JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

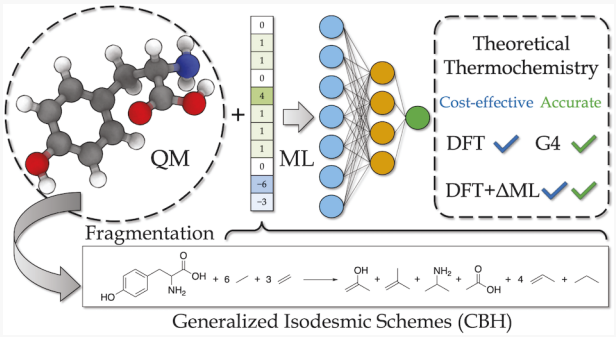
Effective Molecular Descriptors for Chemical Accuracy at DFT Cost: Fragmentation, Error-Cancellation, and Machine Learning

Eric M. Collins and Krishnan Raghavachari*

Cite This: J. Chem. Theory Comput. 2020, 16, 4938−4950 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Recent advances in theoretical thermochemistry have allowed the study of small organic and bio-organic molecules with high accuracy. However, applications to larger molecules are still impeded by the steep scaling problem of highly accurate quantum mechanical (QM) methods, forcing the use of approximate, more cost-effective methods at a greatly reduced accuracy. One of the most successful strategies to mitigate this error is the use of systematic error-cancellation schemes, in which highly accurate QM calculations can be performed on small portions of the molecule to construct corrections to an approximate method. Herein, we build on ideas from fragmentation and error-cancellation to introduce a new family of molecular descriptors for machine learning modeled after the Connectivity-Based Hierarchy (CBH) of generalized isodesmic reaction schemes. The best performing descriptor ML(CBH-2) is constructed from fragments preserving only the immediate connectivity of all heavy (non-H) atoms of a molecule along with overlapping regions of fragments in accordance with the inclusion−exclusion principle. Our proposed approach offers a simple, chemically intuitive grouping of atoms, tuned with an optimal amount of error-cancellation, and outperforms previous structure-based descriptors using a much smaller input vector length. For a wide variety of density functionals, DFT+ΔML(CBH-2) models, trained on a set of small- to medium-sized organic HCNOSCl-containing molecules, achieved an out-of-sample MAE within 0.5 kcal/mol and 2σ (95%) confidence interval of <1.5 kcal/mol compared to accurate G4 reference values at DFT cost.



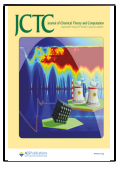
1. INTRODUCTION

Chemical accuracy has been long sought after in theoretical thermochemistry.[1−19] Unfortunately, highly accurate quantum mechanical (QM) methods to achieve this accuracy, such as complete basis set CCSD(T) or G4 composite method,[6] are too computationally demanding for most medium- to large-sized molecules. Despite the new developments in cost-effective density functional theory (DFT) methods, the accuracy standards (∼1 kcal/mol) are not close to being met. Alternative computational strategies have been developed for these methods, exploiting systematic errors in computational thermochemistry (*viz.*, error-cancellation schemes),[7,18,19] and are often used to achieve acceptable accuracies from inexpensive methods (e.g., DFT). Of particular interest are the use of *generalized isodesmic reaction schemes*, based around extensions of the original isodesmic bond separation scheme.[18] The central idea of the bond separation scheme is to calculate the enthalpy of formation of a larger molecule by first extracting all heavy-atom bonds into their simplest stable forms and then combining the corresponding calculated change in energy, in conjunction with their highly accurate experimental enthalpies of formation, through Hess's Law. Isodesmic schemes, particularly used in their more recent generalizations, have been shown to be relatively method-independent and can yield high quality reaction energies from basic methods such as DFT. These generalized isodesmic schemes are constructed in such a way that the environment of atoms and bonds is consistent on both sides of a reaction, effectively canceling out any intrinsic systematic errors from more approximate methods.[19]

2. MACHINE LEARNING MODELS IN THERMOCHEMISTRY

Simultaneously, machine learning (ML) has been added to the quantum chemical toolbox,[8,11,14,16,20−33] leading to a significant decrease in the computational cost and/or increase in the accuracy of the corresponding calculated properties. The success of a given ML model depends on its chosen set of molecular descriptors, as the representation must fully describe patterns in the desired output values. Choosing a good representation for a molecule can be fairly challenging as these patterns are not

Received: March 10, 2020
Published: July 17, 2020



https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

known and may be difficult for humans to recognize. Nonetheless, a plethora of molecular descriptors has been previously developed (*vide infra*), and new developments in models now scan over thousands of available descriptors to find the best combination for the problem at hand.

2.1. Machine Learning − ΔML Models. Through ideas somewhat analogous to those from error-cancellation schemes, neural network-based machine learning correction (ΔML) models have been developed, combining low cost QM calculations with the machine learning framework.[14] Additionally, these ideas have been generalized into a multilevel combination technique in the quantum machine learning realm referred to as CQML.[33] In this scheme, any number of levels of theory can be used together with varying training data, aiming to learn the patterns between many different levels of theory.

Almost all currently available molecular descriptors, however, have been created to represent patterns in the *absolute values* of the properties of molecules. In the ΔML model and generalized CQML models, such patterns are not necessarily present in the *difference between two levels of theory*. In comparison, recognizing and correcting these patterns are routinely done with the error-cancellation of fragmentation-based methods. Thus, well established techniques in fragmentation can be applied directly into the ΔML framework.

ML models have shown promise for applications in large scale drug design screening,[34−36] retrosynthetic planning,[37,38] and designing new materials,[39−41] though large and chemically diverse data sets are necessary to ensure model generality. Typically, experimentally derived data sets are either too small or too sparse, so ML models are often trained to reproduce QM calculations. With the currently available computing power, moderately accurate properties of thousands of molecules can be calculated in a reasonable time frame with DFT. From these databases, ML models have been developed to predict properties within a few kcal/mol of DFT. However, the accuracy of DFT depends on the functional employed, typically ranging from 3 to 15 kcal/mol for the calculation of thermochemical properties.[42] For many thermochemical applications, such as the evaluation of combustion processes for the design of new fuels[43,44] or detection of new structure−property relationships (SAR), this accuracy is often inadequate when "chemical accuracy" (∼1 kcal/mol) is the ultimate target. Additionally, as pointed out in ref 14, a high accuracy is often needed for the determination of reaction rates, for processes such as catalysis, since they depend exponentially on energy differences.[14]

Achieving the chemical accuracy threshold with ML is a nontrivial task. First of all, a large enough data set of reliable data must be collected to cover the scope of the problem. Databases of highly accurate experimental data have been developed[6,15,45] but are currently insufficient for ML models to effectively learn patterns from the data. Alternatively, these databases could be purely computational, but gold-standard ab initio methods such as CCSD(T)/CBS, required to achieve such accuracy, are still too computationally prohibitive to construct a database of comparable size to that of DFT.

Herein, we propose a hierarchy of molecular descriptors tuned for systematic error cancellation, based solely on the local connectivity within the molecule, and using the generalized isodesmic reaction framework from the Connectivity-Based Hierarchy (CBH, *vide infra*).

2.2. Machine Learning − Related Work. ML-based thermochemical predictions have been of substantial interest in the past decade.[8,27,30−32,46,47] Large data sets for training ML models are typically required to ensure that a model is general and not overfitted. Data sets covering all of chemical space have gained popularity in this context, e.g., the General Database of every hypothetically *feasible* HCNOSCl-containing molecules with up to 13 heavy atoms (GDB-13)[46] spanning 977 million molecules, though no associated properties are included. Two very popular subsets of this database are the GDB-7 data set of 7k molecules and the GDB-9 data set of 134k molecules, featuring GDB-13 molecules containing up to seven or nine heavy atoms, respectively. Later work aimed at standardizing these data sets for training ML models resulted in a collection called QM9,[32] including a variety of DFT calculated properties of the GDB-9 set and a more recent expansion to include G4(MP2) computed thermochemical properties.[47] Various molecular descriptors have been developed based on these data sets, including Coulomb Matrix (GDB-7 MAE = 10 kcal/mol),[8] Bag of Bonds (GDB-7 MAE = 1.5 kcal/mol, GDB-9 MAE = 2.0 kcal/mol),[27] Encoded Bonds (GDB-9 MAE = 1.5 kcal/mol),[30] and Bonds in Molecules (GDB-9 MAE = 0.94 kcal/mol).[31] Although these models are seemingly approaching the "target accuracy" of ±1 kcal/mol, the accuracy of the DFT calculated the QM9 values was estimated in the original paper, through a subset of 100 randomly drawn molecules, giving an average error of 4.9 kcal/mol compared to G4 reference values.[32] Thus, models trained on QM9 values in their current state only achieve a target 4 to 5 times larger than the chemical accuracy threshold and are insufficient for experimental quality thermochemical properties.

