

# RMG Database for Chemical Property Prediction

Matthew S. Johnson, Xiaorui Dong, Alon Grinberg Dana, Yunsie Chung, David Farina, Jr., Ryan J. Gillis, Mengjie Liu, Nathan W. Yee, Katrin Blondal, Emily Mazeau, Colin A. Grambow, A. Mark Payne, Kevin A. Spiekermann, Hao-Wei Pang, C. Franklin Goldsmith, Richard H. West, and William H. Green\*



Cite This: *J. Chem. Inf. Model.* 2022, 62, 4906–4915



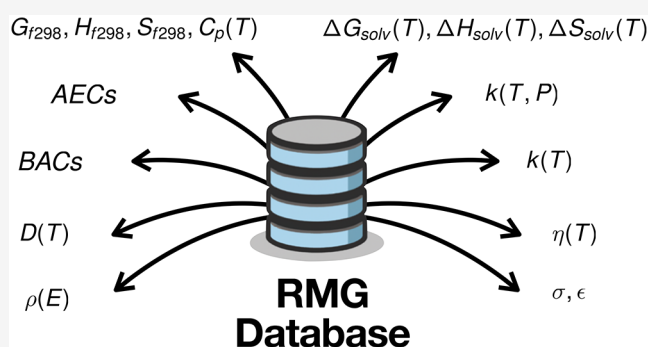
Read Online

ACCESS |

Metrics & More

Article Recommendations

**ABSTRACT:** The Reaction Mechanism Generator (RMG) database for chemical property prediction is presented. The RMG database consists of curated datasets and estimators for accurately predicting the parameters necessary for constructing a wide variety of chemical kinetic mechanisms. These datasets and estimators are mostly published and enable prediction of thermodynamics, kinetics, solvation effects, and transport properties. For thermochemistry prediction, the RMG database contains 45 libraries of thermochemical parameters with a combination of 4564 entries and a group additivity scheme with 9 types of corrections including radical, polycyclic, and surface absorption corrections with 1580 total curated groups and parameters for a graph convolutional neural network trained using transfer learning from a set of >130 000 DFT calculations to 10 000 high-quality values. Correction schemes for solvent–solute effects, important for thermochemistry in the liquid phase, are available. They include tabulated values for 195 pure solvents and 152 common solutes and a group additivity scheme for predicting the properties of arbitrary solutes. For kinetics estimation, the database contains 92 libraries of kinetic parameters containing a combined 21 000 reactions and contains rate rule schemes for 87 reaction classes trained on 8655 curated training reactions. Additional libraries and estimators are available for transport properties. All of this information is easily accessible through the graphical user interface at <https://rmg.mit.edu>. Bulk or on-the-fly use can be facilitated by interfacing directly with the RMG Python package which can be installed from Anaconda. The RMG database provides kineticists with easy access to estimates of the many parameters they need to model and analyze kinetic systems. This helps to speed up and facilitate kinetic analysis by enabling easy hypothesis testing on pathways, by providing parameters for model construction, and by providing checks on kinetic parameters from other sources.



## INTRODUCTION

Understanding and optimization of many important chemical processes such as combustion, polymerization, electrochemistry, pyrolysis, and oxidation can benefit from detailed and predictive chemical kinetic mechanisms. In many of these cases, to accurately represent the chemistry involved, mechanisms need to include hundreds to thousands of species and tens to hundreds of thousands of reactions.

In these cases, we need to assign a rate coefficient to each reaction in the model and often even to many reactions outside the model to determine whether they should remain outside the model. Additionally, a reverse rate needs to be available for most reactions either by explicitly specifying it or by defining the thermochemistry of the species involved. For larger models, the latter is typically preferred: it guarantees thermodynamic consistency, thermochemistry is easier to estimate than kinetics, and there are typically far fewer species in a model than reactions.

The stakes for these estimations are highly variable. Some reactions that are non-limiting and do not impact any important branching can tolerate several orders of magnitude of error, while others may drastically change the chemistry when their rate coefficient is adjusted by as little as 25%. Sensitivity to thermochemical parameters may also vary significantly, although large underestimates of any species' Gibbs free energy will almost always have a drastic effect.

Typically, databases, property estimators, and quantum chemistry methods are used to estimate these parameters. While many advances have been made recently in on-the-fly

**Received:** July 27, 2022

**Published:** October 12, 2022



quantum chemistry calculations, in most cases, these systems are too computationally expensive or not robust enough to use for every parameter in a mechanism. For this reason, we will focus our review on databases and property estimators.

