JOURNAL OF CHEMICAL INFORMATION AND MODELING

Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds

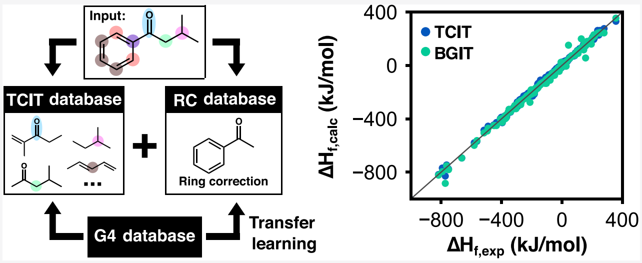
Qiyuan Zhao, Nicolae C. Iovanac, and Brett M. Savoie*

Cite This: J. Chem. Inf. Model. 2021, 61, 2798−2805

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computational predictions of the thermodynamic properties of molecules and materials play a central role in contemporary reaction prediction and kinetic modeling. Due to the lack of experimental data and computational cost of high-level quantum chemistry methods, approximate methods based on additivity schemes and more recently machine learning are currently the only approaches capable of supplying the chemical coverage and throughput necessary for such applications. For both approaches, ring-containing molecules pose a challenge to transferability due to the nonlocal interactions associated with conjugation and strain that significantly impact thermodynamic properties. Here, we report the development of a self-consistent approach for parameterizing transferable ring corrections based on high-level quantum chemistry. The method is benchmarked against both the Pedley−Naylor−Kline experimental dataset for C-, H-, O-, N-, S-, and halogen-containing cyclic molecules and a dataset of Gaussian-4 quantum chemistry calculations. The prescribed approach is demonstrated to be superior to existing ring corrections while maintaining extensibility to arbitrary chemistries. We have also compared this ring-correction scheme against a novel machine learning approach and demonstrate that the latter is capable of exceeding the performance of physics-based ring corrections.



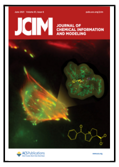
1. INTRODUCTION

Thermochemistry prediction plays an important role in many emerging applications including automated synthetic planning, reaction prediction, and process safety.[1−6] For these applications, transferable (i.e., applicable to a broad range of molecular structures and functional groups), accurate (i.e., absolute prediction errors within 1 kcal/mol), and inexpensive (i.e., <1 min for on-the-fly applications) methods are required that can supply on-demand predictions for diverse chemical structures. Group additivity models have historically filled this capability gap, albeit with limited chemical coverage.[7−10] In additivity schemes, the thermochemical property of interest (denoted as P) is assumed to be decomposable into a sum of the contributions from different parts of the molecule

$$P = \sum_{i} p_i \tag{1}$$

where p_i is the contribution of each part and is defined as the additivity value. Additivity schemes have been under continuous development since their formalization by Benson and Buss.[7−16] In the Benson group increment theory (BGIT), thermochemistry properties are decomposed into "groups", which are defined for each nonterminal atom (i.e., atoms bonded to at least two neighbors), and the group additivity values (GAVs) are defined to be specific to the central atom and its bonded neighbors. GAVs are typically fit in a least-squares manner to a combination of experimental data or high-accuracy quantum chemistry data, after which the GAVs can be used to predict properties for any molecule that exhibits the corresponding groups. We recently reported a novel additivity framework for predicting the enthalpy of formation (ΔH_f) of molecular species, called the TAFFI component increment theory (TCIT).[17] Among the unique features of this approach is that the additivity values are chemically specific out to two bonds from the central atom (i.e., it is a "component" theory in the hierarchy defined by Benson) and that it prescribes a self-consistent and automated framework for generating model compounds and parameterization data. In combination, these features of TCIT make it arbitrarily extensible to new chemistries. Additionally, benchmarking results show that TCIT outperforms state-of-the-art implementations of BGIT in ΔH_f predictions of linear compounds. Despite this promising initial demonstration, the treatment of ring-containing compounds was omitted due to the unique challenges conjugation and ring strain pose to the accurate

