JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

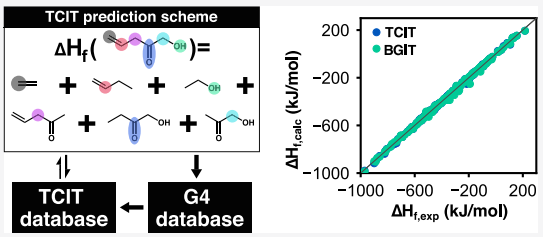
Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation

Qiyuan Zhao and Brett M. Savoie*

Cite This: J. Chem. Inf. Model. 2020, 60, 2199−2207 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The gas-phase enthalpy of formation (ΔH_f) plays a fundamental role in predicting reaction thermodynamics and constructing kinetic models. With advances in computational power and method development, chemically accurate quantum chemistry methods that can predict ΔH_f values for small molecules are available; however, large molecules are still out of reach. Increment theories provide a means of extending the prediction capability of high-level methods by decomposing the molecular ΔH_f into the additive contributions from individual atoms, bonds, groups, or components. Here, we introduce a novel component increment theory, topology-automated force−field interaction component increment theory (TCIT), in which all component contributions are derived exclusively from Gaussian-4 (G4) results for algorithmically generated model compounds. In a benchmark evaluation of noncyclic compounds from the Pedley, Naylor, and Kline experimental ΔH_f dataset, TCIT exhibits consistently lower signed and absolute errors compared with the conventional Benson group increment theory (BGIT). These results pave the way for future extensions of TCIT to ring-containing, ionic, and radical species for which experimental data scarcity currently limits the application of BGIT.



1. INTRODUCTION

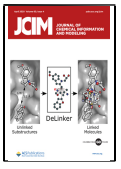
The standard enthalpy of formation ΔH_f plays a fundamental role in determining reaction equilibria, exothermicity, and molecular stability.[1−4] Given this centrality, there has been a concerted 80 year effort (going back at least to Pauling[5]) to inexpensively calculate ΔH_f for new molecules. An enduring contribution to the canon of ΔH_f calculation methods was supplied by Benson and Buss in 1958,[6] with their proposal to decompose the molecular ΔH_f into the sum of contributions from individual groups

$$\Delta H_f = \sum_i h_i \tag{1}$$

where h_i values are the individual group contributions, which are called group additivity values (GAVs). In eq 1, the individual h_i is independent of its position within the molecule. This is a strong locality assumption with notable limitations, but one that works surprisingly well in practice and enables the reuse of the parameters and broad chemical coverage in predicting ΔH_f.[7] This type of model is known more generally as a group increment theory or group additivity theory. The corresponding groups in Bensons terminology are defined as every nonterminal atom (i.e., an atom bonded to at least two other atoms) and its bonded neighbors. Although many subsequent group increment theories have been developed,[8−10] that developed by Benson throughout his career (hereafter referred to as Benson group increment theory, BGIT) and continued now through others remains the de facto standard.[6,7,11] BGIT has proven remarkably durable since its original development and includes coverage of most neutral organic functional groups, many ring-containing compounds, and to a lesser degree, ionic species.[11−14] Although modern quantum chemical methods are capable of chemical accuracy (i.e., prediction errors less than 1 kcal/mol) for a useful range of molecular sizes,[15,16] BGIT is still unmatched with respect to chemical coverage, accuracy, and trivial cost. For applications in molecular discovery and reaction prediction, on-the-fly predictions of ΔH_f for thousands to millions of compounds are not unusual, and BGIT is currently the only viable general method for calculations on this scale. It is foreseeable that fully empirical machine learning (ML) approaches will overtake BGIT in the near future;[17,18] however, at the time of publication, ML approaches to ΔH_f are still incipient.

Despite the successes of group decomposition, the circumstances that have led to the reliance on Benson groups have changed. First, in Benson's original formulation, GAVs were parameterized to carefully curated experimental data; whereas, in modern extensions of BGIT, groups are regularly derived from quantum chemistry data.[2,19−22] This mixed provenance of experimental and quantum chemistry values leads to potential issues when extending BGIT to new chemistries while

Received: January 25, 2020
Published: March 11, 2020



https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

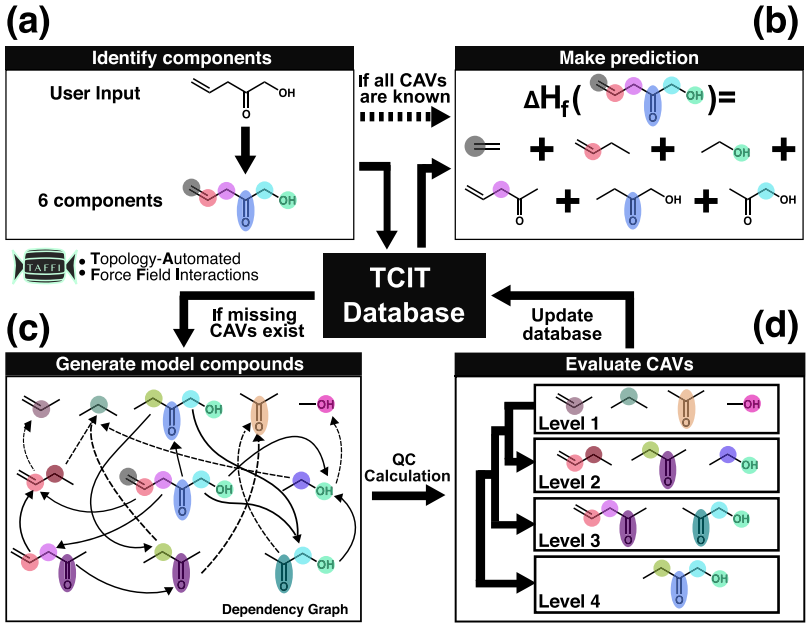


Figure 1. Overview of the TCIT component decomposition, model compound generation, CAV parameterization, and prediction. (a) Molecules are decomposed into components for each nonterminal atom. Components are uniquely defined based on a central nonterminal atom, its bonded neighbors, and next nearest-bonded neighbors. (b) Predictions are based on eq 1 when all requisite CAVs exist within the TCIT database. (c) Each CAV is derived from a single algorithmically generated model compound. Model compounds are recursively generated until all required components are represented. (d) Conformational sampling and quantum chemistry calculations of ΔH_f are performed on all model compounds. One unique component is derived from each model compound by sequentially solving eq 1 for the unknown CAVs. CAV parameterizations are ordered by dependency and subject to constraints to address rank deficiency. All CAVs and quantum chemistry calculations are stored in the TCIT database for future use.