One of the top performing approaches for more accurate ML-based thermochemistry, the Δ_b^t-ML model, predicts the difference between two levels of theory referred to as a baseline (b) and target (t).[14] If successful, the target accuracy (typically that from a more expensive calculation) can be achieved at the cost of the (cheaper) baseline. For example, the initial application $\Delta_{B3LYP}^{G4(MP2)}$-ML employing the Coulomb Matrix descriptor achieved an average accuracy of <1 kcal/mol for a set of 6k constitutional isomers of $C_7H_{10}O_2$ with a training size of 1k molecules.[14] The similar model Δ_{PM7}^{B3LYP}-ML improved the accuracy of the semiempirical method PM7 from 7.2 kcal/mol to ∼3 kcal/mol for 134k HCNOF molecules with 9 or fewer heavy atoms. Recently, newer, but more complicated, $\Delta_{B3LYP}^{G4(MP2)}$-ML models employing state-of-the-art ML strategies, *viz.*, SchNet[48] and FCHL,[49] have been used to obtain impressive MAEs (within 0.3 kcal/mol) for molecules containing up to 14 heavy atoms, though only with the use of millions of parameters trained using the three-dimensional structure of the molecule.[50]

Our proposed framework strives to more efficiently encode the chemical environments present in a molecule specifically designed to perform well in the Δ-ML regime with a *much simpler neural network benefiting from the ideas from error-cancellation and fragmentation*. The present study could serve as a starting point for further developments to more complex models utilizing both advances in cost-effective quantum chemistry calculations as well as machine learning. Herein, our models target experimental quality results offering a few significant improvements over most current thermochemical ML models:

1. A data set built from experimentally known (HCNOSCl) molecules
2. More accurate target values (G4 opposed to DFT)

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

3. Lower average errors (<0.5 kcal/mol) for a wider set of density functionals

2.3. Machine Learning − Data Set. The present work focuses on producing accurate thermochemical values, *viz.*, standard enthalpies of formation, ΔH_f(298 K), of real, neutral, organic HCNOSCl-containing molecules for which experimental values are available. Unfortunately, several of the largest collections of experimental properties have large or unknown uncertainties associated with them. The Active Thermochemical Tables (ATcT)[45] has made an effort to construct an accurate compilation of thermochemical data into a self-consistent network by means of a critical evaluation of competing measures and all reaction pathways that interrelate chemical species. While the ATcT provides the best currently available enthalpies of formation, in its current version, it only has 33 neutral organic molecules with 4 heavy atoms or larger. One of the most popular databases, the NIST Chemistry WebBook,[51] contains information for over 6000 species. However, it gives an overview of relevant properties found in the literature rather than recommending a certain value, many times reporting multiple values from competing sources for the same molecule.[17] Thus, these values cannot be blindly trusted, and, therefore, in this work, reference values have been computed using the composite method Gaussian-4 (G4).[6]

All ΔH_f(298 K) were calculated using the atomization method.[1] First the total atomization energy at 0 K (TAE_0) is computed using calculated zero-point-corrected energies of neutral isolated atoms and the full molecule, as illustrated in eq 1. Note: all atomic properties in eqs 1−3 correspond to neutral isolated atoms at their ground state multiplicities.

$$\mathrm{TAE}_0(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 0\ \mathrm{K}) = x\mathrm{E}_0(\mathrm{C}, 0\ \mathrm{K}) + y\mathrm{E}_0(\mathrm{H}, 0\ \mathrm{K}) + z\mathrm{E}_0(\mathrm{O}, 0\ \mathrm{K}) - \mathrm{E}_0(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 0\ \mathrm{K}) \quad (1)$$

Atomization energies can then be converted directly to ΔH_f(0 K) and ΔH_f(298 K) via eqs 2 and 3, using well-known experimental atomic enthalpies of formation and thermal corrections $H^{298K}-H^{0K}$.

$$\Delta H_f(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 0\ \mathrm{K}) = x\Delta H_f(\mathrm{C}, 0\ \mathrm{K}) + y\Delta H_f(\mathrm{H}, 0\ \mathrm{K}) + z\Delta H_f(\mathrm{O}, 0\ \mathrm{K}) - \mathrm{TAE}(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 0\ \mathrm{K}) \quad (2)$$

$$\Delta H_f(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 298\ \mathrm{K}) = \Delta H_f(\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z, 0\ \mathrm{K}) + [\mathrm{H}^{298K} - \mathrm{H}^{0K}](\mathrm{C}_x\mathrm{H}_y\mathrm{O}_z) - x[\mathrm{H}^{298K} - \mathrm{H}^{0\,K}](\mathrm{C}) - y[\mathrm{H}^{298K} - \mathrm{H}^{0\,K}](\mathrm{H}) - z[\mathrm{H}^{298K} - \mathrm{H}^{0\,K}](\mathrm{O}) \quad (3)$$

Atomization energies are the most elementary quantity computationally as a molecule is broken into the corresponding isolated atoms. Accurate calculated TAE_0 values are challenging for approximate methods due to the influence of secondary effects such as core−valence correlation, scalar relativistic effects, and spin−orbit coupling.[3] Highly accurate composite methods have been developed to capture these effects but are much more computationally expensive. To overcome these limitations, alternative strategies to the atomization approach have been developed, utilizing error-cancellation inherent in the isodesmic reaction schemes, to achieve a more acceptable accuracy for enthalpies of formation. In these approaches, the reaction enthalpy of a given isodesmic reaction is calculated at a low level of theory, and then the ΔH_f of the full molecule is obtained by using experimental values for all the reactant and product fragment species. Since this method relies heavily on accurate experimental fragment ΔH_f, only reactions with small product fragments can be used.

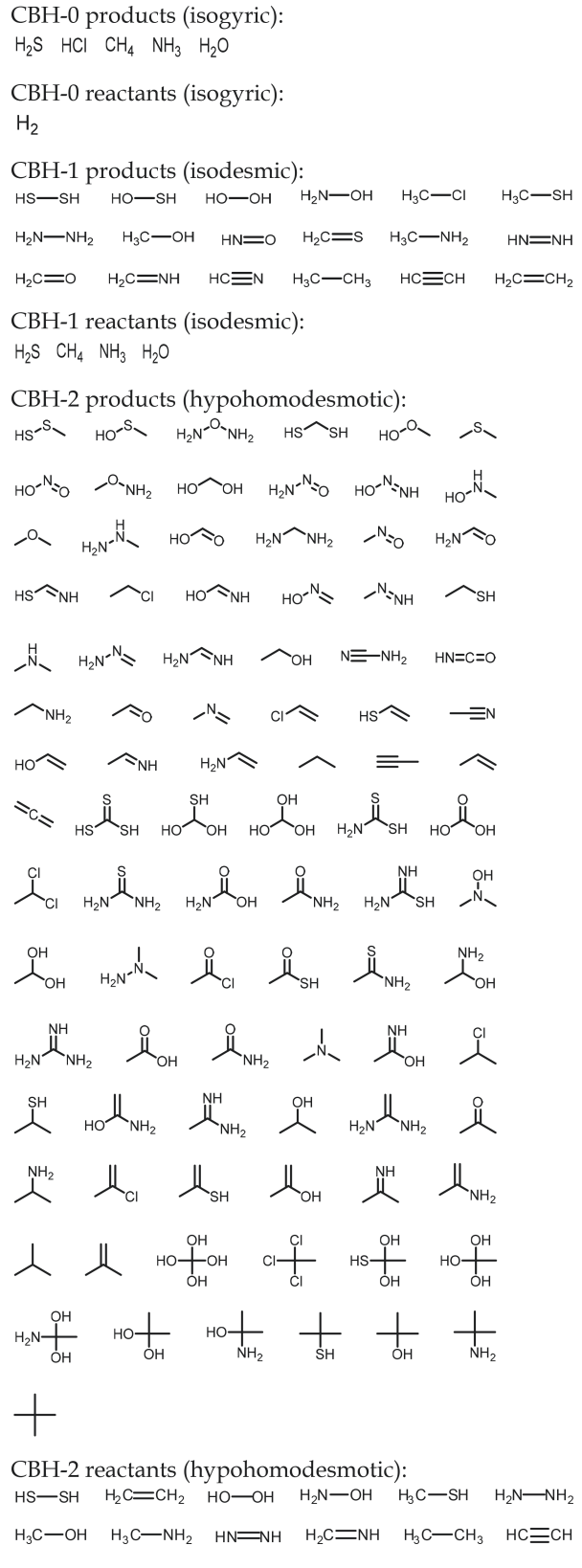


Figure 1. All possible CBH-0, -1, -2 fragments for the 1k-G4-C9 data set.

