A Subgraph Isomorphic Decision Tree to Predict Radical Thermochemistry with Bounded Uncertainty Estimation

Hao-Wei Pang, Xiaorui Dong, Matthew S. Johnson,* and William H. Green*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

E-mail: mjohnson541@gmail.com; whgreen@mit.edu

Abstract

Detailed chemical kinetic models offer valuable mechanistic insights into industrial applications. Automatic generation of a reliable kinetic model requires fast and accurate radical thermochemistry estimation. Kineticists often prefer hydrogen bond increment (HBI) corrections from a closed shell molecule to the corresponding radical for their interpretability, physical meaning, and facilitation of error cancellation as a relative quantity. Tree estimators, used due to limited data, rely on expert knowledge and manual construction currently, posing challenges in maintenance and improvement. In this work, we extend the subgraph isomorphic decision tree (SIDT) algorithm originally developed for rate estimation, to estimate HBI corrections. We introduce a physics-aware splitting criterion, explore a bounded weighted uncertainty estimation method, and evaluate aleatoric uncertainty-based and model variance reduction-based pre-pruning methods. Moreover, we compile a dataset of thermochemical parameters for 2,210 radicals involving C, O, and H based on quantum chemical calculations from recently published works. We leverage the collected dataset to train the SIDT model.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Compared to existing empirical tree estimators, the SIDT model (1) offers an automatic approach to generating and extending tree estimator for thermochemistry, (2) has better accuracy and R^2, (3) provides significantly more realistic uncertainty estimates, and (4) has a tree structure much more advantageous in descent speed. Overall, the SIDT estimator marks a great leap in kinetic modeling, offering more precise, reliable, and scalable predictions for radical thermochemistry.

1 Introduction

Detailed chemical kinetic models are valuable tools for many industrial applications, including investigating the fundamentals of polymer fouling,[1,2] fuel pyrolysis and oxidation,[3–7] and active pharmaceutical ingredient oxidative degradation.[8,9] Radicals are vital to the chemical kinetic models in all the abovementioned and many other applications. During the process of automatic mechanism generation, hundreds of thousands of potential molecules (including radicals) are generated and assessed for their importance under given reacting conditions.[10] Some assessment algorithms utilize the thermochemistry of all involved molecules[11] and thus necessitate the fast estimation of thermochemistry.

The thermochemical properties often include standard enthalpy of formation ($\Delta H^\circ_{\mathrm{f},298}$), standard entropy of formation (S°_{298}), and heat capacity ($C_\mathrm{p}(T)$) at various temperatures in order to account for the temperature dependency. There exist experimental measurements for some radicals,[12] but more often than not we need thermochemical properties of radicals that have not been measured experimentally. One can perform quantum chemical calculations to obtain thermochemistry for radicals. However, it is computationally expensive to perform calculations for all involved radicals, as it requires a thorough molecular conformation search, geometry optimization, frequency calculation, and single-point calculation at a trustworthy level of theory. Rotor scans are particularly important to obtain accurate entropies and heat capacities.

Estimators are often built to provide fast estimation for radical thermochemistry, while

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

the accuracy of estimates and uncertainties often vary. Deep Neural Network (DNN)-based estimators for thermochemistry have gained much interest in recent years.[13–15] While transfer learning has shown promise as a mitigation strategy,[16] obtaining a reliable DNN-based estimator for radical thermochemistry can still be challenging due to the scarcity of high-quality thermochemical data for diverse sets of radical species. Moreover, DNN-based estimators often need extra efforts[17–19] to improve the model's interpretability. Even with augmented explainability, it is difficult for a kineticist to incorporate qualitative chemical knowledge to improve erroneous predictions for radicals of interest. DNN-based estimators also often require a non-negligible amount of memory usage even during inference due to millions of parameters and the use of ensemble models for uncertainty estimation, competing for resources with the already memory-intensive process of generating detailed kinetic models.[20,21]

A widely-used approach to estimate thermochemistry for radicals is to combine the Benson-type group additivity (GA) method[22] with hydrogen bond increments (HBIs).[23] The Benson-type group additivity method assumes that the thermochemical properties of a closed-shell molecule can be estimated as a linear combination of contributions from its atom-centered groups. The HBI method assumes that the thermochemical property of a radical can be estimated by applying an HBI correction to the corresponding (hydrogen) saturated closed shell molecule.

Reaction Mechanism Generator (RMG), an open-source software package for the automatic generation of detailed chemical kinetic mechanisms, uses tree structures to estimate hydrogen bond increments. The search descends from the top of the tree estimator until it finds the most specific node matching the specific radical. The tree estimator is very lightweight due to the small number of parameters (hundreds to thousands, not millions). Historically, the structure of the tree and HBI groups have been defined manually based on heuristics. This heavily relies on expert knowledge along with manual optimization and does not guarantee truly optimized performance. Moreover, the heuristic tree provides limited uncertainty estimates[24] so a representative uncertainty is often assigned.[25]

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In this work, we develop a subgraph isomorphic decision tree (SIDT) estimator for HBI corrections by extending the preceding work developed for rate estimation by Johnson et al.[26] The SIDT method offers various advantages, making it valuable for automated mechanism generation. It is straightforward to extend and (re)train, doesn't necessitate GPU access, features a hierarchical structure that facilitates interpretation, enables uncertainty estimation, allows for the easy integration of expert knowledge, and gives reasonable results despite smaller dataset sizes. We collect thermochemical parameters derived from quantum chemistry calculations from past publications of detailed kinetic models.[1–3,5–7,27] We consider equivalent resonance structures of each radical from the collected data and derive the HBI corrections. We develop a SIDT model to predict the HBI corrections, implement an uncertainty estimation from a bounded weighted standard deviation, and implement and evaluate pre-pruning methods. We compare the prediction performance with RMG's empirical tree and analyze the quality of the uncertainty estimates.

2 Methods

2.1 Dataset

Thermochemical data for 2,210 radical molecules excluding duplicates are collected from past publications of detailed kinetic models.[3,5–7,27] For duplicate calculations, the one with lower energy is kept. Among them, 1,978 radicals are collected from Dong et al.,[3,7] Pio et al.,[6] and Liang et al.,[27] 187 are collected from Pang et al.,[1,2] and 45 are collected from Johnson et al.[5] The distribution of collected radical thermochemical parameters can be found in Fig. S1 in the Supporting Information (SI).

Most of the collected radical thermochemical parameters were obtained using ARC[28] using the following procedure. Conformers were embedded stochastically using the ETKDG algorithm[29] and optimized using MMFF94s force field[30,31] implemented in RDKit[32] in an attempt to obtain the lowest energy conformer. Geometry optimization and harmonic vibra-

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

tional frequencies and torsional scans were performed using density functional theory (DFT) at B3LYP/CBSB7, and the single point energy was calculated using the CBS-QB3 composite method[33] implemented in Gaussian.[34,35] If a lower energy conformer was found during the torsional scan, the above steps were repeated with the newly found lowest-energy conformer. The frequencies are corrected using a factor of 0.99.[33] The rigid rotor harmonic oscillator approximation with 1D hindered rotor corrections was used to compute the $\Delta H^{\circ}_{\mathrm{f},298}$, S°_{298}, and $C_{\mathrm{p}}(T)$'s from the quantum chemical calculation results using Arkane.[36] Atom energy corrections (AECs) and bond additivity corrections (BACs) are applied to improve the energies obtained from calculations.[37]

36 out of 45 from Johnson et al. were obtained using the following procedure. The conformers were selected manually. The geometries and vibrational frequencies were calculated using DFT with B3LYP/6-31G(2df,p), and single point energies were calculated with the G4 composite method[38] using Gaussian.[34] Torsional scans were performed in an attempt to find the lowest energy conformer. If found, the above steps were repeated with the newly found lowest-energy conformer. The rigid rotor harmonic oscillator approximation with 1D hindered rotor corrections was used to compute the thermochemical parameters using Arkane[36] with AECs and BACs applied.[37]

2.1.1 HBI corrections and data augmentation

Benson-type group additivity values are often fitted with multivariate least squares regression because there are multiple groups, and therefore, multiple molecules need to be fit simultaneously to determine the group values. In contrast, the HBI of a radical group can be derived from a single radical ($\mathrm{R}\cdot$), by taking the difference with the closed shell molecule (RH) and hydrogen atom contribution using a general relation of

$$\Delta B(\mathrm{HBI}) = B(\mathrm{R}\cdot) - B(\mathrm{RH}) + \mathrm{const.} \tag{1}$$

where B refers to a thermochemical property. The specific relations are[23]

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$\Delta\Delta H^\circ_{\mathrm{f},298}(\mathrm{R\cdots H}) = \Delta H^\circ_{\mathrm{f},298}(\mathrm{R}\cdot) - \Delta H^\circ_{\mathrm{f},298}(\mathrm{RH}) + 52.1\mathrm{kcal/mol} \tag{2}$$

$$\Delta S^\circ_{298}(\mathrm{HBI}) = S^\circ_{\mathrm{int},298}(\mathrm{R}\cdot) - S^\circ_{\mathrm{int},298}(\mathrm{RH}) \tag{3}$$

$$\Delta C_\mathrm{p}(T, \mathrm{HBI}) = C_\mathrm{p}(T, \mathrm{R}\cdot) - C_\mathrm{p}(T, \mathrm{RH}) \quad \forall T \in \underline{T} \tag{4}$$

The derived HBI corrections contain 9 terms: $\Delta\Delta H^\circ_{\mathrm{f},298}(\mathrm{R\cdots H})$, $\Delta S^\circ_{298}(\mathrm{HBI})$, and $\Delta C_\mathrm{p}(T, \mathrm{HBI})$ at 7 different temperatures ($\underline{T}$): 300, 400, 500, 600, 800, 1000, 1500 K. Eqs. (2) to (4) adhere to the standard symbol nomenclature prevalent in chemical literature, while S°_int refers to intrinsic entropy excluding symmetry and 52.1kcal/mol[23] is $\Delta H^\circ_{\mathrm{f},298}(\mathrm{H}\cdot)$. Note that $\Delta\Delta H^\circ_{\mathrm{f},298}(\mathrm{R\cdots H})$ is the bond dissociation energy, while $\Delta S^\circ_{298}(\mathrm{HBI})$ and $\Delta C_\mathrm{p}(T, \mathrm{HBI})$ implicitly contain the contribution from the hydrogen atom radical ($\mathrm{H}\cdot$). While $\Delta H^\circ_{\mathrm{f},298}(\mathrm{R}\cdot)$, $S^\circ_{\mathrm{int},298}(\mathrm{R}\cdot)$, and $C_\mathrm{p}(T, \mathrm{R}\cdot)$ are extracted from the collected data (excluding symmetry for entropy), since not all radicals come in a pair with its closed shell molecule in the collected data, $\Delta H^\circ_{\mathrm{f},298}(\mathrm{RH})$, $S^\circ_{\mathrm{int},298}(\mathrm{RH})$, and $C_\mathrm{p}(T, \mathrm{RH})$ have a mixture of sources. If the closed-shell molecules are in the collected data, we use the calculated parameters; otherwise, we use Benson-type group additivity estimates. This introduces some uncertainties, as discussed in Sec. 2.1.2.

As discussed below, the same HBI might be used to estimate the thermochemistry of many radicals, depending on how specifically one defines the HBI group template. In that situation, more than one radical could be used to estimate the same HBI, e.g., by averaging the values derived using Eqs. (2) to (4).

Additionally, multiple HBI corrections can be derived from a single radical if it has resonance structures. E.g., the radical in Fig. 1 can be used to derive the HBI needed to estimate its thermo from the closed-shell molecule, 2-butene, or a different HBI needed to estimate its thermo from 1-butene. These HBIs would be useful for estimating the thermo

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

of radicals derived from larger alkene, e.g., if one had thermo data on 1-decene or 2-decene, but no data on the dec-1-ene-3-yl radical.

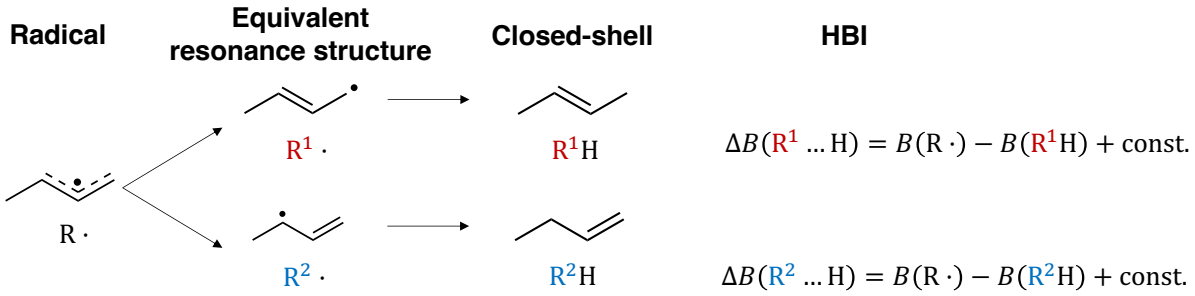


Figure 1: Example of deriving multiple hydrogen bond increment corrections from equivalent resonance structures of a radical. B refers to $\Delta H^{\circ}_{\mathrm{f},298}$, ΔS°_{298}, or $C_{\mathrm{p}}(T)$.

We augment the data by deriving multiple HBI corrections to help the final predicted radical thermochemistry to be invariant for all resonance structures. We generate radical resonance structures using RMG's resonance algorithm.[39] The distribution of the number of equivalent resonance structures found for each radical is shown in Fig. S2. After resonance structure augmentation, the number of radical data points increases to 2,805. For thermochemical parameters of closed-shell molecules, 1,824 of them are Benson-type group additivity estimates, 959 of them were computed at CBS-QB3 level of theory, and 22 of them were computed at G4 level of theory.

