).
56. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, **15**(3), 1652–1671, <https://doi.org/10.1021/acs.jctc.8b01176> (2019).
57. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **11**(2), e1493, <https://doi.org/10.1002/wcms.1493> (2021).
58. Grimme, S., Hansen, A., Ehlert, S. & Mewes, J.-M. r2SCAN-3c: A “Swiss army knife” composite electronic-structure method. *The Journal of Chemical Physics* **154**(6), 064103, <https://doi.org/10.1063/1.50040021> (2021).
59. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98**(7), 5648–5652, <https://doi.org/10.1063/1.464913> (1993).
60. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **98**(45), 11623–11627, <https://doi.org/10.1021/j100096a001> (1994).
61. Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Computational Molecular Science* **12**(5), e1606, <https://doi.org/10.1002/wcms.1606> (2022).
62. Spicher, S. & Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angewandte Chemie International Edition* **59**(36), 15665–15673, <https://doi.org/10.1002/anie.202004239> (2020).
63. Grimme, S., Bannwarth, C. & Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation* **13**(5), 1989–2009, <https://doi.org/10.1021/acs.jctc.7b00118> (2017).
64. Grimme, S., Brandenburg, J. G., Bannwarth, C. & Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *The Journal of Chemical Physics* **143**(5), 054107, <https://doi.org/10.1063/1.4927476> (2015).
65. Brandenburg, J. G., Bannwarth, C., Hansen, A. & Grimme, S. B97-3c: A revised low-cost variant of the B97-D density functional method. *The Journal of Chemical Physics* **148**(6), 064104, <https://doi.org/10.1063/1.5012601> (2018).
66. Mardirossian, N. & Head-Gordon, M. Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V. *The Journal of Chemical Physics* **142**(7), 074111, <https://doi.org/10.1063/1.4907719> (2015).
67. Ehlert, S. *et al.* r2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications. *The Journal of Chemical Physics* **154**(6), 061101, <https://doi.org/10.1063/5.0041008> (2021).
68. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **120**(1), 215–241, <https://doi.org/10.1007/s00214-007-0310-x> (2008).
69. Mardirossian, N. & Head-Gordon, M. ω B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics* **16**(21), 9904–9924, <https://doi.org/10.1039/C3CP54374A> (2014).
70. Mardirossian, N. & Head-Gordon, M. ω B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *The Journal of Chemical Physics* **144**(21), 214110, <https://doi.org/10.1063/1.4952647> (2016).
71. Santra, G., Sylvetsky, N. & Martin, J. M. L. Minimally Empirical Double-Hybrid Functionals Trained against the GMTKN55 Database: revDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4. *The Journal of Physical Chemistry A* **123**(24), 5129–5143, <https://doi.org/10.1021/acs.jpca.9b03157> (2019).
72. Chai, J.-D. & Head-Gordon, M. Long-range corrected double-hybrid density functionals. *The Journal of Chemical Physics* **131**(17), 174105, <https://doi.org/10.1063/1.3244209> (2009).
73. Klopper, W., Manby, F. R., Ten-no, S. & Valeev, E. F. methods in explicitly correlated molecular electronic structure theory. *Int. Rev. Phys. Chem.* **25**, 427, <https://doi.org/10.1080/01442350600799921> (2006).
74. Ten-no, S. & Noga, J. Explicitly correlated electronic structure theory from R12/F12 ansätze. *WIREs Comput. Mol. Sci.* **2**, 114–125, <https://doi.org/10.1002/wcms.68> (2012).
75. Hättig, C., Klopper, W., Köhn, A. & Tew, D. P. Explicitly correlated electrons in molecules. *Chem. Rev.* **112**, 4–74, <https://doi.org/10.1021/cr200168z> (2012).
76. Peterson, K. A., Adler, T. B. & Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al-Ar. *The Journal of Chemical Physics* **128**(8), 084102, <https://doi.org/10.1063/1.2831537> (2008).
77. Noga, J., Kedzuch, S. & Simunek, J. Second order explicitly correlated R12 theory revisited: A second quantization framework for treatment of the operators’ partitionings. *J. Chem. Phys.* **127**, 034106 (2007).
78. Knizia, G. & Werner, H.-J. Explicitly correlated RMP2 for high-spin open-shell reference states. *J. Chem. Phys.* **128**, 154103 (2008).
79. Adler, T. B., Knizia, G. & Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **127**, 221106 (2007).
80. Hill, G. J., Mazumder, S. & Peterson, K. A. Correlation consistent basis sets for molecular core-valence effects with explicitly correlated wave functions: The atoms b-ne and al-ar. *J. Chem. Phys.* **132**, 054108 (2010).
81. Yousaf, K. E. & Peterson, K. A. Optimized auxiliary basis sets for explicitly correlated methods. *J. Chem. Phys.* **129**, 184108 (2008).
82. 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).
83. 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).
84. 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).