Some of the most popular databases are those archived by NIST containing experimental and quantum chemical rate coefficients and thermodynamic properties.<sup>1</sup> The NIST databases are very extensive, although the data quality can be highly variable. While this makes NIST a great place to search for parameters, it also makes it not ideal for applications where there is no human in the loop. The active thermochemical tables (ATcT) database provides high-accuracy values for a limited number of species.<sup>2</sup> Many databases such as PriMe,<sup>3</sup> ChemKED,<sup>4,5</sup> CloudFlame,<sup>6,7</sup> and ReSpecTh<sup>8</sup> store experimental measurements that may be relevant. The MolSSI QC Archive project is a quantum chemistry-specific database intended to help supply homogeneous datasets.<sup>9</sup>

There are many ways of estimating thermochemical and kinetic parameters. By far, the most common way of estimating thermochemical parameters is the Benson-type group additivity method.<sup>10,11</sup> Implementations of this method are used in THERM and Genesys.<sup>12,13</sup> Graph convolutional neural network methods are beginning to emerge as an alternative.<sup>14–17</sup> Kinetics estimators typically rely on either the use of rate rules assigning specific rates to reactions matching specific templates or reaction group additivity.<sup>18–20</sup> Recent machine learning approaches using hierarchical decision trees have been found to be effective for predicting rate coefficients and reactivity.<sup>21,22</sup>

The RMG database was primarily created to supply the Reaction Mechanism Generator (RMG) software with good estimates for key species and kinetic properties.<sup>23</sup> RMG is a rate-based mechanism generation software. This means that RMG selects species for incorporation in a model based on their computed production rates during simulations. During mechanism generation at each iteration, RMG simulates the set of species and reactions it has already chosen to include in the model, referred to as the “core”. During this simulation, it calculates fluxes through a set of reactions that are under consideration to enter the core, referred to as the “edge”. If the flux toward a species in the edge is high enough, it is added to the core. This means that unlike kinetic models constructed by hand or using rule-based mechanism generation, RMG needs to estimate properties for all the species and reactions in the edge, which are 100× and 10× larger than the core. Furthermore, these estimations have much higher stakes. This is because unlike in other mechanism construction methods, the decisions of whether or not species and reactions are included in the resulting mechanism are directly based on their estimated properties. Poor thermochemistry and kinetic estimation in RMG can therefore cause model generation to miss the real chemical pathways in a kinetic system, highlighting the important role the RMG database has in providing reasonable data estimation schemes.

Accounting for pressure dependence is very important for many gas-phase chemical systems, particularly when the system involves small molecules, low pressures, or high temperatures.<sup>24</sup> As far as the authors are aware, RMG is the only publicly available automatic mechanism generation tool that can automatically handle pressure dependence including chemical activation. It does this by automatically formulating and solving the master equation on the fly using Arkane<sup>25,26,48</sup> for each pressure-dependent reaction network. This requires

the estimation of  $E_0$  and frequencies for each species in the network and  $k(E)$  for each reaction.

To facilitate automated model generation in RMG, the RMG database provides a robust set of estimators for thermochemistry, rate coefficients, and many other relevant properties. It also provides a database of reactions suitable for small-scale machine learning applications and includes many mechanisms and submechanisms from the literature. These estimators, data, and submechanisms have been tempered and refined during the creation of many RMG mechanisms.<sup>27–32</sup>

## THEORY

**Thermo Group Additivity.** RMG's group additivity scheme is broadly based on that developed by Benson et al.<sup>10</sup>  $H_{298}$ ,  $S_{298}$ , and  $C_p(T)$  at a set of temperatures for a given molecule are defined by summing a contribution from each heavy atom in the molecule. In the 20<sup>th</sup> century, most of these group values were derived from the experimental data, but in the last 20 years, a much larger set of group values has been derived from quantum chemical calculations.<sup>33–36</sup> RMG's scheme also integrates several advanced corrections to the ordinary Benson groups:

- radical corrections using the hydrogen bond increment method adopted from Lay et al. 1995.<sup>37</sup>
- ring and polycyclic strain corrections using the method of Han et al. 2018.<sup>38,39</sup>
- long-distance interaction corrections for non-cyclic gauche 1,4 and 1,5 interactions.
- long-distance interaction corrections for halogenated molecules.<sup>36</sup>
- ortho, para, and meta cyclic non-nearest-neighbor interactions using values from Ince et al. 2015 and Ince et al. 2017.<sup>40,41</sup>
- ketene corrections.<sup>33</sup>
- surface adsorption corrections for adsorbed species on nickel and platinum.<sup>42,43</sup>

Each correction is chosen using tree structures such as those shown in Figure 1. Starting from the top of the tree, the algorithm descends to matching child nodes until it finds the most specific correction that matches the actual structure.