The key molecular features of collected radicals after augmentation can be found in Fig. 2. Fig. 2(a) indicates that there are radicals containing H, C, N, and O atoms, while there are only either carbon-centered or oxygen-centered radicals. The collected data cover a range of number of rings, atoms, and rotatable bonds. The distribution of HBI corrections can be found in Fig. 3. As a sanity check, we plot the node values of the empirical tree estimator in Fig. S5, which have a similar range seen in Fig. 3. Fig. 3 aligns with the HBI corrections reported in published works.[40,41]

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

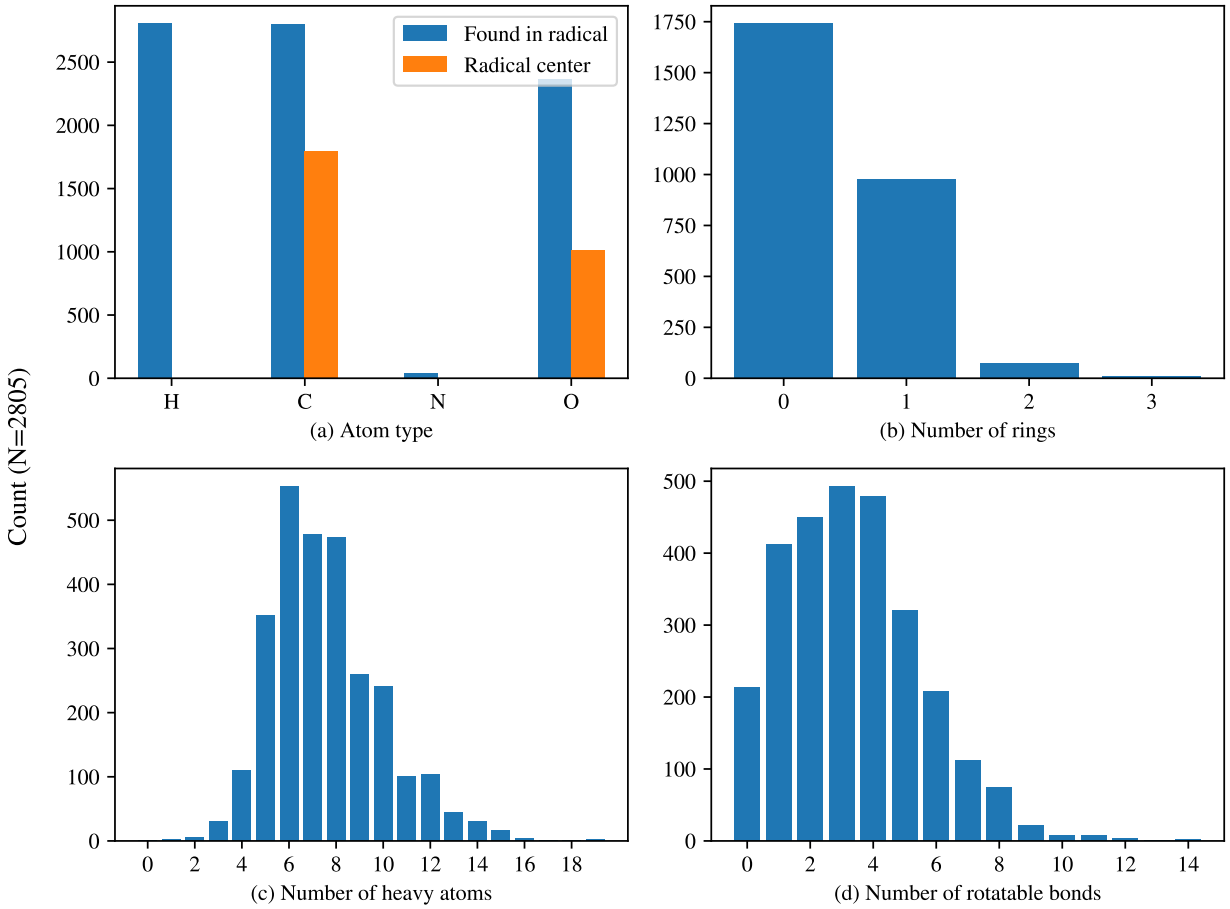


Figure 2: Key molecular features of radicals collected from past publications[1–3,5–7,27] after data augmentation with equivalent resonance radicals: (a) atom types in radicals and of radical centers, (b) number of rings, (c) number of heavy atoms, and (d) number of rotatable bonds.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

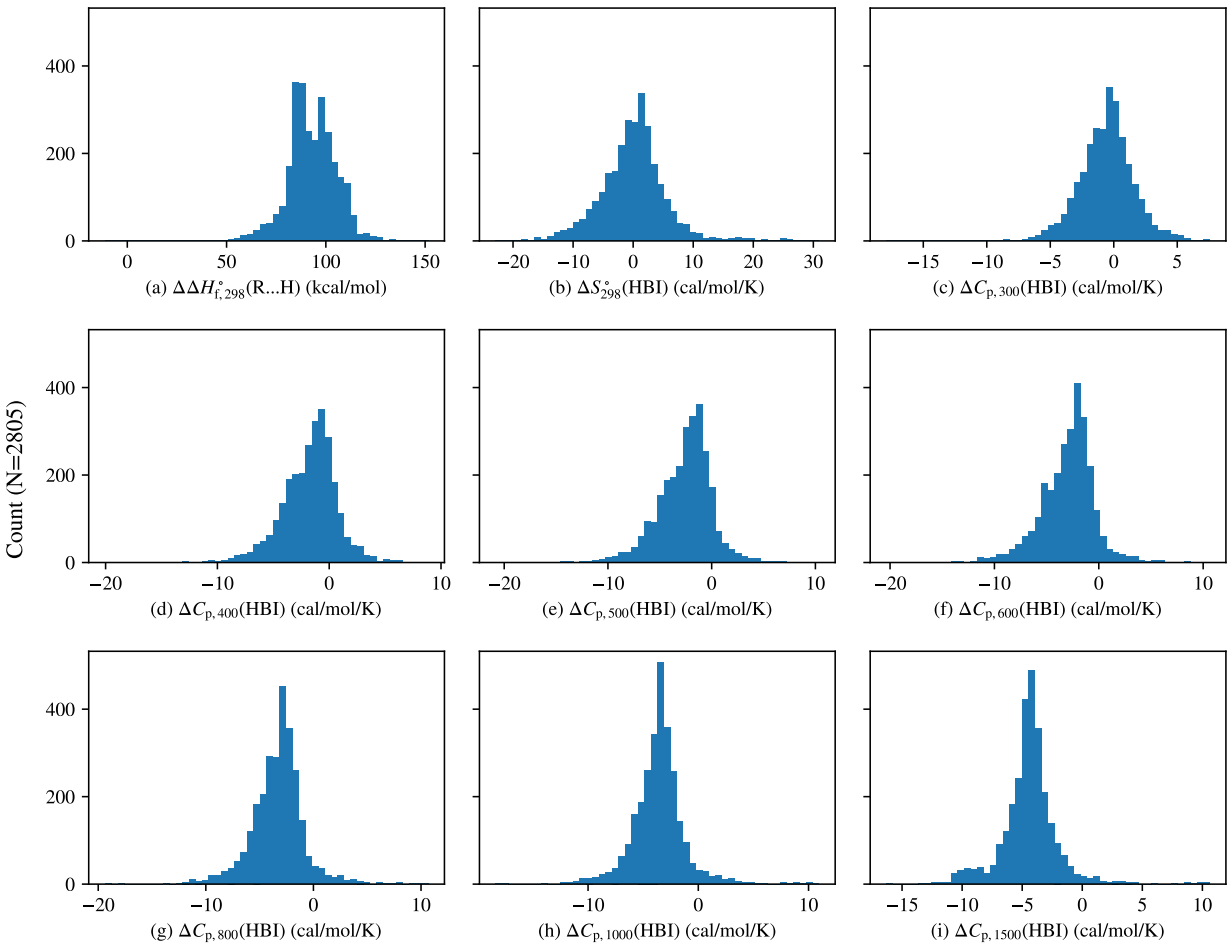


Figure 3: Distribution of HBI corrections for (a) standard enthalpy of formation, (b) intrinsic entropy, (c-i) heat capacities at various temperatures.

2.1.2 Uncertainties in HBI corrections

As delineated by Eqs. (2) to (4), the error in the HBI corrections can come from both the radical and the closed-shell components.

There are several possible sources of error in the collected radical data. First, there are errors associated with the level of theory. For CBS-QB3 level of theory, a root-mean-square deviation (RMSD) of 1.2 kcal/mol in $\Delta H^\circ_{f,298}$, 1.0 cal/mol/K in S°_{298}, and 0.4 cal/mol/K in $C_{p,298}$ are found for radicals by benchmarking with experimental measurements.[42] Somers et al. also reported an RMSD of 1.2 kcal/mol in $\Delta H^\circ_{f,298}$ of radicals.[43]

For G4 level of theory, a mean absolute deviation of 0.83 kcal/mol in $\Delta H^\circ_{f,298}$ was re-

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

ported originally[38], while an RMSD of 1.1 kcal/mol in $\Delta H^{\circ}_{\mathrm{f},298}$, 1.0 cal/mol/K in S°_{298}, and 0.4 cal/mol/K in $C_{\mathrm{p},298}$ were reported for radicals by benchmarking against experimental measurements.[42] Somers et al. reported a smaller RMSD of 0.5 kcal/mol in $\Delta H^{\circ}_{\mathrm{f},298}$ for radicals.[43]

Aside from errors associated with the level of theory mentioned above, the separable rotor approximation (1D hindered rotor) can also introduce errors to the computed thermochemical parameters, particularly to the entropies and the heat capacities. Sharma et al. identified a worst-case scenario, where using the separable rotor approximation introduces errors of 1.1 kcal/mol for $\Delta H^{\circ}_{\mathrm{f},298}$, 5.66 cal/mol/K for S°_{298}, and 15.62 cal/mol/K for $C_{\mathrm{p},298}$ if starting from the ring-shape conformer for hydroperoxyalkylperoxy radicals using CBS-QB3.[44] Large errors like this are most likely if there is intramolecular hydrogen bonding. Additionally, there could be a significant error in $\Delta H^{\circ}_{\mathrm{f},298}$ if a calculation did not converge to the lowest energy conformer.

We use 1.2 kcal/mol as a representative error for radicals calculated at CBS-QB3 level of theory and 1 kcal/mol for those at G4 level of theory. This is a reasonable assumption for $\Delta H^{\circ}_{\mathrm{f},298}$ as long as we find the lowest energy conformer. Radicals with fewer rotatable bonds are more likely to find the lowest energy conformers. We use a representative error of 1 cal/mol/K for S°_{298} and $C_{\mathrm{p}}(T)$'s. This is a reasonable number for the level of theories used in the dataset, similar to what has been assumed in past kinetic model works.[45] We are aware of worst-case scenarios as shown in Sharma et al.,[44] so this may be optimistic, but we have limited ability to verify all collected radical data. Also, we expect partial error cancellation due to the similarity between the radical and the closed-shell molecule, so the HBIs may be more accurate than the total $\Delta H^{\circ}_{\mathrm{f},298}$, S°_{298}, $C_{\mathrm{p}}(T)$'s.

We use the representative errors mentioned above for the closed shells computed using quantum chemistry. We benchmark the group additivity estimated thermochemistry for closed shells against those collected from past publications of detailed kinetic models,[3,5–7,27] as shown in Fig. S4. From that benchmark, we derive a representative error of 4 kcal/mol for

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$\Delta H^{\circ}_{\mathrm{f},298}$, 4 cal/mol/K for S°_{298}, and 2 cal/mol/K for $C_{\mathrm{p}}(T)$'s coming from group additivity.

The uncertainty in the derived HBI corrections is taken to be the root sum of the uncertainty squared in the radical component and the closed-shell component.

2.2 Machine learning methodology

This work extends the subgraph isomorphic decision tree (SIDT) algorithm developed for predicting rates by Johnson et al.[26] to predict HBI corrections. The input format can be a SMILES, InChI, or adjacency list, which is first converted into an RMG molecule.[46,47] An RMG molecule is a 2D graph representing a molecule, storing molecular information such as atom type, bond type, and connectivity. Unlike the conventional regression decision trees that use rules to split the data points, SIDT uses group structures that are subgraph isomorphic to the molecule 2D graphs for molecular splitting. Additionally, SIDT belongs to the Non-Mandatory Leaf-Node Prediction (NMLNP) category,[48] where prediction can be made from both the internal nodes and the leaf nodes.