Received: March 30, 2021
Published: May 25, 2021



ACS Publications

https://doi.org/10.1021/acs.jcim.1c00367

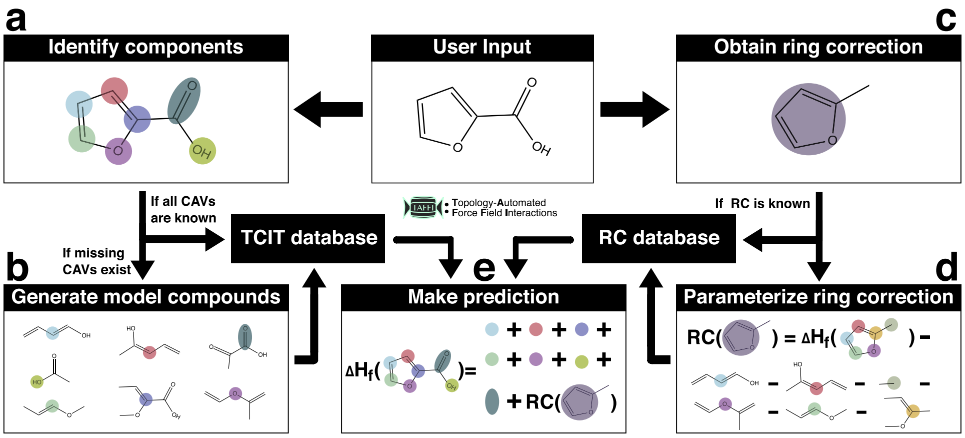


Figure 1. Overview of the TCIT component decomposition, linear and ring model compound generation, CAV and RC parameterization, and final prediction. (a) Components are uniquely defined for each nonterminal atom in target molecules based on acyclic subgraphs of neighbor atoms out to two bonds. (b) Linear model compounds for corresponding components are generated for parameterizing CAVs. (c) When the target molecules contain rings, ring model compounds (RMCs) are automatically generated for each distinct ring structure. (d) G4 calculations of ΔH_f are performed on all linear and ring model compounds. The corresponding RC is parametrized based on eq 4. (e) Predictions are made based on eq 2 when all requisite CAVs and RCs exist within the TCIT database.

prediction of thermodynamic properties. In the current work, we address this omission by implementing self-consistent and transferable ring corrections that can be utilized alongside TCIT predictions.

Additivity schemes are inherently local, meaning that the assumption of decomposability implies that long-range interactions (i.e., beyond two bonds for group schemes and beyond four bonds for component schemes) make a negligible contribution to thermochemistry properties. This assumption is known to break down for ring-containing molecules, where the strain and conformational restriction of the ring substantially impact properties like ΔH_f and molar enthalpy ($S°$).

To address this issue, Benson and others proposed additional additive corrections associated with each ring according to

$$P = \sum_i p_i + \sum_j \mathrm{RC}_j \tag{2}$$

where the index j runs over all distinct rings and RC_j is a ring-specific correction. Within the context of BGIT, distinct ring corrections must be derived for every ring substructure, which substantially increases the training data requirements. Several groups have also derived RCs for BGIT based on quantum chemistry that generally show excellent accuracy.[18,19] Nevertheless, developing transferable RCs with broad chemical coverage represents a long-term challenge for BGIT.

In the present work, we explore two complementary approaches to provide ring corrections to TCIT. In the first approach, we implement a method to systematically generate ring model compounds and parameterize RCs for use with TCIT component additivity values (CAVs). This approach is analogous to the RCs developed by Benson, while addressing the issues with extensibility, selectivity, and provenance that are intrinsic to BGIT. In the second approach, we utilize a transfer learning approach whereby we train neural network (NN)-based models to learn the RCs on the basis of TCIT predictions and the ring structure. We observe that the hybrid ML/TCIT approach shows the best overall performance of all approaches.

2. COMPUTATIONAL METHODS

For a detailed description of the TCIT methodology for noncyclic compounds, we direct readers to our previous publication.[17] Here, we briefly recapitulate the central features of TCIT as they pertain to the present extension for calculating ring corrections. The TCIT methodology consists of separate steps for (i) defining component additivity values (CAVs), (ii) generating model compounds and conformers for high-level quantum chemistry characterization, and (iii) parameterizing CAVs based on that data (Figure 1a,b). In all steps (i–iii), TCIT represents a departure from conventional additivity schemes. In (i), CAVs are unambiguously defined on the basis of local adjacency relationships out to a depth of two bonds. In (ii), model compounds are generated based on the smallest noncyclic exemplary structures that exhibit a given component. In (iii), CAV parameterization is performed with a one-to-one correspondence between CAVs and model compounds, such that new CAVs can be added to TCIT while maintaining backward compatibility. For the present goal of developing ring corrections, we sought to maintain these distinctions while enabling accurate predictions of cyclic structures. This is accomplished by developing additive corrections that are applied on a per-ring basis, rather than developing distinct CAVs for components within rings. In the strategy adopted here, the ΔH_f of a ring-containing molecule is predicted according to

$$\Delta H_f = \sum_i \mathrm{CAV}_i + \sum_j \mathrm{RC}_j \tag{3}$$

where RC_j represents the ring corrections for all distinct rings present in the system. In this approach, the ring correction accounts for the difference between the (noncyclic) TCIT prediction and the ground truth value for the ring-containing structure (Figure 1c,d). Since this approach is additive, it leaves the core extensibility of TCIT unaffected. On the other hand, it necessitates a definition of unique rings, ring model compounds, and parameterizations that conform to (i)−(iii). A detailed description of these aspects is provided in the subsequent sections (Sections 2.1−2.3).