Readily available and reliable values for each group are critical to efficient and accurate estimation of thermodynamic parameters using group additivity schemes. Unfortunately,

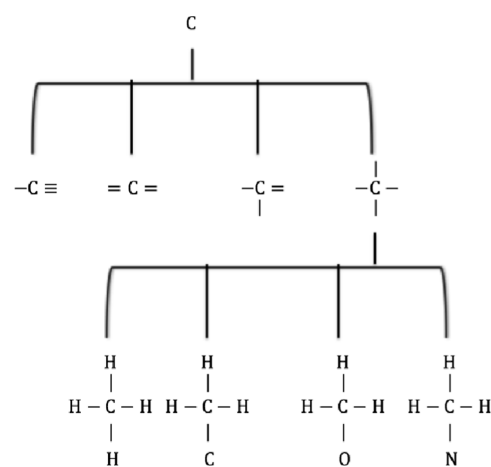


Figure 1. Example segment of a group additivity correction tree.

given the immense diversity of chemical space,<sup>44</sup> it is challenging to construct a scheme that comprehensively covers all chemistries that may be of interest. With this in mind, it is vital to have efficient workflows for continually extending this group additivity scheme. For this purpose, we provide scripts and notebooks for generating new groups from a set of species of interest, proposing new species that share the same chemical motifs to calculate the thermochemistry of a species of interest, and fitting the new groups to the calculated values.

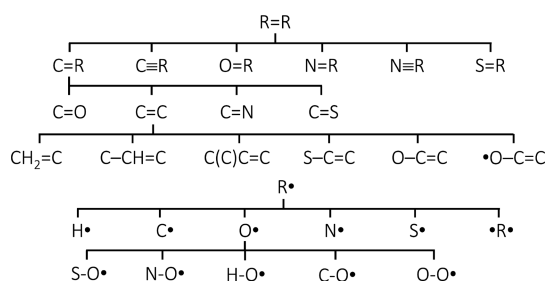
### Thermodynamic Property Neural Network Estimator.

The main thermodynamic property neural network estimator is based on the work from Li et al. 2019 and Grambow et al. 2019.<sup>14,45</sup> It uses a graph convolutional neural network to predict the thermochemistry parameters. The model was trained using transfer learning from 130 000 calculations at the B3LYP/6-31G(2df,p) level to around 10 000 high-quality data points at CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) or better (including accurate experimental values). The low-level calculations were drawn heavily from HCNO species in the QM9 dataset. For  $H_{f298}$  predictions, high-level calculations were selected from the same subset of species plus some molecules with high-accuracy experimental data. The high-level training data for the entropy and heat capacity estimators were a mix of experimental data and 900  $\omega$ B97X-D3/def2-TZVP calculations from the same subset of molecules. The details are available in Grambow et al. 2019.<sup>45</sup>

An additional neural network-based thermochemistry estimator specific for halogenated species was trained off of G4 data from Farina et al. 2021 recently.<sup>17,36,46</sup> While not yet added to the RMG database, this neural network will be available in the near future.

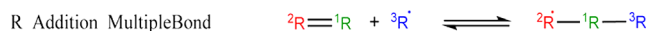
Both the group additivity and neural network approaches give usefully accurate estimates of the thermochemistry of hundreds of thousands of molecules. However, neither is perfectly reliable, and there are no error bars for these estimates. An important future challenge is reliably identifying the uncertainties in the estimates, even for relatively “strange” molecules.

**Rate Coefficient Estimation Rules.** RMG’s rate rule-based estimator uses tree structures similar to those used by group additivity. An example for the “R\_Addition\_MultipleBond” reaction family is shown in Figure 2. The template for



**Figure 2.** Portion of the RMG rate rule trees for the “R\_Addition\_MultipleBond” family, with separate trees for the multiple bond and the attacking radical.

this reaction family is available in Figure 3. Much like traditional rate rule schemes, when a reaction is proposed, it descends down the tree to find the best corresponding rate rule. The rule typically gives modified Arrhenius parameters as  $\log(A)$ ,  $n$ , and  $E_a$  which can be used to compute the desired  $k(T)$ . Note that this is typically a two-dimensional (as shown



**Figure 3.** RMG reaction template for R\_Addition\_MultipleBond family.

in Figure 2) or a three-dimensional tree space. If there is no estimation rule for the rate coefficient where the reaction lands, the algorithm searches the tree towards more specific nodes to find and average the rules closest in Euclidean distance within the tree space.

Unlike traditional rate rule estimators, however, nearly all rules are defined by descending a set of training reactions whose rate parameters are known down the tree and placing those parameters as a rule in the tree. More general rules are then generated by averaging all of the rules one Euclidean distance step below each point in the tree space and so on up the tree. These trees are quite sparsely populated as a result of the combinatorically large tree spaces.