A flow chart of generating a SIDT model for HBI corrections can be found in Fig. 4. There are two stages in this process: the tree extension stage (left-hand side loop in Fig. 4) and the internal and leaf node fitting stage (last step in Fig. 4). During the tree extension stage, potential local extensions are proposed to split radicals based on their substructure, and a splitting criterion is evaluated to determine the local optimal split. Sec. 2.2.1 discusses our choices of the splitting criterion to maintain physical meaning during optimization. Additionally, to prevent an unnecessarily complex tree, we explore two pre-pruning methods based on uncertainty and model variance reduction (Sec. 2.2.2). Once the tree extension completes, methods described in Sec. 2.2.3 are used to fit HBI corrections and uncertainties for each node.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

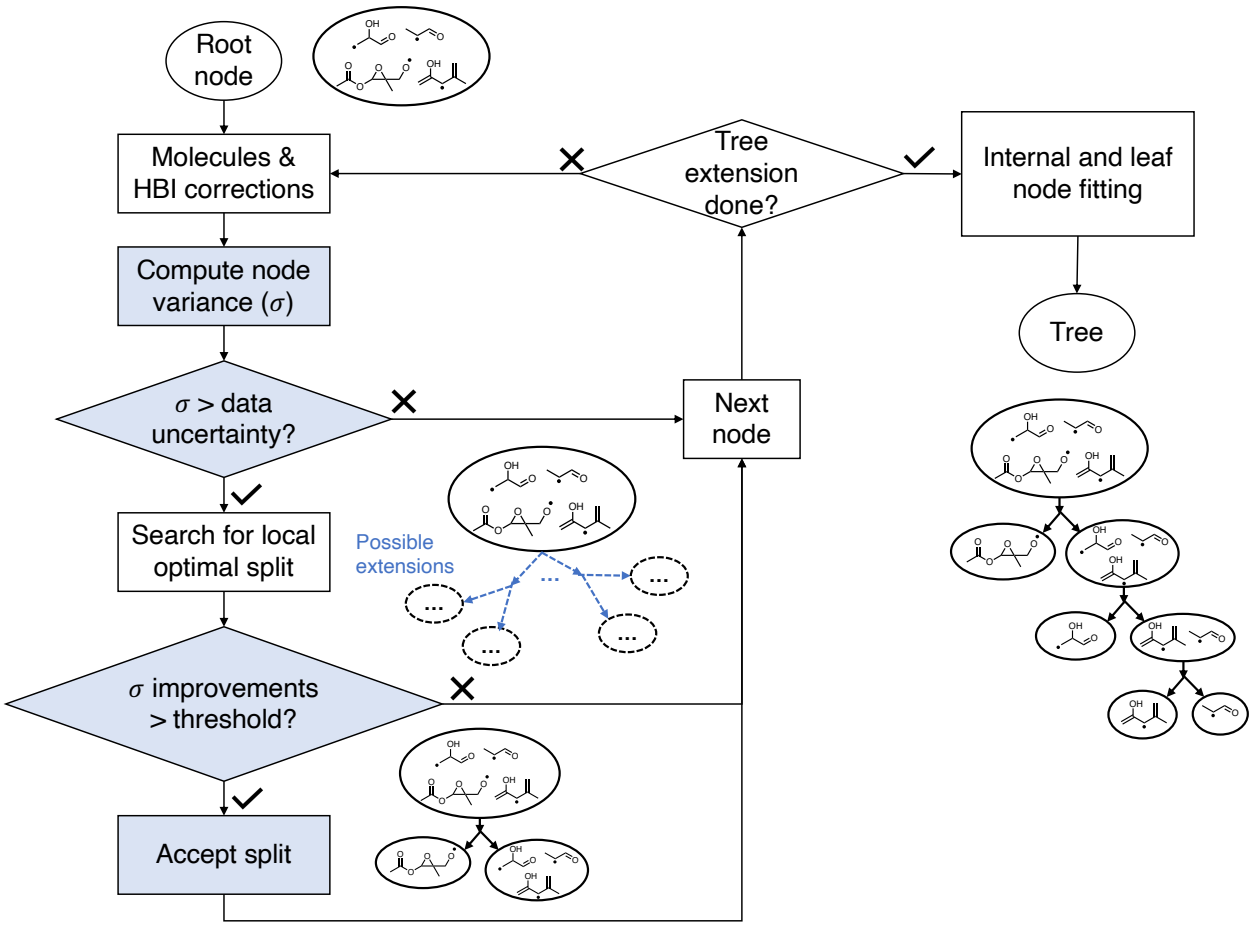


Figure 4: Flow chart of generating a subgraph isomorphic decision tree for HBI corrections. Blocks with light blue shades are associated with pre-pruning.

2.2.1 Splitting criterion

During the tree extension stage, the model takes in radicals as RMG molecules and generates a set of possible extensions ($\mathcal{P} \in \mathbb{P}$) that split the radicals into two partitions based on their substructures. We want to select the extension that puts radicals with the most similar HBI corrections into the same partition. We define the following splitting criterion (Π)

$$\Pi = N(p_1|\mathcal{P})\sigma(p_1|\mathcal{P}) + N(p_2|\mathcal{P})\sigma(p_2|\mathcal{P}) \tag{5}$$

where $N(p_i|\mathcal{P})$ is the number of radicals in partition i given the extension $\mathcal{P}$, and $\sigma(p_i|\mathcal{P})$ is the standard deviation of targets within partition i given the extension $\mathcal{P}$. We then

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

determine the (local) optimal extension ($\mathcal{P}^*$) by

$$\mathcal{P}^* = \arg\min_{\mathcal{P}\in\mathbb{P}} \Pi(\mathcal{P}) \tag{6}$$

Since the model is a multi-target (9 HBI correction targets) regression tree, there are many possible ways to define σ. We can define it directly related to the prediction targets, as shown in Eq. (7). Hyperparameters (α, β, γ) would be needed as $\Delta\Delta H^\circ_{\mathrm{f},298}$, ΔS°_{298}, and $\Delta C_\mathrm{p}(T)$'s are often not on the same scale. We could also scale the raw data using Z-score, but by doing so we lose the interpretability of the physical meaning.

$$\sigma(p_i|\mathcal{P}) = \alpha\,\mathrm{std}(\Delta\Delta\underline{H}^\circ_{\mathrm{f},298}(p_i|\mathcal{P})) + \beta\,\mathrm{std}(\Delta\underline{S}^\circ_{298}(p_i|\mathcal{P})) + \gamma\sum_{T\in\underline{T}}\mathrm{std}(\Delta\underline{C}_\mathrm{p}(T, p_i|\mathcal{P})) \tag{7}$$

Alternatively, we can define it using the HBI correction for Gibbs free energy of formation at temperatures of interest ($\Delta\Delta G^\circ_\mathrm{f}(T)$). There are some advantages to this definition. First, Gibbs free energy of formation ($\Delta G^\circ_\mathrm{f}$) incorporates the information from the enthalpy, entropy, and heat capacity in a way that retains their physical meaning. Second, $\Delta G^\circ_\mathrm{f}$ is what's actually used during kinetic modeling.

We obtain $\Delta\Delta G^\circ_\mathrm{f}(T)$ using the equations below. We bring $\Delta\Delta H^\circ_\mathrm{f}$ and ΔS° to the desired temperature using Eqs. (8) and (9), and then compute $\Delta\Delta G^\circ_\mathrm{f}(T)$ using Eq. (10). For simplicity during node splitting, we use a simple linear interpretation to evaluate the integrals in Eqs. (8) and (9).

$$\Delta\Delta H^\circ_\mathrm{f}(T) = \int_{298}^{T} \Delta C_\mathrm{p}(T)\mathrm{d}T + \Delta\Delta H^\circ_{\mathrm{f},298} \tag{8}$$

$$\Delta S^\circ(T) = \int_{298}^{T} \frac{\Delta C_\mathrm{p}(T)}{T}\mathrm{d}T + \Delta S^\circ_{298} \tag{9}$$

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$\Delta\Delta G_{\mathrm{f}}^{\circ}(T) = \Delta\Delta H_{\mathrm{f}}^{\circ}(T) - T\Delta S^{\circ}(T) \tag{10}$$

We then define the σ using the standard deviation of $\Delta\Delta G_{\mathrm{f}}^{\circ}(T)$ in partition i at selected temperatures ($\underline{T}$)

$$\sigma(p_i|\mathcal{P}) = \sum_{T \in \underline{T}} \mathrm{std}(\Delta\Delta \underline{G}_{\mathrm{f}}^{\circ}(T, p_i|\mathcal{P})) \tag{11}$$

Eqs. (5), (6) and (11) are used for the splitting criterion to extend the tree presented in this work.

2.2.2 Pre-pruning methods

In decision tree learning, a compact tree can often offer more interpretable insights and avoid over-fitting. Pre-pruning and post-pruning methods are developed to learn a small decision tree. In this work, we introduce two pre-pruning methods for the SIDT algorithm.

There are two sources of uncertainty in machine learning for chemistry: the aleatoric uncertainty and the epistemic uncertainty.[49] The aleatoric uncertainty is associated with noises in the training data and is not reducible during training, while the epistemic uncertainty is associated with the model bias and variances and is reducible. The original SIDT algorithm[26] extends the tree until exhaustion. This work implements and evaluates two new pre-pruning methods for the SIDT algorithm to consider the uncertainty during tree extension.

The first pre-pruning criterion is based on the aleatoric uncertainty limit. Let the HBI corrections of molecules matched to a node be $\mathbf{Y} \in \mathbb{R}^{N \times M}$, where N is the number of molecules in the training set and M is the number of HBI corrections for each molecule ($M = 9$). Let the uncertainty in these data points be $\mathbf{Y}_\epsilon \in \mathbb{R}_0^{+N \times M}$, which contains the errors in HBI corrections for each molecule due to reasons described in Sec. 2.1. The first column of $\mathbf{Y}_\epsilon$ contains the uncertainty in $\Delta\Delta H_{\mathrm{f},298}^{\circ}$. The second column in $\mathbf{Y}_\epsilon$ contains the

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

expected error for ΔS°_{298}. The third to ninth columns in $\mathbf{Y}_{\epsilon}$ contain the data uncertainty for $\Delta C_{\mathrm{p}}(T)$ at defined temperatures.

The aleatoric uncertainty-based pre-pruning is shown in Alg. 1. We first compute the standard deviation of HBI corrections for molecules matched to a node ($\underline{\sigma}^T$) weighted by the inverse of their uncertainties squared. We use the minimum expected uncertainty in training data ($\underline{\epsilon}^T$) as a proxy of the aleatoric uncertainty limit. Note that other functions can be used instead of min for $\mathbf{Y}_{\epsilon}$. Using min is more forgiving, while using functions like mean, median, or max can result in more conservative splitting. We only attempt splitting if there is still room for improving model variance ($\mathrm{any}(\underline{\sigma}^T > \underline{\epsilon}^T)$).

Algorithm 1 Aleatoric uncertainty-based pre-pruning

$\underline{\sigma}^T \leftarrow \sqrt{\mathrm{cov}(\mathbf{Y}, \mathrm{axis}=0, \mathrm{weighted{=}True})}$
$\underline{\epsilon}^T \leftarrow \min(\mathbf{Y}_{\epsilon}, \mathrm{axis}=0)$
if $\mathrm{any}(\underline{\sigma}^T > \underline{\epsilon}^T)$ **then**
    Attempt to split
end if

The second pre-pruning criterion is based on the magnitude of model variance reduction, which we use as a proxy of the significance of the splitting. The algorithm is shown in Alg. 2. We first compute the pre-splitting residual (Π_0). The pre-splitting residual is computed using all the molecules matched to the current node, where Π_0 is defined as $N\sigma$, where $N = N(p_1 + p_2|\mathcal{P}^*) = N(p_1|\mathcal{P}^*) + N(p_2|\mathcal{P}^*)$, and $\sigma = \sigma(p_1 + p_2|\mathcal{P}^*)$.

We then determine the local optimal split by evaluating the splitting criterion defined in Eq. (6) and compute the residual of the optimal split (Π^*). Note that the residual after splitting should always be smaller or at least equal to the one prior to splitting. We compare the pre-splitting and post-splitting residuals and only accept the split if the residual improves by a user-defined threshold (λ). In this work, we evaluate the performance of this pre-pruning algorithm with various λ's.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Algorithm 2 Model variance reduction-based pre-pruning

$\Pi_0 \leftarrow N\sigma$
$\mathcal{P}^* \leftarrow \arg\min_{\mathcal{P}\in\mathbb{P}} \Pi(\mathcal{P})$
$\Pi^* \leftarrow N(p_1|\mathcal{P}^*)\sigma(p_1|\mathcal{P}^*) + N(p_2|\mathcal{P}^*)\sigma(p_2|\mathcal{P}^*)$
if $\frac{\Pi_0-\Pi^*}{\Pi_0} > \lambda$ **then**
    Accept split
end if

2.2.3 Internal and leaf node fitting

Once the tree structure is generated, we fit an HBI correction group value for each node on the tree. Let $\mathbf{Y} = [\underline{y}_1, ..., \underline{y}_M]$, where $\underline{y}_m \in \mathbb{R}^{N\times 1}$ refers to the m^{th} column in the data matrix $\mathbf{Y}$. The prediction value for each HBI correction (m) is fitted using the weighted least squares (WLS) of molecules matched to the node using Eq. (12).

$$\hat{\beta}_m = (\underline{X}^T\mathbf{W}_m\underline{X})^{-1}\underline{X}^T\mathbf{W}_m\underline{y}_m \tag{12}$$

where $\underline{X} \in \mathbb{Z}_0^{+N}$ is a column vector containing ones of length number of molecules matched to the node (N). $\hat{\beta}_m \in \mathbb{R}^1$ is a scalar of the fitted group value for HBI correction m. $\mathbf{W}_m$ is a N by N matrix that contains the weights. The diagonal of $\mathbf{W}_m$ is the inverse of the data variance, while the rest of the entries are zeros as shown in Eq. (13). Note that the weights are different for each HBI correction m due to different associated uncertainties.

$$\mathbf{W}_m = \text{diag}\left(\frac{1}{\underline{y}^2_{\epsilon,m}}\right) \tag{13}$$

Eq. (12) is equivalent to computing the weighted average of HBI corrections for molecules matched to the node. The prediction uncertainty for HBI corrections can be estimated from weighted covariance. We report the prediction value for the HBI correction at each node as Eq. (14). The factor of 2 is because it is conventional to use 95% confidence intervals (2 σ) in thermochemistry.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$\hat{\beta}_m \pm 2\sqrt{V(\hat{\beta}_m)} \tag{14}$$

where

$$V(\hat{\beta}_m) = (\underline{X}^\top \mathbf{W}_m \underline{X})^{-1} \hat{\sigma}_m^2 \tag{15}$$

$$\hat{\sigma}_m^2 = \frac{1}{N-1}(\underline{y}_m - \underline{X}\hat{\beta}_m)^\top \mathbf{W}_m (\underline{y}_m - \underline{X}\hat{\beta}_m) \tag{16}$$

As mentioned earlier, there are two sources of uncertainties in machine learning for chemistry: the aleatoric uncertainty, and the epistemic uncertainty. Eq. (14) only indirectly makes use of the aleatoric information from the data points, and the uncertainty estimate is dominated by the model variance. Consequently, Eq. (14) can greatly underestimate the uncertainty for nodes closer to the terminal nodes, due to the smaller amount of data points matched to the node.

