High Accuracy Heats of Formation for Alkane Oxidation: From Small to Large via the Automated CBH-ANL Method

Sarah N. Elliott,[†] Murat Keçeli,[‡] Manik K. Ghosh,[¶] Kieran P. Somers,[¶] Henry J. Curran,[¶] and Stephen J. Klippenstein[*,†]

†*Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, IL 60439, USA*

‡*Computational Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA*

¶*Combustion Chemistry Centre, School of Chemistry, Ryan Institute, MaREI, National University of Ireland, Galway, H91 TK33, Ireland*

E-mail: elliott@anl.gov

Abstract

It is generally challenging to obtain high accuracy predictions for the heat of formation for species with more than a handful of heavy atoms, such as those of importance in standard combustion mechanisms. To this end, we construct the CBH-ANL approach and illustrate that, for a set of 194 alkane oxidation species, it can be used to produce $\Delta H_f(0\ K)$s with 2σ uncertainties of $0.2-0.5$ kcal mol^{-1}. This set includes the alkanes, hydroperoxides, alkyl, peroxy, and hydroperoxy-alkyl radicals for 17 representative hydrocarbon fuels containing up to 10 heavy atoms, and with various degrees of branching in the alkane backbone. The CBH-ANL approach, automated in the QTC and AutoMech software suites, builds balanced chemical equations for the calculation of $\Delta H_f(0\ K)$, in which the reference species may be up to five heavy atoms. The high-level ANL0 and ANL1 reference $\Delta H_f(0\ K)$s are further refined for even the largest of these reference species with a novel laddering approach. We perform a comprehensive quantification of the uncertainties for both the individual reference species, the largest of which is 0.15 kcal mol^{-1}, and the propagation of those uncertainties when used in the calculation of $\Delta H_f(0\ K)$ for the 194 target species. We examine the sensitivity of the predicted $\Delta H_f(0\ K)$s to (i) electronic energies from various methods, including ωB97X-D/cc-pVTZ, B2PLYP-D3/cc-pVTZ, CCSD(T)-F12b/cc-pVDZ-F12//B2PLYP-D3/cc-pVTZ, and CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ, (ii) the zero-point vibrational energies (ZPVEs), where we consider harmonic ZPVES as well as two scaling based estimates of the anharmonic ZPVEs, all implemented for both ωB97X-D/cc-pVTZ and B2PLYP-D3/cc-pVTZ calculations, (iii) the particular CBH-ANL scheme employed, and (iv) the procedure for choosing the reference conformer for the analyses. The discussion concludes with a summary of the estimated overall uncertainty in the predictions and a validation of the predictions for the alkane subset.

1. Introduction

The introduction of focal point analyses (FPA) demonstrated that electronic energies could be computed to subchemical accuracy, approximately 0.1 kcal mol^{-1}.[1] Doing so requires high level treatment of electron correlation in wavefunction methods, large basis sets, and rigorous application of high order corrections like non-adiabatic effects,[2] scalar-relativistic effects,[3,4] and core–valence interactions. Extension to enthalpies of formation further require high accuracy treatments of vibrational zero-point energies including corrections for anharmonicities. Until recently, the large computational requirements limited such high-accuracy to small systems of small molecules.

The accessibility of high-level quantum-mechanical (QM) calculations, however, is ever-advancing. To some extent, these expensive computations are simply more feasible due to increased prevalence and accessibility of supercomputing resources. Method development has also made such calculations more feasible. For instance, parallel algorithms now take advantage of computational resources to address the polynomial scaling of electronic structure methods.[5] Furthermore, explicitly correlated coupled cluster methods, F12-CC,[6] allow for faster convergence to the complete basis set (CBS) limit. Perhaps the widest impact comes from the construction of well designed composite and extrapolation schemes.[7–11] Amplifying the advances in computational chemistry and computational resources, the development of automated workflow codes in recent years have enabled the application of high-level quantum chemistry calculations to sets of chemical species that are beyond the human input/output limit.[12–15]

Despite these advancements, achieving even chemical accuracy, 1 kcal mol^{-1}, still remains a challenge for large systems of medium and larger (6−10 heavy atom) sized molecules. Machine-learning (ML) models provide one means for predicting energies and enthalpies of larger molecules. These methods rely on molecular descriptors that can be as simple as atom counts, or based on connectivity information,[16,17] or transformations of 3D molecular structures.[18,19] The more representative, structural-descriptor sets become prohibitively large

for large molecules, and require QM calculations for geometries. Further, some input vectors require even more QM calculations, such as natural orbital analysis[20] and dipole moments.[21]

ML relies further on existing QM methods in the more successful AI/QM methods, which produce machine learning models that predict high-level QM energies using low-level QM methods. This Δ-learning approach effectively predicts the difference between two levels of theory. Paired on top of structural descriptors, for instance, X3D achieves G4-level accuracy using Δ-learning.[18] The AIQM1[22] method uses Δ-learning on an ODM2 Hamiltonian, a semiempirical QM method that deviates from the more typically used DFT methods, in combination with neural networks potentials from the Accurate NeurAl networK engINe for Molecular Energies (ANAKIN_ME),[23] to achieve a mean-absolute-deviation (MAD) of <1 kcal mol^{-1} on a database of C, H, N, and O containing molecules.

ML methods, however, fundamentally cannot overcome the uncertainties in the datasets they are trained upon. These datasets cannot be constructed from experimental values for unstable species, such as radicals. Moreover, uncertainty is added to experimental measurement when extrapolating to 0 K properties. The training set, then, creates an additional dependence of ML on QM methods. Moreover, even with transfer learning like in the AIQM1 approach, ML models require large datasets to train upon, necessitating the need for high-level but affordable QM results. As such, we look to push the accuracy of affordable quantum chemistry methods for larger species.

Rather than attempting insurmountably expensive high-level electronic structure methods to reduce systematic error in electronic energies, we consider error cancellation strategies when producing the relevant properties, which in this work is the heat of formation at 0 K, $\Delta H_f(0\text{ K})$. The most common, albeit unsophisticated, means to calculate a heat of formation is atomistically, by evaluating the electronic energy of a molecule relative to the energy of the atoms that make it up. An improved prediction of the heat of formation can be obtained by evaluating the energy of a molecule relative to smaller molecules that better describe the bonding environment of the molecule. Such an approach can lead to significant cancellation

because a given electronic structure method produces more or less consistent errors for each atom-atom interaction. The approach does however require accurate values for $\Delta H_f(0\ K)$ for the representative smaller molecules, whereas the atomization approach simply requires accurate $\Delta H_f(0\ K)$ for the individual atoms.

There are many approaches to fragmenting the target species into smaller groups, stemming from Pople's isodesmotic scheme[24–26] and expanding to capture more interactions.[27–29] Several approaches improve the fragments' description of molecular environment hierarchically, like that of Wheeler *et al.*[30] and the connectivity-based hierarchy (CBH) schemes of Raghavachari.[31–33] We focus our attention here on the CBH scheme due to its effective and systematic use of high-level data for a modest and well-defined set of smaller species. Furthermore, chemical informatics allows it to be readily automated.

There are a hierarchy of reference equations within the CBH scheme (CBHn; n = 0, 1, 2, ...) with higher levels employing fragments that more closely represent the parent species of interest (M). At the lowest level of the CBH scheme, CBH0, the fragment species are the constituent heavy atoms of M saturated with just enough hydrogens to replace each bond broken during the fragmentation. CBH1 fragments are each pair of bonded atoms in M, which allows for the preservation of the bond order. Each atom in the bond pair is again saturated with hydrogens for each bond broken in this fragmentation of M. Meanwhile, CBH2 retains each connection to adjacent atoms for each of the atom presents in M, thereby preserving bond order for each of the bonds to a given atom, and now saturates each of these adjacent atoms. As such, the rungs of each CBH ladder alternate between atom and bond-centric approaches, with each rung fragmenting about these centers while retaining increasing levels of adjacent bonds and preserving atom hybridizations. The fragment species form a chemical equation with the parent species, which is balanced with the fragment species from the previous rung. Note that there are several intricacies in this balance pertaining to branching and terminal sites. This intuitive scheme is straightforward to implement using graph representation of molecules in our AutoMol[34] molecular toolkit, which is part of our

AutoMech programming suite.[13]

A CBH scheme presents a chemical equation whose reaction energy is evaluated in terms of the electronic and zero-point energy of a target species relative to the electronic and zero-point energies of its fragment species, which we hereafter refer to as the reference species. Of course, this reaction energy may also be written in terms of the $\Delta H_f(0\ K)$s of the target and reference species. Calculation of $\Delta H_f(0\ K)$ for a parent species, then requires predetermined $\Delta H_f(0\ K)$s for its reference species, which we refer to as reference $\Delta H_f(0\ K)$s. If a parent species is a radical, its reference species set will have radicals as well. Reliable experimental data, then, will generally not be available for all reference species. For a consistent set of reference $\Delta H_f(0\ K)$, we must rely on high-level theoretical chemistry. A lower rung of CBH (*e.g.*, CBH0 or CBH1) will fragment a large parent species into many reference species. The uncertainty in the reference heats of formation will then propagate. While higher rungs of CBH (*e.g.,* CBH2 or CBH3) will create fewer fragment species, for less uncertainty propagation, these fragments will be larger (up to 5 heavy atoms for CBH2 and 8 for CBH3) and thereby cannot be calculated as rigorously. Correspondingly, the uncertainty in the references will generally grow with the order of the CBH scheme. The optimal CBH scheme then depends on a tradeoff between (i) the propagation of uncertainty for many small reference species, (ii) the larger inherent errors in the heats of formation for larger reference species, and (iii) the larger uncertainties in the calculated parent reaction energy for the lower order schemes.

Interestingly, the size of CBH1 reference species is essentially the limit of what can readily be computed with the ANL1 method that was recently introduced in a large scale study of the heat of formation for small combustion relevant species.[11] The CBH2 reference species, moreover, are essentially the limit of what can be calculated with the ANL0 method from the same study. The comparison with reference Active Thermochemical Tables (ATcT) values in[11] indicated that the ANL0 approach yields 2σ error in the heats of formation of about 0.2 kcal mol^{-1}, as long as the CCSDT(Q) correction term is not too large. The ANL1

approach is expected to have somewhat higher accuracy, although the limitation of its ATcT comparison to a much smaller data set, makes it less clear what the uncertainties are for it.

As part of this work, we further examine the uncertainties of the relevant ANL0 and ANL1 ΔH_f(0 K)s. The correlation between the CBH2 references species and the feasibility of ANL0 calculations suggests that a coupling of the CBH2 and ANL0 approaches might provide an optimal scheme for estimating heats of formation. Thus, in this work, we present the CBH-ANL method as a laddered approach that combines ANL1 values for the CBH1 reference species with ANL0 values for the CBH2 reference species. The laddering involves a refinement of the ANL0 energies against those of ANL1 in a CBH1 equation to produce a reliable set of ΔH_f(0 K) values for CBH2 reference species.

As an illustration of the power of this CBH-ANL approach we employ a composite quantum chemistry scheme to produce a large dataset of reliable ΔH_f(0 K) for alkane oxidation with up to 10 heavy atoms. This composite approach builds up to a final B2PLYP-D3/cc-pVTZ geometry and, for the electronic energy, takes advantage of the improved basis set convergence of the explicitly correlated CCSD(T)-F12/F12-cc-pVXZ methods, [X=D,T]. We also consider various approximations to an anharmonic B2PLYP-D3/cc-pVTZ ZPVE. The composite approach is automated through the QTC workflow code,[12] which has now been updated to AutoMech,[13] from starting SMILES representations of species to production of ΔH_f(0 K). An advantage of the hierarchical fragmentation scheme used in the production of ΔH_f(0 K) is that the convergence of a ΔH_f(0 K) along the CBH rungs can elucidate the remaining uncertainty in our composite scheme. We carry out a detailed analysis of this uncertainty by contrasting the rungs of CBH. Further, we compare CBH results across different levels of electronic structure theory.

As such, we examine the computational protocol on a set of 194 species, which is extended to 210 when adding CBH0-CBH2 reference species. This species set is made up of key small to medium-sized (fewer than 10 heavy atoms) alkane oxidation species. These species are chosen due to their relevance to both combustion and atmospheric chemistry.

Further, they constitute a variety of branching and radical substitution patterns, which in turn present various long range interactions. More generally, the methodology described here is directly applicable to the chemistry of biofuels, sustainable aviation fuels, and other problems of relevance to combustion in a sustainable world. Future investigations may use these structures to isolate specific molecular groups and features and elucidate improved ML models as well as the more traditional group additivity methods.

2. Computational Methodology

The chosen alkane oxidation species set, termed $\mathbf{Set}_{\text{target}}$, is compromised of 194 target species that have up to 10 heavy atoms. We additionally carry out electronic structure calculations on 14 smaller reference species. Together, these are 17 alkanes (**RH**), the largest of which is 2,2,4,4-tetramethylpentane, 45 alkyl radicals ($\mathbf{\dot{R}}$), 28 hydroperoxides ($\mathbf{RO_2H}$), 28 peroxy radicals ($\mathbf{R\dot{O}_2}$), and 76 hydroperoxyalkyl radicals ($\mathbf{\dot{Q}OOH}$). The numerous computations and transformations required for a composite approach to achieve high-accuracy heats of formation for each of these necessitates automation – which we carry out through the QTC code.[12] This python code is the prototype to the open-source AutoMech python suite,[13] and is likewise a powerful workflow code. QTC enables a user to input a list of molecules, described by SMILES strings with multiplicities, alongside a list of methods. It then automates many types of QM computations – including various approaches of geometry optimization, frequency analysis, and single point energy – through interfaces to the EStoKTP code[14] and its own calls to electronic structure codes.[35–38] Subsequently, QTC accesses its database of stored QM data and transforms it to $\Delta \mathrm{H}_f(0\ \mathrm{K})$ through its heat of formation module. The composite approach designed for this work advances through several optimization routines and auxiliary computations, which are outlined in Figure 1. Each routine has a plethora of options, some of which are detailed in the following sections.

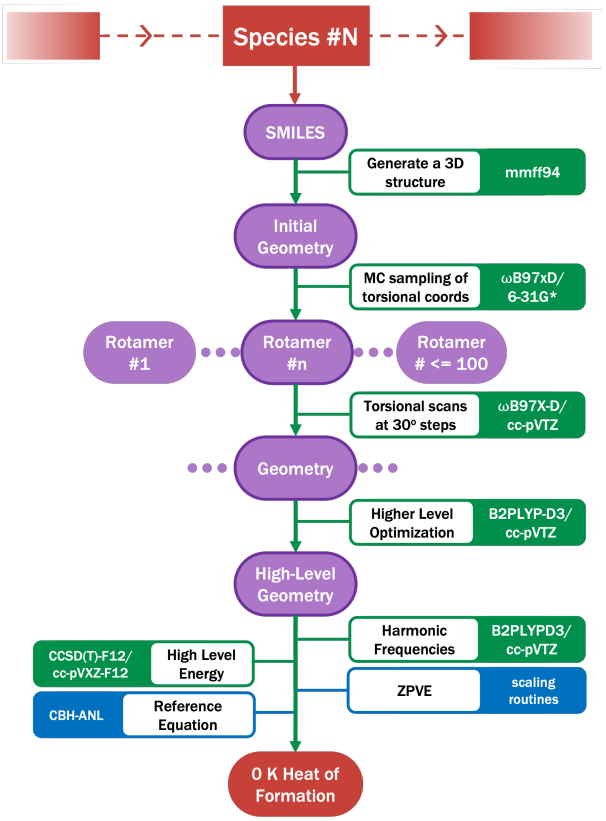

Figure 1: The composite approach, carried out through the QTC software suite, to calculate heats of formation and electronic properties for the 194 species in $\mathbf{Set}_{\text{target}}$, as well as for the species in the reference sets.

2.1 Conformer Sampling

Starting from each SMILES string, QTC calls OpenBabel[39] to generate an initial geometry with the Merck molecular force field (mmff94).[40] QTC identifies key characteristics, *e.g.*, the number of torsional angles, based on this initial geometry. Through parallel calls to the EStoKTP code,[14] it then generates geometries for additional configurations by doing a Monte Carlo sampling over the internal coordinates. EStokTP subsequently optimizes each geometry with a user-defined level of theory. This study uses ωB97X-D/6-31G* in Gaussian09.[41–43] The ωB97X-D functional is a well-motivated range-separated functional that includes empirical dispersion corrections. Each successful optimization is a local minimum on the ωB97X-D/6-31G* potential energy surface (PES). We identify the expected global minimum conformer on this PES by taking an ample number of samples, the lesser

of 100 and $6 + 2 \times 3^N$, where N is the number of non-methyl torsional angles.

2.2 Torsional Scan

In our companion study,[44] one-dimensional torsional scans are needed for the generation of partition functions. These torsional scans provide further information about the minimum energy geometries that is useful for the present analysis. As such, the next module in QTC calls EStokTP for the lowest energy conformer of each of the 210 species to run one-dimensional scans along each of the torsional angles in 30° increments. At each point on the torsional scan, the code builds an initial geometry based on the final geometry of the previous point, updating the torsional angle. It then runs a geometry optimization on the initial geometry, freezing the torsional angle that is currently being scanned. We specified these to run at ωB97X-D/cc-pVTZ[45] in Gaussian09. In several instances, the torsional scan produced a negative potential, meaning the code found a lower energy conformation. For these cases, the saved QTC conformer was updated to correspond to the lower energy ωB97X-D/cc-pVTZ rotamer. QTC then called EStoKTP once more to rerun the torsional scans about the new geometry with ωB97X-D/cc-pVTZ in Gaussian09. Note that, for the majority of these exceptions, the Monte Carlo sampling had acutally discovered this conformer, but it was not the global minimum on the ωB97X-D/6-31G* PES.