maintaining consistency. Second, the expansion of the available parameterization data makes Benson's original designation of group specificity out to one bond no longer necessary. In the nomenclature of increment theories, a theory with specificity out to two bonds (i.e., groups defined based on bonded nearest-neighbors and next nearest-neighbors) is called a component increment theory.[10] Modern BGIT is in fact a hybrid between a group theory and a component theory, with many component-level "groups" being defined when data availability permits.[7] This mixed specificity also leads to complications in systematically extending BGIT. Third, the definition of a model compound (i.e., the molecular species used to derive GAVs) in BGIT has never been formalized. When defining GAVs for new groups, the standard procedure is to find a molecule, or ideally molecules, that exhibits the unknown group, determines its ΔH_f (either through experiment or computation), and then define the missing value through eq 1. The selection of model compounds for this procedure is nontrivial because errors in the known groups that are used in evaluating eq 1 will propagate to the newly derived group. In its early development, the choice of model compounds was determined based on experimental availability, but the formalization of this procedure is now possible using computational algorithms. In combination, these three issues—provenance, specificity, and model compound selection—currently limit the extensibility of BGIT, despite its many otherwise attractive features.

In the present work, we develop a fully consistent component increment theory that overcomes the issues of data provenance, group specificity, and model compound selection in BGIT. This method is based on our group's previous work on topology-automated force−field interactions (TAFFI), which provides a systematic method for specifying atom types and model compounds for deriving molecular dynamics force fields.[23] Although the previous TAFFI framework was not developed for property prediction, the problems of component definition and model compound selection are identical to the situation in force field development, and we name the resulting theory TAFFI component increment theory (TCIT). In TCIT, ΔH_f is decomposed into components that are chemically specific out to two bonds. Model compounds for determining the component additivity values (CAVs) are generated on an algorithmic basis, and the ΔH_f of these model compounds is determined from Gaussian-4 (G4) calculations.[24] To demonstrate the generality of this approach, we present TCIT results for all 626 acyclic compounds in the Pedley, Naylor, and Klein (PNK) dataset.[10] The treatment of ring corrections and the extension to ionic and radical species is left for separate publication.

2. COMPUTATIONAL METHODS

A high-level overview of the TCIT method is provided here and in Figure 1 to assist the reader. A detailed description of each step is provided in the subsequent sections (Sections 2.1 to 2.4).

The first step in TCIT is to determine the component types in a molecule of interest (Section 2.1). The components are automatically determined from a user-supplied molecular geometry or SMILES string, based on the bonding structure of the compound. TCIT adopts a machine-readable syntax for naming and databasing CAVs. In the case that all CAVs for a target compound already exist in the TCIT database, the calculation of ΔH_f is performed via eq 1 (Figure 1b). If any of the required CAVs do not exist, TCIT proceeds with model compound generation and quantum chemistry calculations to parameterize the missing CAVs.

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

For CAV parameterization, TCIT algorithmically generates the smallest acyclic molecule that both exhibits the required component and conserves its Lewis structure to use as a model compound (Section 2.2, and shown in Figure 1c). It may happen that the model compounds exhibit new components whose CAVs are also missing from the TCIT database. Thus, model compound generation proceeds recursively until all model compounds for all required components have been generated. Because model compounds decrease in size at each step of the recursion, the procedure eventually terminates with a finite number of small model compounds that can be characterized using G4 calculations.

Conformer searching is performed for every model compound, and all conformers within an energy threshold are utilized for G4 calculations (Section 2.3). The ΔH_f value for each model compound is then calculated as a Boltzmann weighted average over all of the remaining conformers. After ΔH_f has been calculated for all of the model compounds, the missing CAVs are parameterized by solving eq 1 (Section 2.4). During CAV parameterization, two subtleties can occur. First, the model compounds are potentially dependent on one another (i.e., they share components). This is addressed by using a topological sort on the dependency graph such that only the components associated with the model compound are fit for each evaluation of eq 1 (Figure 1d). Second, when two or more components share the same model compound, the system of equations is rank-deficient. This is addressed through the use of three auxiliary constraints. After CAV parameterization, all G4 calculations and CAVs are stored in a database for reuse in future ΔH_f predictions and CAV parameterizations.

2.1. TAFFI Component Definition. TCIT components are defined for every nonterminal atom in a molecule. Here, nonterminal is defined as any atom bonded to two or more atoms, and each nonterminal atom is the central atom of exactly one component. No component is associated with terminal atoms because their influence on increment theory predictions is accounted for by the component that they are bonded to.

Following the terminology of Benson, the central atom of the component is combined with the nearest and next-nearest neighbor atoms to form a component. Compared with Benson groups, TAFFI components contain more chemically specific information, and the CAVs reflect contributions from leading order non-nearest interactions.[1,2,25]

The determination of components in TCIT is based on the TAFFI methodology[23] and can be divided into following steps. First, the chemical topology is determined from the adjacency matrix, **A**, an N by N matrix, where N is the number of atoms in the molecule. The elements of **A** are defined as

$$A_{ij} = \begin{cases} 1 & \text{if a bond exists between atom } i \text{ and atom } j \\ 0 & \text{if a bond does not exist between atom } i \text{ and atom } j \end{cases} \quad (2)$$

Second, each component is determined by the subgraph of the adjacency matrix obtained by keeping all atoms within two bonds of the central atom in the component. This is algorithmically carried out by recursive walks of the rows in the adjacency matrix while keeping track of bonded atoms, out to a recursion depth of two.