Instead, we believe an ideal uncertainty estimate should account for both aleatoric uncertainty and model variance. We can propagate the data uncertainty through Eq. (12), assuming negligible covariance between each $\underline{y}_m$

$$\begin{aligned} V(\underline{y}_{\epsilon,m}) &\approx \left| \frac{\partial \hat{\beta}_m}{\partial \underline{y}_m} \right|^{\circ 2} \underline{y}_{\epsilon,m}^{\circ 2} \\ &= \left((\underline{X}^T \mathbf{W}_m \underline{X})^{-1} \underline{X}^T \mathbf{W}_m \right)^{\circ 2} \underline{y}_{\epsilon,m}^{\circ 2} \end{aligned} \tag{17}$$

where $\circ$ denotes the Hadamard power.

We introduce a new uncertainty estimation equation (Eq. (18)). When the model variance is small, the uncertainty estimate has a lower bound set by the uncertainty propagated from the data points. We evaluate the effects of this uncertainty estimation in Sec. 3.1.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$\hat{\beta}_m \pm 2\sqrt{V(\hat{\beta}_m) + V(\underline{y}_{\epsilon,m})} \tag{18}$$

3 Results and Discussion

The results and discussion will unfold as follows. We first show the effectiveness of the proposed bounded uncertainty estimation method in Sec. 3.1. Then, we assess the effects of the aleatoric uncertainty-based and model variance-based pre-pruning methods in Sec. 3.2. The names of the models and their descriptions are shown in Table 1 Finally, in Sec. 3.3, we present a comprehensive comparison between the empirical tree model in RMG[50] and the new SIDT model.

Table 1: Names and description of SIDT models presented in this work.

Model name	Description
BC	Base case model with Eq. (14) as uncertainty estimation
BC_UB	Base case model with Eq. (18) as bounded uncertainty estimation
AP_UB	Model with aleatoric uncertainty-based pre-pruning and Eq. (18) as bounded uncertainty estimation
MPλ_UB	Model with model variance reduction-based pre-pruning using λ as reduction threshold and Eq. (18) as bounded uncertainty estimation

3.1 Uncertainty estimation evaluation

Accurate uncertainty estimates are vital for uncertainty analysis for kinetic modeling, making it important to examine how well the uncertainty estimates represent the true errors. To assess this, we use the fraction of test error bounded by the uncertainty estimate as a metric of uncertainty estimation performance.

We use 9:1 random splitting as the training and test set and train the SIDT model with various fractions of the training set (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). We compare the uncertainty estimation by weighted standard deviation (Eq. (14)) and the bounded equation (Eq. (18)) using this metric. Fig. 5 shows the fraction of test error bounded

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

by the predicted uncertainty as a function of the number of training data used, along with either Eq. (14) or Eq. (18). Using Eq. (18) improves the bounded fraction, suggesting that Eq. (18) reports a more realistic uncertainty estimate.

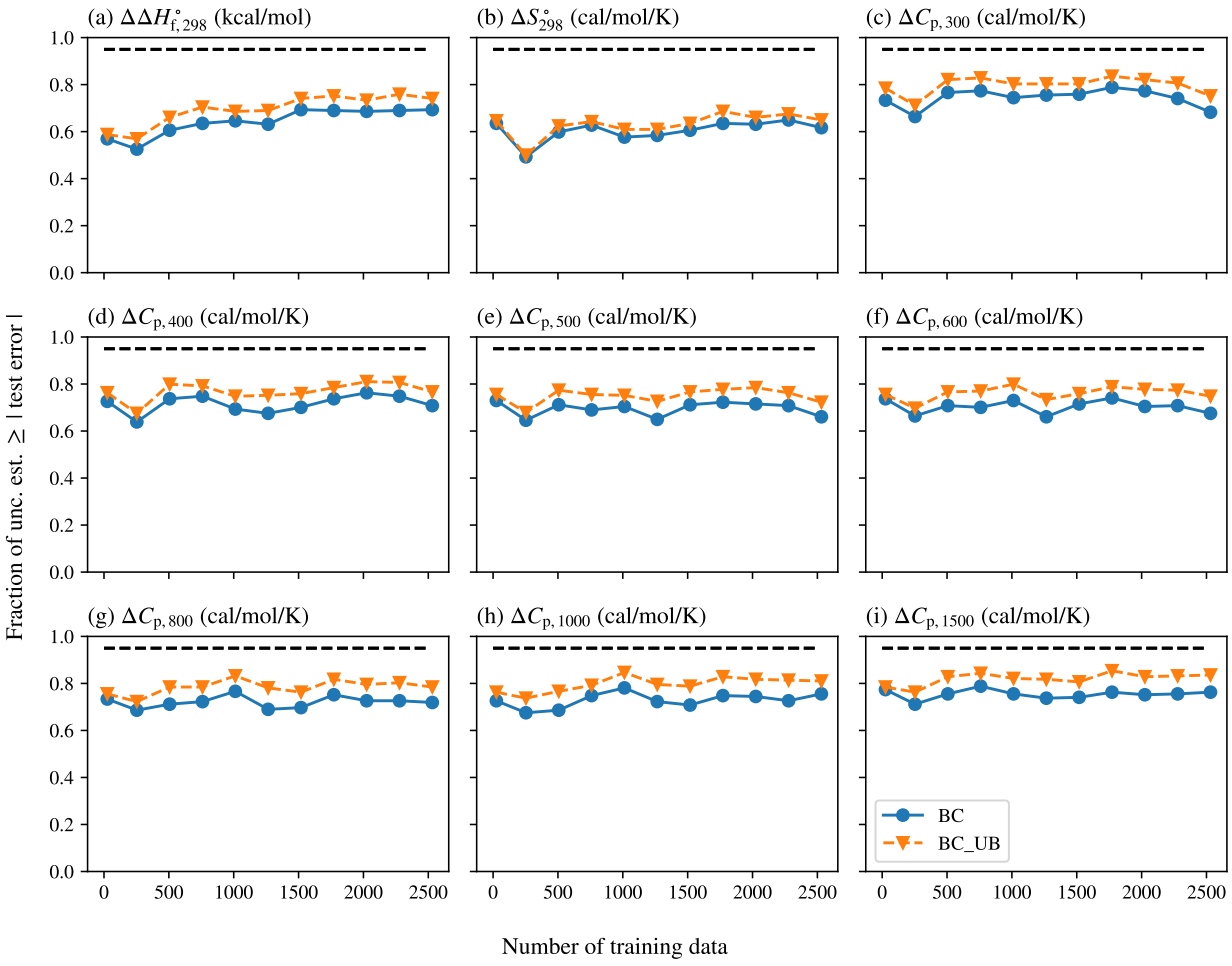


Figure 5: Fraction of test error bounded by uncertainty estimate as a function of various number of training data. The dashed line at 0.95 indicates the ideality (95% confidence level from using 2 times weighted standard deviations as uncertainty estimates). BC refers to the base case SIDT with a weighted standard deviation for the uncertainty estimate, while BC_UB uses a bounded weighted standard deviation.

Additionally, in the discussion of uncertainty estimation, there are three terms at play. The first is the prediction error, which we measure using the difference between the model prediction and the hold-out test set (also referred to as test error). The second is the data uncertainty in the test set. The third is the uncertainty estimate, which is the uncertainty

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

in model prediction estimated by the model. If the data has no noise, the ideal uncertainty estimate should be similar to the test error. However, since both the data have noises, the uncertainty estimate should also reflect both the epistemic components (error caused by the model variance and biases) and the aleatoric components (noises in the data).

Fig. 6 shows the distribution of normalized uncertainty estimates from the BC_UB model trained on the full training set and tested on the hold-out test set to assess the quality of uncertainty estimates with the noisy dataset. We normalize the uncertainty estimate using the root sum squared of test errors and the data noises. The bounded uncertainty estimation shifts the long tail of underestimated uncertainty towards the more reasonable range of estimates (within the region between the dashed lines).

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

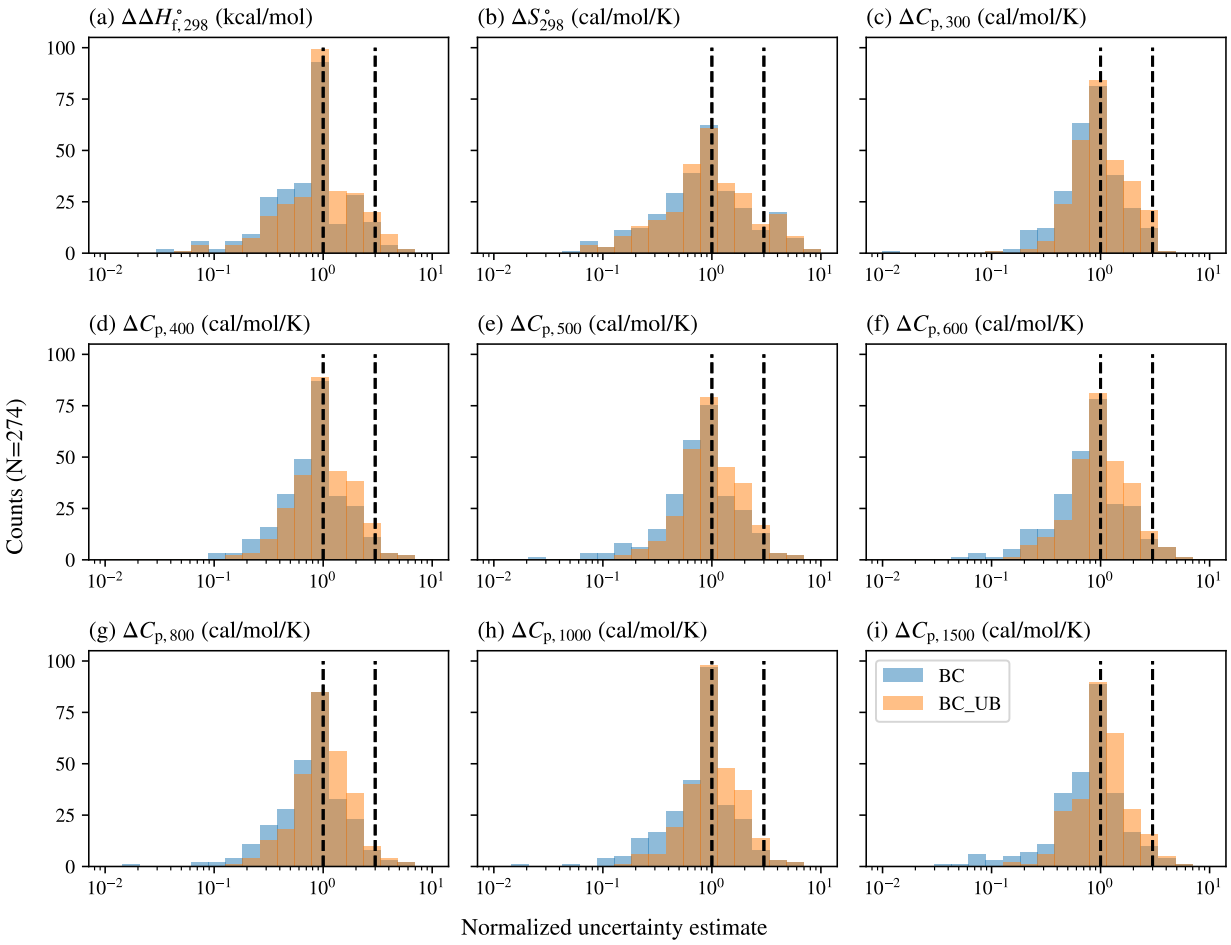


Figure 6: Distribution of normalized uncertainty estimate for BC and BC_UB. The left-side vertical dashed line shows accurate uncertainty, and the right-side vertical line indicates a 3x overestimation. The space in between is the reasonable uncertainty region. BC refers to the base case SIDT with a weighted standard deviation for uncertainty estimate, while BC_UB uses a bounded weighted standard deviation.

Additionally, Fig. 7 shows the normalized uncertainty estimate for BC_UB only, distinguishing between data points where test errors dominate and those where data uncertainties dominate. If the model learns the molecules well, it tends to estimate more reliable uncertainties, while uncertainty can be underestimated when the test error dominates.