These trees described in this section are designed to estimate the “high-pressure limit” gas-phase rate coefficients  $k(T)$ , the quantities typically computed using transition-state theory from quantum chemical calculations. Due to fall-off and chemical activation, the actual pressure-dependent rate coefficients for unimolecular reactions in the gas phase can be very different. Also, in the liquid phase, the rate coefficients are different due to solvent effects. The way RMG implements the corrections needed to obtain the actual rate coefficients in the environment of interest is discussed in later sections.

**Automatic Tree Generation.** Automated tree generation methods in the RMG database can automatically generate the nodes and rules from the training data using the Subgraph Isomorphic Decision Tree training algorithm developed by Johnson and Green 2021.<sup>21</sup> Many of the current RMG Database reaction families already use these automatically generated trees. Work is in progress on switching all families to use automatically generated trees.

**Vibrational Frequency Estimation for Estimating  $\rho(E)$ .** RMG estimates the vibrational frequencies of each molecule as one step in predicting the density of states,  $\rho(E)$ , for pressure-dependent rate calculations.<sup>25,47</sup> For this purpose, the RMG database contains parameters for a group contribution scheme for estimating the vibrational frequencies of functional groups in a molecule. The details of this scheme are available in Grinberg Dana et al. 2022.<sup>47</sup> Note that these frequencies are only intended for and validated for the estimation of  $\rho(E)$  for pressure-dependent rate calculations within this scheme and are likely not suitable for other purposes. A different scheme, outside of RMG, is available for those who are interested in predicting IR spectra.<sup>48</sup>

**Liquid-Phase Diffusivity Estimation.** In the liquid phase, it is common for some rate coefficients to be diffusion-controlled. To estimate these rate coefficients, one needs estimates of molecular diffusivities. These diffusivities are also needed in liquid-phase reaction-diffusion and reacting flow simulations. The RMG database can estimate the liquid-phase diffusion coefficient  $D$  of a molecule using the Stokes–Einstein equation based on the method of Zhao et al. 2003

$$D = \frac{k_B T}{6\pi\eta r} \quad (1)$$

with

$$r = \left( \frac{0.75V}{\pi N_A} \right)^{1/3} \quad (2)$$

$$V = \sum_i^{N_{\text{atoms}}} (V_i) - 6.56 \times 10^{-6} N_{\text{bonds}} \quad (3)$$

where  $V_i$  is the McGowan volume<sup>49</sup> of a particular atom in  $\text{m}^3/\text{mol}$ ,  $\eta$  is the viscosity of the solvent, and  $N_{\text{bonds}}$  is the number of bonds in the molecule.<sup>50</sup> The database also contains sets of parameters for calculating the pure component viscosity of 150 solvents using the relation

$$\ln(\eta) = A + \frac{B}{T} + C \log(T) + DT^E \quad (4)$$

where  $\eta$  is the viscosity,  $T$  is the temperature, and the rest are solvent-specific viscosity parameters. The majority of the viscosity parameters are obtained from the work by Viswanath et al.,<sup>51</sup> and some are obtained from DIPPR.<sup>52</sup>

**Estimating the Lennard-Jones Parameters.** Gas-phase transport properties are important in many kinetics simulations, and they can be easily calculated by most simulation software from the Lennard-Jones parameters for the associated species. These are calculated by first estimating the critical temperature and pressure using the Joback group additivity method

$$\Sigma_{T_c} = \sum_i \text{GAV}_{T_c,i} \quad (5)$$

$$\Sigma_{P_c} = \sum_i \text{GAV}_{P_c,i} \quad (6)$$

$$T_b = 198.2 + \Sigma_{T_c} \quad (7)$$

$$T_c = \frac{T_b}{0.584 + 0.965 \Sigma_{T_c} - \Sigma_{T_c}^2} \quad (8)$$

$$P_c = \frac{1}{(0.113 + 0.0032 N_{\text{atoms}} - \Sigma_{P_c})^2} \quad (9)$$

where the summations are over the groups associated with the heavy atoms in the molecule,  $T_b$  is the boiling point in K,  $T_c$  is the critical temperature in K,  $P_c$  is the critical pressure in bar, and  $N_{\text{atoms}}$  is the number of atoms in the molecule.<sup>53,54</sup> For halogenated molecules, a boiling point correction derived by Devotta and Pendyala<sup>55</sup> is applied. We are then able to estimate the Lennard-Jones parameters using empirical correlations from Tee et al.

$$\sigma = 2.44 \left( \frac{T_c}{P_c} \right)^{1/3} \quad (10)$$

$$\epsilon = 0.77 k_B T_c \quad (11)$$

where  $\sigma$  and  $\epsilon$  are the Lennard-Jones parameters

$$V(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \quad (12)$$

in angstroms and joules, respectively, where  $k_B$  is the Boltzmann constant.<sup>56</sup> As a fallback, the Lennard-Jones parameters can be estimated simply based on the number of atoms in the molecule.