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

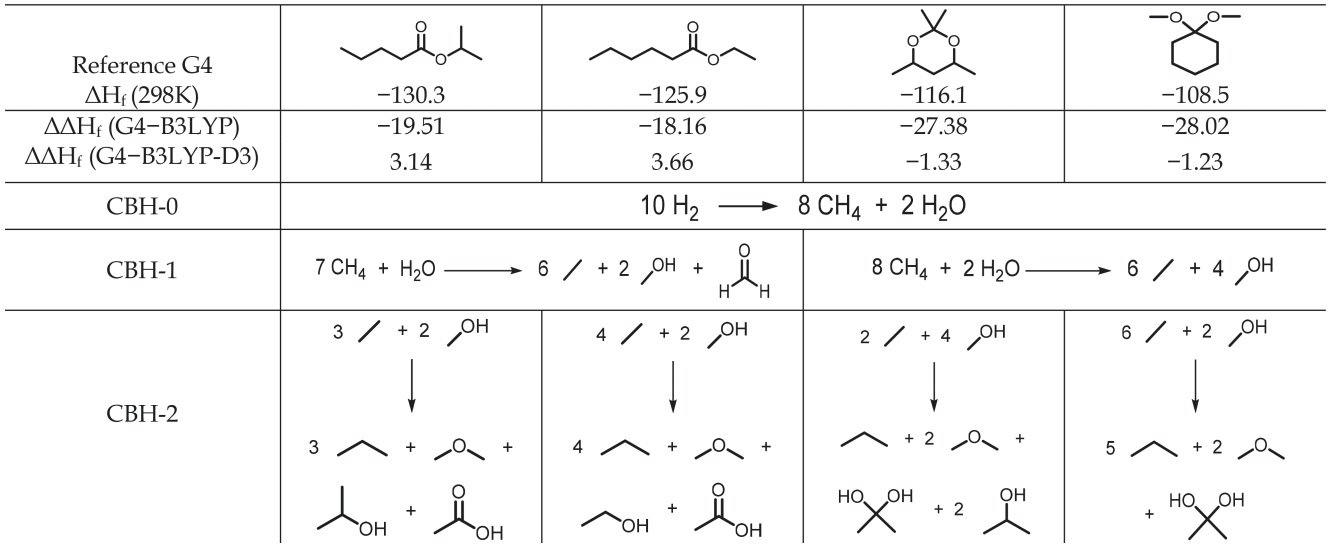


Figure 2. CBH-0, -1, and -2 fragmentation schemes for four compositionally similar molecules along with corresponding G4 ΔH_f(298 K) and baseline errors for B3LYP and B3LYP-D3(BJ).

Table 1. Comparison of Full 1k-G4-C9 Mean Absolute Errors (MAE) in kcal mol^{-1} for Both ML and B3LYP+ΔML Models with Various Molecular Descriptors

molecular descriptor (length of input vector)	ML	B3LYP+ΔML
null model	58.47	14.98
geometry-based		
Coulomb Matrix (33)	31.87	2.77
Bag of Bonds (896)	3.82	0.81
molecular graph-based		
atom counts (5)	16.40	3.69
bond counts (18)	14.53	1.80
rank-3 (92)	10.97	1.43
connectivity counts (92)	7.89	1.19
isodesmic (CBH)-based		
CBH-0 (isogyric) (6)	12.63	2.75
CBH-1 (isodesmic) (22)	6.99	1.65
CBH-2 (hypoHD) (103)	4.02	0.62

Reference ΔH_f(298 K) values calculated with the G4 composite method have been shown to be within 1 kcal/mol of experiment on average.[6] The data set used in this work was adapted from the training set used to develop the Atom Pair Contribution (APC) method,[52] which was compiled from various sources listed in the NIST Chemistry WebBook.[51] G4 computed values are an approximation to the "gold standard" CCSD(T)/CBS level of theory but are significantly more computationally intensive for larger molecules compared to DFT. Another key factor is that most previous ML training sets have typically been truncated based on the number of heavy atoms. However, in an attempt to group molecules with similar (carbon) backbone structures, but varying numbers of heavy atoms (from different functional groups), we have organized the full APC database[52] by the number of carbon atoms rather than by the number of heavy atoms. This can be beneficial as there will be more overlap between the molecular descriptors of the full data set. For demonstrative purposes, the full data set was made more computationally feasible by creating a subset termed "1k-G4-C9" consisting of 1051 HCNOSCl-containing molecules with 9 or fewer carbon atoms.

2.4. Machine Learning − Baseline. In the ΔML regime, the ML model is trained on the difference between a high level (e.g., G4) and a low level of theory (DFT). The corresponding ΔML correction should improve the performance of the baseline (low level of theory), and a correction of zero (null model) corresponds to the performance of the low level of theory. Since the molecular descriptors employed here are based on the connectivity in the molecule, nonbonded interactions are not represented, and the low level is fully responsible for modeling these effects. In addition, relative conformer energies mirror the performance of the low level of theory, i.e., our model will return the same correction for two conformers with the same connectivity. Within these limitations, a wide variety of DFT methods were tested as the baseline using the 6-311++G(3df,2p) basis set. All low-level and G4 calculations were performed with the Gaussian 16 program suite.[53]

2.5. Machine Learning − Neural Network Architecture. Since this study is focused on the utility of our molecular descriptors on a simple architecture, the standard scikit-learn[54] implementation of a feed-forward neural network with one hidden layer was used for all models. Hyperparameter values for L2-regularization and the initial learning rate for the ADAM optimizer were chosen through the standard 5-fold cross-validation after an 80:20 split of the full 1k-G4-C9 data set. Hyperparameters and number of nodes in the hidden layer were kept constant across all models using the same input variables for a direct comparison of various low levels. Further details about the hyperparameter search, cross-validation, and other calculations can be found in the Supporting Information.

2.6. Machine Learning − Features. As pointed out in previous studies on constant size descriptors,[30] the length of the input vector for neural networks employing popular geometry-based molecular descriptors, such as Coulomb Matrix (CM) and Bag of Bonds (BoB), scales with the size of the system, becoming too prohibitive for large molecules. These molecular descriptors are training set-dependent, i.e., the largest molecule in the training set will dictate the size of the input vector, with smaller molecules appending zeroes to their input vectors to match sizes. Molecular graph-based descriptors have been previously proposed, such as atom counts (rank-1), bond counts (rank-2), and larger graph-based connectivity counts (rank-3, rank-4,

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

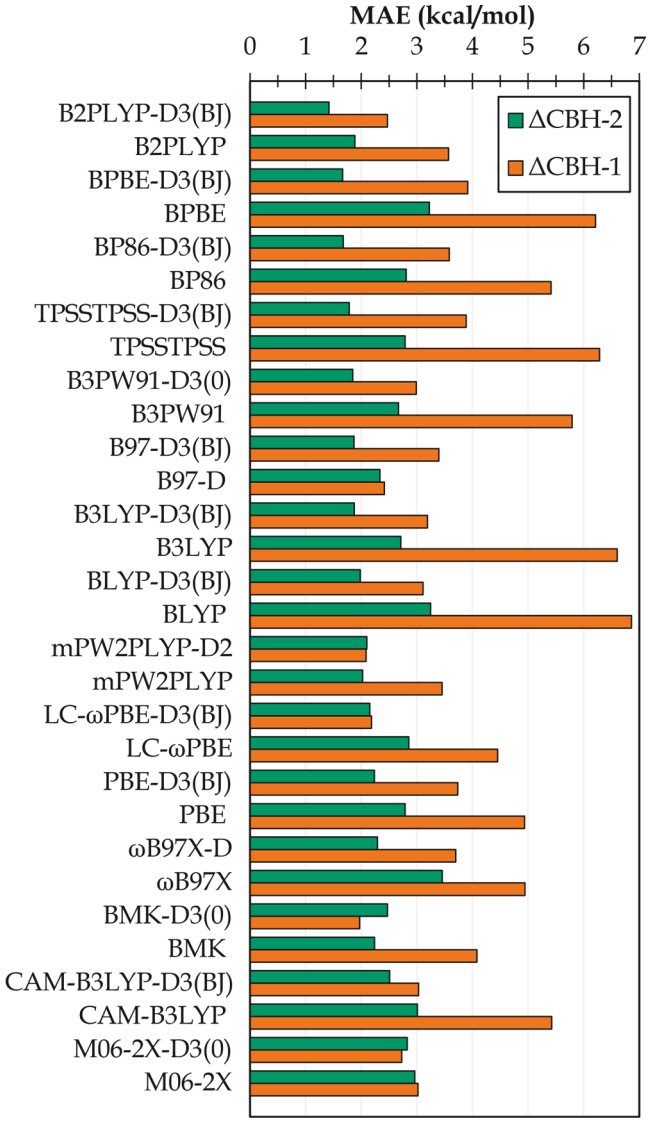


Figure 3. Performance of traditional ΔCBH-1 and ΔCBH-2 corrections on the full 1k-G4-C9 data set in kcal mol^{-1}.

rank-5, etc.), where the molecule is broken into fragments and the corresponding length of the feature vector is every possible atom, bond, or combination of 3+ connected atoms. The new set of molecular descriptors introduced here offers a different approach to these molecular graph-based groupings, wherein similar amounts of the environment are preserved for each heavy atom for optimal error-cancellation while maintaining a *constant input length independent of system size*.

2.7. Machine Learning – Descriptors from Connectivity-Based Hierarchy (CBH). Isodesmic bond-separation schemes and their generalizations have a long history in quantum chemistry as error-balancing reactions that can be utilized for theoretical thermochemistry to calculate more accurate enthalpies of formation from computationally inexpensive methods. The idea behind these schemes is to form a reaction for a given molecule in which the local environment is preserved on the product and reactant sides. In doing so, the inherent systematic errors are mostly canceled out, allowing higher accuracy at a reduced computational cost. The original isodesmic bond-separation scheme, proposed in the 1970s,[18] preserves the numbers of each type of bond on both sides of the reaction, where the product side is a compilation of fragmented heavy-atom bonds (rank-2 type groups) in a molecule, and the reactant side contains the full molecule along with overlapping fragments of the bonds (rank-1 type groups) to balance the reaction. For larger molecules, more sophisticated schemes are required for acceptable accuracies. Many such methods have subsequently been developed including the homodesmotic,[9,55] semihomodesmotic,[56] isogeitonic,[5] and homoplesitoic[57] schemes. Each of these reactions were developed by manually matching of bond and hybridization types and have varying amounts of chemical environment preservation on the product and reactant sides of the reaction. However, these developments did not include a generalizable method for all functional groups nor an automated way to generate the reactions. To overcome such limitations and to make such schemes applicable for all organic species, Raghavachari and co-workers developed the Generalized Connectivity-Based Hierarchy (CBH) of error-cancellation schemes.[7,19] The hierarchy overcomes the disadvantages of other related schemes by providing a simple, yet reliable approach, bypassing any manual balancing of bond and hybridization types. CBH organizes these types of reactions into rungs (referred as CBH-*n*) of alternating atom- and bond-centered reactions starting from the isogyric (CBH-0) and isodesmic (CBH-1) reaction schemes, preserving the number of each atom and the number of each bond, respectively. Subsequent CBH-*n* rungs include immediate connectivity to all CBH-(n−2) type fragments, e.g., CBH-2 (hypohomodesmotic) includes all connectivity to each heavy atom (CBH-0 fragments), and CBH-3 (hyperhomodesmotic) includes all connectivity to all heavy-atom bonds (CBH-1 fragments). Additionally, the overlapping fragments on the reactant side are fragments from the previous rung, establishing an extended relationship between the rungs and allowing for easy automation of the hierarchical reactions.