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

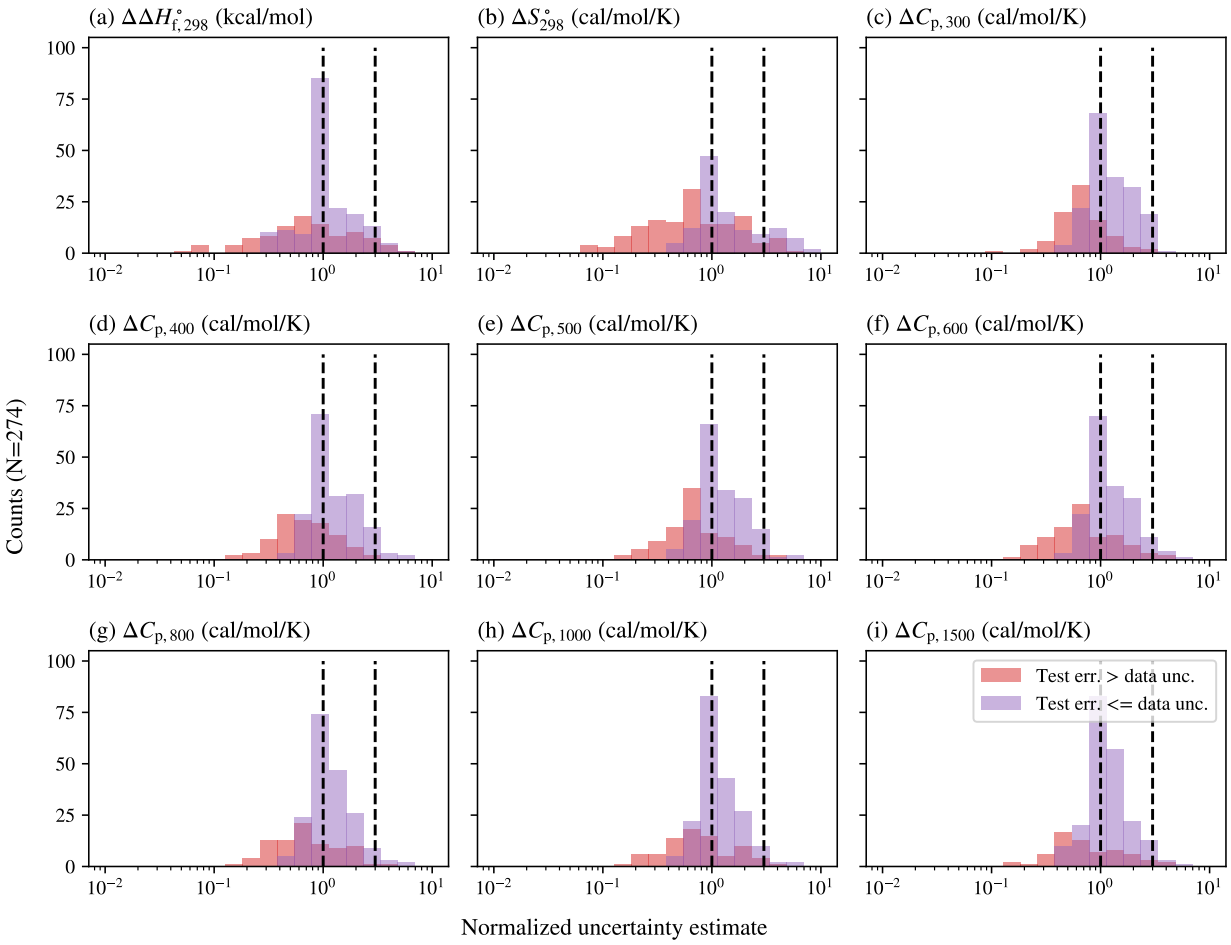


Figure 7: Distribution of normalized uncertainty estimates for BC_UB only for cases where the test error dominates (red) or the data uncertainty dominates (purple).

3.2 Effects of pre-pruning

Fully-expanded trees can overfit the noise in the data and decrease model interpretability, unless special algorithms are implemented, such as the ascending scheme used in the original SIDT for rate estimation[26]. Thus, a compact tree without the loss of prediction accuracy (tree regularization) is often desired,[51] motivating us to explore new pre-pruning methods for our application. Here, we evaluate the effects of aleatoric uncertainty-based pre-pruning on BC_UB, measuring prediction accuracy through test errors and evaluating tree compactness based on tree sizes. We use 9:1 training:test random splitting.

Fig. 8 shows the reduction in tree sizes in terms of the number of nodes by comparing

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

the base case model (BC_UB) and the model using aleatoric uncertainty-based pre-pruning (AP_UB). Additionally, Figs. S6 and S7 show the root-mean-square error (RMSE) and mean-absolute error (MAE) as a function of the training data size. It is clear that AP_UB has smaller tree sizes compared to BC_UB while achieving almost the same model performances. Moreover, the degree of tree size reduction becomes more apparent when the training size becomes larger by preventing unnecessary exhausted splitting.

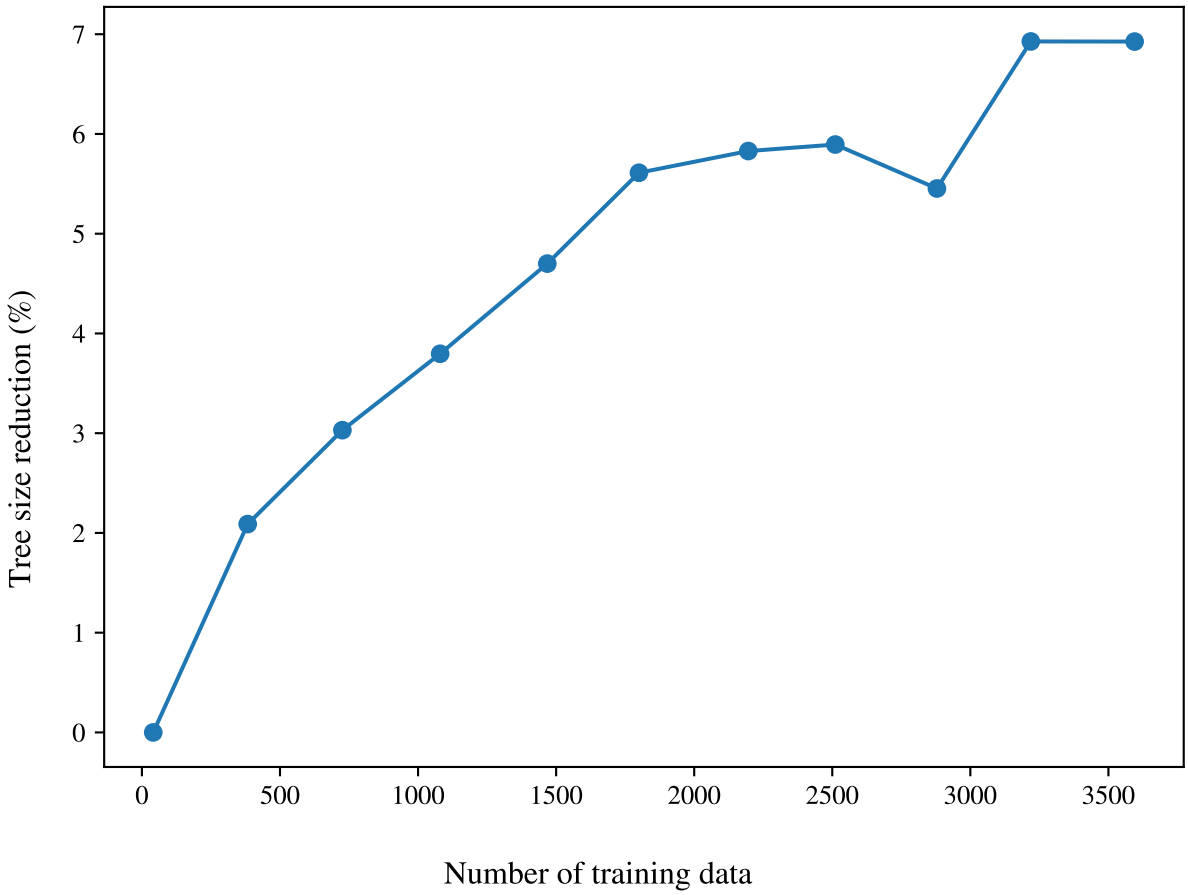


Figure 8: Tree size reduction (%) as a function of various training data sizes by adding aleatoric pre-pruning to BC_UB. Tree size reduction (%) is computed as $(1 - N_{\text{node,AP\_UB}}/N_{\text{node,BC\_UB}}) \times 100\%$.

The model variance reduction-based pre-pruning requires a hyperparameter for the model variance reduction threshold (λ). We further split the abovementioned training set by 8:1 for training:validation, and use the validation set to search for the optimal λ. Figs. S8 to S10

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

show the validation errors and tree size as a function of λ.

We also evaluate the effects of model variance reduction-based pre-pruning compared to the aleatoric uncertainty-based pre-pruning. Figs. 9 and S11 show the test errors as a function of various number of training data for the model with aleatoric uncertainty-based pre-pruning (AP_UB) and model variance reduction-based pre-pruning (MP_UB).

The model variance reduction-based pre-pruning is highly sensitive to the choice of model variance reduction threshold (λ). The size of λ can depend on the dataset and the learning target, and a large λ can result in an underfitting tree.

For our HBI correction dataset, we find that model variance reduction-based pre-pruning has a mix of impacts on the model performance. As shown in Figs. S9 and S10, a large λ decreases the RMSE for ΔS°_{298} and many of $\Delta C_{\mathrm{p}}(T)$, but significantly increases the RMSE and MAE for $\Delta\Delta H^{\circ}_{\mathrm{f},298}$. We notice that at the root node, the SIDT algorithm often attempts to split the radicals based on whether the atom with the radical is in a ring. This splitting decision makes chemical sense, but it doesn't necessarily lead to a large reduction in the model variance, as the radicals are still diverse. Instead, the larger reduction in the model variance often occurs at nodes closer to the terminal nodes. Consequently, the SIDT learning can be greatly hindered, as this pre-pruning method tends to prune nodes closer to the root.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

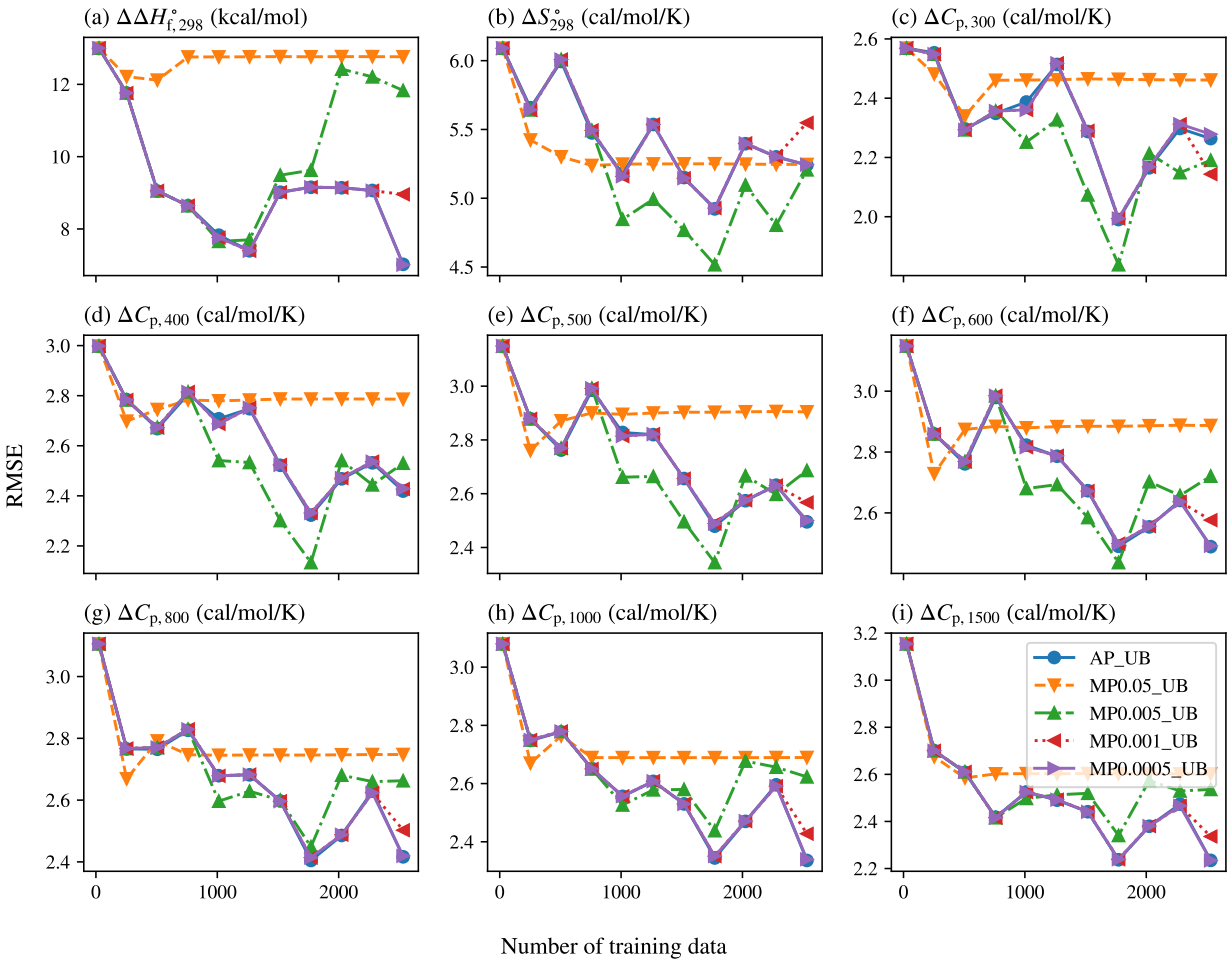


Figure 9: Test root-mean-square error (RMSE) as a function of various fractions of training data used for the model with aleatoric uncertainty-based pre-pruning (AP_UB) or model variance reduction-based pre-pruning with various model variance reduction thresholds (λ) (MPλ_UB).