2.1. TCIT Ring-Component Definition. The procedure for defining component types for ring-containing molecules is identical to that for acyclic molecules.[17] Target molecules are first converted to a chemical graph described by an adjacency matrix. Components are defined for each nonterminal atom (i.e., atoms with more than one bonded neighbor) based on the acyclic subgraph of its neighbors out to two bonds. These acyclic subgraphs are generated by recursively growing the subgraph based on the adjacency matrix of the target molecule (Figure 2). Each subgraph is seeded with the nonterminal atom, its bonded neighbors are added to the subgraph, then their bonded neighbors are added to the subgraph. Critically, this growth procedure generates an *acyclic* subgraph that corresponds to an equivalent component derived from a linear compound. For small rings with less than five atoms, this distinction is important because simply taking the subgraph from the original adjacency matrix would retain the ring in the subgraph. For components within rings larger than four atoms, this definition is identical to the definition for acyclic components. Model compounds are generated from these subgraphs and CAVs are parameterized using the procedures reported in the original TCIT paper.[17]

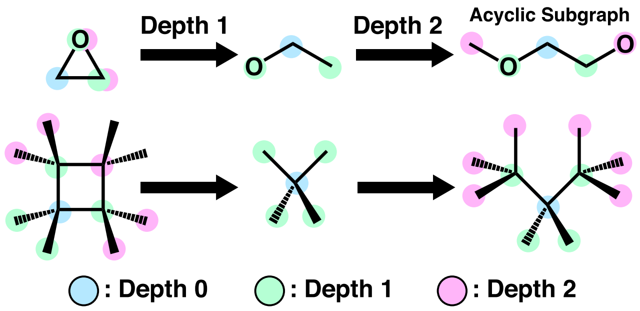


Figure 2. Illustration of acyclic subgraph generation starting at the nonterminal atoms indicated in blue (depth = 0) for a three-membered ring (top) and four-membered ring (bottom). Atoms are added to the subgraph starting with bonded neighbors (depth = 1) and then next-nearest neighbors (depth = 2). Adding the atoms and bonds at each recursion results in an acyclic graph with an equivalent number and type of atoms at each neighbor depth from the central atom in the subgraph.

2.2. TCIT Ring Corrections. Because the CAVs in TCIT are derived from acyclic model compounds, they do not capture ring strain or aromaticity contributions to the thermodynamic properties.[19] To address this problem, Benson and Cohen introduced ring corrections to describe missing thermodynamic contributions for specific rings.[9] For a molecule with known ΔH_f and containing a single ring, the ring correction can be calculated by rearranging eq 3

$$\mathrm{RC} = \Delta H_\mathrm{f}(\mathrm{ring}) - \sum_i \mathrm{CAV}_i \tag{4}$$

As originally proposed by Benson and Cohen, RC is assumed to be independent of any groups attached to the ring. This definition is widely accepted in many group additivity schemes due to its high transferability and the limited experimental data available for parameterizing RCs. However, only considering the ring structure itself neglects potentially substantial substitution effects that can modulate strain energy and, where applicable, aromaticity.[18] In the context of TCIT, we designate a ring correction that only includes the ring, and no substituents, as a "depth = 0" correction, RC_0, meaning that it contains no information on ring substitutions (Figure 3). In keeping with the TCIT component definitions, it would be ideal to define ring corrections to be specific to rings and all of their substituents out to two bonds (i.e., RC_2). However, even with modern quantum chemistry throughput, such a definition would have extremely limited transferability and would require large model compounds for parameterization. To address the limited accuracy of ring corrections that neglect substitution effects, we have implemented and benchmarked two complementary strategies.

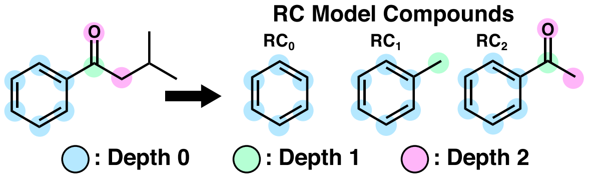


Figure 3. Illustration of ring corrections with varying substituent specificity (left) and model compounds for the corresponding ring corrections (right).

In the first approach, TCIT-RC1, we derive a set of ring corrections that are specific to each ring and its nearest bonded atoms (i.e., RC_1, Figure 3). This reflects a compromise between transferability and accuracy, such that some of the effects of substitution can be captured while maintaining manageable parameterization costs. In this approach, each ring correction is parameterized on the basis of eq 4 using a ring model compound (RMC) composed of the ring and its nearest bonded atoms, with undercoordinated atoms hydrogenated to a level consistent with the Lewis structure of the ring (Figure 3). In eq 4, ΔH_f of the model compound is determined based on high-level quantum chemistry (*vide infra*) and the CAVs are derived from the corresponding linear model compounds, as described previously.

In the second approach, TCIT-RC2, we have trained a graph convolutional neural network[20] to predict ring corrections that are specific to each ring and its substituents out to two bonds (i.e., the depth = 2 ring correction, RC_2 Figure 3). This approach achieves the targeted specificity, albeit at the expense of a well-defined error estimate for performance on rings that significantly deviate from the training structures. Our graph neural network is trained to predict the difference $\mathrm{RC}_2-\mathrm{RC}_0$ (Figure 3). The model uses the RC_0 and RC_2 RMCs as inputs for the prediction task. We note that the RC_0 correction contains the leading order contribution to RC_2 (e.g., $|\mathrm{RC}_0| \gg |\mathrm{RC}_0 - \mathrm{RC}_2|$ for the majority of cases). By predicting the difference, we are leveraging a transfer learning (TL) strategy, whereby the model only needs to learn the substituent effects and not the leading order contribution from RC_0. To make predictions on new structures, this requires that RC_0 is on hand, but these are typically derived from small structures and it is thus manageable to calculate them as needed and store them in a database. Implementation details and network architecture are provided in the SI.