2.3 High-Level Geometry

The workflow arrives at a final geometry for each species by taking the lowest electronic energy ωB97X-D/6-31G* rotamer and optimizing it at B2PLYP-D3/cc-pVTZ[46] in Gaussian09. The double-hybrid method B2PLYP-D3/cc-pVTZ method provides CCSD(T) quality geometries and frequencies at a small fraction of the cost.[47] The next module carries out a B2PLYP-D3/cc-pVTZ vibrational frequency analysis with analytical second derivatives in Gaussian09. The harmonic frequencies confirm that the geometry is at a local minimum on the PES. In the discussion we will describe the effectiveness of using the ωB97X-D/6-31G*

minimum electronic energy rotamer for the production of high-level ΔH_f(0 K)s.

2.4 High-Level Energy

The QTC program carried out single point energy computations for each of the final geometries in Molpro.[35] It ran explicitly correlated coupled cluster singles, doubles, and perturbative triples CCSD(T)-F12b[6] computations with the specially optimized correlation consistent F12 double zeta basis set cc-pVDZ-F12. For all species with fewer than 9 heavy atoms, and for several 9 heavy atom species, the single point was also run with the triple zeta basis set to achieve CCSD(T)-F12b/cc-pVTZ-F12 energies.

2.5 ZPVE

The heat of formation at 0 Kelvin, ΔH_f(0 K), of a species is the sum of its electronic energy and zero-point vibrational energies (ZPVE) relative to the corresponding sum for its constituent elements in their standard state. Half the sum of the B2PLYP-D3/cc-pVTZ harmonic vibrational frequencies, which are produced during the composite approach, is the harmonic ZPVE (E_{harm}). Neglecting anharmonic effects, however, can cause significant errors to properties even at zero Kelvin, proportionate to the corresponding error in ZPVE. For kinetic properties, the anharmonicity in the zero point vibrational energy (ZPVE) is often canceled out between transition states and reactants. Isolated species, however, do not benefit from this error cancellation, reinforcing the importance of anharmonicity to thermochemical properties. We also note that the CBH schemes are an alternative way to introduce a chemical equation capable of balancing anharmonic effects, which we explore in Section 3.3−4. Here we examine the effects for individual species.

2.5.1 Scaled-Frequency ZPVE

Within second order vibrational perturbation theory (VPT2) the anharmonic ZPVE, E_0, is written as[48]

$$E_0 = G_0 + \frac{1}{2}\sum_i \omega_i + \frac{1}{4}\sum_{i \leq j} \chi_{ij} \tag{1}$$

where χ_{ij} are the anharmonicity constants and the G_0 term is assumed negligible in comparison to the scaling approximation for estimating the effect of anharmonicity. The central term of the right-hand-side (RHS) is equivalent to a harmonic ZPVE, with ω_i as the harmonic frequency of the ith vibrational mode. Unfortunately, it is challenging to implement VPT2, which requires quartic force fields, for systems as large as those considered here. Thus, we consider two approximate schemes for estimating the anharmonic contribution to the ZPVE.

First we consider an approximate scheme based on a scaling of the frequencies designed to roughly reproduce the corresponding anharmonic frequencies. In companion work, we developed a frequency dependent scaling factor to estimate B2PLYP-D3/cc-pVTZ fundamental frequencies, ν, from the B2PLYP-D3/cc-pVTZ harmonic frequencies, ω, with MAD of 0.51%.[44] This scale factor, s, scales each harmonic frequency, ω, according to the relation $\mathrm{s}(\omega) = \mathrm{a} - (\mathrm{b} * \omega^{\mathrm{c}})$, where a=1.045, b=0.00851, and c=0.292 and the scaled frequency $\omega_s = s(\omega)\omega$. The scaling factor was intended to mitigate the error in the partition function that was introduced by taking the harmonic approximation. But, we can also try to use it to estimate the effect of anharmonicities on the ZPVE. In particular, we can estimate anharmonic constants by relating the scaled frequencies, ω_s, to the equation for the fundamental frequency, ν, from VPT2.

$$\omega_{s,i} \approx \nu_i = \omega_i + 2\chi_{i,i} + \frac{1}{2}\sum_{i \neq j} \chi_{i,j} \tag{2}$$

Assuming that the off-diagonal elements of the anharmonic constant matrix, which constitute the third term of Equation 2, are negligible, the remaining anharmonic constants are approximately half the difference of the scaled and harmonic frequencies. Substituting this relation into the third term of Equation 1, under the additional assumption that G_0 is

negligible, yields a scaled-frequency based expression for the ZPVE (sf-ZPVE):

$$\mathrm{E_{sf}} = \sum_i \left(\frac{\omega_i}{2} + \frac{\omega_{\mathrm{s},i} - \omega_i}{8} \right) \tag{3}$$

Table S1.1 gives the B2PLYP-D3/cc-pVTZ harmonic and scaled zero-point vibrational energies for the species of $\mathbf{Set}_{\mathrm{target}}$, which we determine with Equation 3. On average, the ZPVE is scaled by 3.64%. Consequently the ZPVE diminishes by up to 5.68 kcal mol^{-1}, which is for our largest alkane 2,2,4,4-tetramethylpentane. Aside, the coefficients in Equation 3, which are 3/8 and 1/8 for ω and ω_s, allow for the scaled ZPVE to be seen as a weighting of 3/4 the harmonic ZPVE and 1/4 of a fundamental ZPVE. Perdew and coworkers[49] derive a comparable equation to determine an anharmonic ZPVE from harmonic and fundamental frequencies. They suggest updating these weights to 5/8 of the harmonic ZPVE and 3/8 of the fundamental ZPVE to empirically account for some contribution from the off-diagonal elements of the anharmonic constant matrix. To analyze the sensitivity of $\mathrm{E_{sf}}$ to electronic structure method, we also consider the scaling of ωB97X-D/cc-pVTZ harmonic frequencies. Using the set of anharmonic B2PLYP-D3/cc-pVTZ frequencies built for companion work as well as the minimization condition, we optimize $\mathrm{s}(\omega) = \mathrm{c} - (\mathrm{d} * \omega^{\mathrm{e}})$ to obtain c=1.0535, d=0.01186, and e=0.2617. We find that the max difference between the B2PLYP-D3/cc-pVTZ and ωB97X-D/cc-pVTZ based $\mathrm{E_{sf}}$ values is 0.80 kcal mol^{-1}, which is for $\mathrm{HOOC(CH_3)_2CH_2C(CH_3)_3}$. On trend, B2PLYP-D3/cc-pVTZ for this species results in the smaller $\mathrm{E_{sf}}$. The mean average deviation (MAD) between the $\mathrm{E_{sf}}$ for the B2PLYP-D3/cc-pVTZ and ωB97X-D/cc-pVTZ methods is 0.29 kcal mol^{-1}.

2.5.2 Directly-Scaled ZPVE

An alternative approach is to directly scale the ZPVE (ds-ZPVE) to approximate the anharmonic ZPVE. Here,

$$\mathrm{E_{ds}} = s_{\mathrm{ds}} \mathrm{E_{harm}} = s_{\mathrm{ds}} \left(\frac{1}{2} \sum_i \omega_i \right) \tag{4}$$

where the right-hand-side is a static scaling factor, s_{ds}, and the harmonic ZPVE, $\mathrm{E_{harm}}$. The B2PLYP-D3/cc-pVTZ scaling factor, s_{ds}, is determined by fitting scaled $\mathrm{E_{harm}}$(B2PLYP-D3/cc-pVTZ) to VPT2 determined anharmonic $\mathrm{E_0}$(B2PLYP-D3/cc-pVTZ) for 45 C, H, and O containing species with fewer than 5 heavy atoms. The optimization minimizes the standard deviation between $\mathrm{E_{ds}}$ and $\mathrm{E_0}$. This results in $s_{\mathrm{ds}} = 0.9863$, with a MAD of 0.06 kcal mol^{-1} and a 2σ error of 0.17 kcal mol^{-1}. Martin and coworkers[50] determined scaling factors on an array of double-hybrid functionals with various basis sets. For the most comparable B2PLYP/cc-pV(T+d)Z method, they fit a scaling factor of 0.9822, with a 0.06 kcal mol^{-1} RMSD for a set of 20 experimental ZPVEs.

With two approaches to scaling the ZPVE, which produce $\mathrm{E_{sf}}$ and $\mathrm{E_{ds}}$, we consider both accuracy, through the MAD, and precision, through the root-mean-square deviation (RMSD), from the anharmonic B2PLYP-D3/cc-pVTZ ZPVE, $\mathrm{E_0}$. Figure 2 displays the difference of the two scaled ZPVEs from $\mathrm{E_0}$ ($\Delta\mathrm{E_{ds}} = \mathrm{E_{ds}} - \mathrm{E_0}$ and $\Delta\mathrm{E_{sf}} = \mathrm{E_{sf}} - \mathrm{E_0}$) for the 45 species species used in the fit. They have MADs of 0.06 kcal mol^{-1} and 0.20 kcal mol^{-1}, respectively, making the $\mathrm{E_{ds}}$ a factor of three more consistent than $\mathrm{E_{sf}}$. The $\Delta\mathrm{E_{sf}}$, further, scales with the size of $\mathrm{E_0}$ in Figure 2. The observation that species with greater number of modes have greater failure suggests that off-diagonal terms on the anharmonic constant matrix are non-negligible, contradicting the key assumption of the scaled-frequency ZPVE approach.

The ωB97X-D/cc-pVTZ method produces consistent trends with the B2PLYP-D3/cc-pVTZ method. For such a comparison, we fit ωB97X-D/cc-pVTZ ZPVEs on the original 45 B2PLYP-D3/cc-pVTZ $\mathrm{E_0}$ to produce a $s_{\mathrm{ds}} = 0.9864$. For ωB97X-D/cc-pVTZ, $\Delta\mathrm{E_{ds}}$ and $\Delta\mathrm{E_{sf}}$ have MADs of 0.14 and 0.30 kcal mol^{-1}, respectively. The more expensive, B2PLYP-D3/cc-pVTZ, method is only 0.1 kcal mol^{-1} closer to $\mathrm{E_0}$, on average, for both scaling approaches. The directly-scaled approach, however, remains much closer to $\mathrm{E_0}$ than the scaled-frequency approach, even for ωB97X-D/cc-pVTZ. With these considerations in mind, we select $\mathrm{E_{ds}}$(B2PLYP-D3/cc-pVTZ), which differs by an average 0.52% from the

anharmonic B2PLYP-D3/cc-pVTZ ZPVE, for the heat of formation computation.

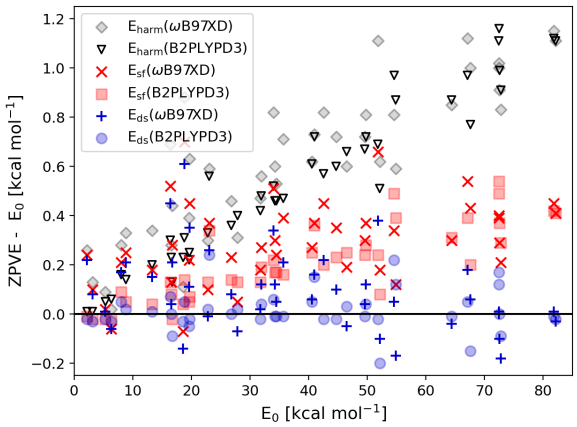

Figure 2: The difference between scaled zero-point vibrational energy (ZPVE) and anharmonic ZPVE for 45 species where the scaling approach is, in red-squares: scaled-frequency (SF) B2PLYP-D3/cc-pVTZ, in blue-circles: directly-scaled (DS) B2PLYP-D3/cc-pVTZ, in red x marks: SF-ωB97X-D/cc-pVTZ, and in blue + marks: DS-ωB97X-D/cc-pVTZ. The harmonic ZPVE is shown in black triangles for the B2PLYP-D3/cc-pVTZ method.

2.6 CBH-ANL ΔH_f(0 K)

By definition, the heat of formation of a species represents its electronic energy and ZPVE relative to its constituent elements in their standard form. In practice, however, it is calculated by evaluating its electronic energy and ZPVE relative to those from some specific balanced chemical equation, with the remaining components of the chemical equation consisting of species with known heats of formation. Well designed chemical equations provide significant error cancellation in the relative energy computation. We here describe the coupling of the CBH scheme for carefully tailoring the reference species with a rigorous ANL based determination of reliable reference energies for each of them.

2.6.1 Reference sets

The CBH-ANL approach builds chemical equations for the heats of formation according to the Connectivity Based Hierarchy (CBH) of Raghavachari.[31–33] It builds from the CBH0 scheme, on to the CBH1 scheme, and finally the CBH2 scheme. The advantage of this

approach is its hierarchical nature, or laddering, which facilitates extension of high-level small-molecule data to large species. The ladders alternate between atom centric and bond centric fragmentation approaches, increasingly preserving the surrounding environment of each atom or bond and retaining its hybridization. Take for example the ethyl peroxy radical, where CBH[x] is balanced with components from the RHS of CBH[x-1]:

$$\begin{aligned}
\text{CBH0}: & \ \mathrm{CH_3CH_2O\dot{O}} + 3\,\mathrm{H_2} \\
& = 2\,\mathrm{CH_4} + \mathrm{H_2O} + \mathrm{H\dot{O}} \\
\text{CBH1}: & \ \mathrm{CH_3CH_2O\dot{O}} + \mathrm{CH_4} + \mathrm{H_2O} \\
& = \mathrm{CH_3CH_3} + \mathrm{CH_3OH} + \mathrm{H\dot{O}_2} \\
\text{CBH2}: & \ \mathrm{CH_3CH_2O\dot{O}} + \mathrm{CH_3OH} \\
& = \mathrm{CH_3CH_2OH} + \mathrm{CH_3O\dot{O}}
\end{aligned} \tag{5}$$

The chemical equation for CBH1, here, is able to capture the peroxy functional group that the CBH0 equation does not. If a specific electronic structure method is particularly erroneous on a vibration of the peroxy stretch, then the CBH1 equation will be able to cancel the error in a larger peroxy species with that in $\mathrm{H\dot{O}_2}$. The CBH2 equation improves on the CBH1 result by also capturing the electron withdrawing effects of a carbon backbone. CBH2 fragments are sufficiently large, with up to five heavy atoms, to have medium range effects, such as induction. In this work we do not investigate higher rungs than CBH2 because the fragment size for these becomes too expensive for the ANL0 style high-level reference energies. We do, however, evaluate the convergence of CBH0-CBH2 results in the discussion section, to predict the shift that might be expected from a CBH3 equation.

Tables S2.1−5 contain the chemical equation coefficients for the $\mathbf{Set}_{\mathrm{target}}$ molecules fragmented according to each of the CBHx [x=0, 1, 2], schemes, termed $\mathrm{Set_{CBH0}}$, $\mathrm{Set_{CBH1}}$, and $\mathrm{Set_{CBH2}}$, and two simple schemes: $\mathrm{Set_{H_2O}}$, which contains only $\mathrm{H_2O}$, $\mathrm{CH_4}$, and $\mathrm{H_2}$, and $\mathrm{Set_{O_2}}$, which contains only $\mathrm{O_2}$, $\mathrm{CH_4}$, and $\mathrm{H_2}$. The latter two sets were explored in the earlier ANL work as a simplistic approach for improving upon atomic reference species.[11]

2.6.2 Reference energy computation

The accuracy of the heat of formation of the parent species is, inherently, reliant on the heats of formation of the reference species. Moreover, the uncertainty in each reference species heat of formation additively propagates according to its coefficient for a target species, *i.e.*, the number of times that reference species is a fragment of the target species. Uncertainty propagation, then, scales with the size of the parent molecule, because they will have more fragments. As part of this work, we aim to mitigate and interpret the uncertainties in the the heats of formation of the reference species as we build up a database for the present and future CBH calculations.

The heats of formation for the reference molecules that are present in the CBH0 reaction equations [CH_4, H_2, H_2O, $H\dot{O}$, $\dot{C}H_3$] are known to within 0.02 kcal mol^{-1}.[51] As a result, the uncertainty propagation is typically negligible for CBH0 calculations compared to the uncertainty in the electronic structure based evaluation of the CBH0 reaction enthalpy. Note that the error cancellation in the CBH0 reaction enthalpy calculation, however, pales in comparison to that for the higher rungs of CBH. The art of reference species selection comes in choosing fragments that are sufficiently large and chemically similar to achieve effective error cancellation in the computation of the reaction enthalpy, but that are small enough and/or well known enough to not add large, propagating, uncertainties from the reference heats of formation.