Based on the above results, we recommend the AP approach, given its ability to achieve the same performance with reduced tree sizes. The MP approach is not recommended for this application.

3.3 Comparison with empirical tree

This work aims to replace the hand-made empirical tree that requires expert knowledge to improve and update in RMG-database.[50] On one hand, the empirical tree was trained historically over decades by many generations of researchers covering a wide scope of chem-

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

istry.[24,41,52–54] On the other hand, the empirical tree could have errors and missing values due to human errors, non-systematic construction, and implicit assumptions that successor researchers are not aware of. Both the empirical tree estimator and the SIDT estimator can suffer from the scarcity of high-quality data. The SIDT model presented below uses the aleatoric pre-pruning and bounded uncertainty estimation, following the same settings as AP_UB.

3.3.1 Comparison with original empirical tree

We compare the empirical tree model and the SIDT model by examining their performance on the test set from 9:1 training:test random splitting. It is worth noting that we are comparing with RMG's current empirical tree with no modifications or refitting. The current empirical tree model has 1994 nodes. Among them, 1828 have group values, while the rest borrow the group values from their parent, ancestor, or other nodes. 1473 of them are recently added for halogenated radicals with uncertainty estimates,[24] 107 of them are recently added for radicals involving H, C, N, O with uncertainty estimates,[24] while 248 of them have no uncertainty estimates. A representative value of 0.1 kcal/mol per group in $\Delta\Delta H^\circ_{\mathrm{f},298}$ is often assumed for them.[25]

Fig. 10 shows the error distribution of the empirical tree and SIDT models. Surprisingly, although the two models are constructed using different approaches with different training data, they have similar error distributions. Fig. S12 shows a parity plot of the absolute test errors of SIDT model compared to the empirical tree model. Despite the two models having similar error distribution, molecules with larger errors are different for the two models for most HBI corrections.

As shown in Fig. 10, SIDT improves the estimation of $\Delta\Delta H^\circ_{\mathrm{f},298}$(R···H) by 1.96/1.35 kcal/mol in terms of RMSE/MAE. The improvements in other HBI corrections are smaller due to the use of $\Delta\Delta G^\circ_{\mathrm{f}}(T)$ in the splitting criterion, which can place more weight on the performance for $\Delta\Delta H^\circ_{\mathrm{f},298}$. However, we note that the improvements in the HBI correction

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

estimation are limited by the error in the closed shell thermochemistry estimated using the group additivity method (see Fig. S4).

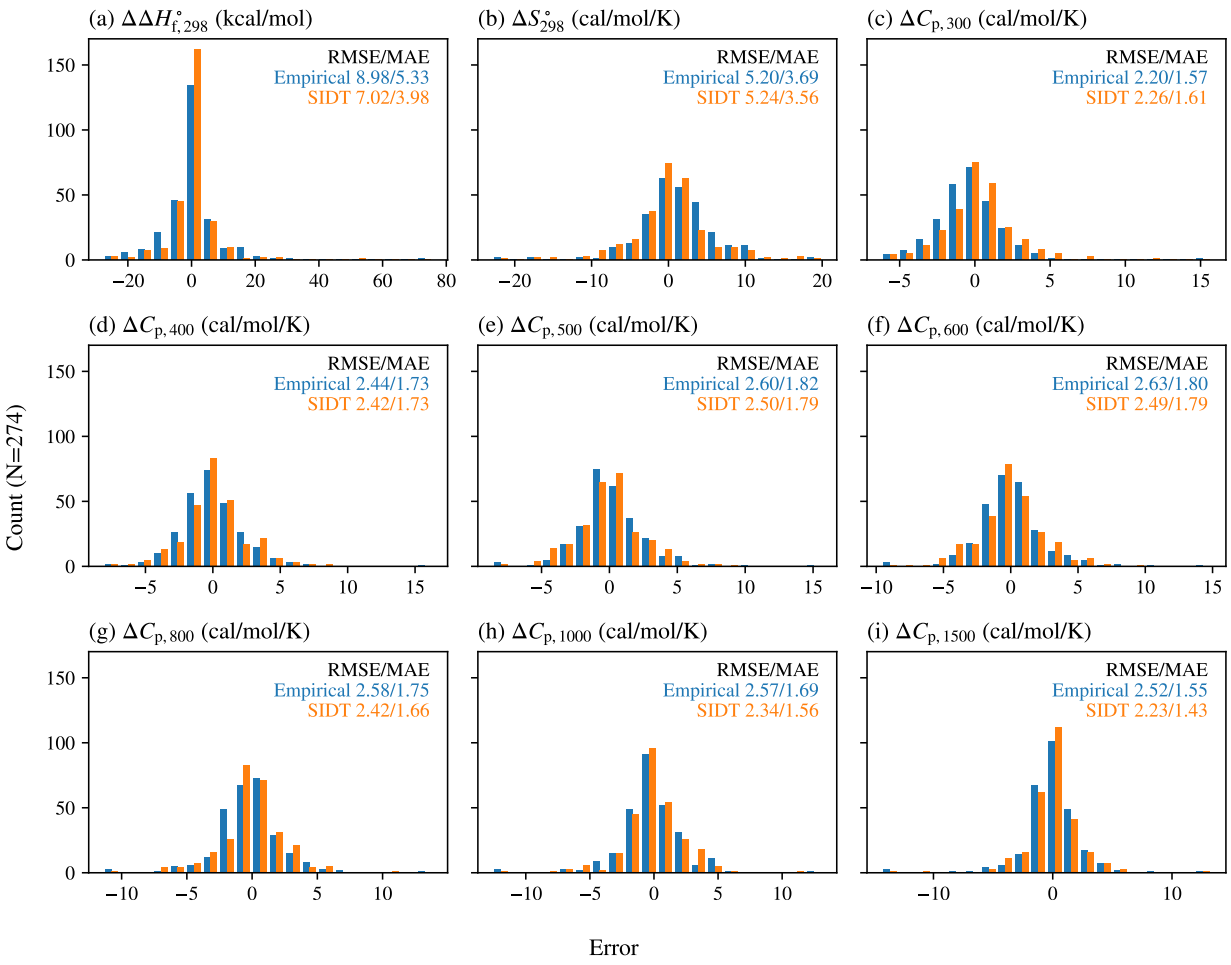


Figure 10: Histogram of error in HBI corrections estimated by the hand-made empirical tree model and the SIDT model.

Figs. 11 and 12 show the parity plots of the empirical tree and SIDT models. The SIDT estimator improves the parity of HBI corrections compared to the empirical tree, particularly for enthalpy. In Fig. 11, many dots are forming horizontal lines. This suggests underdeveloped branches in the empirical tree. Some terminal nodes match molecules with a wide range of HBI corrections, leading to overgeneralization. The SIDT does not have this problem.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

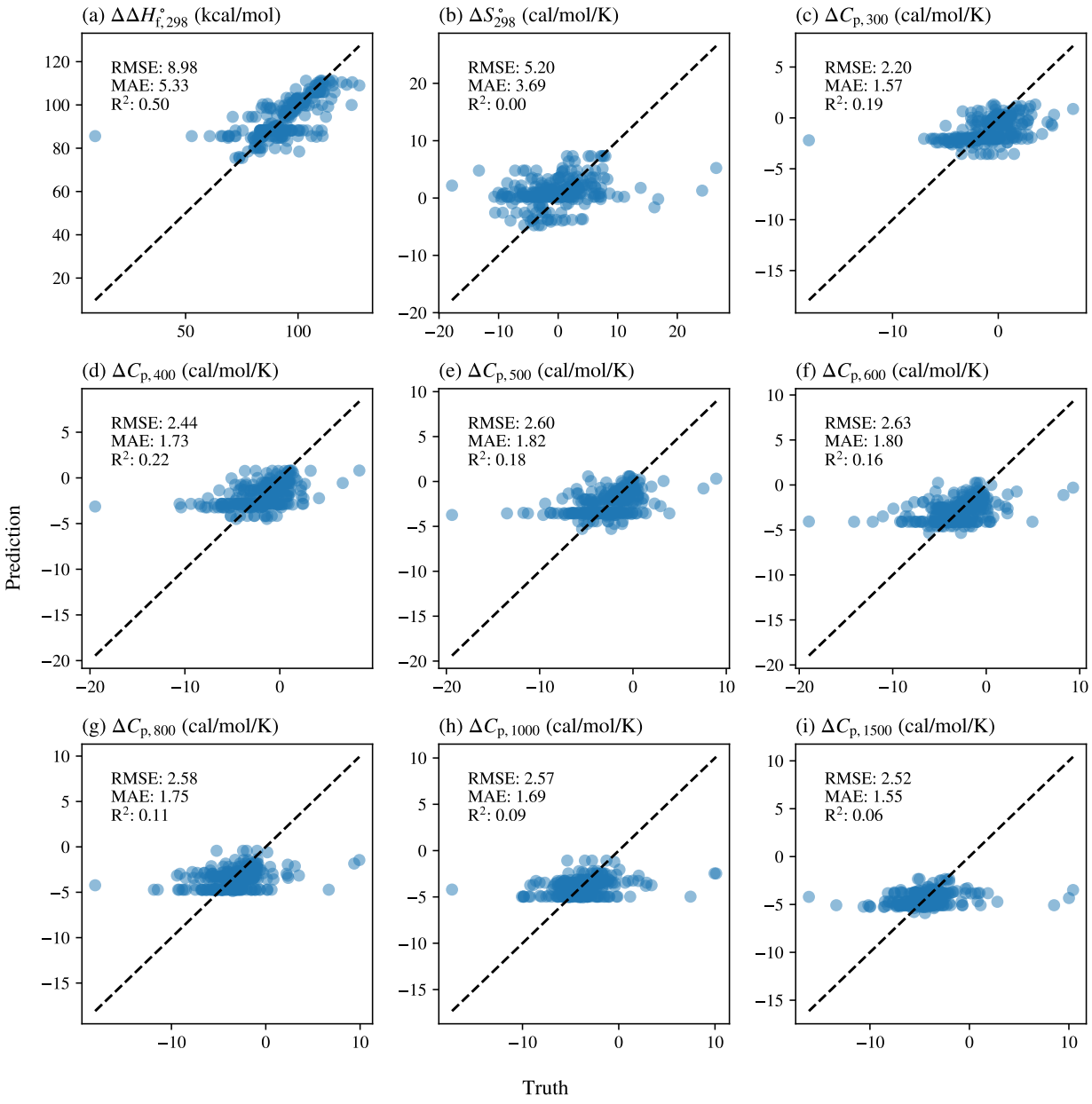


Figure 11: Parity plot of HBI corrections predicted by the empirical tree compared to the true values.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

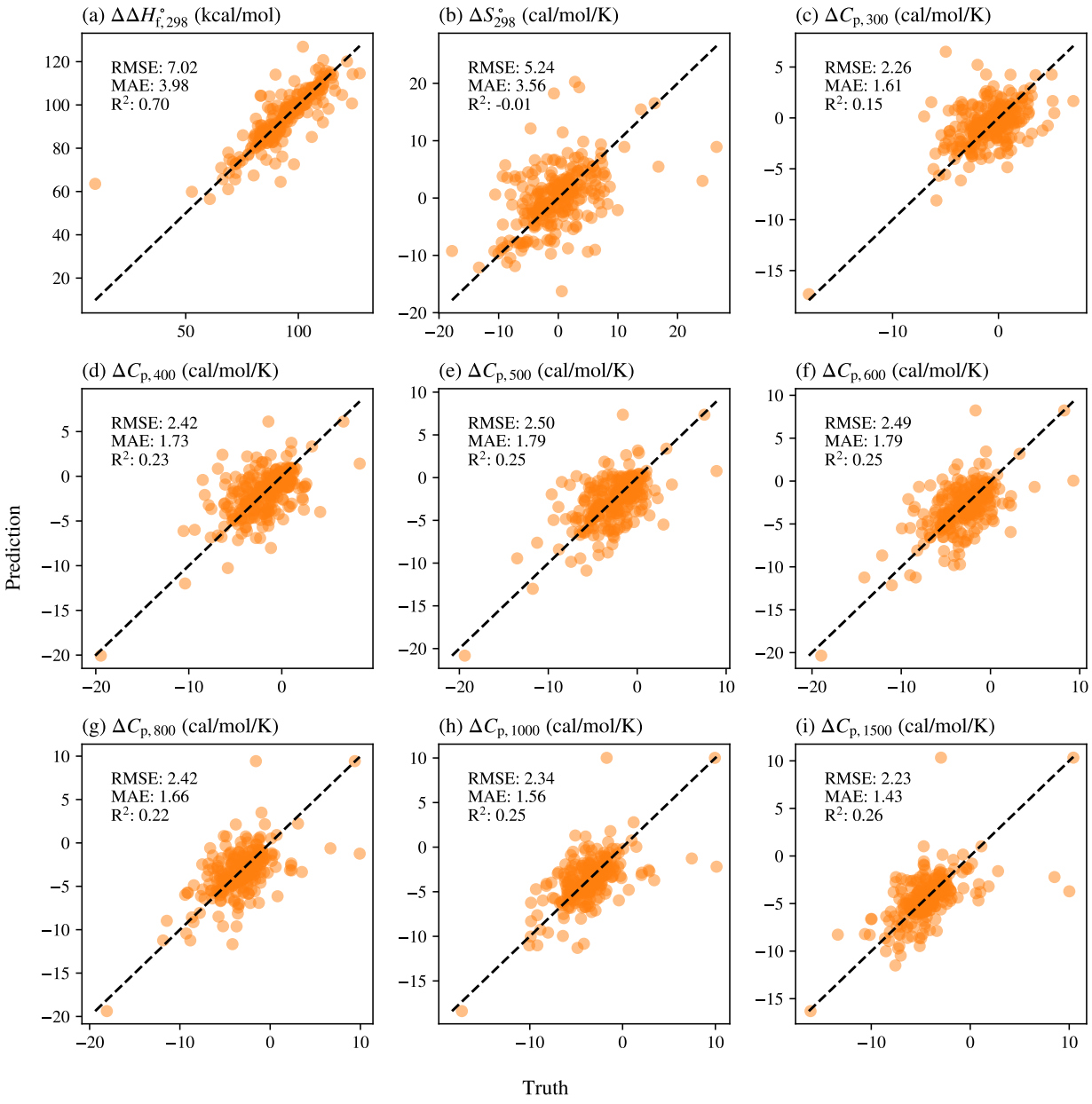


Figure 12: Parity plot of HBI corrections predicted by the SIDT compared to the true values.