2.3. Quantum Chemistry Methods. As previously reported, CAVs are parameterized in TCIT on the basis of high-level quantum chemistry data for small acyclic model compounds containing each component.[17] We presently adopt Gaussian-4 (G4)[21] as implemented in the Gaussian16 software package[22] as our high-level method, although the parameterization protocol is independent of the chosen level of theory. Each model compound is subjected to conformer searching, geometry optimization, frequency calculations, G4 characterization, and Boltzmann weighted averaging to obtain reference ΔH_f values for each structure. Other than the exception noted below, these characterization steps are identical for the model compounds associated with our two ring-correction schemes. One difference concerns conformer generation for the RMCs. For acyclic compounds, we use the "confab" algorithm in Open Babel to generate conformers.[23,24] This algorithm applies specific torsion rules to every rotatable bond in the structure; however, this is not suitable for exploring conformers of cyclic compounds. To circumvent this problem, bond-breaking steps are performed to open the cyclic structure before performing conformer generation. After generating the acyclic conformers, the broken bond is reformed and the geometry is optimized into the nearest conformational basin using the MMFF94 force field.[25] We note that conformer generation algorithms that are compatible with rings have recently been published that could improve on this approach.[26] All conformers generated by this process are subjected to semiempirical Geometry, Frequency, Non-covalent, eXtended Tight Binding method (GFN2-xTB)[27,28] geometry optimization, and single-point energy calculations. Consistent with the previous TCIT procedure, conformers with energy differences larger than 2 kJ/mol are considered as distinct, conformers with energies greater than 5 kcal/mol of the minimum energy conformer are discarded, and (up to) 6 of the most energetically stable conformers are selected for further quantum chemistry characterization and use in Boltzmann weighted averaging of ΔH_f.

2.4. Benchmark Data. The selected cyclic testing compounds come from the Pedley Naylor Kline (PNK) ΔH_f database,[16] which is one of the commonly used databases for parameterizing group increment methods. From 1417 C-, H-, O-, N-, S-, and halogen-containing chemical species with gaseous-phase ΔH_f available, we retained the 635 cyclic compounds for this study. Those cyclic compounds can be further divided into two categories: simple and complex ring systems. If one compound contains a bond or an atom shared by at least two rings, it is defined as a complex ring system (i.e., a fused ring), otherwise it falls into the simple ring system category. In this work, we first focus on 475 simple cyclic compounds and leave complex cyclic compounds for future consideration. In particular, several strategies are feasible for decomposing complex rings to avoid parameterizing them whole. Among the 475 simple cyclic compounds, 15 nitro group-containing compounds were removed from the TCIT testing set because nitro group decomposition requires a treatment of formal charges that goes beyond the scope of the current work. Additionally, to constrain computational costs, 15 compounds containing larger than 10-membered rings or with corresponding ring model compounds (RMC) containing more than 12 heavy atoms were excluded from the database. Out of the remaining compounds, 232 are so small that they are RMCs for the RC_1 model, thus these are only used to validate the accuracy of G4 in comparison with the experimental data. TCIT validation is performed on the remaining 213 relatively large compounds that are not directly utilized in RC parameterization. We have also performed G4 calculations on the non-RMC structures exhibiting less than 12 heavy atoms (133 compounds), which is used as a validation set for comparing the performance of TCIT against G4. The NIST Chemistry WebBook[29] was used to make a judgment when experimental data (PNK) and TCIT predictions (or G4 calculations) have a difference of more than 20 kJ/mol. For 32 compounds (22 from TCIT and 10 from G4), NIST provided no data and these compounds were excluded from the benchmark evaluation. BGIT predictions were also performed using the CHETAH software for comparison with TCIT.[30] During BGIT evaluation, we discovered 11 compounds that could not be predicted by CHETAH due to missing BGIT GAVs or missing corresponding ring corrections, and these were excluded from the BGIT/TCIT comparison dataset. Additionally, due to the prevalence of benzyl structures in many applications, BGIT adopts a special group type for benzene carbon atoms (C_B) that breaks the general group definition. For 7 compounds containing two or more benzene rings, BGIT exhibits perfect predictions, which we interpret as these compounds being reference structures for parameterizations involving the C_B group. To better reflect general performance, these 7 structures were excluded from the BGIT/TCIT comparisons and separate comparisons are performed between substituted benzene structures and other structures. Lists of all excluded molecules are shown in Tables S1 and S2.

The foregoing description of the data curation is summarized by the following partially overlapping data slices.