Each subgraph obtained in this way uniquely defines each TCIT component in the molecule. TCIT utilizes a string syntax for canonicalizing these subgraphs and expressing them in a machine-readable format. In the TCIT syntax, all numbers refer to atomic numbers (i.e., 1 corresponds to hydrogen, and 6 to carbon), open brackets ([) designate bonds, and closed brackets (]) designate the end of bonded groups.[23] For instance, the component corresponding to the central carbon atom in propane is encoded as [6[6[1][1][1]][6[1][1][1]][1][1]], where the first 6 refers to the central carbon atom itself, the two [6[1][1][1]] connections refer to the two bonded methyl groups, and the final [1][1] are the two hydrogens directly bonded to the central carbon. To resolve the ambiguity associated with graph isomorphism, the ordering of groups within each component label is determined by the mass of the bonded atoms. In the case of ambiguity, the mass and number of next-nearest bonded atoms are utilized (similar to Cahn−Ingold−Prelog priority rules). Many more examples are provided in the CAVs distributed with this work.

2.2. Model Compound Generation. In TCIT, all CAVs are derived from a deterministic and systematically generated set of model compounds for which G4 calculations can be performed. For a given component, the model compound is defined as the smallest acyclic molecule that both exhibits the required component and conserves its Lewis structure. Starting with the target compound supplied by the user, these model compounds are generated in two steps. First, all atoms more than two bonds away from the central atom of each component are removed to form a preliminary compound. Second, any undercoordinated atoms that result from this truncation are hydrogenated to a level that is consistent with their hybridization within the component. This definition leads to ambiguity in cases involving double bonds between nearest and next-nearest neighbors (e.g., keto−enol tautomers). In these cases, double bonds are preferentially formed with the highest bond energy.[26,27] For example, the model compound for [6[6[8][6]][6[1][1][1]][1][1] is 2-butanone rather than 1-buten-2-ol (i.e., the ketone as opposed to the alcohol, consistent with the Erlenmeyer rule). This rule substantially improves the prediction accuracy for compounds involving such components (not shown).

After generating a model compound for an unknown component, the model compound may exhibit additional unknown components. Thus, model compound generation is recursively performed for these new components until all model compounds have been generated for all unknown components. Because each model compound is smaller than its parent, this recursion will eventually terminate with model compounds containing only one unknown component (see Section 2.4 for additional details on the model compound dependencies for CAV parameterization). This procedure yields model compounds that are generally small and amenable to high-level quantum chemistry calculations. More than 90% of model compounds generated in this study had no more than eight heavy atoms (mode is five), and no model compound had more than twelve heavy atoms (Figure S1).

2.3. Quantum Chemistry Methods. Within TCIT, all CAVs are parameterized to ΔH_f values calculated at the G4 level for small model compounds (generated as described above). To obtain reliable ΔH_f predictions, it is necessary to include contributions from all conformers with significant Boltzmann probability at standard conditions and implement robust geometry optimization protocols.

For each model compound, up to 100 different conformers are generated by open-babel (not all model compounds are complex enough to encounter the threshold of 100 conformers).[28,29] The single-point energy of these conformers is

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

then calculated using the semiempirical geometry, frequency, noncovalent, and eXtended tight binding method (GFN2-xTB) developed by Grimme.[30,31] Based on the GFN2-xTB single-point energy calculation results, conformers with energies greater than 6 kcal/mol of the minimum energy conformer are discarded, and then (up to) 10 of the most energetically most stable conformers are then selected for further quantum chemistry characterization.

After conformer generation and selection, geometry optimization and frequency calculations are performed on the remaining conformers. To robustly converge the geometries, geometry optimizations are performed at incrementally increasing levels of theory. Initial optimization is performed at the B3LYP/6-31G level (functional and basis set, respectively), followed by the B3LYP/6-31G(d,p) level of theory, and final optimization and frequency calculations are performed at the B3LYP/6-31G(2df,p) level, which is the same as the functional and basis set used by the G4 method.[24]

ΔH_f calculations are performed on the optimized geometries of the retained conformers at the G4 level of theory as implemented in Gaussian 16.[32] G4 is based on a sequence of single-point energy calculations with basis set extrapolation that achieves an overall average absolute deviation of 0.8 kcal/mol for the reported test set of 270 experimental enthalpies of formation.[24]

According to the definition, $\Delta H_f(0\ \mathrm{K})$ can be calculated by subtracting nonrelativistic atomization energies $\sum D_0$ from known enthalpies of formation of the isolated atoms. For instance, the $\Delta H_f(0\ \mathrm{K})$ of a molecule, $A_xB_yC_z$, can be calculated as

$$\Delta H_f(A_xB_yC_z,\ 0\ \mathrm{K}) = x\Delta H_f(A,\ 0\ \mathrm{K}) + y\Delta H_f(B,\ 0\ \mathrm{K}) + z\Delta H_f(C,\ 0\ \mathrm{K}) - \sum D_0 \quad (3)$$

The atomic $\Delta H_f(0\ \mathrm{K})$ is taken from Curtiss et al.,[33] and $\sum D_0$ is defined as

$$\sum D_0 = xE_0(A) + yE_0(B) + zE_0(C) - E_0(A_xB_yC_z) \quad (4)$$

where E_0 is the zero-point energy calculated at the G4 level of theory. To calculate the ΔH_f at 298 K, a temperature correction is applied

$$\Delta H_f(A_xB_yC_z,\ 298\ \mathrm{K}) = \Delta H_f(A_xB_yC_z,\ 0\ \mathrm{K}) + \Delta H^0(A_xB_yC_z,\ 298\ \mathrm{K}) - x\Delta H^0(A,\ 298\ \mathrm{K}) - y\Delta H^0(B,\ 298\ \mathrm{K}) - z\Delta H^0(C,\ 298\ \mathrm{K}) \quad (5)$$

where $\Delta H^0(\mathrm{M},\ 298\ \mathrm{K}) = H^0(\mathrm{M},\ 298\ \mathrm{K}) - H^0(\mathrm{M},\ 0\ \mathrm{K})$ is the enthalpy difference between 298 and 0 K, and M can refer to either atoms or molecules.