Fig. 13 shows the distribution of normalized uncertainty estimates by the SIDT model (AP_UB) and the empirical tree model. In general, the SIDT model has realistic uncertainty estimates that center around 1, while the uncertainty estimates from the empirical tree model in RMG, which stem from previously published works,[24,25] are too optimistic. Additionally, Fig. S13 show the distribution of normalized uncertainty estimates by the SIDT model

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

(AP_UB) only, separating test data whose closed-shell components are derived from GAV and QM. The quality of uncertainty estimates for test data with closed-shell components from QM is generally better than its GAV counterpart.

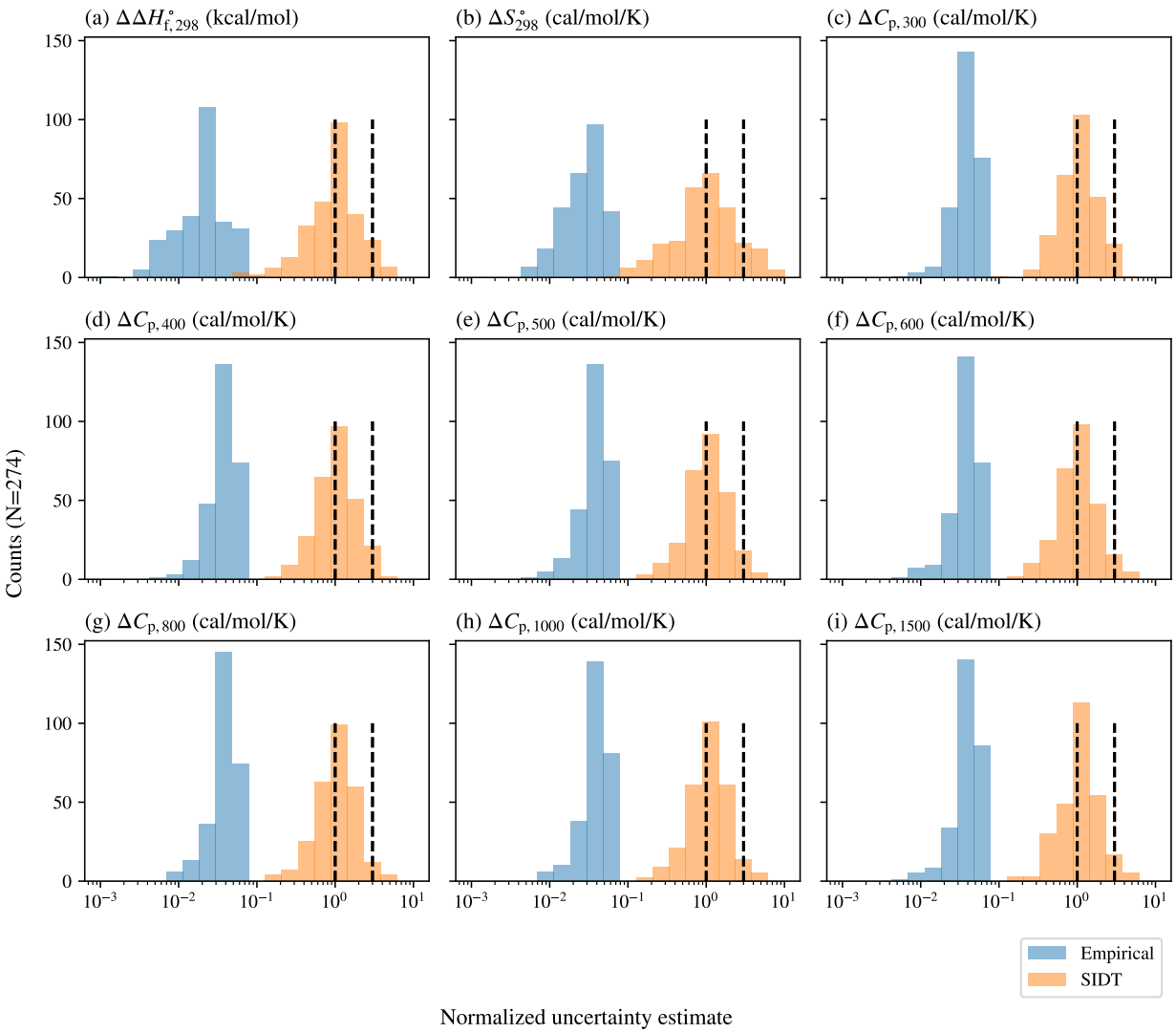


Figure 13: Distribution of normalized uncertainty estimates for the empirical tree model in RMG and the SIDT model (AP_UB). The left-side vertical dashed line shows accurate uncertainty, and the right-side vertical line indicates a 3x overestimation. The space in between is the reasonable uncertainty region.

3.3.2 Comparison with re-fitted empirical tree

As mentioned in Sec. 2.2, there are two stages to construct a SIDT model: the tree extension stage, and the internal and leaf node fitting stage. For the empirical tree model, the

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

researchers do the tree extension manually based on domain knowledge, and then they fit small subbranches of the tree to small specialized sets of training data. For the SIDT model, the algorithm learns the tree structure from the given training set and fits node values from the same set of training data. In light of this, we perform additional comparisons to assess the empirical tree model and the SIDT model, as discussed below.

We refit the empirical tree model with the same training set as the SIDT model, while keeping the empirical tree model's structure. We use the same 9:1 training:test random splitting employed previously. The error distributions of the retrained empirical tree model and the SIDT model can be found in Fig. S14. Small improvements can be found in the RMSE and MAE of the 9 HBI correction targets. We also re-estimate the uncertainties for the empirical tree model using Eq. (18), which significantly improves its quality of uncertainty estimation as shown in Fig. S16.

3.3.3 Comparison with empirical tree using advanced data splitting

The random splitting approach employed above assesses the interpolation ability of the tree models.[55] Here, we assess the extrapolation ability of the empirical tree model and the SIDT model with a more challenging data split, i.e., cluster split. The details on how we perform cluster split can be found in Sec. S3.3.3. We split the dataset using a 9:1 training:test cluster split. The error distribution of the SIDT model compared to the empirical tree model and the re-fitted empirical tree model can be found in Figs. S18 and S19, respectively. Both empirical tree models outperform the SIDT model with the cluster split test set.

However, this is not a fair comparison due to two reasons. First, as we don't know what datasets were used to train the empirical model, the test set likely partially overlaps with the training set of the empirical model. Second, the chemical knowledge embedded in the empirical tree acts as a "learned" representation even when relevant data are missing. Nonetheless, this signals the importance of being exposed to a wide range of chemistry during the tree extension stage, as the empirical tree's handmade tree structure aids its performance

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

in the extrapolation task. It's possible that an expert-informed design of the tree structure near the root node, while filling the rest of the tree out with the SIDT algorithm to solve the known issue of underdeveloped branches (as discussed in connection with Fig. 11), could be beneficial.

3.3.4 Tree structure comparison

Fig. 14 shows the number of branches for each internal node, the depth of individual leaf nodes, and the number of subgraph isomorphic comparisons needed when descending from the root node to leaf node in the ordinary empirical tree in RMG and the SIDT model. From Figs. 14(a), (b), (d), and (e), the empirical tree model is wider but shallower, while the SIDT model is narrower but deeper. Figs. 14(c) and (f) show that SIDT requires fewer subgraph isomorphic comparisons when descending from the root node to the leaf node than the empirical tree, signaling a more efficient descent during parameter estimation.

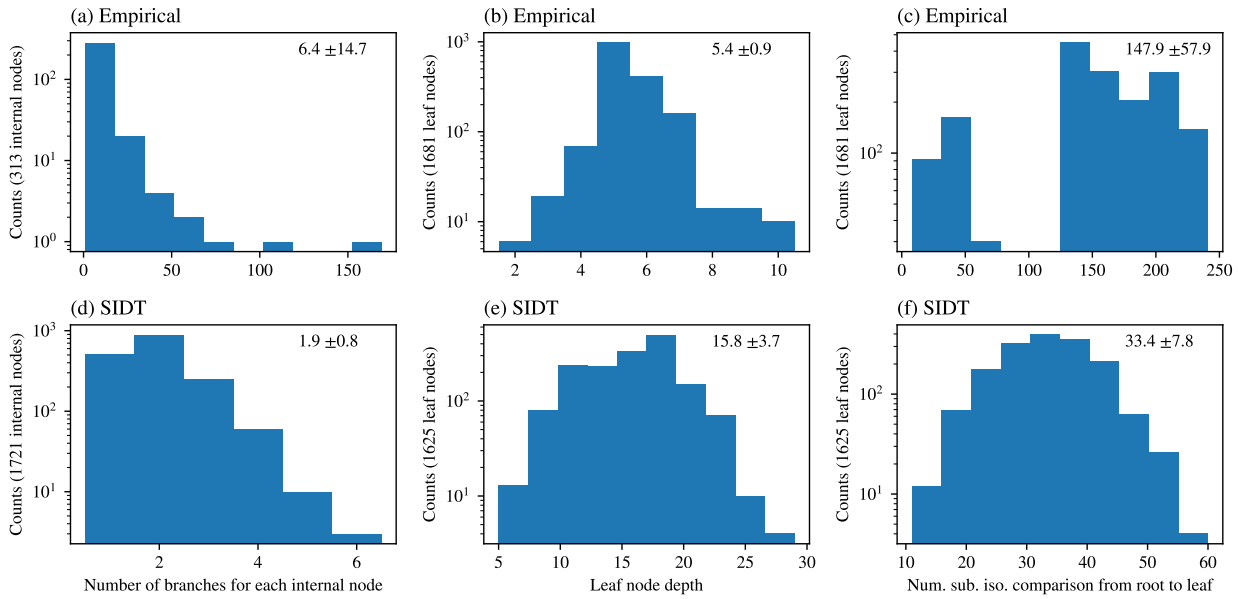


Figure 14: Number of branches for each internal node in the empirical tree (a) and SIDT (d), depth of individual leaf nodes for the empirical tree (b) and SIDT (e), and number of subgraph isomorphic comparisons needed when descending from the root to leaf node for the empirical tree (c) and SIDT (f). The number of needed subgraph isomorphic comparisons is estimated by the sum of the number of children for nodes on the route descending from the root node to the leaf node; note that this is an upper-bound estimation.

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Overall, SIDT provides a method to easily extend, (re-)train, and improve empirical tree-based methods that have previously required intensive manual care. As shown in Sec. 3.1, SIDT provides reasonable uncertainty estimates. The SIDT model has great computational and memory efficiency both for training and inferencing. Specifically, the SIDT model takes around 6 min to train, has around 3000 nodes, and has around 3000 parameters for each of the 9 HBI correction targets.

4 Conclusions

We extended the subgraph isomorphic decision tree (SIDT) algorithm originally developed for rate estimation[26] to estimate hydrogen bond increment (HBI) corrections in the context of automated mechanism generation. We evaluated different uncertainty estimation methods and pre-pruning, and compared the performance of SIDT model with the empirical tree model. For uncertainty estimation evaluation, we showed that using a weighted standard deviation bounded by aleatoric uncertainty consistently improves the quality of uncertainty estimates. Additionally, we explored and evaluated two pre-pruning methods. Aleatoric uncertainty-based pre-pruning demonstrated significant improvements in tree compactness without compromising prediction accuracy. However, model variance reduction-based pre-pruning showed sensitivity to the choice of hyperparameter λ and did not grant meaningful improvements when applied to the noisy HBI correction data. Eventually, we compared the SIDT model with the existing hand-made empirical tree model in the Reaction Mechanism Generator (RMG) database. While the empirical tree was constructed manually over time by many researchers to cover a wider range of chemistry, it can have errors and missing values due to human error. Compared to the empirical tree model, the SIDT model (1) is much easier to generate and extend, (2) improves the accuracy and R^2 for HBI correction prediction, (3) provides significantly more reliable uncertainty estimates, and (4) has a more advantageous tree structure for descending speed. We also compared the SIDT model with

https://doi.org/10.26434/chemrxiv-2024-20zjc ORCID: https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

a re-fitted empirical tree and assessed its extrapolation ability through cluster splitting, highlighting that a tree structure learned from a diverse chemical space could potentially improve the robustness of the tree estimator. In conclusion, the SIDT algorithm provides a promising alternative for estimating HBI corrections, offering an automated way to construct lightweight and easy-to-update estimators for kinecticists.