- 365 compounds with no more than 12 heavy atoms were used for testing G4 accuracy with respect to the experimental data.
- 191 compounds that are validated by the NIST database and are not TCIT RMCs were used for testing the accuracy of TCIT with ring corrections.
- Among the 191 TCIT testing compounds, 133 compounds with no more than 12 heavy atoms were used to compare the predictions of TCIT with G4 and 173 BGIT predictable compounds were used for comparing the predictions of TCIT and BGIT with experimental values.
- To clarify distinct error contributions, the 191 TCIT testing compounds were further divided into subsets of 72 and 119 compounds based on whether the corresponding RMC for RC_1 breaks conjugation between the substitution and the ring or not, respectively.
- Since benzene is treated as a special ring in BGIT, the 173 BGIT-TCIT comparison compounds were divided into two subsets that include 117 substituted benzenes and 56 other cyclic compounds.

2.5. Machine Learning Training Dataset. For training the ML-based ring corrections, it is necessary to assemble a training dataset that significantly exceeds the data available in the PNK database. In particular, a training dataset for the ML correction should have (i) examples of a broad range of ring types, (ii) examples of a broad range of substitutions on each ring type, and (iii) well-balanced data for each ring class and set of substitutions. To generate a training dataset that fulfills these criteria, we identified all of the unique ring types (i.e., unique RC_0 structures) in the PNK database and unique substituents on those rings (i.e., the RC_1 and RC_2 components that differentiated specific rings from RC_0). Out of the 635 distinct cyclic compounds in PNK, this evaluation identified 91 distinct ring types (RC_0) and approximately 7 classes of substitutions, including alkyl, alkenyl, alcohol, ketone, amide, imine, and halogen (fluoro, chloro, and bromo). To generate representative training and testing data, we applied an algorithm to sample model compounds based on random combinations of the distinct ring types and substitutions observed in PNK. This algorithm consisted of selecting a random set of substitution sites on each ring and applying randomly chosen substitutions while avoiding duplicates. In total, 1939 depth = 1 and 2889 depth = 2 RMCs were generated and their corresponding ring corrections were evaluated via eq 4 at the G4 level. The experimental PNK data, which was not used during training, was employed as a test set for ML models.

3. RESULTS AND DISCUSSION

In the result section, benchmark results are presented to evaluate the TCIT performance for cyclic molecules where the ring corrections are calculated by the physics-based RC_1 and TL-based RC_2 models. The TCIT results are compared with the experimental data, G4 calculations, and BGIT predictions, as implemented in CHETAH.[30] Mean signed errors (MSE) that represent systematic bias and mean absolute errors (MAE) that account for an average accuracy are reported for each comparison.

3.1. G4 Benchmarking. TCIT CAVs and ring corrections are parameterized to G4 results for linear and ring model compounds, respectively. Thus, the errors in TCIT prediction are strongly correlated with the G4 performance. In the previous work, we tested G4 performance on 572 acyclic compounds and observed an MSE of −0.06 kJ/mol and an MAE of 4.19 kJ/mol (~1 kcal/mol).[17] Here, we selected 365 reference cyclic compounds with no more than 12 heavy atoms to compare G4 calculations with the experimental data (Figure 4a). An MSE of −0.46 kJ/mol is observed for cyclic compounds, which is slightly larger than the MSE for acyclic compounds but can still be considered as having negligible systematic bias. The observed MAE is 5.97 kJ/mol (~1.5 kcal/mol), which is also larger than the MAE for acyclic compounds and the experimental uncertainty.[9,31] There are a few factors that potentially contribute to the increased errors for G4 predictions of cyclic structures, including the additional challenge of sampling conformers and potentially higher errors in experimental data for cyclic structures, which are being used to evaluate errors. Nevertheless, the G4 predictions are still very close to the 1 kcal/mol target for chemical accuracy and serve as an accurate basis for parameterizing ring corrections.