**Solvent and Solute Property Estimation.** Within the RMG database, liquids are largely assumed to be dilute. The Gibbs free energy of a solute in solution ( $\Delta G_{\text{liq}}^*$ ) is defined using the standard state of a solute dissolved in an ideal solution at a concentration of 1 mol/L and can be calculated from the relationship

$$\Delta G_{\text{liq}}^* = \Delta G_{f,\text{gas}}^o + \Delta G_{\text{solv}}^* + \Delta G^{o \rightarrow *} \quad (13)$$

with

$$\Delta G^{o \rightarrow *} = -RT \ln \frac{V^*}{V^o} = 1.9 \text{ kcal/mol at } 298 \text{ K} \quad (14)$$

where  $\Delta G_{f,\text{gas}}^o$  is derived from gas-phase values with the standard state of an ideal gas at 1 bar,  $\Delta G_{\text{solv}}^*$  is the solvation free energy with the standard state of an ideal gas at a concentration of 1 mol/L dissolving into an ideal solution at a concentration of 1 mol/L,  $V^* = 1 \text{ L/mol}$ , and  $V^o = \frac{RT}{1 \text{ bar}}$ .  $\Delta G^{o \rightarrow *}$  corrects for the difference between the gas-phase and liquid-phase standard states. The partition coefficient for a given solute and solvent can be defined as

$$K = \frac{c_{\text{liq}}}{c_{\text{gas}}} \quad (15)$$

where  $c$  is the concentration of the solute in the appropriate phase at equilibrium. This  $K$  is directly related to  $\Delta G_{\text{solv}}^*$  by

$$\Delta G_{\text{solv}}^* = -RT \ln(K) \quad (16)$$

Thus,  $\Delta G_{\text{solv},298\text{K}}^*$  can be calculated using the Abraham linear solvent energy relationship (LSER)

$$\log_{10}(K_{298\text{K}}) = c + eE + sS + aA + bB + lL \quad (17)$$

where the lower case parameters are Abraham solvent parameters and the upper case parameters are solute parameters.<sup>57</sup> The solvent and solute parameters are independent of each other. The similar Mintz LSER can be used to calculate the enthalpy of solvation at 298 K

$$\Delta H_{\text{solv},298\text{K}}^* (\text{kJ/mol}) = c' + e'E + s'S + a'A + b'B + l'L \quad (18)$$

where the lower case are Mintz solvent parameters and the upper case are the same solute parameters as those used in eq 17.<sup>58</sup> The RMG database can then calculate  $\Delta S_{\text{solv},298\text{K}}^*$  using

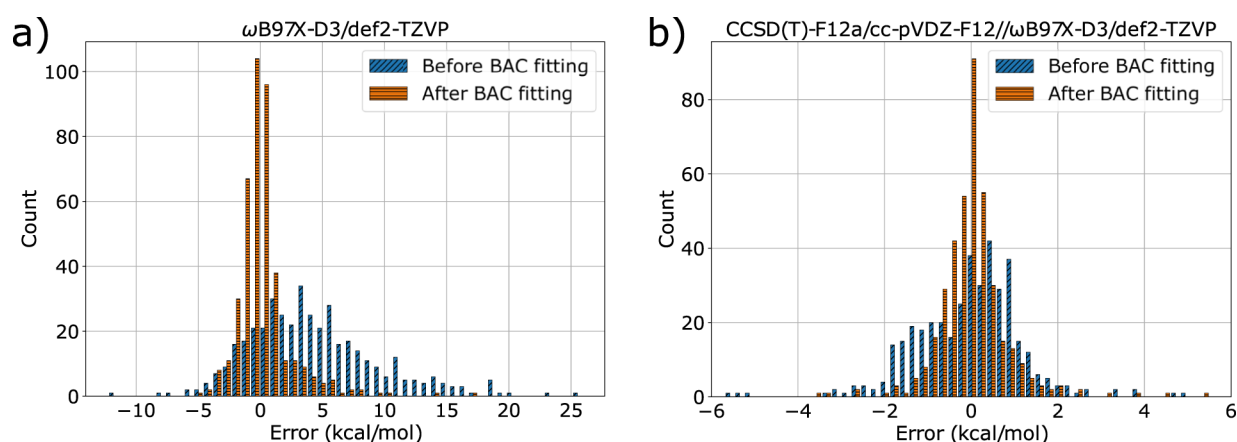
$$\Delta S_{\text{solv},298\text{K}}^* = \frac{\Delta H_{\text{solv},298\text{K}}^* - \Delta G_{\text{solv},298\text{K}}^*}{298\text{K}} \quad (19)$$