To maximize the error cancellation in a heat of formation, accordingly, without introducing large margins of uncertainty, we must achieve reliable heats of formation for the CBH1 and CBH2 reference molecules. The **ANL0** composite method is on the threshold of feasibility for CBH2 sized reference species.[11] Meanwhile CBH1 species can undergo the more rigorous **ANL1** calculations.[11] The ANL energies are shown in Equations 6 and 7, where the largest contribution comes from the extrapolation of the CCSD(T)/aug′-cc-pVnZ methods (n= Q, 5, or 6) towards the complete-basis-set [CBS] limit on CCSD(T)/cc-pVnZ (n=T,Q) geometries, and there are corrections for anharmonic contributions to the ZPVE, higher-

order excitations in the coupled cluster theory [CCSDT(Q), CCSDTQ(P)], core-valence interactions [Full], Douglass Kroll Hamiltonian relativistic effects [DKH],[4] diagonal Born-Oppenheimer effects [DBOC],[2] and spin-orbit effects [SO].

$$
\begin{aligned}
E_{\mathrm{ANL0}} =& E_{\mathrm{CCSD(T)/CBS(a'QZ,a'5Z)//CCSD(T)/TZ}} \\
&+ (E_{\mathrm{CCSDT(Q)/DZ}} - E_{\mathrm{CCSD(T)/DZ}}) \\
&+ (E_{\mathrm{CCSD(T,Full)/CBS(cTZ,cQZ)}} - E_{\mathrm{CCSD(T)/CBS(cTZ,cQZ)}}) \\
&+ \Delta E_{\mathrm{DKH/CCSD(T)/a'cTZ}} + \Delta E_{\mathrm{DBOC/HF/TZ}} + \Delta E_{\mathrm{SO}} \\
&+ E^{\mathrm{ZPVE,har}}_{\mathrm{CCSD(T)/TZ}} + (E^{\mathrm{ZPVE,anh}}_{\mathrm{B3LYP/TZ}} - E^{\mathrm{ZPVE,har}}_{\mathrm{B3LYP/TZ}})
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
E_{\mathrm{ANL1}} =& E_{\mathrm{CCSD(T)/CBS(a'5Z,a'6Z)//CCSD(T)/QZ}} \\
&+ (E_{\mathrm{CCSDT(Q)/TZ}} - E_{\mathrm{CCSD(T)/TZ}}) \\
&+ (E_{\mathrm{CCSDTQ(P)/DZ}} - E_{\mathrm{CCSDT(Q)/DZ}}) \\
&+ (E_{\mathrm{CCSD(T,Full)/CBS(cTZ,cQZ)}} - E_{\mathrm{CCSD(T)/CBS(cTZ,cQZ)}}) \\
&+ \Delta E_{\mathrm{DKH/CCSD(T)/acTZ}} + \Delta E_{\mathrm{DBOC/HF/TZ}} + \Delta E_{\mathrm{SO}} \\
&+ E^{\mathrm{ZPVE,har}}_{\mathrm{CCSD(T)/CBS(TZ,QZ)}} + (E^{\mathrm{ZPVE,anh}}_{\mathrm{B3LYP/TZ}} - E^{\mathrm{ZPVE,har}}_{\mathrm{B3LYP/TZ}})
\end{aligned}
\tag{7}
$$

Alongside the presentation of the ANL methods, is their application to a database of small species.[11] In it are all CBH1 reference species required for the $\mathbf{Set}_{\mathrm{target}}$ molecules, computed with the **ANL1** method, and the majority of CBH2 reference species, with **ANL0**. We carry out, as part of this work, **ANL0** calculations for several remaining species – $C(CH_3)_4$, $\dot{C}H_2C(CH_3)_3$, $\dot{C}H_2C(OH)(CH_3)_2$, and $OHC(CH_3)_3$.

2.6.3 Reference energy laddering

The ANL energies, in their original work,[11] are converted to heats of formation using H_2, CH_4, H_2O, and NH_3 as reference species. This is equivalent to the $\mathrm{Set}_{\mathrm{H_2O}}$ used in this work. We here improve upon the **ANL** heats of formation by updating their reference species. By applying a CBH0 equation to the **ANL1** heats of formation we provide better references for radical molecules. Each updated heat of formation for a species, M, is $\Delta \mathrm{H}_{f,\mathbf{ANL1'},M}$

$$\Delta H_{f,\mathbf{ANL1'},M} = \Delta \mathrm{H}_{f,\mathbf{ANL1},M} - \sum_i c_{M,i} \Delta \mathrm{H}_{f,\mathbf{ANL1},i} + \sum_i c_{M,i} \Delta \mathrm{H}_{f,\mathbf{ATcT},i} \tag{8}$$

Here, the reference energies for each fragment species, i, are heats of formation, $\Delta\ \mathrm{H}_{f,\mathbf{ATcT},i}$, from the Active Thermochemical Tables (ATcT).[51] The coefficient, $c_{M,i}$ is the number of times M breaks into fragment i in the CBH0 fragmentation. Equation 9 provides an illustrative application of Equation 8 for the ethyl radical. For the CBH0 sized species, the maximum ATcT uncertainty is 0.01 kcal mol^{-1} for the methyl radical.[51]

$$\begin{aligned}
&\Delta H_{f,\mathbf{ANL1'},\mathrm{CH_3CH_2\cdot}} = \\
&\quad \Delta \mathrm{H}_{f,\mathbf{ANL1},\mathrm{CH_3CH_2\cdot}} \\
&\quad - \left(\Delta \mathrm{H}_{f,\mathbf{ANL1},\mathrm{CH_4}} + \Delta \mathrm{H}_{f,\mathbf{ANL1},\mathrm{CH_3\cdot}} - \Delta \mathrm{H}_{f,\mathbf{ANL1},\mathrm{H_2}}\right) \\
&\quad + \left(\Delta \mathrm{H}_{f,\mathbf{ATcT},\mathrm{CH_4}} + \Delta \mathrm{H}_{f,\mathbf{ATcT},\mathrm{CH_3\cdot}} - \Delta \mathrm{H}_{f,\mathbf{ATcT},\mathrm{H_2}}\right)
\end{aligned} \tag{9}$$

Next we produce **ANL0**† heats of formation, an improvement upon the **ANL0** energies by laddering them with **ANL1'**. We do this by building CBH1 chemical equations for species that have **ANL0** energies and use **ANL1'** heats of formation as reference values.

$$\Delta H_{f,\mathbf{ANL0}^\dagger,M} = \Delta \mathrm{H}_{f,\mathbf{ANL0},M} - \sum_j c_{M,j} \Delta \mathrm{H}_{f,\mathbf{ANL0},j} + \sum_j c_{M,j} \Delta \mathrm{H}_{f,\mathrm{ANL1'},j} \tag{10}$$

The coefficient, $c_{M,j}$ is the number of times M breaks into fragment j in the CBH1 fragmentation. An example of this conversion is provided for the propyl radical in Equation 11, which is fragmented to ethane, methane, and ethyl radicals in the CBH1 scheme. In Section 3.7 we examine the effect of substituting other sources of reference energy into Equation 10 rather than **ANL1′**. First, we examine the uncertainties within the **ANL1′** and **ANL0**† approaches.

$$\begin{aligned}
&\Delta H_{f,\mathbf{ANL0}^{\dagger},\mathrm{CH_3CH_2CH_2\cdot}} = \\
&\quad \Delta \mathrm{H}_{f,\mathbf{ANL0},\mathrm{CH_3CH_2CH_2\cdot}} \\
&\quad - \left(\Delta \mathrm{H}_{f,\mathbf{ANL0},\mathrm{CH_3CH_3}} + \Delta \mathrm{H}_{f,\mathbf{ANL0},\mathrm{CH_3CH_2\cdot}} - \Delta \mathrm{H}_{f,\mathbf{ANL0},\mathrm{CH_4}}\right) \\
&\quad + \left(\Delta \mathrm{H}_{f,\mathbf{ANL1'},\mathrm{CH_3CH_3}} + \Delta \mathrm{H}_{f,\mathbf{ANL1'},\mathrm{CH_3CH_2\cdot}} - \Delta \mathrm{H}_{f,\mathbf{ANL1'},\mathrm{CH_4}}\right)
\end{aligned} \tag{11}$$

2.6.4 Reference energy uncertainty

The laddered **ANL1′** $\Delta \mathrm{H}_f(0\ \mathrm{K})$s are based off of ANL1 computations, see Equation 7. The terms of Equation 7 are provided in Table 1, and are collated from their original work.[11] They are, then, in reference to CH_4, H_2O and H_2. We can break down the individual contributions of the **ANL1** terms to the **ANL1′** $\Delta \mathrm{H}_f(0\ \mathrm{K})$. For example, methanol has a CBH0 formula of $CH_3OH + H_2 \rightleftharpoons CH_4 + H_2O$ and so its high-level (HL) correction term for **ANL′** is:

$$\begin{aligned}
&\Delta'_{\mathrm{HL,\ CH_3OH}} = \\
&\Delta_{\mathrm{HL,\ CH_3OH}} + \Delta_{\mathrm{HL,\ H_2}} - \Delta_{\mathrm{HL,\ CH_4}} - \Delta_{\mathrm{HL,\ H_2O}}
\end{aligned}$$

The **ANL1′** terms, computed in this manner, are given in Table 2. The uncertainty of the **ANL1′** heat of formation for each CBH1 reference species is evaluated in two parts: (1) the uncertainty in each term of Equation 7, and (2) the uncertainty from the **ATcT** reference heats of formation used in the CBH0 equation.

Each term of the **ANL1** and **ANL1′** equations have uncertainty (ϵ) to their calculations, excepting the empirical spin-orbit (SO) correction. Common to FPA,[10] the uncertainty in an electronic energy can be evaluated by how it converges to the CBS limit. Thereby, we estimate the uncertainty of the main component of the electronic energy as:

$$\begin{aligned}\epsilon_{\mathrm{CCSD(T)/CBS(a'nZ,\ a'(n+1)Z)}} = \\ 0.5 \times \big(\mathrm{E}_{\mathrm{CCSD(T)/CBS(a'nZ,\ a'(n+1)Z)}} \\ - \mathrm{E}_{\mathrm{CCSD(T)/a'(n+1)Z)}}\big)\end{aligned} \tag{12}$$

Similarly, the High-Level (HL) correction is a measure of how the treatment of electron excitation converges to the full configuration interaction (FCI) limit and we assign $\epsilon_{\mathrm{HL}} = 0.5\Delta_{\mathrm{HL}}$. For core-valence (CV), relativistic (Rel), and Diagonal Born-Oppenheimer Corrections (DBOC), the uncertainty is assigned to be 10% the size of the correction. Specifically, $\epsilon_{\mathrm{CV}} = 0.1\Delta_{\mathrm{CV}}$, $\epsilon_{\mathrm{Rel}} = 0.1\Delta_{\mathrm{Rel}}$, and $\epsilon_{\mathrm{DBOC}} = 0.1\Delta_{\mathrm{DBOC}}$. By contrasting CCSD(T)/CBS(cc-pVTZ,cc-pVQZ) frequencies with anharmonic correction to experimental frequencies[47] we assign an uncertainty of 1% to the harmonic ZPVE and 10% to the anharmonic correction to achieve $\epsilon_{\mathrm{ZPVE}} = 0.01E_{\mathrm{ZPVE}}$ and $\epsilon_{\mathrm{anh}} = 0.1\Delta_{\mathrm{anh}}$.

We apply the basic uncertainty propagation for additive terms to determine $\epsilon_{\mathrm{ANL1}} = \sqrt{\sum_i \epsilon_i^2}$, the 11th column of Table 2, where ϵ_i are each of the uncertainties corresponding to the terms of the ANL1 equation, discussed in the previous paragraph, and listed in columns 2−10 of Table 2 (*e.g.*, $\epsilon_{\mathrm{HL}} = 0.1\Delta'_{\mathrm{HL}}$). The remaining uncertainty to $\Delta\mathrm{H}_{f,\mathrm{ANL1'}}$ stems from uncertainties in the ATcT reference $\Delta\mathrm{H}_f(0\ \mathrm{K})$, which are listed as ϵ_{ATcT} in Table 1. The $\epsilon'_{\mathrm{ATcT}}$ is the propagated uncertainty of the ϵ_{ATcT} for all of the reference species applied to a $\mathbf{Set}_{\mathrm{target}}$ species through the CBH0 equation. An example is for methanol, where $\epsilon'_{\mathrm{ATcT,CH_3OH}} = \sqrt{\epsilon^2_{\mathrm{ATcT,CH_4}} + \epsilon^2_{\mathrm{ATcT,H_2O}} + \epsilon^2_{\mathrm{ATcT,H_2}}}$. The final ANL1′ uncertainty for each species is $\epsilon_{\mathrm{ANL1'}} = \sqrt{\epsilon^2_{\mathrm{ANL1}} + {\epsilon'_{\mathrm{ATcT}}}^2}$. Both ϵ_{ATcT} and $\epsilon_{\mathrm{ANL1'}}$ are displayed in Table 2.

Ancillary to ANL1′, we are able to compute ANL0† $\Delta\mathrm{H}_f(0\ \mathrm{K})$ (see Equation 10) and its

Table 1: Terms of the ANL1 equation[a], in order of Equation 7, for CBH0 and CBH1 reference species and ATcT uncertainties (ϵ). Energies are in kcal mol^{-1}.

	CCSD(T)/aug′-cc-pVnZ[b]			Δ	Δ	Δ	Δ	Δ	ZPVE	Δ	ϵ
	n=5	n=6	CBS[c]	HL[d]	CV[e]	Rel[f]	DBOC[g]	SO[h]	CBS[i]	anh[j]	ATcT[k]
CBH0 reference species											
H_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\dot{C}H_3$	41.60	41.66	41.72	−0.04	0.18	−0.02	0.03	0.00	−6.29	0.23	0.01
CH_4	−15.91	−15.91	−15.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
$\dot{O}H$	13.49	13.62	13.75	−0.18	0.23	−0.15	0.09	−0.11	−5.01	0.14	0.01
H_2O	−57.11	−57.11	−57.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
CBH1 reference species											
$CH_3\dot{C}H_2$	40.45	40.44	40.42	−0.01	0.21	−0.01	0.02	0.00	−9.53	0.19	
CH_3CH_3	−13.60	−13.66	−13.72	0.04	0.11	0.01	0.00	0.00	−3.10	0.13	
CH_3OH	−42.74	−42.80	−42.87	0.06	0.22	0.01	0.04	0.00	−3.01	0.10	
O_2	11.88	11.98	12.08	−0.05	0.52	−0.35	0.12	0.00	−12.15	0.30	
$H\dot{O}_2$	12.05	12.06	12.07	−0.11	0.50	−0.26	0.16	0.00	−8.64	0.19	
H_2O_2	−26.87	−26.92	−26.97	0.03	0.39	−0.16	0.07	0.00	−4.04	0.09	

[a] Values are as reported in original work, where CH_4, H_2O and H_2 are the reference species.[11]
[b] Calculations use a CCSD(T)/cc-pVQZ geometry
[c] Extrapolation: $E_{\text{CCSD(T)/CBS}} = 2.038E_{\text{CCSD(T)/aug-cc-pV6Z}} - 1.038E_{\text{CCSD(T)/aug-cc-pV5Z}}$
[d] High-level (HL):
$\Delta_{\text{HL}} = E_{\text{CCSDT(Q)/cc-pVTZ}} - E_{\text{CCSD(T)/cc-pVTZ}} + E_{\text{CCSDTQ(P)/cc-pVDZ}} - E_{\text{CCSDT(Q)/cc-pVDZ}}$
[e] Core-Valence (CV): $\Delta_{\text{CV}} = E_{\text{CCSD(T,Full)/CBS(cc-pCVTZ,cc-pVCQZ)}} - E_{\text{CCSD(T)/CBS(cc-pCVTZ,cc-pCVQZ)}}$
[f] Relativistic (Rel): $\Delta_{\text{rel}} = E_{\text{DKH/CCSD(T)/aug-cc-pVCTZ}}$, where DKH is the Douglass-Kroll Hamiltonian
[g] Diagonal Born Oppenheimer Correction (DBOC): $\Delta_{\text{DBOC}} = \Delta E_{\text{DBOC/HF/cc-pVTZ}}$
[h] Spin-Orbit (SO): Δ_{SO}
[i] Zero-Point Vibrational Energy (ZPVE): $\text{ZPVE,har}_{\text{CCSD(T)/CBS(cc-pVTZ,cc-pVQZ)}}$
[j] Anharmonicity (anh): $\Delta_{\text{anh}} = E^{\text{ZPVE,anh}}_{\text{B3LYP/cc-pVTZ}} - E^{\text{ZPVE,har}}_{\text{B3LYP/cc-pVTZ}})$
[k] Values as reported in Active Thermochemical Tables.[51]

uncertainty for the largest reference species that define a CBH2 equation. This is an ANL0 heat of formation relative to the ANL1′ heats of formation in a CBH1 equation. The ANL0† uncertainty is $\epsilon_{\text{ANL0}^\dagger} = \sqrt{\epsilon^2_{\text{ANL0}} + (\epsilon^\dagger_{\text{ANL1}\prime})^2}$. Here, ϵ_{ANL0} is the propagation of uncertainties of the terms of the ANL0 equation, Equation 6. Each term is assigned uncertainty is the same manner as those from the terms of the ANL1 equation, with the exception of the ZPVE. We double the assigned ZPVE uncertainty when using the cc-pVTZ basis in comparison to the CBS extrapolation, which gives $\epsilon_{\text{ZPVE}} = 0.02\text{E}^\dagger_{\text{ZPVE}}$. Tables 3 and 4 define each term and uncertainty for ANL0†.

Final uncertainty estimates for ANL0† are given in the final column of Table 4. The largest uncertainty is for the *tert*-butyl radical, $\dot{C}(CH_3)_3$, with an uncertainty of 0.15 kcal

mol^{-1}. Fortunately, no $\mathbf{Set}_{\text{target}}$ species will have more than one *tert*-butyl in its CBH2 equation, because the radical is on the central carbon. Across our set of 194 target molecules, in fact, the *tert*-butyl radical only has a nonzero coefficient in the CBH2 equation 23 times. The species that occur most commonly in the CBH2 equations are ethane, with 254, methanol, with 130, and propane, with 108 occurrences. The former two are small enough species to have computed ANL1′ $\Delta H_f(0\text{ K})$, with uncertainties of 0.05 and 0.06 kcal mol^{-1}. Propane is calculated with ANL0† $\Delta H_f(0\text{ K})$ and has a 0.07 kcal mol^{-1} uncertainty. We here note that the ANL0† heat of formation for a CBH1 sized molecule is equivalent to its ANL1′ heat of formation. We use the ANL0† to refer to the best laddered energy for a species for the remaining discussion.