For a molecule with N conformers, the final ΔH_f is calculated as a Boltzmann weighted average of the individual conformer ΔH_f values

$$\langle \Delta H_f \rangle = \frac{\sum_{n=1}^{N} \Delta H_f{}^n \exp\left(-\frac{\Delta E_n}{k_B T}\right)}{\sum_{n=1}^{N} \exp\left(-\frac{\Delta E_n}{k_B T}\right)} \quad (6)$$

where $\Delta E_n = E_{G4}{}^n - E_{G4}^{\min}$ $(n = 1, 2 \ldots N)$, and $E_{G4}^{\min}$ is the energy of the most stable conformer.

2.4. CAV Parameterization. After generating all model compounds and their ΔH_f values, it is possible to parameterize the CAVs using eq 1. This can be done in at least two ways. In the first, eq 1 can be applied to each model compound simultaneously to form a system of equations with the CAVs as parameters that can be obtained in a least-squares manner or by minimizing another objective function. However, this system of equations is potentially rank-deficient, which makes such an optimization poorly conditioned. Additionally, in this approach, the CAVs are parameterized to variable amounts of data (i.e., one component may be in many model compounds, another may be in only a single model compound), and when adding additional components, all CAVs would need to be constantly reparameterized. In the alternative approach adopted here, CAVs are only parameterized by solving eq 1 for individual model compounds. In this approach, the system of equations is decomposed into individual applications of eq 1, starting with the model compounds containing only a single component. After a CAV has been parameterized in this way, it is held fixed during subsequent CAV parameterizations. The following sections contain additional details on how these parameterizations are ordered, and how rank deficiency issues are addressed.

2.4.1. Topological Sorting. Because components are present in model compounds besides their own, it is necessary to order the CAV parameterizations such that all CAVs, besides those being parameterized, have been obtained prior to a specific application of eq 1. These dependencies are enumerated during model compound generation and stored in a dependency graph. The dependency graph has nodes for all model compounds and directed connections between all dependent compounds[23] (e.g., pentane depends on butane for the component type [6[6[6][1][1]][6[1][1][1]][1][1]]). Prior to performing CAV parameterizations, a topological sort is applied to the dependency graph such that no dependencies exist within the same level of the sorted graph. The CAVs are then parameterized via eq 1 beginning with model compounds in the bottom level of this graph and working to the top (i.e., level 1 to level 4 in Figure 1d). This addresses the issue of CAVs potentially being missing during parameterization because the CAVs at each level can be directly determined when all of the dependent CAVs in the lower levels are known.

2.4.2. Addressing Rank Deficiency. Whenever a compound is a model compound for more than one component, the system of equations required to fit the CAVs will exhibit rank deficiency. That is, because each model compound supplies one datum (ΔH_f) and one equation, multiple components derived from a single model compound amount to more parameters than equations. For the definition of model compounds utilized by TCIT, all cases of rank deficiency can be addressed by applying three additional constraints to the system of equations. The choice of these constraints is in fact arbitrary in the context of a group or component theory where each additivity value is derived from one model compound (as in the case of TCIT) because the components involved are interdependent, and the effect of the constraints cancels when making predictions. Nevertheless, we have followed earlier work in introducing physically motivated constraints. Constraint I is applied to carbon-based terminal components (i.e., components that are only bonded to one other component); constraint II is applied when two components with underdetermined additivity values

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

are bonded; and constraint III is applied when three or more components with underdetermined additivity values are present in the same model compound. Constraint I is applied before constraint II and III whenever there is overlap in applicability. Constraints II and III are mutually exclusive. Examples of each type of rank deficiency and details on the implementation of each constraint are described below.

Rank deficiencies involving terminal carbon-based components are common for typical organic molecules (e.g., the [6[6[6][1]][1][1]] and [6[6[6][1]][1][1][1]] components of propylene corresponding to the terminal methylene and methyl, respectively, Figure 2a). In this case, we have followed the earlier

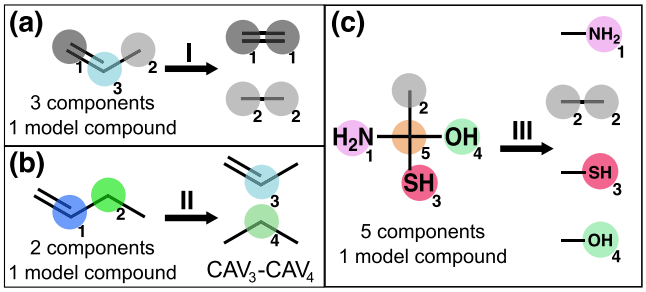


Figure 2. When multiple components share a model compound, the resulting rank deficiency is addressed through the application of three constraints. (a) Terminal $-CH_3$, $=CH_2$, and $\equiv CH$ carbon-based components are set equal to half the ΔH_f of ethane, ethene, or ethyne (constraint I). (b) When two components shared a model compound, their difference is constrained to be equal to the difference of analogous components (constraint II). (c) When three or more components share a model compound, all but one will be terminal. The terminal components are constrained to be equal to analogous carbon-bonded components (constraint III).

work of Pedley et al.[10] and defined three special terminal components: $-CH_3$, $=CH_2$, and $\equiv CH$. The CAVs for all carbon-based terminal components are defined as half of ΔH_f of ethane, ethene, and ethyne, regardless of the other components they are bonded to (constraint I). Applying constraint I in the case of propylene

$$\begin{aligned} \mathrm{CAV}([6[6[6][1]][1][1]]) &= \mathrm{CAV}([6[6[1][1]][1][1]]) \\ \mathrm{CAV}([6[6[6][1]][1][1][1]]) &= \mathrm{CAV}([6[6[1][1][1]][1][1][1]]) \end{aligned} \quad (7)$$

addresses the rank deficiency and allows CAV([6[6[1][1][1]][6[1][1]][1]]) to be parameterized.