and then linearly extrapolate the temperature to estimate  $\Delta G_{\text{solv}}^*$

$$\Delta G_{\text{solv}}^* = \Delta H_{\text{solv},298\text{K}}^* - T \Delta S_{\text{solv},298\text{K}}^* \quad (20)$$

providing an algorithm to calculate  $\Delta G_{\text{solv}}^*$  for solute–solvent pairs with known parameters. More accurate temperature-dependent  $\Delta G_{\text{solv}}^*$  prediction using the method of Chung et al.<sup>59</sup> is available within the RMG database for a limited number of solvents. This method can estimate  $\Delta G_{\text{solv}}^*$  at elevated temperatures along the solvent's saturation pressure based on  $G_{\text{solv},298\text{K}}^*$ ,  $H_{\text{solv},298\text{K}}^*$ , and the solvent's temperature-dependent density and gives a mean absolute error of approximately 0.4





**Figure 4.** Histogram of fitting errors (corrected minus experimental enthalpies) for Petersson-type BACs for (a)  $\omega$ B97X-D3/def2-TZVP and (b) CCSD(T)-F12a/cc-pVDZ-F12// $\omega$ B97X-D3/def2-TZVP. Using a higher level of theory, such as coupled cluster calculations, gives lower errors both before and after applying the fitted BACs.

kcal/mol.<sup>59</sup> The density can be computed for 23 solvents in the RMG database using an open-source package CoolProp.<sup>60</sup>

The Abraham and Mintz solvent parameters are available for 195 and 66 common solvents, respectively, in the RMG database. The solute parameters are calculable for neutral species using the group additivity method from Chung et al. 2022.<sup>61</sup> The group additivity scheme for the solute parameters largely follows that of gas-phase thermochemistry but uses a slightly different approach for halogenated species. If a compound contains any halogens, RMG first computes the solute parameters for a molecule with all the halogens replaced by hydrogen atoms and then adds corrections for each halogen atom. This approach is used because it allows RMG to pull more accurate values from libraries for the dehalogenated species improving the accuracy on the halogenated compounds after correction.

**Forbidden Structures.** The RMG database tabulates the types of species and also specific subtemplates for reaction families that are forbidden. These groups of species are generally either species that RMG's templates may try to create but do not really exist or particular chemistries that RMG cannot represent well yet such as ozonides. The subtemplates are most typically reactions that should not be able to happen or unimportant reactions that RMG severely overestimates. As a result of these forbidden structures, the RMG database may occasionally not predict the existence of certain reactions that a user might expect it to based on the reaction templates. These subtemplates do not consider all forms of non-physical species, which is a topic of on-going research.<sup>62</sup>

**Atom Energy Corrections and Bond Additivity Corrections.** The RMG database includes atom energy corrections (AECs) and bond additivity corrections (BACs). AECs and BACs are empirical corrections to systematic errors in quantum chemistry calculations for specific levels of theory that can be used to improve the accuracy of species thermochemistry calculations. AECs are energy corrections to the atomization energies associated with each atom in a molecule. Most AECs in the RMG database are obtained from fitting to a high-accuracy experimental dataset in the database containing 16 small molecules. The experimental atomization energies come from CCCBDB,<sup>63</sup> and all have uncertainty values less than 0.2 kcal mol<sup>-1</sup>. However, at most levels of theory, AECs alone are insufficient to estimate the enthalpies of formation accurately enough for most kinetics applications.

BACs add an additional correction associated with each bond. They are fit using a dataset of about 400 species with well-known heats of formation, primarily drawn from ATcT<sup>64</sup> and CCCBDB.<sup>63</sup> Figure 4 shows a histogram of the residuals before and after applying BACs; errors are relative to the reliable set of experimental enthalpies used for fitting. After applying BACs, the distribution of errors both centers much more closely to zero and noticeably tightens, representing a significant reduction in both the bias and variance of errors.

The RMG database stores two types of BACs. Petersson-type BACs<sup>65</sup> apply a correction to  $H_{298}$  for each bond type defined by the two atoms and the bond order. Melius-type BACs<sup>66</sup> apply a correction based on bond length integrating information about the 3D geometry. These BAC procedures and the resulting fitted parameters generalize well to new molecules. For example, Grambow et al.<sup>45</sup> applied fitted Melius-type BACs at CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) to a test set of about 400 molecules from NIST. Although the experimental uncertainties were unknown, the resulting RMSE of 1.31 kcal mol<sup>-1</sup> relative to the experimental values is impressive. Similarly, Spiekermann et al.<sup>67</sup> applied fitted Petersson-type BACs at CCSD(T)-F12a/cc-pVDZ-F12// $\omega$ B97X-D3/def2-TZVP to molecules from the Pedley compilation set<sup>68</sup> that have experimental uncertainty of less than 1 kcal mol<sup>-1</sup>. The resulting RMSE of 1.24 kcal mol<sup>-1</sup> demonstrates the accuracy of this approach.