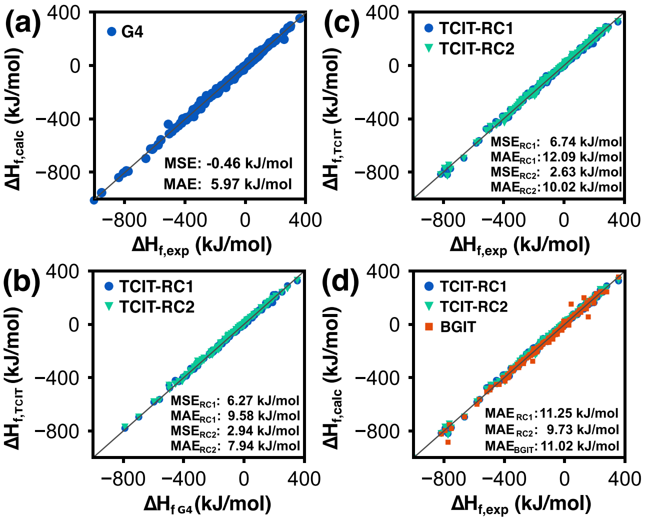


Figure 4. Correlation plots showing comparisons between the benchmark experimental data, G4, TCIT, and BGIT. (a) Comparison between experimental gas-phase ΔH_f values and G4 calculations for 365 compounds in the benchmark dataset. (b) Comparison between G4 and TCIT predictions for the 133 compounds in the benchmark dataset with no more than 12 heavy atoms and excluding RMCs. (c) Comparison between the experimental values and TCIT for the 191 non-RMCs in the benchmark dataset. (d) Comparison between the experimental values, TCIT-RC1 (blue, circles), TCIT-RC2 (green, triangles), and BGIT (red, squares) for the 173 non-RMCs in the benchmark dataset. Mean signed error (MSE) and mean absolute error (MAE) statistics are presented for each comparison.

3.2. G4 and TCIT Comparisons. To characterize the errors associated with applying TCIT to cyclic compounds outside of the training data, a comparison was performed between G4, TCIT-RC1, and TCIT-RC2 for 133 cyclic compounds (Figure 4b). These compounds are large enough not to be RMCs for TCIT-RC1 but contain no more than 12 heavy atoms to make G4 calculations feasible. Errors associated with the TCIT component decomposition for linear compounds were analyzed previously and found to exhibit MSE and MAE of −0.18 and 2.30 kJ/mol, respectively.[17] Here, the resulting MSE and MAE for cyclic compounds are 6.27 and 9.58 kJ/mol for TCIT-RC1 and 2.94 and 7.94 kJ/mol for TCIT-RC2, respectively. We ascribe the larger errors for cyclic compounds to the associated ring corrections. In particular, a systematic overestimation is apparent in the cyclic compounds that was absent in the linear compounds. Additionally, the consistently larger errors for TCIT-RC1 in comparison with TCIT-RC2 reflect the relatively limited transferability of the depth = 1 ring correction. Propagating the three sources of error (i.e., errors from the component decomposition, ring correction, and G4 calculation), we would anticipate that TCIT-RC1 and TCIT-RC2 predictions for cyclic compounds exhibit MAEs of approximately 12 and 10 kJ/mol, respectively, compared with the experimental data.

3.3. Experimental Validation. The most important questions for applications are how well TCIT reproduces experimental data for cyclic molecules and how it performs relative to the state-of-art methodology. To answer these questions, we performed TCIT-RC1 and TCIT-RC2 predictions on the 191 molecules in the PNK dataset that are not TCIT-RC1 RMCs (Figure 4c). Among these testing compounds, 173 of them can be predicted by BGIT, as implemented within CHETAH (Figure 4d).[30]

The MSEs of TCIT-RC1 and TCIT-RC2 with respect to the experimental data are nearly identical to the comparison with G4 (5.63 and 2.29 kJ/mol, respectively), which is consistent with the systematic bias being the result of the ring corrections. The observed MAEs are 12.09 and 10.02 kJ/mol, which are in perfect agreement with the estimate based on the propagation of independent errors. In contrast, BGIT systematically underestimates the experimental ΔH_f (MSE = −4.38 kJ/mol) with a smaller absolute bias than TCIT-RC1 but larger than TCIT-RC2. It is also visually apparent that several distinct outliers occur for BGIT that are not exhibited by either TCIT model. The overall lowest MAE is exhibited by TCIT-RC2.

3.4. Fine-Grained Analysis of TCIT Performance on Cyclic Molecules. The central assumption of increment theories is that certain molecular properties are decomposable into local contributions from distinct parts of the molecular graph. Since rings are topologically nonlocal, they pose a unique challenge for increment theories that must be addressed using nonlocal corrections. Similarly, conjugation