Constraint II is applied when two components that cannot be addressed through constraint I share a model compound. For example, 1-butene is a model compound for both [6[6[6][1][1]][6[1][1]][1]] and [6[6[1][1][1]][6[6][1]][1][1]], resulting in a rank deficiency that cannot be addressed using constraint I (Figure 2b). Because model compounds in TCIT are generated based on two-bond specificity, when exactly two components share a model compound, they will always be bonded to one another. We resolve this deficiency by introducing a constraint that the difference between the CAVs of the bonded components must be equal to the difference between the CAVs of similar components in related model compounds (constraint II). Denoting the two bonded components causing the rank deficiency as 1 and 2 (Figure 2b), two new model compounds with related components are generated by replacing 1 and 2 with the special terminal components (i.e., $-CH_3$, $=CH_2$, or $\equiv CH$) that retain the Lewis structure of 2 and 1, respectively. Denoting the corresponding components in the new model compounds as 3 and 4, constraint II requires

$$\mathrm{CAV}_1 - \mathrm{CAV}_2 = \mathrm{CAV}_3 - \mathrm{CAV}_4 \quad (8)$$

The CAVs for component 3 and 4 can be directly obtained after applying constraint I to the new model compounds. Thus, the second equation supplied by constraint II lifts the rank deficiency in calculating the CAVs for components 1 and 2.

Constraint III is applied when three or more components share a single model compound (e.g., the compound in Figure 2c with four terminal components sharing a single model compound). Because model compounds in TCIT are generated based on two-bond specificity, when three or more components share a model compound, all but one will be terminal components. In these cases, the rank deficiency is addressed using new model compounds generated by bonding each underdetermined terminal component with the corresponding special terminal component (i.e., $-CH_3$, $=CH_2$, or $\equiv CH$) that retains its Lewis structure. CAVs for these terminal components are then defined as identical to the corresponding CAVs in the new model compounds. An example of generalized of constraint III is shown in Figure 2c. After determining the CAVs for the terminal components by constraint III, the remaining non-terminal component can be determined by eq 1. When constraint III is applied to terminal carbon-based components, it is identical to constraint I and can be considered its generalization. Constraint III is only required for highly substituted small compounds, and its application is rare in comparison with constraints I and II.

2.5. Benchmark Data. To generate a CAV database and validate the accuracy of TCIT, we selected testing compounds from the PNK ΔH_f database[10] that is a core dataset for fitting Benson groups.

Initially, we obtained 1417 C, H, O, N, S, and halogen-containing chemical species with the gaseous phase ΔH_f available and 2420 chemical species with the condensed phase ΔH_f available. Because ring-containing compounds require strain corrections that will be addressed in future work, we retained the 666 acyclic compounds with the gaseous phase ΔH_f available for this study. The NIST Chemistry WebBook[34] was used to make a judgment when a difference of more than 10 kJ/mol existed between the TCIT prediction and PNK data. For the cases where no NIST data were available for comparison, the corresponding molecules were removed from the benchmark evaluation. We note that NIST generally adopts PNK values; however, where outliers were found in our predictions, NIST likewise typically excluded those PNK values from the WebBook, preferring instead to present no data for the corresponding compounds or replacing them with alternative references. Ninety compounds exhibited an absolute difference of more than 10 kJ/mol between the TCIT prediction and PNK data. For 24 of these compounds, NIST provided alternative references that were utilized instead of PNK; for 34 compounds, NIST provided only PNK values, and these were retained in the benchmark set; and for 32 compounds, NIST provided no data, and these compounds were removed from the benchmark data. In addition, generating model compounds for nitro-containing molecules requires a treatment of formal charges in TCIT that goes beyond the scope of the current work. Thus, the six acyclic nitro-containing compounds in the PNK dataset were also

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

excluded in the present study. Finally, BGIT predictions were performed using the CHETAH software for comparison with TCIT.[35] During BGIT evaluation, we discovered two compounds that could not be predicted by CHETAH because of missing group values, and these were also removed from the benchmark dataset. A list of all excluded molecules is shown in Table S2. The final dataset consisted of 626 acyclic molecules for generating the CAVs database and validating the accuracy of TCIT.

The benchmark dataset was further split into three partially overlapping subsets based on the comparisons that can be performed in each subset. First, 348 of the compounds in the dataset are small enough to be TCIT model compounds; thus, these were only used to evaluate the accuracy of the G4 method with respect to the experimental data and not for evaluating the accuracy of TCIT. Second, the remaining 278 PNK compounds that are not TCIT model compounds were used for comparing the predictions of TCIT and BGIT, as implemented in CHETAH,[35] with the experimental values. Finally, out of the 278 compounds that are not TCIT model compounds, 224 have no more than 12 heavy (nonhydrogen) atoms, making it feasible to perform G4 predictions. These 224 compounds were used to compare the predictions of TCIT with G4.

TCIT is implemented as a Python package and maintains databases of all G4 calculations and CAVs for future use.

3. RESULTS AND DISCUSSION

In the following sections, benchmark results are presented to evaluate the accuracy of TCIT with respect to experiment, G4 calculations, and BGIT predictions. Mean signed errors (MSEs) are reported for each comparison to evaluate systematic biases, and mean absolute errors (MAE) are reported to evaluate average accuracy. In the final section, additional results are presented on the distribution of errors with respect to the presence of specific functional groups.

3.1. G4 Benchmarking. TCIT is parameterized to G4 results for small model compounds. Thus, the error in TCIT predictions can be decomposed into errors associated with the component theory and errors associated with the underlying G4 parameterization data. Previous work has demonstrated that G4 predictions exhibit chemical accuracy (<1 kcal/mol) for a broad range of chemistries.[24] Here, we investigate this potential source of error with respect to the PNK dataset. Because G4 is expensive (with a scaling of N^7 with respect to basis functions[24]), only a subset of the training compounds can be characterized with this method. Thus, G4 calculations were performed on all training molecules with no more than 12 heavy atoms, resulting in 572 reference values that are compared with corresponding experimental values in Figure 3a. We observe a MSE of −0.06 kJ/mol, which demonstrates that there is no systematic bias in the G4 predictions for these compounds. The observed MAE is 4.19 kJ/mol (~1 kcal/mol), which is comparable with the experimental uncertainty,[7,33] and confirms that the G4 method (including conformational averaging as described above) is accurate enough to support parameterizing CAVs solely from the computational results. Assuming that these errors are representative of how G4 would perform across the whole PNK dataset, we anticipate TCIT would thus also exhibit a MAE of at least 4 kJ/mol because of a combination of G4 errors and experimental uncertainty.