We can determine the uncertainty that the CBH equations build into the $\Delta H_f(0\text{ K})$ of each species, i, through the uncertainties in their ANL0† heats of formation. This quantity is $\epsilon_{i,\text{CBHx}}$, for each of CBH0, CBH1, and CBH2, and is calculated as follows:

$$\begin{aligned}
\epsilon_{M,\text{CBH0}} &= \sum_j |c^{\mathbf{0}}_{M,i}|(\epsilon_{i,\text{ANL0}^{\dagger}})^2 \\
\epsilon_{M,\text{CBH1}} &= \sum_j |c^{\mathbf{1}}_{M,i}|(\epsilon_{i,\text{ANL0}^{\dagger}})^2 \\
\epsilon_{M,\text{CBH2}} &= \sum_j |c^{\mathbf{2}}_{M,i}|(\epsilon_{i,\text{ANL0}^{\dagger}})^2
\end{aligned} \tag{13}$$

The coefficients, $c^{\mathbf{0}}_{M,i}$, $c^{\mathbf{1}}_{M,i}$, and $c^{\mathbf{2}}_{M,i}$, are the number of times each reference species, i, is used in the CBH0, CBH1, and CBH2 equations, respectively, for each $\mathbf{Set}_{\text{target}}$ species, M. This includes negative coefficients for the species used to balance the chemical equation. The $\epsilon_{i,\text{ANL0}^{\dagger}}$ is the ANL0† uncertainty for reference species i. The reference species CBH0, CBH1, and CBH2 coefficients for each $\mathbf{Set}_{\text{target}}$ species are listed in Table S2.1−5 and the ANL0† uncertainties for each reference species are listed in Tables 2 and 4.

Figure 3 shows $\epsilon_{M,\text{CBH0}}$, $\epsilon_{M,\text{CBH1}}$, and $\epsilon_{M,\text{CBH2}}$, for each $\mathbf{Set}_{\text{target}}$ species. Note, these are

not overall uncertainties of $\Delta H_f(0\ \mathrm{K})$, but specifically indicate the uncertainty contributed from employing CBH equations instead of calculating the $\Delta H_f(0\ \mathrm{K})$ relative to the standard state of the elements. As expected, the CBH0 equations contribute the least uncertainty to $\mathbf{Set}_{\mathrm{target}}$ species, because the $\mathrm{Set}_{\mathrm{CBH0}}$ reference species use ANL1′ energies with at most 0.13 kcal mol^{-1} uncertainty, which is for $\dot{\mathrm{O}}\mathrm{H}$ radical.

More remarkable is that the CBH2 equations contribute only modestly more uncertainty than do the CBH1. The max $\epsilon_{i,\mathrm{CBH1}}$ and $\epsilon_{i,\mathrm{CBH2}}$ uncertainties are 0.21 and 0.27 kcal mol^{-1}, respectively. This means we can successfully use the larger, more representative CBH2 reference species without meaningfully increasing our uncertainty. In part, this is because the sum of coefficients for CBH2 equations is smaller than for CBH1, because the terminal fragments often balance out in the CBH2 equation. The observation is also because the uncertainties of $\mathrm{Set}_{\mathrm{CBH2}}$ heats of formation, which are ANL0†, are only moderately greater than the $\mathrm{Set}_{\mathrm{CBH1}}$, which are ANL1′. These range from 0.07−0.15 kcal mol^{-1} in comparison to 0.05−0.18 kcal mol^{-1}. Additionally, no class of species is expected to suffer greater uncertainties than another. From this we anticipate that the error cancellation that CBH2 equations afford to the electronic energies will strongly outweigh any uncertainty they add to the final $\Delta H_f(0\ \mathrm{K})$, at least when implemented with the ANL0† references.

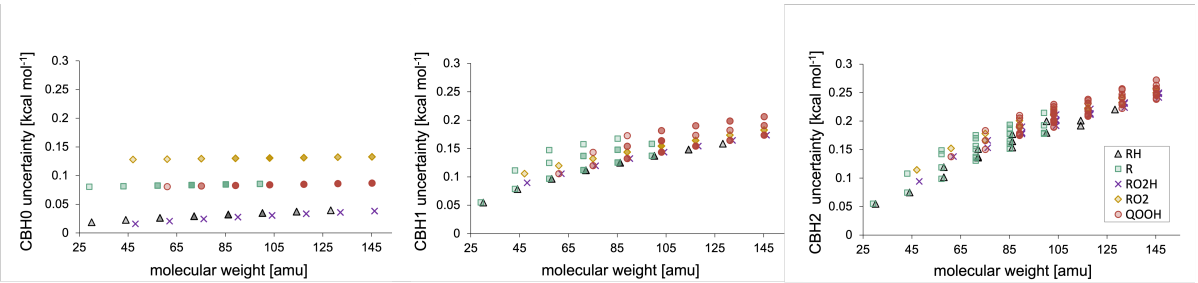

Figure 3: The uncertainty that the CBH schemes build into each $\mathbf{Set}_{\mathrm{target}}$ species plotted against the molecular weight of the $\mathbf{Set}_{\mathrm{target}}$ species for CBH0 (left), CBH1 (middle), and CBH2 (right). This uncertainty for each $\mathbf{Set}_{\mathrm{target}}$ species is the propagated ANL0† uncertainty for each reference species in the CBH equation, see Equation 13.

Table 2: Terms[a] of the ANL1′ equation[b] for CBH1 reference molecules with energies in kcal mol^{-1}.

	CCSD(T)/aug-cc-pVnZ′			Δ'	Δ'	Δ'	Δ'	Δ'	ZPVE′	Δ'	ϵ	ϵ'	ϵ
	n=5	n=6	CBS	HL	CV	Rel	DBOC	SO	CBS	anh	ANL1[c]	ATcT[d]	ANL1′[e]
CH_3OH	−30.27	−30.21	−30.14	−0.06	−0.22	−0.01	−0.04	0.00	3.01	−0.10	0.06	0.01	0.06
$CH_3\dot{C}H_2$	−14.75	−14.68	−14.61	−0.03	−0.04	−0.01	0.01	0.00	3.24	0.04	0.05	0.02	0.06
CH_3CH_3	−18.21	−18.16	−18.10	−0.04	−0.11	−0.01	0.00	0.00	3.10	−0.13	0.05	0.02	0.05
O_2	15.09	15.25	15.42	−0.31	−0.07	0.06	0.05	−0.21	2.14	−0.02	0.18	0.01	0.18
$H\dot{O}_2$	−55.67	−55.55	−55.42	−0.07	−0.28	0.12	−0.07	−0.11	3.64	−0.05	0.09	0.01	0.09
H_2O_2	−87.34	−87.29	−87.24	−0.03	−0.39	0.16	−0.07	0.00	4.04	−0.09	0.07	0.01	0.07

[a] Terms are defined in Table 1.
[b] Values from Table 1 are balanced in a CBH0 chemical equation, which are:

$CH_3OH + H_2 = CH_4 + H_2O$ $\quad$ $CH_3CH_2 + H_2 = CH_4 + CH_3$ $\quad$ $CH_3CH_3 + H_2 = 2\,CH_4$
$O_2 + H_2 = 2\,OH$ $\quad$ $HO_2 + H_2 = H_2O + OH$ $\quad$ $H_2O_2 + H_2 = 2\,H_2O$

[c] $\epsilon_{\rm ANL1} = \sqrt{\sum_i \epsilon_i^2}$ for each of the following ϵ_i: $\epsilon_{\rm HL} = 0.5\Delta'_{\rm HL}$, $\epsilon_{\rm CV} = 0.1\Delta'_{\rm CV}$, $\epsilon_{\rm Rel} = 0.1\Delta'_{\rm Rel}$, $\epsilon_{\rm DBOC} = 0.1\Delta'_{\rm DBOC}$, $\epsilon_{\rm ZPVE} = 0.02{\rm E}'_{\rm ZPVE}$, $\epsilon_{\rm anh} = 0.1\Delta'_{\rm anh}$, and $\epsilon'_{\rm CCSD(T)/CBS(aug\text{-}cc\text{-}pV5Z,\ aug\text{-}cc\text{-}pV6Z)}$, which is defined in equation 12.
[d] $\epsilon'_{\rm ATcT} = \sqrt{\sum_j \epsilon_{{\rm ATcT},j}}$ for each uncertainty for reference species, j in the CBH0 chemical equation.
[e] $\epsilon_{\rm ANL1\prime} = \sqrt{\epsilon^2_{\rm ANL1} + {\epsilon'_{\rm ATcT}}^2}$

3. Results and Discussion

In this work we compute 0 K heats of formation with several *ab initio* methods and with several sets of reference species. The latter are identified by the prefix CBHx, where x=0,1, or 2, for the connectivity based hierarchy species, Set$_{\rm H_2O}$, for the set containing only H_2O, CH_4, and H_2 and Set$_{\rm O_2}$, for the set containing only O_2, CH_4, and H_2. The methods are assigned by levels Ln, where n=1, 2, 3, or 4.

L1: ωB97X-D/cc-pVTZ

L2: B2PLYP-D3/cc-pVTZ

L3: CCSD(T)-F12b/cc-pVDZ-F12//B2PLYP-D3/cc-pVTZ

L4: CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ

Note that L2−L4 use the same B2PLYP-D3/cc-pVTZ geometries and $E_{\rm ds}$(B2PLYP-D3/cc-pVTZ) ZPVEs, while L1 uses ωB97X-D/cc-pVTZ geometries and $E_{\rm ds}$(ωB97X-D/cc-pVTZ) ZPVEs. Comparisons between the heats of formation computed with CBHx-Ln [x=0,1,2] and [n=1,2,3,4] provide insights on accuracy and sensitivity.

Table 3: Terms of the ANL0 equation[a], in order of Equation 6, for CBH0, CBH1, and CBH2 reference species and ANL1′ uncertainties (ϵ). Energies are in kcal mol^{-1}

	CCSD(T)/aug′-cc-pVnZ[b]			Δ	Δ	Δ	Δ	Δ	ZPVE	Δ	ϵ
	n=4	n=5	CBS[c]	HL[d]	CV[e]	Rel[f]	DBOC[g]	SO[h]	TZ[i]	anh[j]	**ANL1′**[k]
CBH0 reference species											
H_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\dot{C}H_3$	41.47	41.60	41.71	0.00	0.17	−0.02	0.03	0.00	−6.28	0.23	0.08
CH_4	−15.91	−15.91	−15.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\dot{O}H$	13.15	13.48	13.74	0.11	0.22	−0.15	0.09	−0.11	−5.01	0.14	0.13
H_2O	−57.11	−57.11	−57.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CBH1 reference species											
$CH_3\dot{C}H_2$	40.48	40.45	40.43	0.01	0.22	−0.01	0.02	0.00	−9.46	0.19	0.05
CH_3CH_3	−13.45	−13.60	−13.71	0.03	0.11	0.01	0.00	0.00	−3.03	0.13	0.05
CH_3OH	−42.57	−42.73	−42.86	0.02	0.22	0.01	0.04	0.00	−2.99	0.10	0.06
O_2	11.58	11.86	12.09	−0.61	0.51	−0.35	0.12	0.00	−12.16	0.30	0.18
$H\dot{O}_2$	12.01	12.06	12.09	−0.47	0.50	−0.26	0.16	0.00	−8.65	0.19	0.09
H_2O_2	−26.77	−26.88	−26.96	−0.26	0.39	−0.16	0.07	0.00	−4.06	0.09	0.07
CBH2 reference species											
$CH_3\dot{C}HCH_3$	38.22	38.03	37.89	0.02	0.26	0.00	0.01	0.00	−13.24	0.19	
$CH_3CH_2\dot{C}H_2$	41.04	40.89	40.76	0.04	0.32	0.00	0.01	0.00	−13.13	0.24	
$CH_3CH_2CH_3$	−13.13	−13.41	−13.62	0.07	0.21	0.02	0.03	0.00	−6.85	0.25	
$\dot{C}H_2CH_2OH$	9.47	9.28	9.14	0.01	0.42	−0.01	0.17	0.00	−12.92	0.29	
CH_3CH_2OH	−45.16	−45.46	−45.70	0.06	0.29	0.01	0.03	0.00	−6.82	0.18	
$CH_3O\dot{O}$	17.70	17.53	17.40	−0.53	0.69	−0.26	0.13	0.00	−12.33	0.31	
CH_3OOH	−19.25	−19.56	−19.81	−0.28	0.60	−0.16	0.09	0.00	−8.10	0.24	
$\dot{C}(CH_3)_3$	35.15	34.83	34.59	0.05	0.31	0.00	−0.03	0.00	−17.20	0.33	
$\dot{C}H_2CH(CH_3)_2$	40.11	39.85	39.64	0.09	0.42	0.00	−0.02	0.00	−17.26	0.50	
$CH(CH_3)_3$	−14.36	−14.73	−15.03	0.13	0.29	0.02	−0.02	0.00	−11.13	0.43	
$\dot{C}H_2CH(CH_3)OH$	6.30	5.99	5.75	0.05	0.51	−0.01	0.02	0.00	−17.10	0.46	
$CH(CH_3)_2OH$	−48.24	−48.66	−48.98	0.10	0.37	0.01	0.00	0.00	−11.16	0.42	
$\dot{C}H_2C(CH_3)_3$	38.69	38.33	38.06	0.06	0.53	0.00	−0.05	0.00	−21.66	0.60	
$C(CH_3)_4$	−16.41	−16.88	−17.24	0.20	0.39	0.02	−0.04	0.00	−15.62	0.57	
$\dot{C}H_2C(CH_3)_2OH$	3.17	2.77	2.47	0.02	0.59	−0.01	−0.02	0.00	−21.74	0.55	
$C(CH_3)_3OH$	−51.89	−52.39	−52.78	0.17	0.44	0.01	−0.02	0.00	−15.75	0.46	

[a] Values are as reported in original work, where CH_4, H_2O and H_2 are the reference species.[11]

[b] Calculations use a CCSD(T)/cc-pVTZ geometry

[c] Extrapolation: $E_{\text{CCSD(T)/CBS}} = 1.779E_{\text{CCSD(T)/aug′-ccpV5Z}} - 0.779E_{\text{CCSD(T)/aug′-cc-pVQZ}}$

[d] High-level (HL): $\Delta_{\text{HL}} = E_{\text{CCSDT(Q)/cc-pVDZ}} - E_{\text{CCSD(T)/cc-pVDZ}}$

[e] Core-Valence (CV): $\Delta_{\text{CV}} = E_{\text{CCSD(T,Full)/CBS(cc-pCVTZ,cc-pVCQZ)}} - E_{\text{CCSD(T)/CBS(cc-pCVTZ,cc-pCVQZ)}}$

[f] Relativistic (Rel): $\Delta_{\text{rel}} = E_{\text{DKH/CCSD(T)/aug-cc-pVCTZ}}$, where DKH is the Douglass-Kroll Hamiltonian

[g] Diagonal Born Oppenheimer Correction (DBOC): $\Delta_{\text{DBOC}} = \Delta E_{\text{DBOC/HF/cc-pVTZ}}$

[h] Spin-Orbit (SO): Δ_{SO}

[i] Zero-Point Vibrational Energy (ZPVE): $\text{ZPVE,har}_{\text{CCSD(T)/cc-pVTZ}}$

[j] Anharmonicity (anh): $\Delta_{\text{anh}} = E^{\text{ZPVE,anh}}_{\text{B3LYP/cc-pVTZ}} - E^{\text{ZPVE,har}}_{\text{B3LYP/cc-pVTZ}})$

[k] Values as reported in Active Thermochemical Tables.[51]

Table 4: Terms[a] of the ANL0† equation[b] for CBH2 reference molecules with energies in kcal mol^{-1}.

	CCSD(T)/aug′-cc-pVnZ†			$\Delta^\dagger$	$\Delta^\dagger$	$\Delta^\dagger$	$\Delta^\dagger$	$\Delta^\dagger$	ZPVE†	$\Delta^\dagger$	ϵ	$\epsilon^\dagger$	ϵ
	n=4	n=5	CBS	HL	CV	Rel	DBOC	SO	TZ	anh	ANL0[c]	ANL1′[d]	ANL0$^{\dagger e}$
$CH_3\dot{C}HCH_3$	1.28	1.27	1.26	−0.01	0.00	0.00	0.01	0.00	0.60	−0.04	0.01	0.11	0.11
$CH_3CH_2\dot{C}H_2$	1.89	1.87	1.86	0.00	0.01	0.01	0.01	0.00	0.64	0.07	0.02	0.07	0.07
$CH_3CH_2CH_3$	2.14	2.12	2.10	−0.01	0.02	0.01	−0.02	0.00	0.78	0.01	0.02	0.07	0.07
$\dot{C}H_2CH_2OH$	4.35	4.34	4.34	0.02	0.01	0.01	−0.11	0.00	0.47	−0.01	0.02	0.08	0.08
CH_3CH_2OH	5.05	5.04	5.04	−0.01	0.04	0.01	0.02	0.00	0.80	0.05	0.02	0.08	0.08
$CH_3O\dot{O}$	8.85	8.90	8.94	0.07	0.03	0.01	0.07	0.00	0.69	−0.02	0.04	0.11	0.11
CH_3OOH	7.03	7.06	7.09	0.03	0.02	0.02	0.02	0.00	1.05	−0.05	0.03	0.09	0.09
$\dot{C}(CH_3)_3$	3.36	3.32	3.28	−0.03	−0.01	0.01	0.04	0.00	1.39	−0.23	0.04	0.14	0.15
$\dot{C}H_2CH(CH_3)_2$	5.27	5.22	5.18	−0.01	0.02	0.01	0.05	0.00	1.73	−0.07	0.04	0.09	0.10
$CH(CH_3)_3$	5.81	5.75	5.71	−0.03	0.04	0.01	0.03	0.00	2.03	−0.05	0.05	0.09	0.10
$\dot{C}H_2CH(CH_3)OH$	9.97	9.95	9.93	0.01	0.04	0.03	0.05	0.00	1.62	−0.05	0.04	0.09	0.10
$CH(CH_3)_2OH$	10.58	10.54	10.51	−0.01	0.08	0.02	0.05	0.00	2.10	−0.07	0.05	0.09	0.11
$\dot{C}H_2C(CH_3)_3$	9.16	9.04	8.96	0.05	0.02	0.03	0.08	0.00	3.10	−0.04	0.08	0.10	0.13
$C(CH_3)_4$	10.32	10.21	10.12	−0.07	0.06	0.03	0.05	0.00	3.49	−0.06	0.09	0.10	0.14
$\dot{C}H_2C(CH_3)_2OH$	15.56	15.47	15.40	0.07	0.07	0.04	0.08	0.00	3.23	−0.02	0.08	0.11	0.13
$C(CH_3)_3OH$	16.68	16.59	16.52	−0.05	0.11	0.04	0.07	0.00	3.66	0.02	0.09	0.11	0.14

[a] Terms are defined in Table 3.