The proposed hierarchy of molecular descriptors is based on the rungs of CBH, including both revised versions of the molecular graph-based counts stemming from the connectivity (corresponding to CBH product fragment coefficients) as well as the full CBH coefficients (both product and reactant fragments). Product counts are similar to previously used molecular graph-based representations and called atom (rank-1), bond (rank-2), and connectivity (mix of rank-3, rank-4, and rank-5) counts. Atom and bond counts are identical to the previous connectivity-based molecular descriptors, without including bonds to hydrogen. However, in connectivity counts, the immediate connectivity of each nonterminal heavy (non-H) atom is preserved and collected as a group. Groups of this type can be as small as 3 heavy atoms (e.g., propane) or as large as 5 heavy atoms (e.g., neopentane) but with a maximum chain length of 3. In contrast, in the rank-*n* approach, every combination of *n* atoms is included, losing information about the immediate connectivity.

Full CBH reaction molecular descriptors are the fragment coefficients from the *product and reactant sides* of the CBH-0, CBH-1, and CBH-2 reactions. Compared to the rank-*n* nomenclature, CBH-0 is a rank-1 type descriptor along with information about the total number of heavy atom bonds (from capped hydrogens balanced in the form of H_2), CBH-1 is a combination of rank-1 and rank-2 type descriptors, and CBH-2 is a combination of *specific* ranks -2, -3, -4, and -5. Each parent molecule is represented by an input vector with the length of every possible fragment of a given CBH rung for the whole data set. Each unique fragment type is one-hot encoded and then

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

Table 2. Out-of-Sample Performance of ΔML(CBH) Models Using the Original 80:20 Train−Test Split Compared to Traditional ΔCBH-2 on the Full 1051 Molecule Set[a]

	mean absolute error in kcal mol^{-1}			
baseline	ΔCBH-2	ΔML(CBH-0)	ΔML(CBH-1)	ΔML(CBH-2)
B2PLYP-D3(BJ)	1.42	1.34	0.76	**0.78**
B2PLYP	1.88	1.56	1.01	**1.04**
BPBE-D3(BJ)	1.67	1.71	0.96	0.87
BPBE	3.22	2.33	1.76	1.20
BP86-D3(BJ)	1.68	1.59	0.93	**1.81**
BP86	2.81	2.20	1.49	**1.68**
TPSSTPSS-D3(BJ)	1.78	2.06	1.15	0.90
TPSSTPSS	2.79	2.59	1.68	0.99
B3PW91-D3(0)	1.85	1.67	0.87	0.63
B3PW91	2.67	2.24	1.53	0.90
B97-D3(BJ)	1.87	2.30	1.14	0.90
B97-D	2.34	2.12	1.22	1.06
B3LYP-D3(BJ)	1.88	1.99	0.83	0.67
B3LYP	2.71	2.61	1.61	1.04
BLYP-D3(BJ)	1.98	2.56	1.12	1.02
BLYP	3.25	3.66	1.93	1.56
mPW2PLYP-D2	2.10	1.20	0.67	**0.90**
mPW2PLYP	2.02	1.55	0.97	0.90
LC-ωPBE-D3(BJ)	2.15	2.51	0.80	0.54
LC-ωPBE	2.86	2.68	1.09	0.89
PBE-D3(BJ)	2.24	2.02	1.07	**1.39**
PBE	2.79	2.20	1.49	**1.70**
ωB97X-D	2.29	1.45	0.83	0.54
ωB97X	3.46	1.49	0.99	0.60
BMK-D3(0)	2.47	1.79	0.75	**0.98**
BMK	2.24	1.71	1.04	0.71
CAM-B3LYP-D3(BJ)	2.51	1.53	0.89	0.64
CAM-B3LYP	3.01	1.86	1.19	0.75
M06-2X-D3(0)	2.83	1.30	0.81	0.56
M06-2X	2.96	1.34	0.80	0.56

[a]Bolded values highlight models which do not follow expected performance trends from the traditional CBH-0 to CBH-2 correction schemes; see text for details.

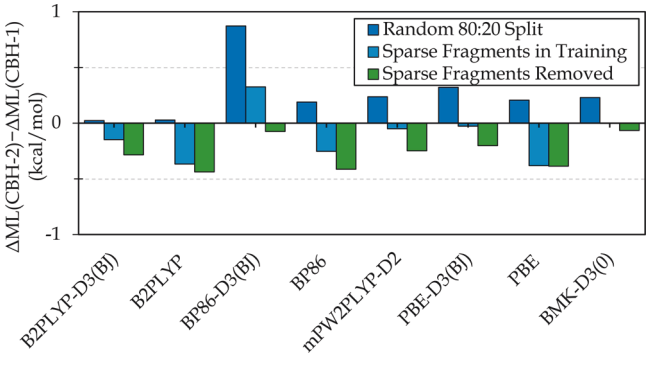


Figure 4. Effect of sparse fragments and unbalanced data on the increase in performance between ΔML(CBH-1) and ΔML(CBH-2).

combined based on the CBH reaction of the parent molecule resulting in a sparse vector of mostly zeroes except for the coefficients of fragments present in the reaction. These fragments are given in Figure 1 for all CBH rungs used here. The uniqueness of input vectors dictates a model's best possible performance (irreducible error). For representations such as the sorted eigenvalues of the Coulomb Matrix or Bag of Bonds, this irreducible error is only nonzero in special cases, e.g., homometric molecules.[8] Conversely, all graph-based descriptors and the coefficients of the CBH reactions are encoded as discrete values and are not completely unique among molecules, especially at lower rungs of CBH.

As an illustration, the CBH-0, -1, and -2 reactions for four molecules are given in Figure 2 along with their enthalpy of formation (ΔH_f) and G4-baseline null error. These molecules have the same number of each heavy atom, as well as the same number of bonds, causing the CBH-0 representations to be identical. Any model will produce identical outputs for identical inputs meaning using the CBH-0 representation introduces some unavoidable error. The predicted ΔH_f for ML models (without a baseline) which will minimize the error is simply the average of the four values (−120.2 kcal/mol), giving >10 kcal/mol deviation for two of the molecules. Moving up one rung, CBH-1 provides uniqueness into two groups, while moving to CBH-2 uniquely identifies all four molecules, decreasing the error further. Additionally, to further point out the utility of the ΔML model, $\Delta\Delta H_f$ (G4−baseline) deviations included in Figure 2 are much closer to each other than the corresponding absolute formation enthalpies. For the B3LYP baseline, the CBH-0 has a maximum deviation of just under 5 kcal/mol from the average value, while the maximum for B3LYP-D3(BJ) is around 2 kcal/mol. Since molecular descriptors are constant between baselines, irreducible errors are a direct representation

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

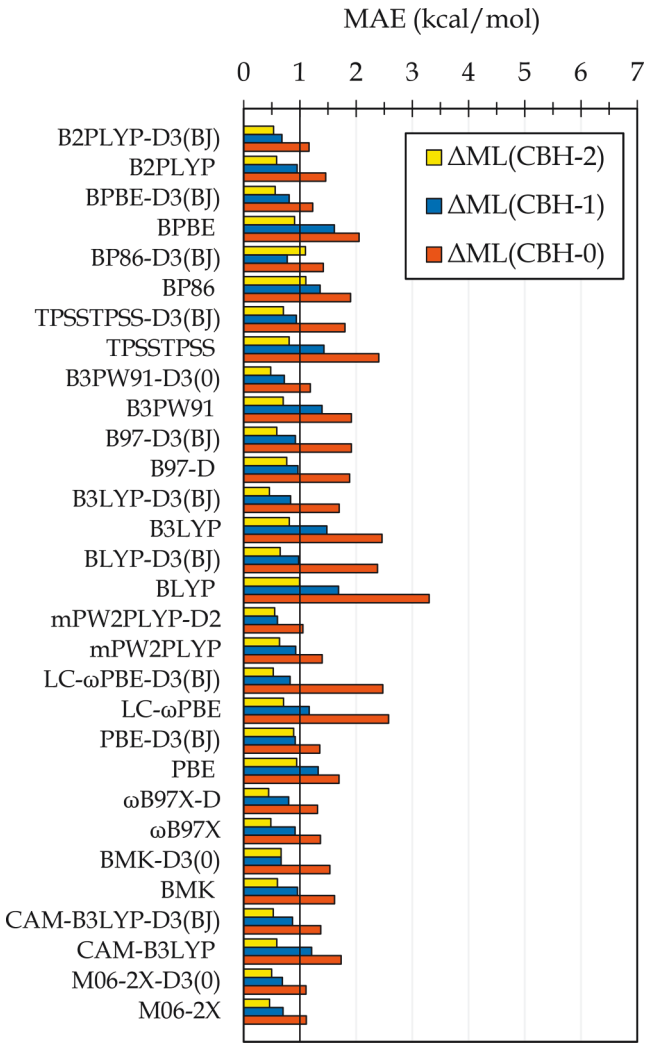


Figure 5. Final out-of-sample performance for all DFT+ΔML(CBH) models across 30 DFT baselines.

of how systematic the errors are between a given baseline and the target values.