3.2. G4 and TCIT Comparisons. To characterize the errors associated with the TCIT component decomposition, we performed a comparison between TCIT and G4 for a 224 compound subset of the PNK dataset that is large enough not to be model compounds for TCIT but still small enough to make G4 calculations feasible. The comparison shown in Figure 3b is thus diagnostic of how well the TCIT approximations reproduce G4 predictions on larger compounds. The resulting MSE of −0.18 kJ/mol indicates that TCIT exhibits negligible systematic bias in predicting $\Delta H_{f,G4}$ while the MAE of 2.30 kJ/mol demonstrates that our new component definition exhibits chemically accurate reproduction of the G4 results. The MAE of TCIT compared with G4 is also significantly smaller than the MAE of G4 compared with the experimental data. Propagating these two sources of error, we would anticipate that TCIT exhibits a MAE of 4.8 kJ/mol with respect to experimental predictions.

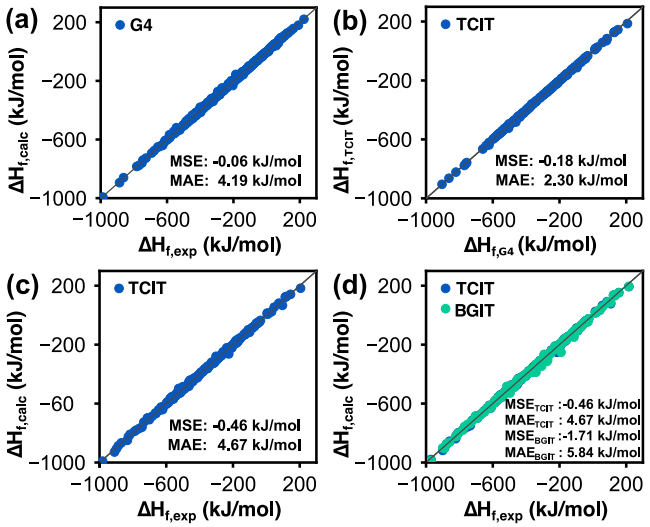


Figure 3. Correlation plots showing comparisons between benchmark experimental data, G4, TCIT, and BGIT. (a) Comparison between experimental gas-phase ΔH_f values and G4 predictions for 572 compounds with 12 or less heavy atoms in the benchmark dataset. (b) Comparison between G4 and TCIT predictions for the 224 compounds in the benchmark dataset with 12 or less heavy atoms and excluding model compounds. (c) Comparison between experiment and TCIT for the 278 nonmodel compounds in the benchmark dataset. (d) Comparison between experiment and TCIT (blue) and BGIT (green) for the 278 nonmodel compounds in the benchmark dataset. MSE and MAE statistics are presented for each comparison.

3.3. TCIT−BGIT and Experimental Comparisons. To characterize the prediction errors for TCIT, we performed TCIT predictions for the 278 molecules in the PNK dataset that is large enough not to be model compounds (i.e., none of these compounds were utilized in parameterizing the CAVs). The comparison between TCIT predictions on these compounds and their experimental ΔH_f is shown in Figure 3c. We observe a MSE of −0.46 kJ/mol, which indicates a negligible systematic bias in TCIT predictions in comparison with the experimental values. The observed MAE is 4.67 kJ/mol is likewise comparable to the experimental uncertainty and consistent with the estimate of 4.8 kJ/mol based on error propagation.

To provide a comparison of TCIT performance with a state-of-the-art implementation of group theory, we also performed BGIT predictions using the CHETAH software[35] for these same compounds (Figure 3d). We note that although none of these compounds were utilized in TCIT parameterization, the PNK dataset has historically been utilized in BGIT parameterization. Thus, many of these compounds are likely training compounds

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

for CHETAH's implementation of BGIT, and this comparison is biased in its favor. Nevertheless, we observe that TCIT exhibits generally superior performance to BGIT across the test set. Specifically, the MSE and MAE for BGIT are −1.71 and 5.84 kJ/mol, respectively, and both larger than the corresponding TCIT results.

3.4. Functional Group Dependence. The results for TCIT and BGIT presented in the previous section were further evaluated on the basis of functional groups, with the corresponding MSE and MAE for molecules exhibiting each group presented in Figure 4. Molecules in the alkanes and

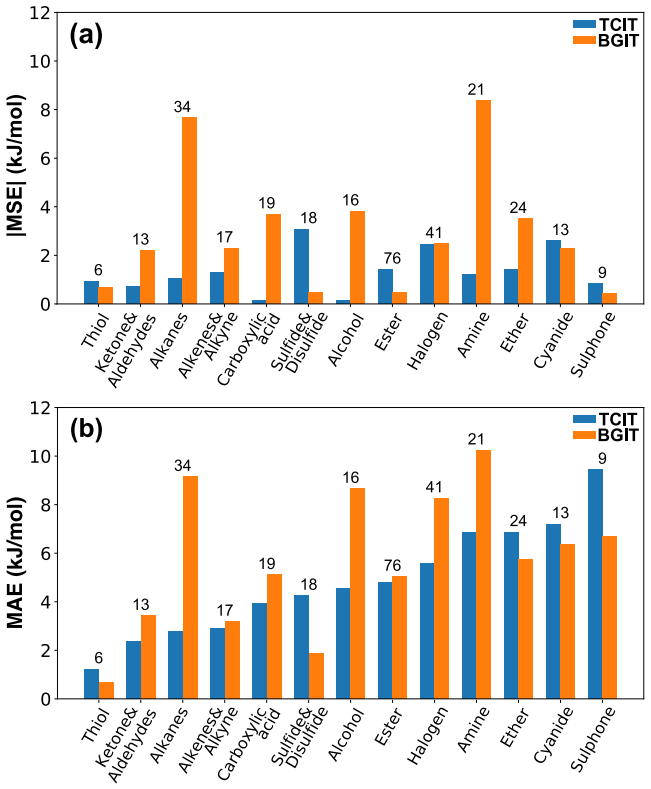


Figure 4. Prediction errors classified by the functional group for TCIT (blue) and BGIT (orange). Benchmark compounds are classified based on if they exhibit the listed functional group(s) and can be in more than one class. Compounds in the alkane, alkene, and alkyne categories are exclusively composed of carbon and hydrogen. (a) Absolute value of the MSE for compounds in each category. (b) MAE for compounds in each category.