[b] Values from Table 3 are balanced in a CBH1 chemical equation, which are:

$CH_3CHCH_3 + CH_3 = 2\,CH_3CH_2$
$CH_3CH_2CH_2 + CH_4 = CH_3CH_3 + CH_3CH_2$
$CH_3CH_2CH_3 + CH_4 = 2\,CH_3CH_3$
$CH_2CH_2OH + CH_4 = CH_3CH_2 + CH_3OH$
$CH_3CH_2OH + CH_4 = CH_3CH_3 + CH_3OH$
$CH_3OO + H_2O = CH_3OH + OOH$
$CH_3OOH + H_2O = CH_3OH + H_2O_2$
$C(CH_3)_3 + 2\,CH_4 = 3\,CH_3CH_2$
$CH_2CH(CH_3)_2 + 2\,CH_4 = 2\,CH_3CH_3 + CH_3CH_2$
$CH(CH_3)_3 + 2\,CH_4 = 3\,CH_3CH_3$
$CH_2CH(CH_3)OH + 2\,CH_4 = CH_3CH_2 + CH_3CH_3 + CH_3OH$
$CH(CH_3)_2OH + 2\,CH_3 = 2\,CH_3CH_3 + CH_3OH$
$CH_2C(CH_3)_3 + 3\,CH_4 = CH_3CH_2 + 3\,CH_3CH_3$
$C(CH_3)_4 + 3\,CH_4 = 4\,CH_3CH_3$
$CH_2C(CH_3)_2OH + 3\,CH_4 = CH_3CH_2 + 2\,CH_3CH_3 + CH_3OH$
$C(CH_3)_3OH + 3\,CH_4 = 3\,CH_3CH_3 + CH_3OH$

[c] $\epsilon_{\mathrm{ANL1}} = \sqrt{\sum_i \epsilon_i^2}$ for each of the following ϵ_i: $\epsilon_{\mathrm{HL}} = 0.5\Delta^\dagger_{\mathrm{HL}}$, $\epsilon_{\mathrm{CV}} = 0.1\Delta^\dagger_{\mathrm{CV}}$, $\epsilon_{\mathrm{Rel}} = 0.1\Delta^\dagger_{\mathrm{Rel}}$, $\epsilon_{\mathrm{DBOC}} = 0.1\Delta^\dagger_{\mathrm{DBOC}}$, $\epsilon_{\mathrm{ZPVE}} = 0.02\mathrm{E}^\dagger_{\mathrm{ZPVE}}$, $\epsilon_{\mathrm{anh}} = 0.1\Delta^\dagger_{\mathrm{anh}}$, and $\epsilon^\dagger_{\mathrm{CCSD(T)/CBS(aug′\text{-}cc\text{-}pV5Z,\ aug′\text{-}cc\text{-}pV6Z)}}$, which is defined in equation 12.

[d] $\epsilon^\dagger_{\mathrm{ATcT}} = \sqrt{\sum_j \epsilon_{\mathrm{ATcT},j}}$ for each uncertainty for reference species, j in the CBH0 chemical equation.

[e] $\epsilon_{\mathrm{ANL1}\dagger} = \sqrt{\epsilon^2_{\mathrm{ANL1}} + \epsilon^{\dagger\,2}_{\mathrm{ATcT}}}$

3.1 Basis Set Sensitivity

A well represented electronic wavefunction assessed by basis set convergence has long been a key criterion for high accuracy electronic structure calculations. The explicitly correlated, F12, methods include interelectronic distance in the functional form of the wavefunction, and improve convergence to the complete basis set (CBS) limit. The difference in the heats of formation computed between L3 and L4, the double-ζ and triple-ζ F12 basis sets, provide a measure of this convergence. The convergence, in turn, is an estimate of the accuracy and the degree of CBH basis set incompleteness error (BSIE) cancellation in the reaction energy calculation. Figure 4 shows the CBHx-L3 heats of formation in reference to the CBH2-L4.

With $\text{Set}_{\text{H}_2\text{O}}$, Set_{O_2}, and CBH0, which provide the least error cancellation, the MAD between the triple-ζ and double-ζ basis sets is 0.61, 1.30, and 0.94 kcal mol^{-1}, respectively. When using the tailored chemical equations of the CBH1 and CBH2 schemes, however, the BSIE effectively vanishes. CBH1-L3 and CBH2-L3 have mean differences from CBH2-L4 of 0.13 and 0.02 kcal mol^{-1}, respectively. Furthermore, the values are within 2σ=0.40 and 0.06 kcal mol^{-1} of that mean. The successful convergence across the CBH schemes suggests negligible BSIE error in the CBH2-L4 results. Because there is no apparent size dependence to the CBH2-L3 deviation, we expect similar $2\sigma = 0.1$ kcal mol^{-1} for the molecules larger than 8 heavy atoms, for which we only compute CBH2-L[1-3] heats of formation.

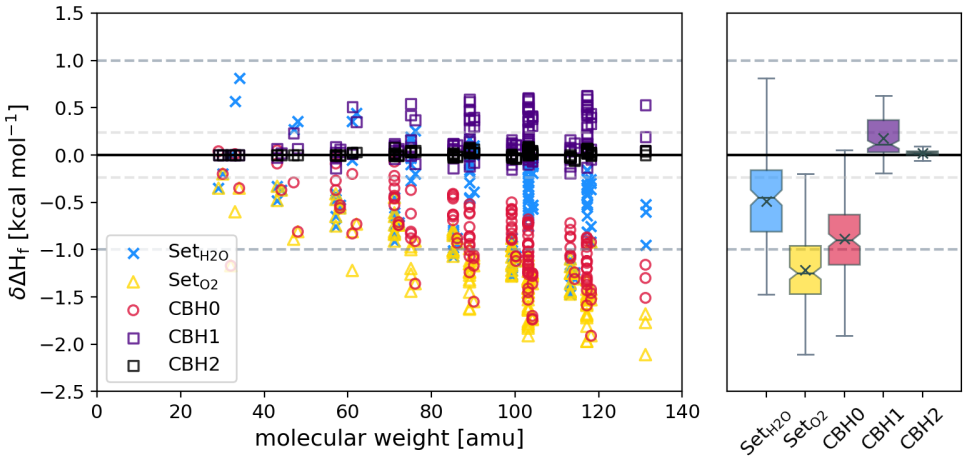

Figure 4: This plot shows the differences in the L3 heat of formations ($\delta\Delta\text{H}_f$(0 K)) for 158 species that are calculated using $\text{Set}_{\text{H}_2\text{O}}$, Set_{O_2}, Set_{CBH0}, Set_{CBH1}, and Set_{CBH2} chemical equations from the CBH2-L4 The blue x marks show $\Delta\text{H}_f(0\text{ K})_{\text{SetH2O-L3}}$ -$\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L4}}$ for each molecule. The yellow triangles show $\Delta\text{H}_f(0\text{ K})_{\text{SetO2-L3}}$ -$\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L4}}$. The red circles are $\Delta\text{H}_f(0\text{ K})_{\text{CBH0-L3}}$ -$\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L4}}$, the purple squares are $\Delta\text{H}_f(0\text{ K})_{\text{CBH1-L3}}$ -$\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L4}}$, and the black squares are $\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L3}}$ -$\Delta\text{H}_f(0\text{ K})_{\text{CBH2-L4}}$. A box and whisker plot is to the right of the individual data. The x in the box represents the mean of the differences, and the median is shown by a line in the center of each box. The perimeters of the box represent the first and third quartiles. The whiskers (vertical lines) extend up and down to the maximum and minimum differences. Horizontal, gray dashed lines help visualize 1 kcal mol^{-1} and 1 kJ mol^{-1} deviations.

3.2 *ab initio* Method Sensitivity

Even our most robust method, L4, does not account for errors due to higher order electronic excitations, core electron correlation, scalar-relativistic effects, or adiabatic effects. The reference species, however, do have auxilliary corrections for these errors. When a chemical equation fully captures the target environment, these high-level corrections from the reference species are effectively applied to the $\mathbf{Set}_{\text{target}}$ species. The success of a given chemical equation in providing such correction is demonstrated by the variance of the heats of formation it produces with varying *ab initio* methods. Take for instance Figure 5(a). Here we see the difference between CBH0-L[1-3] from CBH2-L4 ΔH_f(0 K)s for 158 $\mathbf{Set}_{\text{target}}$ species for which we have L4 computations. CBH0-L1 ΔH_f(0 K)s have deviations of 2.30 ± 7.02 kcal mol^{-1} (i.e., mean values $\pm$ 2σ from that mean) and CBH0-L2 ΔH_f(0 K)s have deviations of 0.16 ± 7.42 kcal mol^{-1} from CBH2-L4. While on average, CBH0-L2 agrees very well with the higher level theory, the variance makes the method quite unreliable.

Figure 5(b), showing CBH1 ΔH_f(0 K)s, is in stark contrast. Here, the low-level DFT methods (L1 and L2) agree with the large basis, coupled cluster method (CBH2-L4) heats of formation to 2.75 ± 2.32 kcal mol^{-1} and 1.19 ± 1.20 kcal mol^{-1}. Their mean absolute deviations are cut by a factor of 1.2 and 2.7, respectively, when compared to the CBH0. At CBH1 the double hybrid L2 method more clearly outperforms the cheapest L1 method. In Figure 5(c) we see that the CBH2-L1 and CBH2-L2 deviations from CBH2-L4 are reduced to -0.19 ± 0.74 kcal mol^{-1} and -0.04 ± 0.38 kcal mol^{-1}, respectively. Surprisingly, with MAD of 0.28 and 0.14, kcal mol^{-1} the DFT CBH2-L1 and CBH2-L2 results are in better agreement with the CBH2-L4 results than are the coupled cluster CBH0-L3 ones. From CBH0 to CBH2, then, the MAD is reduced by a factor of 11.5 and 22.9 for the L1 and L2 methods. Even more, CBH2-L3 has an improvement by a factor of 31.3 over CBH0-L3.

We expect that any improvement to the CCSD(T) method by the inclusion of higher-order corrections would be far less significant than a shift between the unrelated doubled-hybrid B2PLYP-D3 method and the wavefunction-based CCSD(T) method, which have a

2σ deviation of only 0.4 kcal mol^{-1}. Alternatively, we can consider the sizes of higher-order corrections from Table 4. We have already demonstrated near convergence to the CBS limit. The uncertainty in the CCSD(T) method can be estimated by the size of the remaining corrections to the electronic energy, (*i.e.*, $\Delta^{\dagger}_{\rm HL}$, $\Delta^{\dagger}_{\rm CV}$, $\Delta^{\dagger}_{\rm Rel}$, and $\Delta^{\dagger}_{\rm DBOC}$). The largest $\mathbf{Set}_{\rm target}$ species is twice the size of the largest species in Table 4. As such, we double the terms of the largest species in Table 4, even though each term scales sub-linearly with species size. Propagating the resulting $\Delta^{\dagger}_{\rm HL}$=0.14, $\Delta^{\dagger}_{\rm CV}$=0.22, $\Delta^{\dagger}_{\rm Rel}$=0.08, and $\Delta^{\dagger}_{\rm DBOC}$=0.16 kcal mol^{-1} results in 2σ=0.32 kcal mol^{-1}. In the CBH2 scheme these corrections should be considerably smaller than those in Table 4, which employs CBH1 equations. Even so, this approach allows an upper estimate to the uncertainty in the CBH-ANL results from the *ab initio* method.

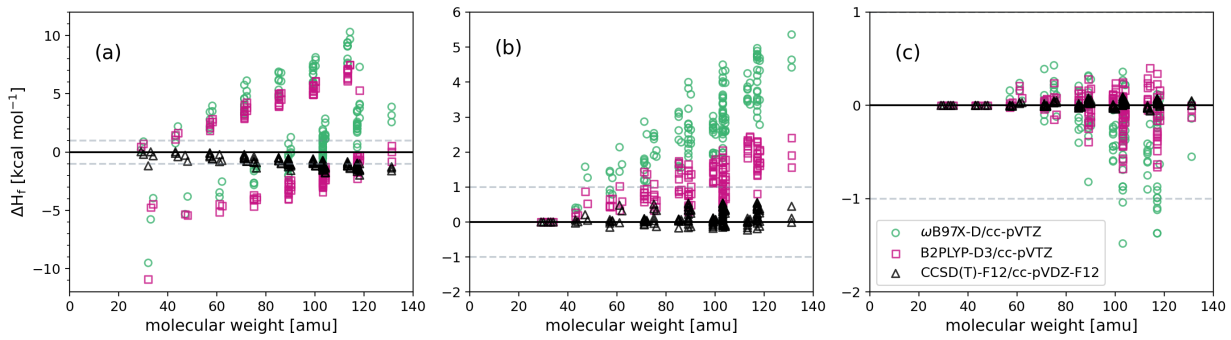

Figure 5: This plot shows the differences in the heat of formations ($\delta\Delta$H$_f$(0 K)) for 158 species that are calculated using various CBHx-Ln schemes from those calculated with CBH2-L4, where x=0, 1, and 2 on the left, middle, and right subplots respectively. Within each subplot the CBHx-Ln ΔH$_f$(0 K)s are computed with L1 in green circles, L2 in pink squares, and L3 in black triangles.

3.3 Reference Set Sensitivity

The connectivity based hierarchy can build up past the CBH2 rung. High accuracy data for CBH3+ reference species, however, is difficult to obtain because even CBH3 reference species can be as large as eight heavy atoms for hydrocarbon species. The motive for including CBH3+ reference species would be to obtain improved error cancellation. By analyzing the

convergence of heats of formation computed with CBH0, CBH1, and CBH2 we can predict the maximum benefit that CBH3 would provide. Figure 6 illustrates the results of this analysis for heats of formation computed with L4. It displays the differences in 158 heats of formation computed with $\mathrm{Set_{H_2O}}$-L4, $\mathrm{Set_{O_2}}$-L4, CBH0-L4, and CBH1-L4 from CBH2-L4. These have MAD, respectively, of 0.96, 1.06, 0.80, 0.14 kcal $\mathrm{mol^{-1}}$. The inclusion of radicals in the reference set, by using CBH0 instead of $\mathrm{Set_{H_2O}}$, improves the difference from our highest level computation, CBH2-L4 by 0.16 kcal $\mathrm{mol^{-1}}$ on average. We see that while CBH1 makes a dramatic improvement over CBH0, a shift from CBH2 to CBH3 should be negligible because CBH2 and CBH1 are already almost in complete agreement. Further, there is a nearly linear size dependence in the differences for $\mathrm{Set_{H_2O}}$, $\mathrm{Set_{O_2}}$, and CBH0. The size of the CBH1 errors, however, stabilize after 60 amu.

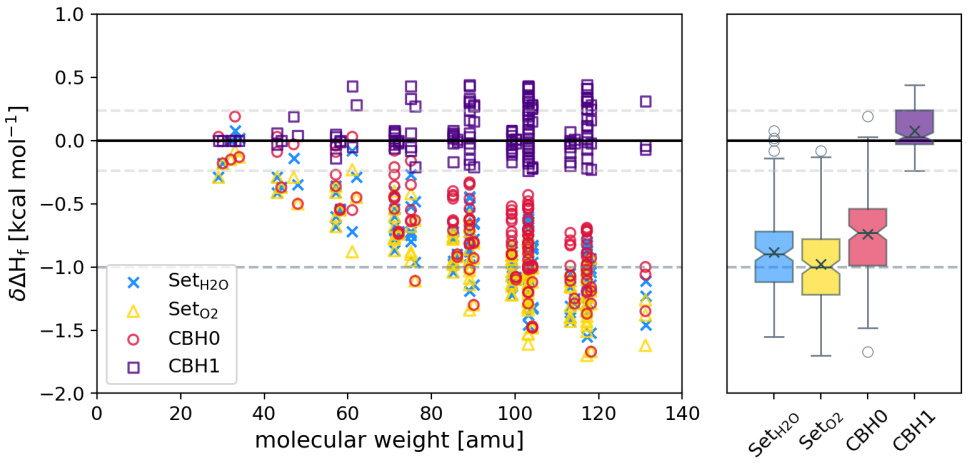

Figure 6: This plot shows the differences in the L4 heat of formations ($\Delta H_f(0\ K)$) for 158 species that are calculated using various reference chemical equations against those calculated according to the Connectivity Based Hierarchy 2 (CBH2). The blue x marks show $\Delta H_f(0\ K)_{\mathrm{L4,SetH2O}}$ -$\Delta H_f(0\ K)_{\mathrm{L4,CBH2}}$ for each molecule. The yellow triangles show $\Delta H_f(0\ K)_{\mathrm{L4,SetO2}}$ -$\Delta H_f(0\ K)_{\mathrm{L4,CBH2}}$. The red circles are $\Delta H_f(0\ K)_{\mathrm{L4,CBH0}}$ -$\Delta H_f(0\ K)_{\mathrm{L4,CBH2}}$ and the purple squares are $\Delta H_f(0\ K)_{\mathrm{L4,CBH1}}$ -$\Delta H_f(0\ K)_{\mathrm{L4,CBH2}}$. A box and whisker plot (as described in the caption to Fig. 4) is to the right of the individual data. The horizontal lines are again provided as a guide to the eye.