3. RESULTS AND DISCUSSION

For a simplified overview of the performance of various molecular descriptors, the mean absolute errors of the full 1051 molecule data set will briefly be compared, while all further error comparisons in this report will be out-of-sample (test set) mean absolute errors (OOS-MAE) in kcal/mol.

Table 1 compares the performance of both ML and B3LYP +ΔML type models for the full 1051 molecule data set. Previously used descriptors Coulomb Matrix (CM),[8] Bag of Bonds (BoB),[27] and rank-3 graph counts were generated using MolML[30] and are included for comparison against the proposed molecular descriptors. Bag of Bonds was the top overall performing standard machine learning model (no baseline) with an MAE of 3.82 kcal/mol, followed by ML(CBH-2), MAE of 4.02 kcal/mol. However, in out-of-sample molecules, BoB had an MAE of 8.48 kcal/mol compared to 5.02 kcal/mol for ML(CBH-2). This increase in variance could be due to the increased length of the input vector, perhaps indicating some amount of overfitting has occurred. Adding a DFT baseline (ΔML) significantly improves the model performance, indicating certain molecular descriptors tested are a function of the systematic errors present in DFT rather than a pure representation of the enthalpy of formation. Significant improvements are also seen on going from graph-based counts to their corresponding CBH-based molecular descriptors (Table 1), though the input vector length is only slightly increased. Since isodesmic reactions are inherently designed to cancel structure-based systematic errors in low levels of theory, namely DFT, the best performing models for ΔML are expected to be the CBH-based ones. Connectivity counts provide the lowest error out of any count-based molecular descriptor with an overall MAE of 1.19 for B3LYP+ΔML. Coincidentally, the rank-3 and connectivity counts descriptors have the same input vector length. The rank-3 descriptor here enumerates all 3 heavy atom combinations taking bond orders (single, double, aromatic, and triple) into account. However, since connectivity counts are based on a cut-and-cap fragmentation scheme, there cannot be a bond order of 1.5, thus all CBH-based approaches are from an aromatic resonance structure of alternating single and double bonds.

Overall, the descriptors based on the isodesmic reactions outperformed their molecular graph-based counterparts, with the ΔML(CBH-2) approach performing best for ΔML (MAE = 0.62 kcal/mol). CBH-based descriptors have the benefit of being derived directly from the heavy-atom connectivity rather than number of atoms in a group and also encode the relationship between heavy-atom fragments through the smaller overlapping fragments, leading to an overall better description of the chemical environments present in a molecule. The CBH-2 approach also benefits from including certain rank-4 and -5 groups. Since full rank-4 and -5 representations are based on every grouping of 4 and 5 atoms, these descriptors scale more harshly, having an input vector size of 238 and 506, respectively, for this 1051 molecule data set. Thus, the improved performance using our proposed isodesmic-based molecular descriptors is due to the more chemically intuitive grouping of atoms based on the structure rather than all possible groupings.

3.1. Traditional Fragmentation through CBH.

Traditionally, the Connectivity-Based Hierarchy has been used as a fragmentation-based correction scheme for theoretical thermochemistry. CBH has been applied to achieve accurate thermochemical values of neutral,[7,58] radical,[12] and cationic[59] organic molecules as well as biomolecular[13,60] systems. Additionally, the method has been applied to reaction energies,[61] bond dissociation energies,[62] pK_a calculations,[63] and redox potentials.[64] The correction method works by first constructing the reaction of a given rung of CBH and then calculating the energy of each of the fragments at a low level baseline (b) and the higher target (t) level of theory. ΔCBH correction terms are then constructed as a sum of the differences in the isodesmic reaction energies (i.e., target−baseline) and added to the full molecule calculation at the low level of theory.

$$E^{t}_{Full} \approx E^{b}_{Full} + \Delta CBH_n = E^{b}_{Full} + \sum (E^{t}_{frag} - E^{b}_{frag}) \quad (4)$$

One of the major benefits of traditional CBH is the ability to construct corrections to a wide range of density functionals. Explicit G4 and DFT fragment calculations were performed to construct CBH-0, CBH-1, and CBH-2 corrections for the full 1k-G4-C9 data set and then added to the enthalpy of formation calculated with the corresponding density functional according to eq 4. CBH-0-corrected enthalpies are conceptually similar to the atomization method as both methods extract all heavy atoms of a molecule with the only difference being the open valence of

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

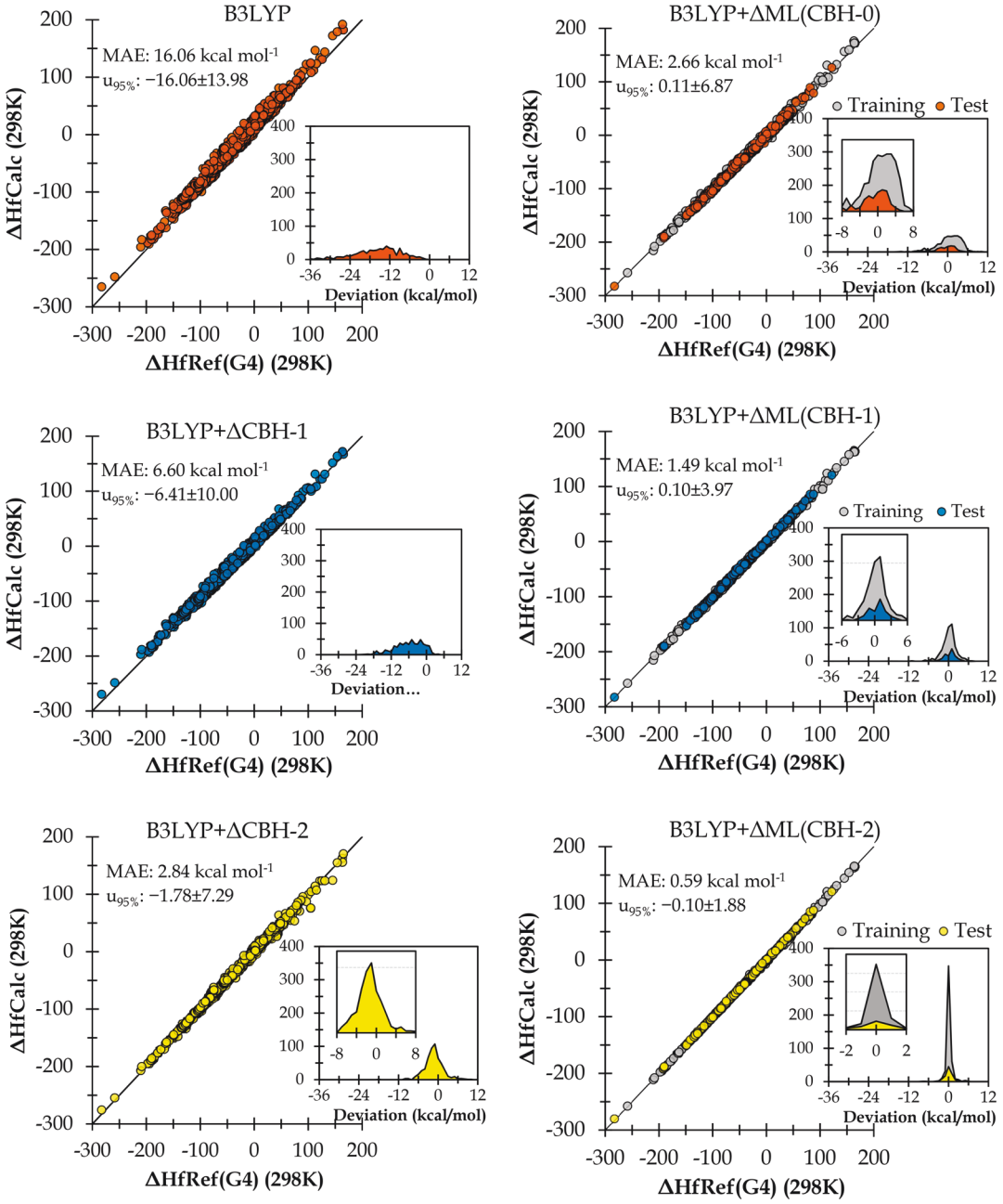


Figure 6. Performance of ΔCBH-corrected and ΔML-corrected ΔH_f(298 K) compared to reference G4 (all values are given in kcal/mol).

heavy atom fragments in CBH-0 is terminated with hydrogens. Thus, CBH-0 provides only a slight improvement over the baseline and is not reported herein. Mean absolute errors for all CBH-1- and CBH-2-corrected enthalpies of formation are summarized in Figure 3.