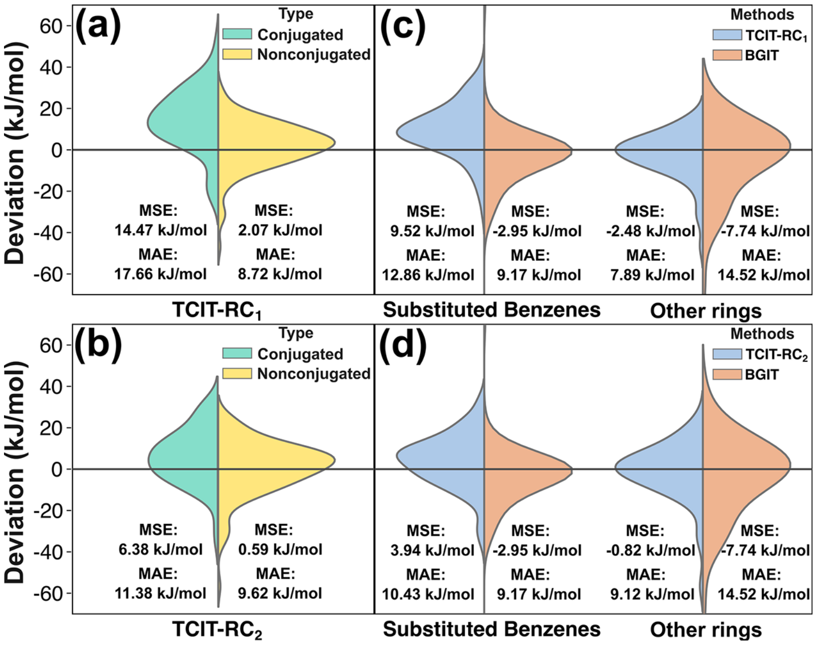


Figure 5. Violin plots showing comparisons between the benchmark experimental data, TCIT, and BGIT on distinct subsets of testing data. (a) Comparison between experimental gas-phase ΔH_f values and TCIT-RC1 calculations on 72 conjugated (left) and 119 nonconjugated compounds (right). (b) Comparison between the experimental values and TCIT-RC2 calculations on the conjugated/nonconjugated testing sets. (c) Comparison between the experimental values, TCIT-RC1 (blue), and BGIT (orange) for 117 substituted benzenes (left) and 56 other simple rings (right) in the benchmark dataset. (d) Comparison between the experimental values, TCIT-RC2 (blue), and BGIT (orange) for the substituted benzenes/other rings classes in the benchmark dataset. Mean signed error (MSE) and mean absolute error (MAE) statistics are presented for each comparison.

creates nonlocal interactions that potentially need to be corrected. Within the current benchmark, we have only attempted to correct ring contributions to ΔH_f, but many of the model compounds also exhibit conjugation. To distinguish the role of conjugation in the TCIT prediction errors, we have split the testing data into compounds for which the corresponding RMC of the RC$_1$ model breaks conjugation between the substitution and the ring (designated "conjugated") and those that do not (designated "nonconjugated"). Comparing the TCIT-RC1 error distributions for these two classes of compounds (Figure 5a), we observe much higher errors for conjugated structures than for nonconjugated structures. In particular, the high systematic bias observed for TCIT-RC1 is driven disproportionately by the conjugated subset of testing compounds (MSE = 14.47 kJ/mol). This reflects the important role played by conjugation and the limited transferability of the depth = 1 ring correction. In contrast, the TL-based TCIT-RC2 model partially captures some conjugation effects by expanding the depth of the graph out to depth = 2, which has the effect of substantially reducing prediction errors in the conjugated subset (MSE decreases to 6.38 kJ/mol, Figure 5b). It is possible that training distinct conjugation corrections and ring corrections is a pathway for further improving accuracy, but this is beyond the scope of the current work.

An additional split of the testing compounds between substituted benzenes and other rings was performed to highlight the role of training data in biasing the assessment of prediction errors (Figure 5c,d). In particular, due to the abundance of experimental data for substituted benzenes, BGIT uses a special group type (C_B) for these species and has much better performance in this subset than in the rest of the structures. In contrast, the TCIT-RC1 model exhibits relatively poor performance on the substituted benzene subset, primarily because these structures also exhibit conjugation (58 out of 117 belong to the conjugated class). Meanwhile, the TCIT-RC2 model exhibits relatively balanced performance across the two splits. This comparison demonstrates that BGIT exhibits much lower accuracy for general ring structures than would be judged from the MAE of the full test set of structures, which has a high representation of substituted benzenes (117 out of 174).

4. CONCLUSIONS

Over the past several decades, increment theories have been widely adopted as cost-effective approaches to predicting molecular thermodynamic properties. However, a long-standing challenge is how to expand these theories to new chemistries encountered in many applications. TCIT is the first component theory derived exclusively from quantum chemistry data, with a systematic protocol for generating model compounds and deriving parameters in an on-the-fly manner.[17] In the current work, we have extended the TCIT framework to include two approaches, TCIT-RC1 and TCIT-RC2, for deriving ring corrections to ΔH_f predictions that are specific to each ring and its bonded substitutions out to one and two bonds, respectively. These models have been implemented into our previously developed TCIT theory and benchmarked against the most commonly used group increment theory, BGIT, on the PNK dataset. We observe that both models have comparable performance to BGIT with respect to overall MAE and MSE but outperform BGIT on rings with limited experimental data. Moreover, upon investigating the performance on distinct subsets of cyclic structures, the TL-based TCIT-RC2 model shows the lowest overall errors and the most consistent performance across distinct rings. This result has encouraged us to implement the TCIT-RC2 model, which has higher transferability and lower computational cost than TCIT-RC1, into an open-source version of TCIT that is available at https://github.com/zhaoqy1996/TCIT-Hf.

Although current results only show the performance of TCIT on molecules with experimental data available, the real potential of TCIT lies in predicting the thermodynamic properties of exploratory species without experimental data. Meanwhile, since the first TCIT work was published, the quantum chemistry training data and CAV database for ΔH_f associated with TCIT have grown substantially. The current database contains more than 20k C-, H-, O-, N-, S-, and halogen-containing components that cover a large swathe of organic chemical space. In particular, the current benchmark includes many structures that BGIT could not predict due to unavailable groups, which TCIT now supports. Extensions to more complex fused rings and to radical and ionic species are currently underway.