Perhaps the clearest indicator for how completely a set of reference species represents the target species, and thereby provides error cancellation, is the reaction enthalpies. The

reaction enthalpy is $\Delta H_f(0\,K)_M - \sum_i c_{M,i}^{[X]} \Delta H_f(0\,K)_i$, where X=[0, 1, or 2] to indicate CBH0, CBH1, and CBH2 coefficients. The $\Delta H_f(0\,K)_M$ is the heat of formation of the target species, M, and $\Delta H_f(0\,K)_i$ is that for each reference species, i. We can also break the reaction enthalpy into the electronic energy and ZPVE components. For each CBH[x] and each electronic energy method, L[n], Figure 7 shows the reaction electronic energies. For all electronic energy methods, the mean absolute CBH2 reaction electronic energies are less than 1 kcal mol^{-1}. Additionally, while CBH0 and CBH1 reaction energies have clear bias, CBH2 has both negative and positive energies, which is a clear reduction in systematic error.

We can also consider the contributions of the the directly-scaled ZPVE (E_{ds}), discussed in Section 2.5.2, to the reaction enthalpies. We show these in Figure 8. The mean absolute reaction ΔE_{ds} for CBH2(B2PLYP-D3/cc-pVTZ) is 0.18 kcal mol^{-1}, which is very minor in comparison to the reaction electronic energies. It is worth noting that during the comparison over 45 species, we found E_{ds}(B2PLYP-D3/cc-pVTZ) differs from E_0(B2PLYP-D3/cc-pVTZ) by 0.5% on average. Doubling this error to account for some size dependence and the slight deviations of E_0(B2PLYP-D3/cc-pVTZ) to a high-level E_0, and applying it to the largest CBH2 reaction ΔE_{ds}, which is 0.7 kcal mol^{-1}, the E_{ds}(B2PLYP-D3/cc-pVTZ) error should be less than 0.01 kcal mol^{-1} when used in a CBH2 scheme. Furthermore, the CBH2 scheme largely removes the linear dependence of the ZPVE reaction energy on the size of the species. It is perhaps worth emphasizing that this dramatic reduction in the ZPVE contribution to the reaction energy is one of the major benefits of the higher order CBH schemes. In Section 3.4 we further explore the sensitivity of the ZPVE within the CBH schemes.

For each total enthalpy, combining the electronic energy and ZPVE, the mean absolute CBH2 reaction enthalpies are more than 2 orders of magnitude less than CBH0, and over a factor of 20 less than CBH1, reaction energies. From the ΔH_f(0 K) MAD of the CBH0-L4 or CBH1-L4 values from the CBH2-L4 values of 0.80 and 0.14 kcal mol^{-1}, we can predict that the mean enthalpy for CBH0-L4 (79.6 kcal mol^{-1}) and for CBH1-L4 (15.5 kcal mol^{-1}) corresponds with ≈1% error to the $\Delta H_f(0\,K)$. As such, the mean absolute reaction energy of

0.51 kcal mol^{-1} for CBH2-L4 would correspond to an ≈0.005 kcal mol^{-1} error from CBH3-L4, which we have not computed. It is, then, unnecessary to introduce the uncertainty from CBH3 reference energies to an ΔH_f(0 K), when we expect negligible additional error cancellation.

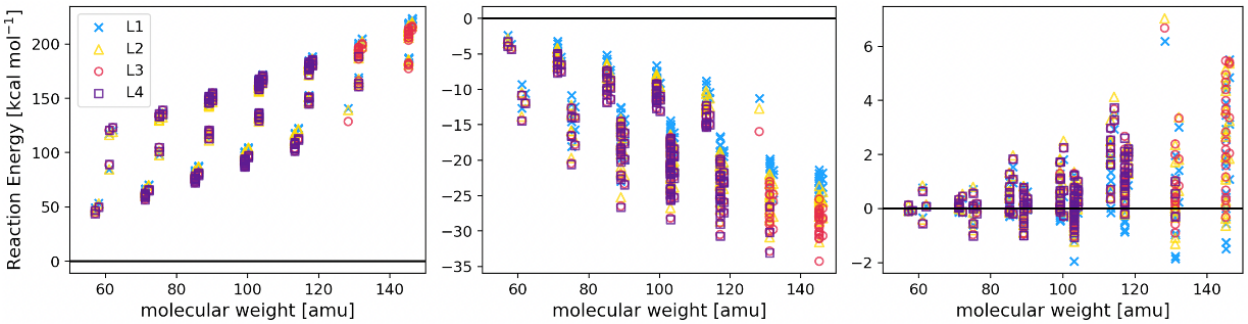

Figure 7: Reaction electronic energies for each species when using CBH0 (left), CBH1 (middle), and CBH2 (right), where within each plot the energies are from the electronic components of the L1, L2 , L3 , and L4 methods.

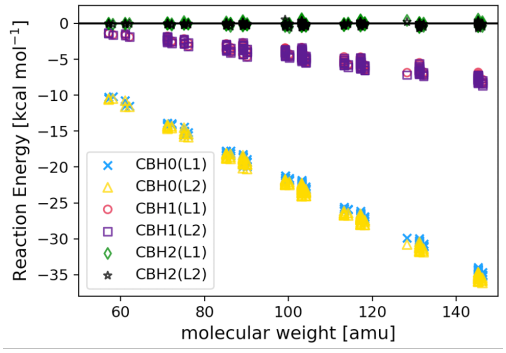

Figure 8: Reaction ZPVEs, using directly-scaled ZPVEs, for each species when using the ZPVE from CBH0(L1) in blue x marks, CBH0(L2) in yellow triangles, CBH1(L1) in red circles, CBH1(L2) in purple squares, CBH2(L1) in green diamonds, and CBH2(L2) in black stars.

3.4 ZPVE sensitivity

In the methodology section we introduce three calculations for ZPVE. Listed in order of their accuracy in comparison to 45 anharmonic ZPVEs (E_0), they are: a harmonic ZPVE (E_{harm}), a ZPVE built from scaled-frequencies (E_{sf}), and E_{ds}, which is a ZPVE that directly

scales E_{harm}. The latter two are scaling approaches that require only harmonic frequencies as input. As part of this work we have carried out a harmonic frequency analysis for all 194 $\mathbf{Set}_{target}$ species with both B2PLYP-D3/cc-pVTZ and ωB97X-D/cc-pVTZ. Note that we exclude six species from this analysis for which the B2PLYP-D3/cc-pVTZ frequencies were evaluated on a different conformer than the ωB97X-D/cc-pVTZ frequencies – a lower energy conformer had been selected during the torsional scan. This presents us with six different sets of 188 ZPVEs to consider in our sensitivity analysis. Placing them relative to the considered best method, the E_{ds}(B2PLYP-D3/cc-pVTZ), in Figure 9, we see that the directly scaled B2PLYP-D3/cc-pVTZ and ωB97X-D/cc-pVTZ ZPVEs agree fairly well across the $\mathbf{Set}_{target}$ species set, with a mean absolute deviation of 0.28 kcal mol^{-1}. This is impressive for two independent DFT methods, especially when one is using a significantly smaller basis set. It would produce, however, heats of formations that vary by the same quantity, if we were to not use an error cancellation scheme, such as the CBH equations. For instance, if using only atomic references the heats of formation would have the full uncertainties of Figure 9. The ZPVEs with the greatest deviation, and which would have the greatest impact on ΔH_f(0 K), are E_{harm}(ωB97X-D/cc-pVTZ), with a MAD of 1.63 kcal mol^{-1} from E_{ds}(B2PLYP-D3/cc-pVTZ).

We next analyze the sensitivity of ΔH_f(0 K) to such change in the ZPVE when we do apply the CBH schemes. We do so by calculating ΔH_f(0 K) with the CBH0, CBH1, and CBH2 reference sets and while using each of E_{harm}(ωB97X-D/cc-pVTZ), E_{sf}(ωB97X-D/cc-pVTZ), E_{ds}(ωB97X-D/cc-pVTZ), E_{harm}(B2PLYP-D3/cc-pVTZ), E_{sf}(B2PLYP-D3/cc-pVTZ), and E_{ds}(B2PLYP-D3/cc-pVTZ). Figure 10 shows a box and whisker plot of the deviation between each combination of these and the CBH2-ΔH_f(0 K) calculated with E_{ds}(B2PLYP-D3/cc-pVTZ). Note that electronic energy cancels out in these comparisons. When employing the CBH0 scheme, the deviations in the CBH0-ΔH_f(0 K) produced with each ZPVE approach are less than half of the deviations in ZPVEs themselves, with the exception of E_{ds}(ωB97X-D/cc-pVTZ).

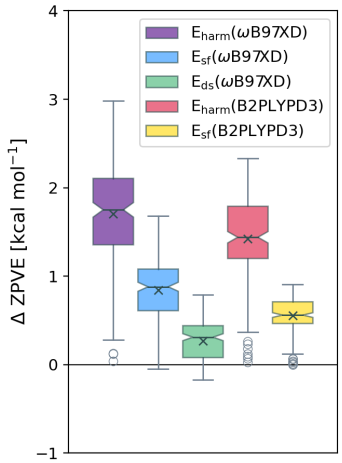

Figure 9: A box plot that shows the ZPVE deviation for 210 species when calculated with the harmonic (E_{harm}), scaled-frequency (E_{sf}), and directly-scaled (E_{ds}) approaches with either ωB97X-D/cc-pVTZ or B2PLYP-D3/cc-pVTZ. They are in reference to E_{ds}(B2PLYP-D3/cc-pVTZ).

For the CBH1-ΔH_f(0 K)s, all approaches to the ZPVE produce a MAD within 0.1 kcal mol^{-1} from the CBH2-ΔH_f(0 K) produced with E_{ds}(B2PLYP-D3/cc-pVTZ). This suggests that CBH1 references should cancel out errors in ZPVE sufficiently for most applications. CBH2 does, however, provide some improvements over CBH1. Notably, the ωB97X-D/cc-pVTZ CBH2-ΔH_f(0 K)s all agree with one another with a MAD within less than 0.01 kcal mol^{-1} and a max deviation of 0.04 kcal mol^{-1}. Similarly the CBH2-ΔH_f(0 K)s calculated with all B2PLYP-D3/cc-pVTZ ZPVE approaches agree with one another with a MAD of 0.01 kcal mol^{-1}.

Clearly, the CBH2 scheme has greatly reduced sensitivity to ZPVE. From the previous examination of reaction ZPVEs, in Section 3.3, we determined that the E_{ds}(B2PLYP-D3/cc-pVTZ) contribution to a CBH2-ΔH_f(0 K) should be accurate to within 0.01 kcal mol^{-1}. Here we see that when using ZPVEs that have an RMSD of even 2.04 kcal mol^{-1} (*i.e.*, E_{harm}(ωB97X-D/cc-pVTZ) and E_{ds}(B2PLYP-D3/cc-pVTZ)) the CBH2-ΔH_f(0 K)s have only an RMSD of 0.11 kcal mol^{-1}. Finally we note that including approximations of anharmonicity has no impact on the CBH2-ΔH_f(0 K)s for either DFT method, and would

expect no impact from a true anharmonic treatment. From this we suggest a $2\sigma = 0.01$ kcal mol^{-1} contribution of uncertainty to the final CBH2-ΔH_f(0 K)s from the E_{ds}(B2PLYP-D3/cc-pVTZ).

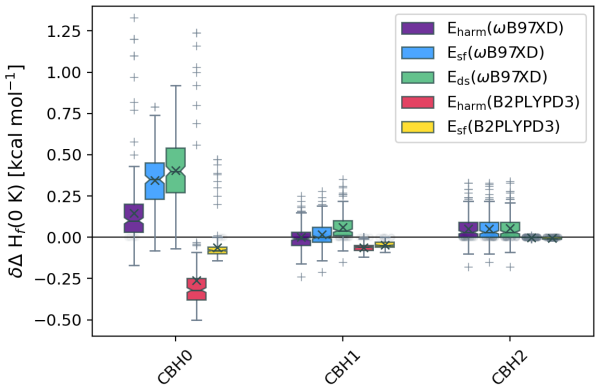

Figure 10: A box plot that shows the CBH0, CBH1, and CBH2 heats of formation deviation, $\delta\Delta H_f$(0 K), for 194 species when calculated with the harmonic (E_{harm}), scaled-frequency (E_{sf}), and directly-scaled (E_{ds}) approaches with either ωB97X-D/cc-pVTZ or B2PLYP-D3/cc-pVTZ. The $\delta\Delta H_f$(0 K) is in reference to the CBH2-ΔH_f(0 K) calculated with E_{ds}(B2PLYP-D3/cc-pVTZ).

3.5 Rotamer Sensitivity

A companion work[44] carried out an extensive conformer analysis on this alkane oxidation set, focusing on a subset of 66 species with fewer than seven atoms. The purpose was to evaluate the rotamer selection at the mid- to high-temperature range. In it we discover that the ZPVE can play a significant role in rotamer energy ordering.[44] Specifically, the global minimum electronic energy structure may have a significantly higher ZPVE than an alternate local minimum, to the extent that its predicted $\Delta H_f(0\ K)$ exceeds that of the alternate structure. We here introduce $\delta\Delta H_f(0\ K)_{\ rot}$ as the difference in heat of formation between the $\Delta H_f(0\ K)$ of the minimum ωB97X-D/6-31G* electronic energy rotamer and the structure that has the lowest $\Delta H_f(0\ K)$. We discover 20 species with $\delta\Delta H_f(0\ K)_{\ rot} < 0$ kcal mol^{-1} by evaluating the CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ $\Delta H_f(0$ K), with the E_{ds}(B2PLYP-D3/cc-pVTZ), for all rotamers of each of the 66 species subset. While four of these have only a $\delta\Delta H_f(0\ K)_{\ rot} > -0.15$ kcal mol^{-1}, some are significantly lower. For instance $CH_3CH(CH_2)CH_2OOH$ has an $\delta\Delta H_f(0\ K)_{\ rot} = -1.16$ kcal mol^{-1}, and four others have $\delta\Delta H_f(0\ K)_{\ rot} < -0.5$ kcal mol^{-1}.

Rotamer selection using the global electronic energy minimum is the most common approach, if any conformer sampling is even attempted. We've here demonstrated that this approach may lead to moderate errors to the $\mathbf{Set}_{\text{target}}$ species. Certainly in sets with even larger species that have many conformers, however, it becomes extremely expensive to evaluate the ZPVE of every conformer. For $\mathbf{Set}_{\text{target}}$, alone, it would require thousands of extra computations. Instead we consider that significant $\delta\Delta H_f(0\ K)$ are most likely when the lowest electronic energy structure is stabilized with a hydrogen bond, the added rigidity of which heightens the ZPVE. We find, consistent with this idea, the greatest energy shifts are in **$\dot{\text{Q}}$OOH** species, which have the strong hydrogen bonding in $\mathbf{Set}_{\text{target}}$. As such, we define a structural hydrogen bond threshold as a distance less than 2.43 Å and an angle greater than 100^o, and select the lowest ωB97X-D/6-31G* electronic energy rotamer that is not within this threshold. We add to our definitions, $\delta\Delta H_f(0\ K)_{\ rot,nHB}$, which is the difference in heat

of formation between the ΔH_f(0 K) of the minimum ωB97X-D/6-31G* electronic energy, non-hydrogen bonded (nHB) rotamer and the structure that has the lowest ΔH_f(0 K). Of the 66 species subset, $\delta\Delta H_f$(0 K) $_{\text{rot,nHB}}$ is within -0.33 kcal mol^{-1} for all but one species. The exception is $CH_3CH_2\dot{C}HCH_2OOH$, with $\delta\Delta H_f$(0 K) $_{\text{rot,nHB}} = -0.54$ kcal mol^{-1}, and does have slight hydrogen bonding character, but not enough to be caught by the hydrogen bond threshold.

This hydrogen bond cutoff is a great improvement in error for a consideration that requires no additional computations in the conformer selection. Accordingly, we apply it to the full $\mathbf{Set}_{\text{target}}$, computing CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ ΔH_f(0 K), with the E_{ds}(B2PLYP-D3/cc-pVTZ), for the lowest ωB97X-D/6-31G* electronic energy structure that is not within the hydrogen bond threshold. In doing so, we find 27 lower ΔH_f(0 K)s across $\mathbf{Set}_{\text{target}}$. Table S4.1 gives the ΔH_f(0 K)s for these species, alongside the higher ΔH_f(0 K)s predicted from the minimum ωB97X-D/6-31G* electronic energy structure. Additionally, Table S4.1 lists the ΔH_f(0 K)s for the remaining 15 species of the 66 species subset that have a less significant $\delta\Delta H_f$(0 K), but were not corrected with the hydrogen bonding consideration.