Acknowledgement

The authors thank the MIT SuperCloud and Lincoln Laboratory Supercomputing Center[56] for providing computing resources that have contributed to the research results reported within this paper. This work at MIT is funded by the Gas Phase Chemical Physics Program of the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences under Award No. DE-SC0014901.

Supporting Information Available

All the data, codes, and models generated for this work can be found in our GitHub Repository at `https://github.com/hwpang/RadicalTree`.

References

(1) Pang, H.-W.; Forsuelo, M.; Dong, X.; Hawtof, R. E.; Ranasinghe, D. S.; Green, W. H. Detailed Multiphase Chemical Kinetic Model for Polymer Fouling in a Distillation Column. *Industrial & Engineering Chemistry Research* **2023**, *62*, 14266–14285.

(2) Pang, H.-W.; Dong, X.; Green, W. H. Oxygen Chemistry in Polymer Fouling: Insights from Multiphase Detailed Kinetic Modeling. *Industrial & Engineering Chemistry Research* **2024**, *63*, 1013–1028.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(3) Dong, X.; Ninnemann, E.; Ranasinghe, D. S.; Laich, A.; Greene, R.; Vasu, S. S.; Green, W. H. Revealing the critical role of radical-involved pathways in high temperature cyclopentanone pyrolysis. *Combustion and Flame* **2020**, *216*, 280–292.

(4) Lai, L.; Pang, H.-W.; Green, W. H. Formation of two-ring aromatics in hexylbenzene pyrolysis. *Energy & Fuels* **2020**, *34*, 1365–1377.

(5) Johnson, M. S.; Nimlos, M. R.; Ninnemann, E.; Laich, A.; Fioroni, G. M.; Kang, D.; Bu, L.; Ranasinghe, D.; Khanniche, S.; Goldsborough, S. S.; others Oxidation and pyrolysis of methyl propyl ether. *International Journal of Chemical Kinetics* **2021**, *53*, 915–938.

(6) Pio, G.; Dong, X.; Salzano, E.; Green, W. H. Automatically generated model for light alkene combustion. *Combustion and Flame* **2022**, *241*, 112080.

(7) Dong, X.; Pio, G.; Arafin, F.; Laich, A.; Baker, J.; Ninnemann, E.; Vasu, S. S.; Green, W. H. Butyl Acetate Pyrolysis and Combustion Chemistry: Mechanism Generation and Shock Tube Experiments. *The Journal of Physical Chemistry A* **2023**, *127*, 3231–3245.

(8) Grinberg Dana, A.; Wu, H.; Ranasinghe, D. S.; Pickard IV, F. C.; Wood, G. P.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation:(1) The AIBN/H2O/CH3OH Radical "Soup". *Molecular Pharmaceutics* **2021**, *18*, 3037–3049.

(9) Wu, H.; Grinberg Dana, A.; Ranasinghe, D. S.; Pickard IV, F. C.; Wood, G. P.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation:(2) Imipramine Stress Testing. *Molecular Pharmaceutics* **2022**, *19*, 1526–1539.

(10) Van de Vijver, R.; Vandewiele, N. M.; Bhoorasingh, P. L.; Slakman, B. L.; Seyedzadeh Khanshan, F.; Carstensen, H.-H.; Reyniers, M.-F.; Marin, G. B.;

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

West, R. H.; Van Geem, K. M. Automatic mechanism and kinetic model generation for gas-and solution-phase processes: a perspective on best practices, recent advances, and future challenges. *International Journal of Chemical Kinetics* **2015**, *47*, 199–231.

(11) Johnson, M. S.; Pang, H.-W.; Liu, M.; Green, W. H. Species Selection for Automatic Chemical Kinetic Mechanism Generation. *ChemRxiv* **2023**, 1–33, `https://doi.org/10.26434/chemrxiv-2023-wwrqf`.

(12) Ruscic, B.; Pinzon, R. E.; Morton, M. L.; Srinivasan, N. K.; Su, M.-C.; Sutherland, J. W.; Michael, J. V. Active thermochemical tables: Accurate enthalpy of formation of hydroperoxyl radical, HO2. *The Journal of Physical Chemistry A* **2006**, *110*, 6592–6601.

(13) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *The Journal of Physical Chemistry A* **2019**, *123*, 5826–5835.

(14) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-evolving machine: A continuously improving model for molecular thermochemistry. *The Journal of Physical Chemistry A* **2019**, *123*, 2142–2152.

(15) Dobbelaere, M. R.; Plehiers, P. P.; Van de Vijver, R.; Stevens, C. V.; Van Geem, K. M. Learning molecular representations for thermochemistry prediction of cyclic hydrocarbons and oxygenates. *The Journal of Physical Chemistry A* **2021**, *125*, 5166–5179.

(16) Ureel, Y.; Vermeire, F. H.; Sabbe, M. K.; Van Geem, K. M. Beyond group additivity: Transfer learning for molecular thermochemistry prediction. *Chemical Engineering Journal* **2023**, *472*, 144874.

(17) Ribeiro, M. T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016; pp 1135–1144.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(18) Goh, G. B.; Hodas, N. O.; Siegel, C.; Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034* **2017**,

(19) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*, 1–10, ISBN: 9781510860964, NeurIPS Proceedings, `https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`.

(20) Han, K.; Green, W. H.; West, R. H. On-the-fly pruning for rate-based reaction mechanism generation. *Computers & Chemical Engineering* **2017**, *100*, 1–8.

(21) Jocher, A.; Vandewiele, N. M.; Han, K.; Liu, M.; Gao, C. W.; Gillis, R. J.; Green, W. H. Scalability strategies for automated reaction mechanism generation. *Computers & Chemical Engineering* **2019**, *131*, 106578.

(22) Benson, S. W.; Buss, J. H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *The Journal of Chemical Physics* **1958**, *29*, 546–572.

(23) Lay, T. H.; Bozzelli, J. W.; Dean, A. M.; Ritter, E. R. Hydrogen atom bond increments for calculation of thermodynamic properties of hydrocarbon radical species. *The Journal of Physical Chemistry* **1995**, *99*, 14514–14527.

(24) Farina Jr, D. S.; Sirumalla, S. K.; Mazeau, E. J.; West, R. H. Extensive high-accuracy thermochemistry and group additivity values for halocarbon combustion modeling. *Industrial & Engineering Chemistry Research* **2021**, *60*, 15492–15501.

(25) Gao, C. W.; Liu, M.; Green, W. H. Uncertainty analysis of correlated parameters in automated reaction mechanism generation. *International Journal of Chemical Kinetics* **2020**, *52*, 266–282.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(26) Johnson, M. S.; Green, W. H. A Machine Learning Based Approach to Reaction Rate Estimation. *ChemRxiv* **2023**, 1–17, https://doi.org/10.26434/chemrxiv-2022-c98gc-v2.

(27) Liang, J.; He, R.; Nagaraja, S. S.; Mohamed, A. A. E.-S.; Lu, H.; Almarzooq, Y. M.; Dong, X.; Mathieu, O.; Green, W. H.; Petersen, E. L.; others A wide range experimental and kinetic modeling study of the oxidation of 2, 3-dimethyl-2-butene: Part 1. *Combustion and Flame* **2023**, *251*, 112731.

(28) Grinberg Dana, A.; Ranasinghe, D.; Wu, H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. ARC - Automated Rate Calculator, version 1.1.0. https://github.com/ReactionMechanismGenerator/ARC, 2019.

(29) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574.

(30) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(31) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of Cheminformatics* **2014**, *6*, 1–4.

(32) RDKit: Open-source cheminformatics. https://www.rdkit.org, (Accessed 2024-01-03).

(33) Montgomery Jr, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A complete basis set model chemistry. VI. Use of density functional geometries and frequencies. *The Journal of Chemical Physics* **1999**, *110*, 2822–2827.

(34) Frisch, M. J. et al. Gaussian 09 Revision E.01. 2009; Gaussian Inc. Wallingford CT.

(35) Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(36) Dana, A. G.; Johnson, M. S.; Allen, J. W.; Sharma, S.; Raman, S.; Liu, M.; Gao, C. W.; Grambow, C. A.; Goldman, M. J.; Ranasinghe, D. S.; others Automated reaction kinetics and network exploration (Arkane): A statistical mechanics, thermodynamics, transition state theory, and master equation software. *International Journal of Chemical Kinetics* **2023**, *55*, 300–323.

(37) Payne, A. M.; Wu, H.; Pang, H.-W.; Grambow, C. A.; Ranasinghe, D. S.; Dong, X.; Dana, A. G.; Green, W. H. Towards Accurate Quantum Mechanical Thermochemistry: (1) Extensible Implementation and Comparison of Bond Additivity Corrections and Isodesmic Reactions. *ChemRxiv* **2023**, 1–41, `https://doi.org/10.26434/chemrxiv-2023-4xlj9-v2`.

(38) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *The Journal of Chemical Physics* **2007**, *127*, 124105, `https://doi.org/10.1063/1.2770701`.

(39) Grinberg Dana, A.; Liu, M.; Green, W. H. Automated chemical resonance generation and structure filtration for kinetic modeling. *International Journal of Chemical Kinetics* **2019**, *51*, 760–776.

(40) Li, Y.; Curran, H. J. Extensive theoretical study of the thermochemical properties of unsaturated hydrocarbons and allylic and super-allylic radicals: the development and optimization of group additivity values. *The Journal of Physical Chemistry A* **2018**, *122*, 4736–4749.

(41) Lai, L.; Khanniche, S.; Green, W. H. Thermochemistry and group additivity values for fused two-ring species and radicals. *The Journal of Physical Chemistry A* **2019**, *123*, 3418–3428.

(42) Ghahremanpour, M. M.; Van Maaren, P. J.; Ditz, J. C.; Lindh, R.; Van der Spoel, D. Large-scale calculations of gas phase thermochemistry: Enthalpy of formation, standard

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

entropy, and heat capacity. *The Journal of Chemical Physics* **2016**, *145*, 114305, `https://doi.org/10.1063/1.4962627`.

(43) Somers, K. P.; Simmie, J. M. Benchmarking compound methods (CBS-QB3, CBS-APNO, G3, G4, W1BD) against the active thermochemical tables: formation enthalpies of radicals. *The Journal of Physical Chemistry A* **2015**, *119*, 8922–8933.

(44) Sharma, S.; Raman, S.; Green, W. H. Intramolecular hydrogen migration in alkylperoxy and hydroperoxyalkylperoxy radicals: accurate treatment of hindered rotors. *The Journal of Physical Chemistry A* **2010**, *114*, 5689–5701.

(45) Merchant, S. S.; Zanoelo, E. F.; Speth, R. L.; Harper, M. R.; Van Geem, K. M.; Green, W. H. Combustion and pyrolysis of iso-butanol: Experimental and chemical kinetic modeling study. *Combustion and Flame* **2013**, *160*, 1907–1929.

(46) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications* **2016**, *203*, 212–225.

(47) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J.; others Reaction mechanism generator v3. 0: advances in automatic mechanism generation. *Journal of Chemical Information and Modeling* **2021**, *61*, 2686–2696.

(48) Silla, C. N.; Freitas, A. A. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery* **2011**, *22*, 31–72.

(49) Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H. Characterizing uncertainty in machine learning for chemistry. *Journal of Chemical Information and Modeling* **2023**, *63*, 4012—-4029.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(50) Johnson, M. S. et al. RMG Database for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2022**, *62*, 4906–4915.

(51) Shi, H. Best-first decision tree learning. Ph.D. thesis, The University of Waikato, 2007.

(52) Sumathi, R.; Green, W. H. Thermodynamic properties of ketenes: Group additivity values from quantum chemical calculations. *The Journal of Physical Chemistry A* **2002**, *106*, 7937–7949.

(53) Class, C. A.; Aguilera-Iparraguirre, J.; Green, W. H. A kinetic and thermochemical database for organic sulfur and oxygen compounds. *Physical Chemistry Chemical Physics* **2015**, *17*, 13625–13639.

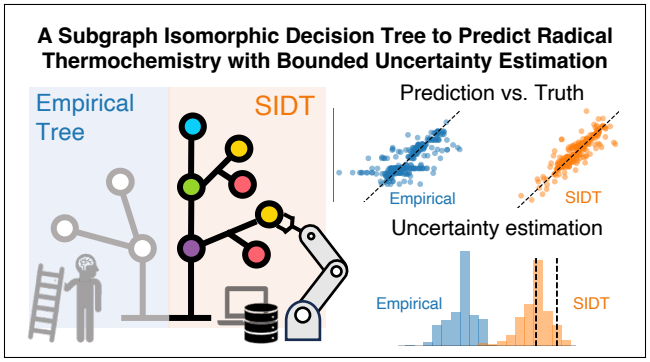
(54) Gillis, R. J.; Green, W. H. Thermochemistry prediction and automatic reaction mechanism generation for oxygenated sulfur systems: A case study of dimethyl sulfide oxidation. *ChemSystemsChem* **2020**, *2*, e1900051.

(55) Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H. Comment on 'Physics-based representations for machine learning properties of chemical reactions'. *Machine Learning: Science and Technology* **2023**, *4*, 048001.

(56) Reuther, A. et al. Interactive supercomputing on 40,000 cores for machine learning and data analysis. 2018 IEEE High Performance Extreme Computing Conference (HPEC). 2018; pp 1–6.

https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

TOC Graphic



https://doi.org/10.26434/chemrxiv-2024-20zjc **ORCID:** https://orcid.org/0000-0001-9381-7500 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0