Dispersion-corrected functionals are grouped with their nondispersion-corrected counterparts and then sorted based on CBH-2 performance of the dispersion-corrected functional. This is the largest CBH benchmark to date, in both number of functionals and number of data points; however, all trends previously reported hold true.

The most prominent trend, seen even before the introduction of the Connectivity-Based Hierarchy correction method, is that excellent performance occurs starting at the hypohomodesmotic (CBH-2) reaction, providing a significant improvement over the isodesmic (CBH-1) reaction scheme. The top performing method using CBH-2 corrections is B2PLYP-D3(BJ) giving an average error below 1.5 kcal/mol, followed by 8 other density functionals with errors under 2 kcal/mol. CBH-1 fragments are at maximum two heavy atoms in size and can cause a severe mismatch of bond types for aromatic systems or in molecules containing nonlocal effects such as hyperconjugation, protobranching,[65] or charge delocalization.[59] Additionally, increasing the fragment size should cause the extrapolated energy to monotonically converge toward the high level of theory. Thus, cases where CBH-1 outperforms CBH-2 are likely due to fortuitous cancellation of errors at the CBH-1 rung. Moreover, results here confirm the recently highlighted importance of the treatment of dispersion in computational thermochemistry. Since CBH provides a local correction based on bonded interactions, long-range effects and intramolecular interactions

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

Table 3. Error Statistics (kcal/mol) for ΔML(CBH) Models for Validation, Training Set, and Test Set

				mean absolute error	
baseline and molecular descriptor	irreducible error	bias	variance	validation	test set
B3LYP					
ΔML(CBH-0)	1.59	1.12	0.25	2.80	2.46
ΔML(CBH-1)	0.75	0.74	0.01	1.85	1.48
ΔML(CBH-2)	0.10	0.44	0.27	1.17	0.81
B3LYP-D3(BJ)					
ΔML(CBH-0)	1.03	0.94	0.27	2.00	1.70
ΔML(CBH-1)	0.37	0.52	0.05	1.08	0.83
ΔML(CBH-2)	0.04	0.30	0.11	0.80	0.46
ωB97X-D					
ΔML(CBH-0)	0.77	0.58	0.03	1.56	1.31
ΔML(CBH-1)	0.40	0.45	0.04	0.97	0.80
ΔML(CBH-2)	0.05	0.33	0.06	0.58	0.44
B2PLYP-D3(BJ)					
ΔML(CBH-0)	0.66	0.66	0.15	1.35	1.16
ΔML(CBH-1)	0.25	0.50	0.06	0.84	0.68
ΔML(CBH-2)	0.03	0.37	0.13	1.19	0.53

must be treated fully at the low level of theory, e.g., with dispersion-corrected DFT.

3.2. Machine Learning Models through CBH. Since both traditional ΔCBH corrections and ΔML(CBH) corrections are based on the same underlying isodesmic reactions, traditional trends are assumed to be carried over in their use as molecular descriptors, i.e., an increase in rungs of CBH should improve overall performance, and density functionals which model dispersion effects should outperform nondispersion counterparts. Out-of-sample errors for ΔML(CBH-0, CBH-1, and CBH-2) models for each of the 30 density functionals along with traditional ΔCBH-2 MAEs are summarized in Table 2. The smallest network class ΔML(CBH-0), where the model is only provided with information about the number of each heavy atom and number of total heavy atom bonds (6 input features), outperforms the traditional ΔCBH-2-corrected values in almost all cases, showcasing the utility of machine learning for recognizing patterns in the systematic errors.

3.3. Effect of Train−Test Split for Sparse Fragments. The ΔML models do not always follow the expected trends from fragmentation methods; such errors are shown in bold in Table 2. One potential issue that could be responsible for this behavior is a byproduct of restricting our data set to real, experimentally known molecules. In other data sets, aimed at covering all of chemical space, every combination of atoms and bonds can be included evenly, allowing for a more general predictive model.

Since our molecular descriptors are based solely on the structure of the molecule, a possible issue could arise from an unbalanced distribution of the fragments. For example, the most common CBH-2 fragment C_3H_8 appears at least once in 54% of the data set, while some uncommon fragments, e.g., SHCH═NH, occur in less than 0.5% of the data set. Due to the nature of the random train−test split, some fragments may only be present in the test set, causing large errors since no information about these structural features has ever been given to the model. The performance of the models in predicting these molecules is then dependent on the magnitude of the error in the null model (uncorrected baseline) and the initialized random state for those weights. To test how large of an effect the unevenness has on the out-of-sample performance, two variations of the original train−test split were evaluated. Fragments which only appear in 5 or fewer molecules were labeled as "sparse fragments", and all molecules containing these fragments (85 in total in this case) were either forced to be in the training set or removed from the training and test sets completely. The difference between the performance of ΔML(CBH-2) and ΔML(CBH-1) for the new splits are shown in Figure 4 for all models which did not follow the correct trends. Here, positive values indicate ΔML(CBH-1) outperforms ΔML(CBH-2), while negative values represent the expected trend.

Upon restricting sparse-fragment molecules to the training set, the difference between the out-of-sample errors for ΔML(CBH-1) and ΔML(CBH-2) decreased significantly and, in all but two cases, now followed the expected performance trends. Removal of the sparse fragment-containing molecules altogether changes the performance to the expected trends even further. Although the overall averaged errors only varied slightly upon removal of the 85 molecules, the performance of the models with the new train−test splits indicates the sparse fragments are contributing significantly to the final trained model, particularly for ΔML(CBH-2) descriptors. Ideally, for a more general model, a larger degree of overlap between input vectors of molecules may be required. Unfortunately, with the unavailability of such experimental data, this is not possible with the current data set. In order to not artificially exclude certain molecule types, these 85 molecules are restricted to the training set for the final trained models.

3.4. Final DFT+ΔML(CBH) Model Performance. Final out-of-sample MAEs for all combinations of functionals and

Table 4. Mean Absolute Errors (kcal/mol) Comparison for ΔCBH-2 and ΔML(CBH-2) for Various Structural Groups of the 1k-G4-C9 Data Set

	B3LYP		B3LYP-D3(BJ)		ωB97X-D		B2PLYP-D3(BJ)	
category	ΔCBH-2	ΔML	ΔCBH-2	ΔML	ΔCBH-2	ΔML	ΔCBH-2	ΔML
all (1051)	2.71	0.59	1.88	0.36	2.29	0.40	1.42	0.43
acyclic (535)	2.54	0.46	1.28	0.23	1.44	0.26	1.03	0.36
conjugated (241)	2.84	0.49	2.91	0.35	4.08	0.31	1.94	0.50
heterocyclic (203)	2.16	0.50	2.50	0.43	2.88	0.28	1.62	0.36
alicyclic (313)	3.28	0.88	2.48	0.56	3.37	0.70	1.95	0.60
hydrocarbon (232)	3.59	0.99	2.33	0.61	2.59	0.78	1.88	0.58
N-containing (270)	2.75	0.47	2.43	0.38	2.92	0.31	1.83	0.38
O-containing (578)	2.56	0.48	1.70	0.28	2.16	0.29	1.27	0.35
S-containing (82)	1.99	0.46	1.31	0.26	1.51	0.21	0.80	0.60
Cl-containing (74)	2.24	0.37	1.28	0.16	2.23	0.24	1.06	0.46

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

CBH-based descriptors are presented in Figure 5. All but two functionals, BP86 and BP86-D3(BJ), had test set mean absolute errors under 1 kcal/mol using ΔML(CBH-2), with 6 functionals achieving MAEs under 0.5 kcal/mol. The best performing functionals were ωB97XD (OOS-MAE = 0.444), B3LYP-D3(BJ) (OOS-MAE = 0.457), and M06-2X (OOS-MAE = 0.462). In general, more advanced density functionals such as double-hybrid density functionals, e.g., B2PLYP-D3(BJ) (OOS-MAE = 0.533) and mPW2PLYP-D2 (OOS-MAE = 0.551), and long-range-corrected hybrids, e.g., LC-ωPBE-D3(BJ) (OOS-MAE = 0.527) and CAM-B3LYP-D3(BJ) (OOS-MAE = 0.528), outperformed the GGA functionals, e.g., BPBE (OOS-MAE = 0.907), TPSS (OOS-MAE = 0.809), or BLYP (OOS-MAE = 0.995). As is well-known, density functionals can be grouped together and organized into the so-called "Jacob's Ladder" based on the complexity of the underlying physics that is included.[66] Errors from functionals higher up on the ladder (double-hybrids) seem to be more systematic with respect to the structure of the molecule. Many of the null errors from double-hybrid models are larger than those of functionals lower on the ladder. Once the corrections are applied from the trained models, the expected trends are now present, indicating a more systematic behavior. In the original train−test split, ωB97X-D+ΔML(CBH-2) significantly outperformed corresponding models with double hybrid functional baselines, due to a smaller null model error (2.08 kcal/mol for ωB97X-D compared to 12.83 and 11.91 kcal/mol for B2PLYP-D3(BJ) and mPW2PLYP-D2, respectively). Upon removal of the 85 molecules, the OOS-MAE for both double-hybrid models decreased by ∼0.35 kcal/mol, while ωB97X-D+ΔML(CBH-2) increased slightly by 0.12 kcal/mol suggesting the original split errors are coming from the magnitude of the null error.