3.6 Summary of uncertainties

In this discussion we have considered the dominant sources of error and uncertainty to the CBH-ANL ΔH_f(0 K)s. The CBH2 scheme allows for a tremendous error cancellation, enabling us to predict a basis set completeness error of $2\sigma = 0.1$ kcal mol^{-1}, errors from neglecting higher-order corrections to have an upper limit of $2\sigma = 0.32$ kcal mol^{-1}, and the uncertainty from approximations to the ZPVE to have a $2\sigma < 0.01$ kcal mol^{-1}. In doing so, the CBH2 schemes introduce their own uncertainty, from the uncertainties in each reference species in the chemical equation of a parent species. We determined that, for the largest $\mathbf{Set}_{\text{target}}$ species, the propagation of reference species uncertainty is $2\sigma = 0.27$ kcal mol^{-1} through careful consideration of each individual ANL0† computation. The cumulation of

the here-mentioned uncertainties gives $2\sigma = 0.43$ kcal mol^{-1}. We additionally consider that this uncertainty only applies to individual rotamers, and that there is an additional 0.33 kcal mol^{-1} uncertainty in our CBH-ANL ΔH_f(0 K)s, for a total of 0.54 kcal mol^{-1}, from having assumed that the rotamer with the lowest ωB97X-D/6-31G* electronic energy, and no hydrogen bonding, will have the lowest ΔH_f(0 K). We also note that the upper limit for many of these uncertainties, which was used in this tabulation, is only relevant for the largest, ten heavy atom, species in $\mathbf{Set}_{\text{target}}$. The average $\mathbf{Set}_{\text{target}}$ species is seven heavy atoms. With these considerations, we conclude that our CBH-ANL ΔH_f(0 K)s have $2\sigma = 0.2 - 0.5$ kcal mol^{-1}.

3.7 Validation

With an estimated 2σ error of only $0.2 - 0.5$ kcal mol^{-1} for the full set of 194 species, the CBH-ANL ΔH_f(0 K)s are highly reliable. As such, they should be consistent with experimental report. We here validate against the API tables of Scott[52] for the 16 alkane species in the target set. The API ΔH_f(0 K)s are assigned uncertainty in the hundredths of a kcal mol^{-1}, where they disclaim that the larger, more branched hydrocarbon uncertainty may be underestimated. The RMSD between API tables and CBH-ANL is 0.22 kcal mol^{-1} when using the ANL0† energies as references for a CBH2 equation, see Table 5. The largest error is for *neo*-pentane, with a 0.54 kcal mol^{-1} deviation from the API ΔH_f(0 K). Neglecting this outlier, the RMSD is 0.17 kcal mol^{-1} across 15 species with up to 8C.

Table 5 also provides CBH2-ΔH_f(0 K) values using 3 alternative sources of reference energies in the CBH2 equation in place of the ANL0†. These are ΔH_f(0 K)s collated from ANL0, ATcT, and calculated with ANL0$^{\text{ATcT}}$ which ladders ANL0 ΔH_f(0 K)s against ATcT ΔH_f(0 K)s in a CBH1 equation analogous to the laddering of ANL0 with ANL1 (*i.e.*, ANL0†). For this set of species, we see a minor, 0.03 kcal mol^{-1}, decrease in the RMSD from using the ANL0† references instead of the ANL0 ones. The ATcT reference energies result in essentially the same RMSD as those from ANL0. We see a more significant advantage of

Table 5: ΔH_f(0 K) collated from the API tables[52] that overlap with species in $\mathbf{Set_{target}}$, and the deviations (δCBH2-ΔH_f(0 K)) between them.

	ΔH_f(0 K)	δCBH2-ΔH_f(0 K) [a]			
SMILES	API	ANL0	ANL0†	ATcT	ANL0$^{\mathrm{ATcT}}$
C	−15.95	0.04	0.04	0.04	0.04
CC	−16.26	−0.20	−0.23	−0.09	−0.09
CCC	−19.74	−0.16	−0.22	−0.03	0.07
CC(C)C	−25.34	0.04	−0.05	0.03	0.38
CCCC	−23.63	0.10	0.01	0.24	0.43
CC(C)(C)C	−32.38	0.65	0.54	0.30	1.10
CCC(C)C	−28.44	0.06	−0.06	0.06	0.50
CCCCC	−27.35	0.19	0.06	0.34	0.63
CCC(C)(C)C	−34.28	0.15	0.00	−0.19	0.71
CC(C(C)C)C	−32.16	−0.18	−0.32	−0.30	0.38
CCCCCC	−31.09	0.26	0.10	0.42	0.80
CC(CC(C)C)C	−37.21	0.09	−0.08	−0.02	0.76
CCC(C(C)C)C	−35.20	−0.16	−0.33	−0.27	0.51
CCCCCCC	−34.83	0.31	0.12	0.48	0.96
CC(CC(C)(C)C)C	−41.23	0.39	0.19	−0.06	1.18
CC(C(C)C)C(C)C	−39.18	−0.09	−0.28	−0.32	0.71
	MAD	0.19	0.16	0.20	0.58
	RMSD	0.24	0.22	0.25	0.67

[a]Columns 2-4 list the deviation of computed CBH2-ΔH_f(0 K)s from the API value, which employ CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ electronic energies with an E_{ds}(B2PLYP-D3/cc-pVTZ) ZPVE held relative to either ANL0, ANL0†, ATcT, or ANL0$^{\mathrm{ATcT}}$ ΔH_f(0 K)s, respectively.

the ANL0† references in that here is visibly no correlation between species size and error of the ΔH_f(0 K)s it produces, while such a trend does exist for the other choices of reference energy.

Interestingly, the ANL0$^{\mathrm{ATcT}}$ reference energies produce significantly worse errors. Numerically, this is largely the accumulation of deviations in the three smallest species, methane, ethane, and propane, which have larger CBH2 coefficients in the larger alkane species. We see that ethane and propane deviate from experiment in the same direction for the first three reference energy sources. The ANL$^{\mathrm{ATcT}}$ energies, however, deviate in opposite directions. This is important because the CBH2 equations employ ethane to balance propane

fragments. Such deviation demonstrates the importance in choosing reliable and consistent reference data.

ATcT additionally reports ΔH_f(0 K)s for several species in our target set that are molecules larger than those contained in Set_{CBH2}. We show these in Table 6 as well as their deviation from CBH2-ΔH_f(0 K)s produced from the four sources of reference energies. The ANL0† reference energies are in impressive agreement, with an RMSD of 0.11 kcal mol^{-1}. Employing ATcT ΔH_f(0 K)s for smaller species in a CBH2 equation, produces ΔH_f(0 K)s that deviate from ATcT values for larger species with an RMSD 0.24 kcal mol^{-1}. If we only consider the largest species in this analysis (*e.g.*, those that are 6 heavy atoms or more), the RMSDs are 0.09 and 0.34 kcal mol^{-1} for the ANL0† and ATcT reference sets, a difference of nearly a factor of 4.

Table 6: ΔH_f(0 K)s collated from ATcT[51] that overlap with species in $\text{Set}_{\text{target}}$, and the deviations ($\delta$CBH2-$\Delta H_f$(0 K)) between them.

	ΔH_f(0 K)	ϵ^a	δCBH2-ΔH_f(0 K) ^{b}			
SMILES	ATcT	ATcT	ANL0	ANL0†	ATcT	ANL0$^{\text{ATcT}}$
CCC[CH2]	24.62	0.17	−0.02	−0.16	0.04	0.22
C[CH]CC	21.82	0.20	−0.07	−0.23	0.07	0.08
CCCC	−23.50	0.06	−0.03	−0.12	0.11	0.30
CCOO	−34.15	0.29	−0.05	−0.01	−0.10	0.06
CCC(C)C	−28.42	0.09	0.03	−0.08	0.04	0.48
CCCCC	−27.29	0.07	0.13	0.00	0.28	0.57
CC(C(C)C)C	−32.62	0.18	0.14	0.00	0.02	0.71
CCCCCC	−31.04	0.08	0.21	0.05	0.37	0.75
CCCCCCC	−34.74	0.11	0.22	0.03	0.39	0.88
CC(CC(C)(C)C)C	−40.87	0.36	0.03	−0.17	−0.42	0.82
		MAD	0.09	0.07	0.15	0.41
		RMSD	0.12	0.11	0.24	0.57

[a] ATcT reported uncertainty [b] Columns 2-4 list the deviation of computed CBH2-ΔH_f(0 K)s from the ATcT value, which employ CCSD(T)-F12b/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ electronic energies with an E_{ds}(B2PLYP-D3/cc-pVTZ) ZPVE held relative to either ANL0, ANL0†, ATcT, or ANL0$^{\text{ATcT}}$ ΔH_f(0 K)s, respectively.

The large and gradually increasing errors when using ANL0$^{\text{ATcT}}$, and, to a lesser extent ATcT, as reference energies for the larger closed shell species in Table 6 are suggestive of some

inconsistency in the ATcT values for the closed shell CBH1 hydrocarbon references species - *i.e.*, CH_4 and C_2H_6. It is not clear what that inconsistency would be, but the increasing contributions from those species in the CBH2 equations leads to a significant amplification of any such inconsistencies. For example and using SMILES strings, for the CC(CC(C)(C)C)C species, the CBH2 equation is CC(CC(C)(C)C)C + 2 CC=CCC + CC(C)C + CC(C)(C)C, and the CBH1 equation used to ladder the $ANL0^{ATcT}$ energy has 4CC for the CC(C)(C)C component, alone. Perhaps the large uncertainties in the ATcT analysis for the larger species leads to too small a constraint on the consistency between the CH4 and C2H6. We here also note that recently updated ATcT reference enthalpies used in the present analysis like that for ethane, -16.35 ± 0.03 kcal mol^{-1} (v1.124)[53] lead to significantly more consistent results in Table 6 than those from previous versions, like -16.28 ± 0.04 kcal mol^{-1} (v1.118).[54] Changes of 0.07 kcal mol^{-1} to a reference species that is used as frequently as ethane can make large impacts on ΔH_f(0 K)s of larger species. Through this observation we see unique potential in the CBH2 schemes, in that the deviation of CBH2-ΔH_f(0 K)s in large species may be used to retroactively improve even small species data. It also showcases the importance of having the well-understood and invariable reference values from ANL0†.

3.8 CBH-ANL ΔH_f(0 K)

We construct the CBH-ANL ΔH_f(0 K)s for 194 $\mathbf{Set}_{\mathrm{target}}$ species, in summary, by determining the lowest ωB97X-D/6-31G* electronic energy conformer, with no hydrogen bonding. After re-optimizing with B2PLYP-D3/cc-pVTZ, we computed the directly-scaled B2PLYP-D3/cc-pVTZ ZPVE for each species. The electronic energies for 158 species were calculated at CCSD(T)-F12/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ, and, for the largest $\mathbf{Set}_{\mathrm{target}}$ species, at CCSD(T)-F12/cc-pVDZ-F12//B2PLYP-D3/cc-pVTZ. These parameters were used in a CBH2 equation, alongside ANL0† reference ΔH_f(0 K)s, to evaluate the final CBH-ANL ΔH_f(0 K)s, which are displayed in Tables 7−9.

Table 7: CBH-ANL ΔH_f(0 K)s, in kcal mol^{-1}, for alkane and alkyl radical species

SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)
C	−15.91	[CH$_3$]	35.80	CC(C(C)C)[CH$_2$]	15.46
CC	−16.49	C[CH$_2$]	31.30	CCCCC[CH$_2$]	17.39
CCC	−19.95	CC[CH$_2$]	28.14	[CH$_2$]CC(C(C)C)C	11.96
CC(C)C	−25.39	C[CH]C	24.99	CC([CH]C(C)C)C	8.91
CCCC	−23.61	C[C](C)C	17.88	CC[CH]CCCC	11.54
CC(C)(C)C	−31.85	CC([CH$_2$])C	23.23	CCC(C(C)C)[CH$_2$]	12.02
CCC(C)C	−28.50	CCC[CH$_2$]	24.46	CCC[CH]CCC	11.52
CCCCC	−27.27	C[CH]CC	21.59	C[C](CC(C)C)C	5.57
CC(C(C)C)C	−32.61	[CH$_2$]CC(C)C	19.15	C[CH]C(C(C)C)C	9.03
CCC(C)(C)C	−34.28	CC([CH$_2$])(C)C	17.37	C[CH]CCCCC	11.10
CCCCCC	−30.97	CC[C](C)C	14.90	CC(CC(C)C)[CH$_2$]	11.23
CC(CC(C)C)C	−37.29	CC[CH]CC	18.19	CC[C](C(C)C)C	7.58
CCC(C(C)C)C	−35.57	CCC([CH$_2$])C	20.01	CCC([C](C)C)C	6.72
CCCCCCC	−34.69	CCCC[CH$_2$]	20.80	CCC(C(C)[CH$_2$])C	12.43
CC(C(C)C)C(C)C	−39.46	C[CH]C(C)C	16.94	CCCCCC[CH$_2$]	13.66
CC(CC(C)(C)C)C	−41.04	C[CH]CCC	17.91	[CH$_2$]C(C(C)C)C(C)C	7.66
CC(CC(C)(C)C)(C)C	*−43.80*	[CH$_2$]CC(C)(C)C	13.24	C[C](C(C)C)C(C)C	3.09
		CC[CH]CCC	14.45	C[C](CC(C)(C)C)C	−0.15
		CCC([CH$_2$])(C)C	14.43	CC([C](C)C)C(C)C	2.22
		C[C](C(C)C)C	10.53	CC([CH]C(C)(C)C)C	3.00
		C[CH]C(C)(C)C	11.34	CC(C([CH$_2$])C)C(C)C	8.33
		C[CH]CCCC	14.50	CC(CC([CH$_2$])(C)C)C	6.87
				CC(CC(C)(C)C)[CH$_2$]	6.81

Values in italics use the cc-pVDZ-F12 basis set rather than cc-pVTZ-F12.

4. Conclusion

We have built reliable ΔH_f(0 K)s for 194 alkane oxidation species by pairing a composite quantum chemistry scheme with highly accurate and representative reference species. The composite scheme determined a geometry with a ωB97X-D/6-31G* conformer sampling, considering non-hydrogen bonding species when relevant. It then improved upon that geometry with ωB97X-D/cc-pVTZ 1D torsional scans followed by B2PLYP-D3/cc-pVTZ geometry optimization and harmonic frequency analysis. We calculated CCSD(T)-F12/cc-pVXZ-F12 single point energies, where X is the triple-ζ basis for species with fewer than 9 heavy atoms and the double-ζ for all species.

Table 8: CBH-ANL ΔH_f(0 K)s, in kcal mol^{-1}, for hydroperoxides and alkylperoxy radicals

SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)
COO	−27.35	OOCC(C(C)C)C	−49.99	CCCCCO[O]	−12.67
CCOO	−34.16	OOCCC(C)(C)C	−51.26	[O]OC(C(C)C)C	−17.52
CCCOO	−37.63	OOC(C(C)C)C(C)C	*−55.84*	[O]OCC(C)(C)C	−16.54
OOC(C)C	−41.44	OOC(CC(C)C)(C)C	*−60.10*	[O]OCCC(C)C	−13.92
CCCCOO	−41.23	OOCC(CC(C)C)C	*−54.80*	CC(O[O])CCC	−16.80
CC(OO)CC	−44.86	OOC(C(C)(C)C)C(C)C	*−60.16*	CCC(O[O])(C)C	−20.90
OOC(C)(C)C	−49.16	OOC(CC(C)(C)C)(C)C	*−62.90*	CCC(O[O])CC	−16.50
OOCC(C)C	−43.12	OOCC(CC(C)(C)C)C	*−59.75*	[O]OC(C(C)(C)C)C	−23.04
CC(CC)COO	−45.90	OOCC(CC(C)C)(C)C	*−58.89*	[O]OC(C(C)C)(C)C	−24.88
CCCCCOO	−44.67			[O]OCC(C(C)C)C	−17.92
CC(OO)CCC	−48.19	CO[O]	5.33	[O]OCCC(C)(C)C	−19.06
CCC(OO)(C)C	−52.40	CCO[O]	−1.76	CCC(CO[O])(C)C	−19.21
CCC(OO)CC	−48.21	[O]OC(C)C	−9.55	[O]OC(C(C)C)C(C)C	−25.43
OOC(C(C)C)C	−49.16	CCCO[O]	−5.28	[O]OC(CC(C)C)(C)C	−28.75
OOCC(C)(C)C	−48.91	CCCCO[O]	−8.95	[O]OCC(CC(C)C)C	*−22.79*
OOCCC(C)C	−46.01	[O]OC(C)(C)C	−17.84	[O]OC(C(C)(C)C)C(C)C	*−28.36*
CCC(COO)(C)C	−51.67	[O]OCC(C)C	−10.73	[O]OC(CC(C)(C)C)(C)C	*−31.92*
OOC(C(C)(C)C)C	−54.55	CC(O[O])CC	−13.10	[O]OCC(CC(C)(C)C)C	*−27.26*
OOC(C(C)C)(C)C	−56.17	CC(CC)CO[O]	−13.82	[O]OCC(CC(C)C)(C)C	*−26.28*

Values in italics use the cc-pVDZ-F12 basis set rather than cc-pVTZ-F12.