## CONTENT

**Thermodynamic Parameters.** Broadly, RMG's thermochemical parameter estimation has been designed to accurately handle chemistry containing the elements H, C, O, N, and S. Performance on halogens and metal adsorbates is rapidly improving as well. However, H, C, and O molecules are by far the best covered.

**Libraries.** RMG has many libraries of accurate experimental and quantum chemistry calculations for important species. The libraries DFT\_QCI\_thermo, CBS\_QB3\_1dHR, and thermo\_DFT\_CCSDTF12\_BAC are curated calculated libraries of thermochemistry at the particular level of theory. For H<sub>2</sub> combustion chemistry, BurkeH2O2 is recommended, while for C2 chemistry, Klippenstein\_Glarborg2016 is recommended for general purpose use. These libraries cover common C3 and smaller species quite well.<sup>69,70</sup> Many nitrogen species are

available within primaryNS, NitrogenCurran, and NOx2018 libraries.<sup>71,72</sup> Many sulfur species are available in primaryNS, SulfurGlarborgH2S, and SulfurLibrary.<sup>73</sup> Species involving both nitrogen and sulfur are available in primaryNS, and many important aromatic species are available in SABIC\_aromatics, SABIC\_aromatics\_1dHR, and SABIC\_aromatics\_1dHR\_extended. For halogen chemistry, the CHO<sub>X</sub>\_G4 (X = F, Cl, Br, FCl, ClBr, and FClBr) libraries are recommended as they cover a significant portion of their respective chemical spaces with up to four carbons and oxygen atoms.<sup>36</sup> Catalyst adsorbates are available in SurfaceThermoPt111 and SurfaceThermoNi111.<sup>42,43</sup> Many other specialized thermo libraries are also available within the RMG database.

**Group Additivity.** The RMG database's group additivity method for estimating gas-phase thermochemistry performs very well on C/H/O species. With some exceptions for heavily oxygenated species, the overall accuracy is around on par with that of B3LYP DFT calculations. Nitrogen and sulfur species are represented within the groups, and estimates are reasonable, but accuracy is variable. For halogens, the groups perform well for sparsely halogenated non-cyclic species but often underpredict enthalpies for cyclics and more heavily halogenated molecules. For C/H/O cyclics, the mean absolute error of  $H_{f298}$  for small cyclics is about 3 kcal/mol, while it is about 5 kcal/mol for large linear cyclics and 10 kcal/mol for fused cyclics.<sup>38</sup>

**Neural Network.** The neural network thermochemistry estimator performs well on polycyclics and C/H/O/N species. In practice, it is usually superior to the group additivity method on polycyclics and nitrogen species. Sulfur species were not included in the training set. Aliphatic and monocyclic oxygenated and hydrocarbon species tend to be better estimated by RMG's group additivity method.

**Kinetics. Libraries.** The RMG database includes a wide variety of literature mechanisms for different purposes. For general purpose unfitted chemistries, we recommend Primary-H2O2 and Klippenstein\_Glarborg2016 for C/H/O molecules, primaryNitrogenLibrary, and Nitrogen\_Dean\_and\_Bozzelli for nitrogen species and First\_to\_Second\_Aromatic\_Ring for aromatic species. The database also includes the CurranPentane, JetSurF1.0, JetSurF2.0, and FFCM1(−) libraries. The content of each of RMG's kinetic libraries is specified and kept up to date in the RMG documentation online.

**Families.** RMG contains an extendable set of 87 reaction families. Reaction types common to combustion and pyrolysis are very well covered. Common types important for low-temperature chemistry and gas surface chemistry on metal catalysts are covered.

**Training Reactions.** About half of the RMG database families have four or fewer training reactions. The number of training reactions for each of the larger families, with more than 20 training reactions, is available in Table 1. Except for "H\_Abstraction" and "R\_Addition\_MultipleBond", the training reactions for families without loose transition states tend to be mostly CBS-QB3<sup>74</sup> or a similar level of theory calculations. However, there are experimental values and estimates mixed in for most of the families (Table 1). The largest families, "H\_Abstraction" and "R\_Addition\_MultipleBond", have about 834 and 430 reactions, respectively, at CBS-QB3 or similar levels of theory and are supplemented by 2274 and 2521 reactions, respectively, estimated from reaction group

**Table 1. Number of Training Reactions in Families with More Than 20 Training Reactions**

reaction family	number of reactions
H_Abstraction	3107
R_Addition_MultipleBond	2894
Intra_R_Add_Endocyclic	843
intra_H_migration	431
Intra_R_Add_Exocyclic	371
Cl_Abstraction	238
substitutionS	148
R_Recombination	145
disproportionation	137
substitution_O	128
Br_Abstraction	100
F_Abstraction	90
retroene	65
disproportionation-Y	42
Cyclic_Ether_Formation	37
XY_Addition_MultipleBond	33
1,3_Insertion_ROR	22

additivity schemes inherited from RMG's original set of rate rules.