"alkenes and alkynes" categories are exclusively composed of carbon and hydrogen; for all other categories, molecules are included if they exhibit the specified functional group. Thus, molecules may be included in more than one category. We have also combined similar functional groups in some cases so each category has at least five compounds. Evaluating the error distributions at this level of granularity enables us to identify several salient differences between BGIT and TCIT.

First, BGIT exhibits large systematic errors for several functional groups, including alkanes and amines. For comparison, we have calculated the standard deviation in MSE for 20 molecule subsets (i.e., approximately the average category size) using bootstrap resampling. The resulting bootstrapped MSE means are −0.49 and −1.71 kJ/mol, and the standard deviations are 1.56 kJ/mol and 2.22 kJ/mol for TCIT and BGIT, respectively. The BGIT MSE for alkanes and amines is more than three standard deviations from the bootstrap mean, indicating that these groups exhibit anomalously poor performance. In contrast, for TCIT, no functional group exhibits an MSE beyond two standard deviations. The systematic errors associated with alkanes and amines are potentially addressable through reparameterization, or they may reflect comprises made by the CHETAH developers that are not reflected in the PNK dataset. Regardless, the more consistent distribution of MSE across functional groups in TCIT reflects positively on the low systematic bias associated with the model compound generation and G4 parameterization data that underlie the method.

Second, the large BGIT errors associated with alkanes are initially surprising, considering the large amount of parameterization data available for these species. However, we observe that the errors are largest for branched alkanes, which is consistent with the absence of non-nearest neighbor effects in BGIT. In contrast, TCIT errors for alkanes are among the lowest of all functional groups. In the component theory, methylene groups adjacent to branch points are distinct and can reflect steric contributions from 1,2 and 1,3 substitutions.

Third, we note that BGIT exhibits lower MAE than TCIT for all sulfur-containing functional groups (i.e., thiol, sulfides, and sulphones). To identify the source of TCIT error in these compounds, we have also presented the G4 errors with respect to the experimental values, analyzed according to the functional group, in Figure S2. We observe that the G4 model errors are closely correlated with the TCIT errors, which suggests that the TCIT errors on the benchmark compounds are because of the underlying G4 results or experimental uncertainty, rather than the component theory itself.

4. CONCLUSIONS

We have introduced a self-consistent component increment theory, TCIT, for calculating gas-phase ΔH_f and have evaluated its performance against the conventional group increment theory, BGIT. In direct comparisons on the PNK benchmark dataset, we observe that TCIT outperforms BGIT with respect to overall MAE, MSE, and the distribution of errors with respect to functional groups. The improved performance of TCIT compared to BGIT is remarkable, considering that TCIT is parameterized only to quantum chemistry data and that the PNK dataset has historically been used to parameterize BGIT implementations. On the test set of compounds, TCIT exhibits a MAE of 4.67 kJ/mol, which is commonly considered chemical accuracy and within the experimental uncertainty of these comparisons.[7,33] Moreover, upon investigating the origin of these small errors in TCIT, we observe that they are strongly correlated with the prediction errors of the underlying quantum chemistry calculations (G4). Thus, we observe that TCIT is capable of extending the predictions of G4 theory with near-perfect fidelity, and we anticipate that it could be applied to other quantum chemistry methods, should demonstrably better methods become available or more accurate experimental data motivate a reevaluation of this choice.[36−40]

TCIT is the first component theory derived exclusively from quantum chemistry data, with a systematic protocol for generating model compounds and deriving CAVs. In particular, these model compounds are relatively small, making them amenable to high-level quantum chemistry characterization, and each component is derived from one model compound. In combination, these features of TCIT make it arbitrarily extensible to new chemistries while maintaining self-consistency. In the current work, we supply over 1600 CAVs applicable

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207

to acyclic C, H, O, N, S, and halogen-containing molecules (Table S5). These CAVs include coverage for most common organic functional groups involving these elements and can be utilized by readers in their own work either through the CAV tables or a Python script supplied in the Supporting Information.

The current study is limited to gas-phase ΔH_f predictions of acyclic linear compounds, but several extensions to other properties and classes of molecules are obvious and underway. In terms of properties, TCIT is primarily limited by the accuracy and availability of the underlying quantum chemistry data. For instance, condensed-phase ΔH_f predictions will require the use of solvation free energy models that could affect accuracy and generality. In terms of new classes of molecules, the presented implementation of TCIT supports deriving CAVs for elements beyond the second row, but this will require separate benchmarking. To extend TCIT to cyclic molecules, ring corrections are required to account for missing strain contributions to ΔH_f. Because TCIT is based solely on quantum chemistry data, it can likewise be extended to radical and ionic species that are poorly represented in conventional BGIT because of limited experimental data. The extension to ions and radicals will require a more general treatment of formal charge propagation in the model compounds than has been presented here.

In their 1993 review, Cohen and Benson wrote that "it is our opinion that, while component additivity as a method certainly has no conceptual flaws; nevertheless, the database in general is neither refined nor extensive enough to provide the enormous number of components that are necessary to calculate thermochemical properties for all organic compounds."[7] More than 20 years later, circumstances have changed to make component theories viable, including the maturation of quantum chemistry methods for accurately predicting many thermodynamic properties, as well as the vastly increased calculation throughput enabled by processor availability.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00092.

Molecular size distribution of model compounds, G4 prediction errors classified by the functional group, tables with ΔH_f reference values and predictions from all applicable methods for PNK molecules, and tables with component increment values (PDF)

A python implementation of TCIT ΔH_f prediction for reproducing the data in this manuscript (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Brett M. Savoie − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States;* orcid.org/0000-0002-7039-4039; Email: bsavoie@purdue.edu

Author

Qiyuan Zhao − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00092

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work performed by Q.Z. was made possible through support of the Purdue Process Safety and Assurance Center. The work performed by B.M.S. was made possible through the Air Force Office of Scientific Research (AFOSR) under support provided by the Organic Materials Chemistry Program (grant number: FA9550-18-S-0003, Program Manager: Dr Kenneth Caster). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant no. ACI-1548562. Simulations were performed on the Comet supercomputer at the University of California, San Diego, under the Allocation no. TG-CHE190014.