Unsigned average errors (MAE or RMSE), although commonly reported and compared when discussing model performance, can sometimes underrepresent the actual errors during applications using the model.[17] A better indication of the expected error (and the accepted convention) for the uncertainty in thermochemistry is given by its 95% confidence interval ($u_{95\%}$), or twice the standard deviation (2σ), meaning only one out of every 20 values should be expected outside of the given uncertainty.[67] Therefore, three main error statistics are needed to describe the spread of the errors of a method, mean absolute error (MAE), mean signed error (MSE), and 95% confidence interval ($u_{95\%}$). MAE is a pure average of the magnitude of the difference between the measured (or calculated) value and its corresponding benchmark and is not easily skewed by outliers. MSE is an indication of the systematic error and is commonly reported as MSE ± 2σ since MSE is associated with the center of the 95% confidence interval. All three error statistics for the full 1051 molecule data set are given along with plots comparing traditional CBH corrections with the CBH-based ΔML corrections in Figure 6.

One major benefit of using a regression-based correction method, such as machine learning, is statistical systematic errors will be effectively canceled out. For example, the simplest model ΔML(CBH-0) features an MSE of 0.11 kcal/mol, while the corresponding value for ΔCBH-0 is −16.06 kcal/mol. Fragmentation-based corrections derived from quantum chemical calculations have no reason to minimize the MSE. Similar MSEs are present for all three ΔML models shown here. Of particular interest is the sharp decrease in the spread of the errors by increasing size of the fragments (both in traditional CBH and ML-CBH). Note: B3LYP+ΔCBH-0 is statistically identical to the baseline of B3LYP calculated ΔH_f(298 K). The difference between the two is the difference in performance of the atomization method of calculating the ΔH_f and an isogyric-based approach (*vide supra*). The uncertainty ($u_{95\%}$) associated with the B3LYP+ΔML(CBH-2) model is −0.10 ± 1.88 kcal/mol with a standard deviation of σ = 0.94 approaching the chemical accuracy threshold. Top performing models in terms of OOS-MAE B2PLYP-D3(BJ), ωB97X-D, and M06-2X have uncertainties of 0.15 ± 1.34, −0.04 ± 2.05, and −0.03 ± 2.28 kcal mol^{-1}, respectively. ωB97X-D and M06-2X gave significantly larger standard deviations (σ = 1.03 and σ = 1.14) compared to B2PLYP-D3(BJ) (σ = 0.67) suggesting double hybrid models are more robust even though they feature a slightly larger MAE.

Typically in machine learning, a trade-off between bias and variance is observed, i.e., as the error on the training set (bias) decreases toward the irreducible error, the overall error on the test set (variance) increases. Additionally, models with a low bias and high variance are thought to be overfitted to the training data, since they feature low errors on the training data but high errors on the test set, while large biases indicate underfitting. An ideal model, neither under- nor overfit, will produce low errors for both the training and test sets, though minimizing these two errors simultaneously is not a simple task. The test set error is only calculated once, after the model is completely trained on the training data, and model hyperparameters cannot be tuned based on the performance of the test set. Cross-validation helps mitigate this problem by choosing the hyperparameters which minimize the overall validation set errors. If the data is balanced, and both the test and validation sets are representations of the full data set, this will minimize the error on the unseen test set. Analysis of these errors, shown in Table 3, gives more insight into model performance than the final out-of-sample MAE. The irreducible error is defined as the best possible performance a network can attain. This is calculated from the nonunique input representations. For many molecules, more noticeably at lower rungs of CBH, their reaction is not unique from similar molecules, as shown in Figure 2. Thus, the irreducible error is a hypothetical quantity calculated from minimizing the error a model would produce when trained on each group of nonunique inputs with different outputs.

Bias is calculated as the difference between the mean absolute error of the training set and the irreducible error. Variance is taken as the difference between the test set and training set errors. Each validation MAE is an average of five validation MAEs, where the model is trained on four of the k-fold groups and tested on one group. Due to the sparse fragment molecules now being involved in cross-validation, the validation set errors are larger than the final test set errors in some cases. For the original train−test split, validation errors matched test set errors almost exactly due to sparse fragment molecules being present in both the validation and test sets. However, upon the new split, some of the CV folds now contain many of these molecules, returning a large validation error and skewing the average validation error. This is not necessarily problematic since hyperparameters were chosen from the minimization of the averaged validation MAE over all functionals and folds. A higher variance is observed in ΔML(CBH-2) models compared to ΔML(CBH-1); however, the magnitude of the variance is <0.20 kcal/mol for the best performing models, indicating the models are not overfit.

In order to assess the generality of the models with respect to molecule type, the full data set was split into 8 groups based on

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

structural features (acyclic, conjugated, heterocyclic, and alicyclic) and chemical composition (hydrocarbons, N-, O-, S-, and Cl-containing) and compared to traditional ΔCBH-2 corrections, all shown in Table 4.

Certain medium- to long-range structural features, e.g., conjugation, are not fully represented by calculated corrections based on CBH-2 fragments; however, since the ΔML(CBH-2) model is not based on a fragment energy calculation, systematic errors can be learned, and the model can overcome this deficiency. Conversely, the largest errors were present in hydrocarbons and alicyclic molecules (containing carbon-only rings) for both ΔCBH-2 and ΔML(CBH-2). This trend is most likely due to the noncovalent intramolecular interactions in cyclic molecules. Some CBH-2 fragments, more specifically propane (C_3H_8), are common between both acyclic and alicyclic molecules. For example, the hypohomodesmotic (CBH-2) reactions for both cyclohexane and hexane are comprised solely of propane and ethane fragments, causing a mismatch of systematic error if acyclic and alicyclic molecules require vastly different corrections to the low-level treatment of the full molecule. The final correction given to these molecules will be closer to that of the acyclic molecule, as there are simply more of this group than alicyclic molecules in the full data set (535 compared to 313). Ideally, the low-level would capture all important noncovalent interactions, but this is not always the case. Dispersion-corrected density functionals perform better in this aspect as the gap between acyclic and alicyclic is slightly smaller for B3LYP-D3(BJ) compared to B3LYP.

4. COMPARISON TO OTHER CONNECTIVITY-BASED REPRESENTATIONS

As stated earlier, the molecular descriptors most similar to the isodesmic-based ones developed in this work are of the rank-based family. A brief comparison was done previously, showing connectivity counts outperforming rank-3 descriptors with a similar input vector size. This improvement can at first glance be attributed to the inclusion of certain rank-4 and rank-5 type fragments in the CBH-2 representation. The number of rank-4 and rank-5 descriptors grows much faster than the rungs of CBH, since *n*-atom groups are formed for rank-*n*, compared to the CBH-*n* rungs which preserve a max chain length of *n*+1. The chain length constraint allows a more chemically intuitive representation of the molecular structure. To show this, the layers of rank-based descriptors were concatenated and trained to produce 1k-G4-C9. Since the best performing model, CBH-2, is comprised of fragments of comparable size to rank-2 through rank-5 groups, these rank counts were all used together as the input vector, referred to as rank-2:5. These descriptors lead to an input vector of size $N = 854$ compared to CBH-2 of $N = 103$. DFT+ΔML(Rank-2:5) models did not outperform corresponding CBH-2 models, giving OOS-MAEs of 0.83, 0.52, and 0.82 kcal/mol compared to 0.81, 0.44, and 0.53 kcal/mol for DFT = B3LYP, ωB97X-D, and B2PLYP-D3(BJ), some of the top performing combinations found in this study. Associated 2σ uncertainties ($u_{95\%}$) were also larger with the rank-2:5 for the three low levels mentioned above giving 2.41, 2.08, and 1.91 kcal/mol compared to 1.88, 2.05, and 1.34 kcal/mol, respectively. Additionally, the size of the input vector for rank-2:5, $N = 854$, is much larger than CBH-2, $N = 103$, implying CBH-2 more efficiently encodes the structure of the molecule. The good performance of CBH-2 is most likely from the satisfaction of the inclusion−exclusion principle, by which, the overlap of different groups of atoms is subtracted out, such that every atom of the molecule gets counted only once.

5. CONCLUSIONS

A new hierarchy of molecular descriptors has been presented based on the ideas from the fragmentation-based methodology and the error cancellation nature of isodesmic reactions along with the ΔML model in computational thermochemistry. These molecular descriptors provide a significant improvement over commonly used geometry- and molecular graph-based descriptors such as Coulomb Matrix, Bag of Bonds, and Rank-n connectivity representations as the isodesmic-based descriptors are constant size for large and small molecules and relatively small in input length, minimizing the number of degrees of freedom. The framework laid out herein provides a foundation for a newly developed series of molecular representations based around the ΔML regime. Corrections obtained using these models are analogous to the Connectivity-Based Hierarchy (CBH) correction method offering local, fragment-based corrections to the energy of a molecule. Unlike other fragmentation-based correction approaches, CBH combines fragments with the same connectivity allowing the use of one-hot encoding based on reaction coefficients. This procedure could be expanded to larger fragments, such as amino acids, or based on similarity of fragments. The ΔML corrections achieve an averaged out-of-sample error within 0.5 kcal/mol and $u_{95\%} < 1.5$ kcal/mol for a wide variety of density functionals. This approach can be used to correct for the overall intrinsic errors of DFT in thermochemistry with no expensive calculations required. Further developments that use the detailed structure of the molecule (e.g., bond lengths, etc.) in addition to connectivity-based representations could potentially offer additional improvements and will be pursued in the future.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00236.