85. Papajak, E. *et al.* Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *Journal of Chemical Theory and Computation* **7**(10), 3027–3034, <https://doi.org/10.1021/ct200106a> (2011).
86. Peterson, K. A. & Dunning Jr, T. H. 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).
87. Werner, H.-J. *et al.* *WIRES Comput Mol Sci* **2**, 242–253, <https://doi.org/10.1002/wcms.82> (2012).
88. Werner, H.-J. *et al.* *J. Chem. Phys.* **152**, 144107, <https://doi.org/10.1063/5.0005081> (2020).
89. Werner, H.-J. *et al.* MOLPRO, version 1, a package of ab initio programs. see <https://www.molpro.net> (2024).
90. Karton, A. Post-ccsd(t) contributions to total atomization energies in multireference systems. *J. Chem. Phys.* **149**, 034102 (2018).
91. Feller, D., Peterson, K. A. & Ruscic, B. Improved accuracy benchmarks of small molecules using correlation consistent basis sets. *Theor. Chem. Acc.* **133**, 1407 (2014).
92. Peterson, K. A., Feller, D. & Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: Current strategies and future challenges. *Theor. Chem. Acc.* **131**, 1079 (2012).
93. Feller, D., Peterson, K. A. & Dixon, D. A. A survey of factors contributing to accurate theoretical predictions of atomization energies and molecular structures. *J. Chem. Phys.* **129**, 204105 (2008).
94. Feller, D. & Peterson, K. A. Probing the limits of accuracy in electronic structure calculations: Is theory capable of results uniformly better than “chemical accuracy”? *J. Chem. Phys.* **126**, 114105 (2007).
95. Ruscic, B. *et al.* Introduction to active thermochemical tables: Several key examples. *The Journal of Physical Chemistry A* **108**(45), 9979–9997 (2004).
96. Ruscic, B. Active thermochemical tables. In *2005 Yearbook of Science and Technology*, pages 3–7. McGraw-Hill, New York, (2004).
97. Ruscic, B. *et al.* Active thermochemical tables: A posteriori validation of enthalpies of formation of key radicals from atomization energies. *The Journal of Physical Chemistry A* **110**(21), 6592–6601 (2006).

Acknowledgements

We thank J. Zhou (EPFL) for bringing our attention to missing information in earlier versions of the manuscript and dataset, as well as A. Pedro Fontes (UCLouvain) for reporting to us inconsistencies in earlier versions of the dataset. We acknowledge Microsoft Accelerating Foundation Models Research Program for access to Microsoft Azure cloud compute.

Author contributions

S.E. designed the structure refinement pipeline. J.H., M.S. and S.E. designed the graph sampling routines. A.K. performed the W1-F12 calculations. V.G.S. implemented the GPT-2 SMILES model. T.V., S.L., D.P.K., C.-W.H., K.J.H.G., and G.L. tested the generated data. S.E., T.V. and S.L. designed and built the software infrastructure. J.H., K.T., R.v.d.B., P.G.-G. and A.K. supervised the research. A.K., S.E., and J.H. wrote the manuscript.

Competing interests

All authors declare employment by Microsoft. All authors declare a filed patent application for a data generation system described in this article.

Additional information

Correspondence and requests for materials should be addressed to S.E., J.H. or A.K.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026