Considering opportunities for the future, it is foreseeable that some combination of ML and physics-based approaches like TCIT will yield a valuable balance between interpretability, accuracy, and computational cost for predicting molecular thermodynamic properties. In this regard, the TCIT-RC2 model is a useful case study where transfer learning was used to improve the predictions of a physics-based model. Many other opportunities can be envisioned in this vein, including further improving the accuracy of TCIT predictions or possibly enabling training data to be generated from lower-cost quantum chemistry sources like the density functional theory. In addition, the automatic data generation associated with TCIT as new chemistries are encountered is also potentially valuable for training ML-only approaches to thermodynamic property predictions while avoiding extrapolation errors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00367.

> Implementation details for the neural fingerprint model, performance of purely machine learning approaches, lists of excluded compounds, experimental values and G4 calculation results of model compounds (PDF); full list of the data referenced in the "Results and Discussion" section, including Experimental values, G4, TCIT-RC1, TCIT-RC2, and BGIT calculations (CSV) (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

Authors

Qiyuan Zhao − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States;* orcid.org/0000-0003-3228-8160

Nicolae C. Iovanac − *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.1c00367

Notes

The authors declare no competing financial interest.

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. The version of TCIT used in this study is available at https://github.com/zhaoqy1996/TCIT-Hf under the MIT license.

■ ACKNOWLEDGMENTS

The work performed by Q.Z. and B.M.S was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant number: N00014-21-1-2476, Program Manager: Dr. Chad Stoltz). The work performed by N.C.I. was supported by the National Science Foundation under Grant No. DGE-1333468. B.M.S. also acknowledges partial support from the the Purdue Process Safety and Assurance Center and the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. 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.

■ ABBREVIATIONS USED

TAFFI, Topology Automated Force-Field Interactions; PNK, Pedley−Naylor−Kline; G4, Gaussian-4; GAV, group additivity values; CAV, component additivity values; BGIT, Benson group increment theory; TCIT, TAFFI Component Increment Theory; RC, ring correction; ML, machine learning; TL, transfer learning; NN, neural network; RMC, ring model compound; RC1/2, TCIT ring corrections specific to rings and all of their substituents out to one/two bond; MSE, mean signed error; MAE, mean absolute error

■ REFERENCES

(1) Green, W.; Barton, P.; Bhattacharjee, B.; Matheu, D.; Schwer, D.; Song, J.; Sumathi, R.; Carstensen, H.-H.; Dean, A.; Grenda, J. Computer Construction of Detailed Chemical Kinetic Models for Gas-Phase Reactors. *Ind. Eng. Chem. Res.* **2001**, *40*, 5362−5370.

(2) 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.

(3) Zimmerman, P. M. Navigating Molecular Space for Reaction Mechanisms: an Efficient, Automated Procedure. *Mol. Simul.* **2015**, *41*, 43−54.

(4) Qiu, Y.; Collin, F.; Hurt, R. H.; Külaots, I. Thermochemistry and Kinetics of Graphite Oxide Exothermic Decomposition for Safety in Large-Scale Storage and Processing. *Carbon* **2016**, *96*, 20−28.

(5) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1354.

(6) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23*, 1203−1218.

(7) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic properties. *J. Chem. Phys.* **1958**, *29*, 546−572.

(8) Benson, S.; Cruickshank, F.; Golden, D.; Haugen, G.; O'Neal, H.; Rodgers, A.; Shaw, R.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69*, 279−324.

(9) Cohen, N.; Benson, S. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419−2438.

(10) Cohen, N. Revised Group Additivity Values for Enthalpies of Formation (at 298 K) of Carbon-Hydrogen and Carbon-Hydrogen-Oxygen Compounds. *J. Phys. Chem. Ref. Data* **1996**, *25*, 1411−1481.

(11) 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.

(12) 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.

(13) 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.

(14) Wang, H.; Castillo, A.; 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.

(15) 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.

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

(17) Zhao, Q.; Savoie, B. M. Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation. *J. Chem. Inf. Model.* **2020**, *60*, 2199−2207.

(18) Lay, T. H.; Yamada, T.; Tsai, P.-L.; Bozzelli, J. W. Thermodynamic Parameters and Group Additivity Ring Corrections for Three-to Six-Membered Oxygen Heterocyclic Hydrocarbons. *J. Phys. Chem. A* **1997**, *101*, 2471−2477.

(19) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation. *Int. J. Chem. Kinet.* **2018**, *50*, 294−303.

(20) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Networks on Graphs for Learning Molecular Fingerprints. *CoRR* 2015, abs/1509.09292.

(21) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126*, No. 084108.

(22) 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.

(23) 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*, No. P32.

(24) 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*, No. 33.

(25) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(26) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169−7192.

(27) 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.

(28) 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.

(29) Linstrom, P.; Mallard, W.. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, 2018.

(30) Seaton, W. H. *CHETAH-The ASTM Chemical Thermodynamic and Energy Release Evaluation Program*; American Society for Testing and Materials, 1974.

(31) 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.