■ REFERENCES

(1) Sabbe, M. K.; Saeys, M.; Reyniers, M.-F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. Group additive values for the gas phase standard enthalpy of formation of hydrocarbons and hydrocarbon radicals. *J. Phys. Chem. A* **2005**, *109*, 7466−7480.

(2) Sabbe, M. K.; De Vleeschouwer, F.; Reyniers, M.-F.; Waroquier, M.; Marin, G. B. First principles based group additive values for the gas phase standard entropy and heat capacity of hydrocarbons and hydrocarbon radicals. *J. Phys. Chem. A* **2008**, *112*, 12235−12251.

(3) Suleimanov, Y. V.; Green, W. H. Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248−4259.

(4) Green, W. H.; Barton, P. I.; Bhattacharjee, B.; Matheu, D. M.; Schwer, D. A.; Song, J.; Sumathi, R.; Carstensen, H.-H.; Dean, A. M.; Grenda, J. M. Computer construction of detailed chemical kinetic models for gas-phase reactors. *Ind. Eng. Chem. Res.* **2001**, *40*, 5362−5370.

(5) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press, 1960; p 260.

(6) Benson, S. W.; Buss, J. H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *J. Chem. Phys.* **1958**, *29*, 546−572.

(7) Cohen, N.; Benson, S. W. Estimation of heats of formation of organic compounds by additivity methods. *Chem. Rev.* **1993**, *93*, 2419−2438.

(8) Joback, K. G.; Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **1987**, *57*, 233−243.

(9) Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE J.* **1994**, *40*, 1697−1710.

(10) Pedley, J.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*; Springer Science & Business Media, 1986.

(11) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279−324.

(12) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. Revised group additivity parameters for the enthalpies of formation of oxygen-containing organic compounds. *J. Phys. Chem.* **1973**, *77*, 1687−1691.

(13) Holmes, J. L.; Aubry, C. Group additivity values for estimating the enthalpy of formation of organic compounds: an update and reappraisal. 1. C, H, and O. *J. Phys. Chem. A* **2011**, *115*, 10576−10586.

(14) Holmes, J. L.; Aubry, C. Group additivity values for estimating the enthalpy of formation of organic compounds: an update and reappraisal. 2. C, H, N, O, S, and halogens. *J. Phys. Chem. A* **2012**, *116*, 7196−7209.

(15) Minenkov, Y.; Wang, H.; Wang, Z.; Sarathy, S. M.; Cavallo, L. Heats of formation of medium-sized organic compounds from

contemporary electronic structure methods. *J. Chem. Theory Comput.* **2017**, *13*, 3537−3560.
(16) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gn theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 810−825.
(17) Iovanac, N. C.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**, *123*, 4295−4302.
(18) 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.
(19) Sumathi, R.; Green, W. H. Missing thermochemical groups for large unsaturated hydrocarbons: Contrasting predictions of G2 and CBS-Q. *J. Phys. Chem. A* **2002**, *106*, 11141−11149.
(20) Khan, S. S.; Yu, X.; Wade, J. R.; Malmgren, R. D.; Broadbelt, L. J. Thermochemistry of radicals and molecules relevant to atmospheric chemistry: determination of group additivity values using G3//B3LYP theory. *J. Phys. Chem. A* **2009**, *113*, 5176−5194.
(21) Ince, A.; Carstensen, H.-H.; Reyniers, M.-F.; Marin, G. B. First-principles based group additivity values for thermochemical properties of substituted aromatic compounds. *AIChE J.* **2015**, *61*, 3858−3870.
(22) Wang, H.; Castillo, Á.; Bozzelli, J. W. Thermochemical Properties Enthalpy, Entropy, and Heat Capacity of C1-C4 Fluorinated Hydrocarbons: Fluorocarbon Group Additivity. *J. Phys. Chem. A* **2015**, *119*, 8202−8215.
(23) Savoie, B. M.; Webb, M. A.; Miller, T. F., III Enhancing cation diffusion and suppressing anion diffusion via Lewis-acidic polymer electrolytes. *J. Phys. Chem. Lett.* **2017**, *8*, 641−646.
(24) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.
(25) Paraskevas, P. D.; Sabbe, M. K.; Reyniers, M.-F.; Papayannakos, N.; Marin, G. B. Group additive values for the gas-phase standard enthalpy of formation, entropy and heat capacity of oxygenates. *Chem.—Eur. J.* **2013**, *19*, 16431−16452.
(26) Sanderson, R. T. Electronegativity and bond energy. *J. Am. Chem. Soc.* **1983**, *105*, 2259−2261.
(27) Sanderson, R. *Chemical Bonds and Bonds Energy*; Elsevier, 2012; Vol. *21*.
(28) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(29) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab-Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 8.
(30) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(31) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(32) 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-Young, D.; 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*, Revision C.01.; Gaussian Inc.: Wallingford CT, 2016.
(33) 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.
(34) Linstrom, P.; Mallard, W. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology, 2018.
(35) Seaton, W. H. *CHETAH-The ASTM Chemical Thermodynamic and Energy Release Evaluation Program*; American Society for Testing and Materials, 1974.
(36) Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges. *Theor. Chem. Acc.* **2012**, *131*, 1079.
(37) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions. *J. Chem. Phys.* **2006**, *125*, 144108.
(38) 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.
(39) Feller, D.; Peterson, K. A.; Dixon, D. A. A survey of factors contributing to accurate theoretical predictions of atomization energies and molecular structures. *J. Chem. Phys.* **2008**, *129*, 204105.
(40) Feller, D.; Peterson, K. A.; Grant Hill, J. On the effectiveness of CCSD(T) complete basis set extrapolations for atomization energies. *J. Chem. Phys.* **2011**, *135*, 044102.

https://dx.doi.org/10.1021/acs.jcim.0c00092
J. Chem. Inf. Model. 2020, 60, 2199−2207