Summary of 1k-G4-C9 data set, along with computational details, technical information for ML models, hyperparameter search, and training curves for various levels of theory (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Author

Eric M. Collins − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0002-9113-1705

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00236

Funding

This work was supported by funding from the National Science Foundation Grant CHE-1665427 at Indiana University.

Notes

The authors declare no competing financial interest.

Additional supporting research data of the sample code for regression models and input representation generation along with the 1k-G4-C9 data set for this article may be accessed at https://github.com/colliner/MLCBH.

REFERENCES

(1) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.

(2) Martin, J. M. L.; de Oliveira, G. Towards standard methods for benchmark quality ab initio thermochemistry—W1 and W2 theory. *J. Chem. Phys.* **1999**, *111*, 1843−1856.

(3) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. W3 theory: Robust computational thermochemistry in the kJ/mol accuracy range. *J. Chem. Phys.* **2004**, *120*, 4129−4141.

(4) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High accuracy extrapolated ab initio thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599−11613.

(5) El-Nahas, A. M.; Bozzelli, J. W.; Simmie, J. M.; Navarro, M. V.; Black, G.; Curran, H. J. Thermochemistry of Acetonyl and Related Radicals. *J. Phys. Chem. A* **2006**, *110*, 13618−13623.

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

(7) 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.

(8) Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(9) Wheeler, S. E. Homodesmotic reactions for thermochemistry. *Wires Comput. Mol. Sci.* **2012**, *2*, 204−220.

(10) Dixon, D. A.; Feller, D.; Peterson, K. A. Chapter One - A Practical Guide to Reliable First Principles Computational Thermochemistry Predictions Across the Periodic Table. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Ed.; Elsevier: 2012; Vol. *8*, pp 1−28, DOI: 10.1016/B978-0-444-59440-2.00001-6.

(11) Sun, J.; Wu, J.; Song, T.; Hu, L. H.; Shan, K. L.; Chen, G. H. Alternative Approach to Chemical Accuracy: A Neural Networks-Based First-Principles Method for Heat of Formation of Molecules Made of H, C, N, O F, S, and Cl. *J. Phys. Chem. A* **2014**, *118*, 9120−9131.

(12) 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.

(13) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. Accurate and Computationally Efficient Prediction of Thermochemical Properties of Biomolecules Using the Generalized Connectivity-Based Hierarchy. *J. Phys. Chem. B* **2014**, *118*, 9631−9643.

(14) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.

(15) Karton, A.; Sylvetsky, N.; Martin, J. M. L. W4−17: A Diverse and High-Confidence Dataset of Atomization Energies for Benchmarking High-Level Electronic Structure Methods. *J. Comput. Chem.* **2017**, *38*, 2063−2075.

(16) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

(17) Ruscic, B.; Bross, D. H. Chapter 1 - Thermochemistry. In *Computer Aided Chemical Engineering*; Faravelli, T., Manenti, F., Ranzi, E., Eds.; Elsevier: 2019; Vol. *45*, pp 3−114, DOI: 10.1016/B978-0-444-64087-1.00001-2.

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

(19) Ramabhadran, R. O.; Raghavachari, K. The Successful Merger of Theoretical Thermochemistry with Fragment-Based Methods in Quantum Chemistry. *Acc. Chem. Res.* **2014**, *47*, 3596−3604.

(20) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255−260.

(21) Gil, Y.; Greaves, M.; Hendler, J.; Hirsh, H. Amplify scientific discovery with artificial intelligence. *Science* **2014**, *346*, 171−172.

(22) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.

(23) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(24) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(25) Wang, X. J.; Wong, L. H.; Hu, L. H.; Chan, C. Y.; Su, Z. M.; Chen, G. H. Improving the accuracy of density-functional theory calculation: The statistical correction approach. *J. Phys. Chem. A* **2004**, *108*, 8514−8525.

(26) Grambow, C. A.; Li, Y. P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826−5835.

(27) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(28) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem., Int. Ed.* **2018**, *57*, 4164−4169.

(29) Yang, G.; Wu, J.; Chen, S.; Zhou, W.; Sun, J.; Chen, G. Size-independent neural networks based first-principles method for accurate prediction of heat of formation of fuels. *J. Chem. Phys.* **2018**, *148*, 241738.

(30) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(31) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689−2694.

(32) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(33) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546−1559.

(34) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.

(35) Chen, H. M.; Engkvist, O.; Wang, Y. H.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(36) Ekins, S. The Next Era: Deep Learning in Pharmaceutical Research. *Pharm. Res.* **2016**, *33*, 2594−2603.

(37) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(38) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589.

(39) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **2017**, *3*, 54.

(40) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73.

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950

(41) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(42) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(43) Sarathy, S. M.; Vranckx, S.; Yasunaga, K.; Mehl, M.; Osswald, P.; Metcalfe, W. K.; Westbrook, C. K.; Pitz, W. J.; Kohse-Hoinghaus, K.; Fernandes, R. X.; Curran, H. J. A comprehensive chemical kinetic combustion model for the four butanol isomers. *Combust. Flame* **2012**, *159*, 2028−2055.
(44) Merchant, S. S.; Zanoelo, E. F.; Speth, R. L.; Harper, M. R.; Van Geem, K. M.; Green, W. H. Combustion and pyrolysis of iso-butanol: Experimental and chemical kinetic modeling study. *Combust. Flame* **2013**, *160*, 1907−1929.
(45) Ruscic, B.; Pinzon, R. E.; von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoya, D.; Wagner, A. F. Active Thermochemical Tables: thermochemistry for the 21st century. *J. Phys.: Conf. Ser.* **2005**, *16*, 561−570.
(46) Blum, L. C.; Reymond, J. L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(47) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate quantum chemical energies for 133 000 organic molecules. *Chem. Sci.* **2019**, *10*, 7449−7455.
(48) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(49) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.
(50) Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Narayanan, B.; Curtiss, L. Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations. *MRS Commun.* **2019**, *9*, 891−899.
(51) Linstrom, P. J.; Mallard, W. G. The NIST Chemistry WebBook: A chemical data resource on the internet. *J. Chem. Eng. Data* **2001**, *46*, 1059−1063.
(52) Mathieu, D. Atom Pair Contribution Method: Fast and General Procedure To Predict Molecular Formation Enthalpies. *J. Chem. Inf. Model.* **2018**, *58*, 12−26.
(53) 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.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Rev. C.01; Wallingford, CT, 2016.
(54) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach Learn Res.* **2011**, *12*, 2825−2830.
(55) 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.
(56) Gilbert, T. M. Application of a Semi-homodesmotic Approach in Estimating Ring Strain Energies (RSEs) of Highly Substituted Cyclobutanes: RSEs for c-C4R8 That Make Sense. *J. Phys. Chem. A* **2014**, *118*, 6060−6067.
(57) George, P.; Bock, C. W.; Trachtman, M. The matching of structural elements in reactions for evaluating stabilization energies for benzene and monosilabenzene. *Theor Chim Acta* **1987**, *71*, 289−298.
(58) 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.
(59) Collins, E. M.; Sengupta, A.; AbuSalim, D. I.; Raghavachari, K. Accurate Thermochemistry for Organic Cations via Error Cancellation using Connectivity-Based Hierarchy. *J. Phys. Chem. A* **2018**, *122*, 1807−1812.
(60) Ramabhadran, R. O.; Sengupta, A.; Raghavachari, K. Application of the Generalized Connectivity-Based Hierarchy to Biomonomers: Enthalpies of Formation of Cysteine and Methionine. *J. Phys. Chem. A* **2013**, *117*, 4973−4980.
(61) Sengupta, A.; Raghavachari, K. Solving the Density Functional Conundrum: Elimination of Systematic Errors To Derive Accurate Reaction Enthalpies of Complex Organic Reactions. *Org. Lett.* **2017**, *19*, 2576−2579.
(62) Debnath, S.; Sengupta, A.; Raghavachari, K. Eliminating Systematic Errors in DFT via Connectivity-Based Hierarchy: Accurate Bond Dissociation Energies of Biodiesel Methyl Esters. *J. Phys. Chem. A* **2019**, *123*, 3543−3550.
(63) Thapa, B.; Raghavachari, K. Accurate pKa Evaluations for Complex Bio-Organic Molecules in Aqueous Media. *J. Chem. Theory Comput.* **2019**, *15*, 6025−6035.
(64) Maier, S.; Thapa, B.; Raghavachari, K. G4 accuracy at DFT cost: unlocking accurate redox potentials for organic molecules using systematic error cancellation. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4439−4452.
(65) Wodrich, M. D.; Wannere, C. S.; Mo, Y.; Jarowski, P. D.; Houk, K. N.; Schleyer, P. V. R. The concept of protobranching and its many paradigm shifting implications for energy evaluations. *Chem. - Eur. J.* **2007**, *13*, 7731−7744.
(66) Perdew, J. P.; Ruzsinszky, A.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.* **2005**, *123*, 062201.
(67) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *Int. J. Quantum Chem.* **2014**, *114*, 1097−1101.

https://dx.doi.org/10.1021/acs.jctc.0c00236
J. Chem. Theory Comput. 2020, 16, 4938−4950