We emphasize, in this work, the error cancellation to ΔH_f(0 K) that can be achieved by determining it with a well-designed chemical equation. We have examined 5 different approaches, forming reference species sets that are simply H_2O, CH_4, and H_2 and one that is O_2, CH_4 and H_2. The remaining three approaches are the first three rungs of the connectivity based hierarchy of Raghavachari[31–33] (*i.e.*, CBH0, CBH1, and CBH2). We find increasing levels of error cancellation across these sets, seeing that, for the 158 species that are fewer than 9 heavy atoms, the MAD between ΔH_f(0 K)s produced with CCSD(T)-F12/cc-pVTZ-F12 and CCSD(T)-F12/cc-pVDZ-F12 diminishes from 0.94 kcal mol^{-1}, when using CBH0, to 0.02, when using CBH2. Moreover, again compared to ΔH_f(0 K)s produced with CBH2-CCSD(T)-F12/cc-pVTZ-F12//B2PLYP-D3/cc-pVTZ, those produced with ωB97X-D/6-31G* improve in deviation from 2.30 ± 7.02 to -0.19 ± 0.74 kcal mol^{-1} when moving

from CBH0 to CBH2 and the same comparision, for B2PLYP-D3/cc-pVTZ, improve from 0.16 ± 7.42 to -0.04 ± 0.38 kcal mol^{-1}.

The ZPVE is also key in the determination of ΔH_f(0 K). We here have analyzed the impact of using a harmonic ZPVE compared to two different scaled ZPVEs. The scaled-frequency approach approximates an anharmonic ZPVE by scaling frequencies to approximate anharmonic frequencies. In contrast, the directly-scaled ZPVE scales the harmonic ZPVE directly to approximate the anharmonic ZPVE. The scaling parameters are determined by fitting against the B2PLYP-D3/cc-pVTZ VPT2 anharmonic frequencies and ZPVEs for 45 species. The directly-scaled B2PLYP-D3/cc-pVTZ ZPVEs for this set have $2\sigma = 0.17$ kcal mol^{-1} error from the anharmonic ZPVE. We then consider the sensitivity that ΔH_f(0 K) will have to this error within a CBH scheme. Across the 210 species, which include $\mathbf{Set}_{\text{target}}$ and the reference species, the RMSD between harmonic ωB97X-D/6-31G* ZPVEs and directly-scaled B2PLYP-D3/cc-pVTZ ZPVEs is 2.04 kcal mol^{-1}, but the ΔH_f(0 K) produced with them only have RMSD of 0.11 kcal mol^{-1}.

The CBH2 scheme is capable of tremendous error cancellation, shown both through the diminished deviation between ΔH_f(0 K)s produced with it when using ωB97X-D/6-31G* energies or harmonic frequencies and ΔH_f(0 K)s produced with our higher-level composite scheme. It however, can introduce significant uncertainties through the propagation of uncertainties of the reference species. As such we have designed a CBH-ANL approach, which ladders high-level ANL0 energies with higher-level ANL1 energies. We are able to quantify the uncertainties of each reference species, the highest of which is 0.15 kcal mol^{-1} and their propagation into the target species, the largest of which has 0.27 kcal mol^{-1} introduced uncertainty. The improved accuracy of CBH2 ΔH_f(0 K)s, then, strongly triumphs over the introduced uncertainty. Our final prediction of uncertainty, after considering uncertainty and errors in electronic method, ZPVE, rotamer selection, and CBH2 reference species, is $2\sigma = 0.2 - 0.5$ kcal mol^{-1} for the 194 alkane oxidation species.

Supporting Information

- Additional data and descriptions of Section S1: harmonic, anharmonic, and scaled ZPVE, S2: reference species coefficients for each set and species, S3: energies of reference species, and S4: energy tables for alternative rotamers (SI.pdf)
- Species dictionary of InChI, SMILES, and IUPAC names with 2D images (dicitonary.pdf)
- Excel formatted heats of formation, listed in Tables 7-9 (enthalpies.xlsx)

Acknowledgements

Sarah N. Elliott, Murat Keçeli, and Stephen J. Klippenstein are supported in this work by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two U.S. Department of Energy (DOE) organizations, the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem including software, applications, hardware, advanced system engineering, and early test bed platforms to support the nation's exascale computing imperative. Part of this material is based on work at Argonne supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences, under Contract No. DE-AC02-06CH11357 as part of the Gas Phase Chemical Physics program. The authors at NUI Galway acknowledge funding from Science Foundation Ireland via project numbers 15/IA/3177 and the Irish Centre for High-End Computing (ICHEC).

References

(1) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. Toward Subchemical Accuracy in Computational Thermochemistry: Focal Point Analysis of the Heat of Formation of NCO and [H,N,C,O] Isomers. *J. Chem. Phys.* **2004**, *120*, 11586–11599.

(2) Sellers, H.; Pulay, P. The adiabatic correction to molecular potential surfaces in the SCF approximation. *Chem. Phys. Lett.* **1984**, *103*, 463 – 465.

(3) Ilias, M.; Saue, T. An infinite-order two-component relativistic Hamiltonian by a simple one-step transformation. *J. Chem. Phys.* **2007**, *126*, 064102.

(4) Wolf, A.; Reiher, M.; Hess, B. A. The generalized Douglas-Kroll transformation. *J. Chem. Phys.* **2002**, *117*, 9215–9226.

(5) Aprà, E.; Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, H. J. J.; Alexeev, Y.; Anchell, J. et al. NWChem: Past, present, and future. *J. Chem. Phys.* **2020**, *152*, 184102.

(6) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(7) Karton, A.; Daon, S.; Martin, J. M. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **2011**, *510*, 165 – 178.

(8) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(9) Harding, M. E.; Vázquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. High-accuracy extrapolated ab initio thermochemistry. III. Additional improvements and overview. *J. Chem. Phys.* **2008**, *128*, 114111.

(10) Jaeger, H. M.; Schaefer, H. F.; Demaison, J.; Császár, A. G.; Allen, W. D. Lowest-lying conformers of alanine: Pushing theory to ascertain precise energetics and semiexperimental R_e-structures. *J. Chem. Theory Comput.* **2010**, *6*, 3066–3078.

(11) Klippenstein, S. J.; Harding, L. B.; Ruscic, B. Ab initio computations and active thermochemical tables hand in hand: Heats of formation of core combustion species. *J. Phys. Chem. A* **2017**, *121*, 6580–6602.

(12) Keçeli, M.; Elliott, S. N.; Li, Y.-P.; Johnson, M. S.; Cavallotti, C.; Georgievskii, Y.; Green, W. H.; Pelucchi, M.; Wozniak, J. M.; Jasper, A. W. et al. Automated computational thermochemistry for butane oxidation: A prelude to predictive automated combustion kinetics. *Proc. Combust. Inst.* **2019**, *37*, 363–371.

(13) Elliott, S. N.; Moore, K. B.; Copan, A. V.; Keçeli, M.; Cavallotti, C.; Georgievskii, Y.; Schaefer, H. F.; Klippenstein, S. J. Automated theoretical chemical kinetics: Predicting the kinetics for the initial stages of pyrolysis. *Proc. Combust. Inst.* **2021**, *38*, 375–384.

(14) Cavallotti, C.; Pelucchi, M.; Georgievskii, Y.; Klippenstein, S. J. EStokTP: Electronic structure to temperature- and pressure-dependent rate constants—A code for automatically predicting the thermal kinetics of reactions. *J. Chem. Theory Comput.* **2019**, *15*, 1122–1145.

(15) Smith, D. G. A.; Lolinco, A. T.; Glick, Z. L.; Lee, J.; Alenaizan, A.; Barnes, T. A.; Borca, C. H.; Di Remigio, R.; Dotson, D. L.; Ehlert, S. et al. Quantum chemistry common driver and databases (QCDB) and quantum chemistry engine (QCEngine): Automation and interoperability among computational chemistry programs. *J. Chem. Phys.* **2021**, *155*, 204801.

(16) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties:

Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331, PMID: 26113956.

(17) Collins, E. M.; Raghavachari, K. Effective Molecular Descriptors for Chemical Accuracy at DFT Cost: Fragmentation, Error-Cancellation, and Machine Learning. *Journal of Chemical Theory and Computation* **2020**, *16*, 4938–4950, PMID: 32678593.

(18) Zhou, Y.; Wu, J.; Xu, X. Improving B3LYP heats of formation with three-dimensional molecular descriptors. *J. Comp. Chem.* **2016**, *37*, 1175–1190.

(19) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

(20) Duan, X.-M.; Li, Z.-H.; Song, G.-L.; Wang, W.-N.; Chen, G.-H.; Fan, K.-N. Neural network correction for heats of formation with a larger experimental training set and new descriptors. *Chem. Phys. Lett.* **2005**, *410*, 125 – 130.

(21) Balabin, R. M.; Lomakina, E. I. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.* **2009**, *131*, 074104.

(22) Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O. Artificial intelligence-enhanced quantum chemical method with broad applicability. *Nat. Commun.* **2021**, *12*, 1–13.

(23) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(24) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular orbital theory of the electronic structure of organic compounds. V. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.

(25) Pople, J. A.; Radom, L.; Hehre, W. J. Molecular orbital theory of the electronic structure of organic compounds. VII. Systematic study of energies, conformations, and bond interactions. *J. Am. Chem. Soc.* **1971**, *93*, 289–300.

(26) Hehre, W.; Radom, L.; Schleyer, P. v. R.; Pople, J. *Molecular Orbital theory*; Wiley—Interscience New York, 1986.

(27) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. *Theoretica chimica acta* **1975**, *38*, 121–129.

(28) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. An alternative approach to the problem of assessing destabilization energies (strain energies) in cyclic hydrocarbons. **1976**, *32*, 317–323.

(29) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. Homodesmotic reactions for the assessment of stabilization energies in benzenoid and other conjugated cyclic hydrocarbons. *Journal of the Chemical Society, Perkin Transactions 2* **1976**, 1222–1227.

(30) Wheeler, S. E.; Houk, K. N.; Schleyer, P. v. R.; Allen, W. D. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *J. Am. Chem. Soc.* **2009**, *131*, 2547–2560.

(31) Ramabhadran, R. O.; Raghavachari, K. Theoretical thermochemistry for organic molecules: Development of the generalized connectivity-based hierarchy. *J. Chem. Theory Comput.* **2011**, *7*, 2094–2103.

(32) Ramabhadran, R. O.; Raghavachari, K. Connectivity-based hierarchy for theoretical thermochemistry: Assessment using wave function-based methods. *J. Phys. Chem. A* **2012**, *116*, 7531–7537.

(33) Sengupta, A.; Raghavachari, K. Prediction of accurate thermochemistry of medium

and large sized radicals using connectivity-based hierarchy (CBH). *J. Chem. Theory Comput.* **2014**, *10*, 4342–4350.

(34) Copan, A. V.; Elliott, S. N.; Moore, K. B.; Klippenstein, S. J. Automol. `https://github.com/AutoMech/autochem`, 2022; Accessed: 2022-11-16.

(35) Werner, H.; Knowles, P. Getting Started with Molpro Version 2015.1.

(36) Valiev, M.; Bylaska, E.; Govind, N.; Kowalski, K.; Straatsma, T.; Dam, H. V.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. et al. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477 – 1489.

(37) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. et al. Gaussian˜16 Revision B.01. 2016; Gaussian Inc. Wallingford CT.

(38) Shao, Y.; . . . ,; III, H. F. S.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184–215.

(39) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011**, *3*, 33.

(40) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comp. Chem.* **1996**, *17*, 553–586.

(41) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(42) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.;

Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(43) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. et al. Gaussian 09 Revision D. 01, 2009, Gaussian Inc. *Wallingford CT* **2009**, *93*.

(44) Elliott, S. N.; Moore, K. B.; Copan, A. V.; Georgievskii, Y.; Keceli, M.; Somers, K.; Ghosh, M. K.; Curran, H. J.; Klippenstein, S. J. Systematically Derived Thermodynamic Properties for Alkane Oxidation. *Combust. Flame* **2022 (in press)**,

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

(46) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(47) Elliott, S. N. High-level automated thermochemistry: building from small to large. Work presented at 5th Annual Flame Chemistry Workshop, Virtual, 2020.

(48) Nielsen, H. H. The vibration-rotation energies of polyatomic molecules. *Phys. Rev.* **1941**, *60*, 794–810.

(49) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P. Estimation, computation, and experimental correction of molecular zero-point vibrational energies. *J. Phys. Chem. A* **2005**, *109*, 6779–6789.

(50) Kesharwani, M. K.; Brauer, B.; Martin, J. M. Frequency and zero-point vibrational energy scale factors for double-hybrid density functionals (and other selected methods): can anharmonic force fields be avoided? *J. Phys. Chem. A* **2015**, *119*, 1701–1714.

(51) Ruscic, B.; Pinzon, R. E.; von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoy, D.; Wagner, A. F. Active thermochemical tables: Thermochemistry for the 21st century. *J. Phys.: Conference Series* **2005**, *16*, 561–570.

(52) Scott, D. *Chemical thermodynamic properties of hydrocarbons and related substances: properties of the alkane hydrocarbons, C1 through C10, in the ideal gas state from 0 to 1500 K*; 1974.

(53) Bross, D. H.; Ruscic, B. ATcT enthalpies of formation based on version 1.124 of the Thermochemical Network. 2021; `ATcT.anl.gov`.

(54) Bross, D. H.; Ruscic, B. ATcT enthalpies of formation based on version 1.118 of the Thermochemical Network. 2015; `ATcT.anl.gov`.

Table 9: CBH-ANL ΔH_f(0 K)s, in kcal mol^{-1}, for hydroperoxy-alkyl radicals

SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)	SMILES	ΔH_f(0 K)
[CH$_2$]COO	15.59	CC(OO)[CH]CC	−2.29	OOC([C](C)C)C(C)C	−15.64
[CH$_2$]CCOO	10.69	CC(OO)C[CH]C	−3.01	OOC([CH]C(C)C)(C)C	*−14.84*
OOC([CH$_2$])C	7.96	CC(OO)CC[CH$_2$]	−0.30	OOC(C([CH$_2$])C)C(C)C	*−8.32*
C[CH]COO	8.58	CCC(OO)([CH$_2$])C	−3.04	OOC(C[C](C)C)(C)C	*−18.12*
[CH$_2$]CCCOO	6.89	CCC(OO)C[CH$_2$]	−0.37	OOC(CC([CH$_2$])C)(C)C	*−12.25*
[CH$_2$]C(OO)CC	4.47	OOC([C](C)C)C	−6.64	OOC(CC(C)C)([CH$_2$])C	*−10.52*
C[CH]CCOO	4.09	OOC(C([CH$_2$])C)C	−1.07	OOC[C](CC(C)C)C	*−11.60*
CC[CH]COO	5.23	OOC(C(C)C)[CH$_2$]	−0.29	OOCC([CH]C(C)C)C	*−9.52*
OOC([CH$_2$])(C)C	0.25	OOCC([CH$_2$])(C)C	0.33	OOCC(C[C](C)C)C	*−12.50*
CC(OO)[CH]C	0.93	OOCC[C](C)C	−3.20	OOCC(CC([CH$_2$])C)C	*−6.72*
CC(OO)C[CH$_2$]	3.11	OOCCC([CH$_2$])C	2.16	OOCC(CC(C)C)[CH$_2$]	*−6.57*
OOC[C](C)C	0.70	[CH$_2$]CC(COO)(C)C	−4.09	OOC([CH]C(C)(C)C)(C)C	*−18.20*
OOCC([CH$_2$])C	5.40	C[CH]C(COO)(C)C	−6.15	OOC(C([CH$_2$])(C)C)C(C)C	*−11.45*
[CH$_2$]C(CC)COO	2.34	CCC(COO)([CH$_2$])C	−2.80	OOC(C(C)(C)C)[C](C)C	*−20.66*
[CH$_2$]C(OO)CCC	0.95	OOC([C](C)C)(C)C	−13.25	OOC(C(C)(C)C)C(C)[CH$_2$]	*−11.47*
[CH$_2$]CC(OO)(C)C	−4.78	OOC(C([CH$_2$])(C)C)C	−5.76	OOC(CC([CH$_2$])(C)C)(C)C	*−15.07*
[CH$_2$]CCCCOO	3.31	OOC(C(C)(C)C)[CH$_2$]	−6.12	OOC(CC(C)(C)C)([CH$_2$])C	*−14.36*
C[C](CC)COO	−2.16	OOC(C(C)[CH$_2$])(C)C	−8.42	OOC[C](CC(C)(C)C)C	*−17.30*
C[CH]CCCOO	0.52	OOC(C(C)C)([CH$_2$])C	−7.30	OOCC([CH]C(C)(C)C)C	*−15.11*
CC([CH]C)COO	−0.67	OOC[C](C(C)C)C	−6.65	OOCC([CH]C(C)C)(C)C	*−14.45*
CC(C[CH$_2$])COO	1.47	OOC[CH]C(C)(C)C	−5.36	OOCC(C[C](C)C)(C)C	*−18.13*
CC[CH]CCOO	0.68	OOCC([C](C)C)C	−7.39	OOCC(CC([CH$_2$])(C)C)C	*−11.54*
CCC[CH]COO	1.40	OOCC(C([CH$_2$])C)C	−2.32	OOCC(CC([CH$_2$])C)(C)C	*−11.09*
OOC[CH]C(C)C	0.68	OOCC(C(C)C)[CH$_2$]	−2.30	OOCC(CC(C)(C)C)[CH$_2$]	*−11.69*
C[CH]C(OO)(C)C	−6.72	OOCCC(C)(C)[CH$_2$]	−2.87	OOCC(CC(C)C)([CH$_2$])C	*−10.46*
C[CH]C(OO)CC	−2.57				

Values in italics use the cc-pVDZ-F12 basis set rather than cc-pVTZ-F12.