**Transport.** The RMG database contains four literature libraries of transport parameters: GRI-Mech, NOx2018, OneDMinN2, and NIST\_Fluorine and a curated library of species not well accounted for by the groups called primaryTransportLibrary. The Joback groups cover C/H/O/N/S and halogens fairly well.<sup>53,55</sup> In cases where the Joback groups are missing, RMG makes very rough estimates based on the number of atoms in the molecule.

**Solvent and Solute Properties.** RMG's solvent library contains 195 pure solvents with known Abraham solvent parameters, 66 of which have Mintz solvent parameters and 150 of which have viscosity parameters. It also has Abraham solvent parameters for four co-solvents. RMG's solute library holds parameters for 310 common solutes, and the groups for estimating non-radical neutral solute parameters cover elements H, C, O, N, S, P, F, Cl, Br, and I. The solute parameters predicted from the group additivity scheme and the solvent parameters in the library together give a mean absolute error of approximately 0.6 kcal/mol for both solvation free energy and solvation enthalpy estimates at 298 K when evaluated on a 10% test set containing randomly chosen out-of-sample solute compounds.<sup>61</sup> The groups needed to estimate solvation of radicals are a bit more limited and do not include any groups containing S, P, or halogen atoms.

**Atom Energy Corrections and Bond Additivity Corrections.** RMG's database of AECs and BACs covers the elements H, C, O, N, and S quite consistently. Newer and more recently updated corrections also cover halogens. Some AECs in the database that are drawn from the literature cover phosphorous.

## DISCUSSION

**Combining Data from Many Different Sources.** One major challenge when constructing chemical kinetic mechanisms is combining data from many different sources. Carelessly combining parameters from independently accurate mechanisms in many cases can result in an inaccurate mechanism. This occurs primarily due to two different reasons. The first is that many literature mechanism parameters are fit

in bulk to experimental data. Doing this can greatly improve the accuracy on a set of targets for a specific kinetic model. However, these fits are often not unique. In many cases, this results in parameters that are far from their physical values and are strongly correlated with each other. When these correlated parameters are used out of context, or worse — with other independent sources of fitted parameters, they often contribute to significant inaccuracies in constructed mechanisms. Apart from libraries based on popular literature mechanisms and named appropriately, the thermochemistry and kinetic data within the RMG Database is entirely free from fitted parameters. Great care should be taken when combining fitted parameters with any other data sources.

The second challenge is a result of a convenient cancellation of errors that occurs in the thermochemical parameters of chemical kinetic mechanisms. Kinetic mechanism simulations are only dependent on the heats of reaction and Gibbs free energies of reaction, not on the absolute values of individual species thermodynamic properties. Quantum chemical calculations in particular have many correlated errors that cancel out conveniently in these cases. However, when mixed with other thermochemistry data from other sources, these convenient error cancellations can disappear, resulting in inaccurate estimates of reaction thermochemical properties. The RMG database handles this situation by ensuring that all *ab initio* thermochemical parameters in curated libraries are corrected using AECs and either BACs or other cancellation schemes that reference to high-accuracy experimental values. All of RMG's thermochemical estimators are trained on corrected *ab initio* or experimental data.

**What Is the RMG Database Good for Doing?** The RMG database has predictors for many properties as discussed above, but it is worth keeping in mind that the RMG database has been primarily developed to improve the accuracy of predictive chemical kinetic mechanisms produced by the RMG software. This means that one should expect the RMG database to be particularly good at predicting reaction rate coefficients and species thermochemistry and any properties they are sensitive to. In general, the existence of a published, experimentally-validated RMG mechanism involving a specific chemistry means some care has been taken inside the RMG database with regard to those chemistries and that at least the most relevant corrections for those reaction conditions are present. For example, Grinberg Dana et al. 2018 demonstrated RMG's capabilities on nitrogen compounds, Class et al. 2019 demonstrated RMG's capabilities on some sulfur chemistries, and Chu et al. 2019 demonstrated RMG's capabilities for small aromatic species.<sup>28,30,75</sup> However, the user should keep in mind that all large databases, including the RMG database, likely include some erroneous values — perfection is hard to achieve, and it is challenging to check tens of thousands of numbers.

The RMG database has been repeatedly shown to be sufficiently accurate and that it can be used to build satisfactory high temperature (>650 K) chemical kinetic models on non-aromatic C/H/O chemistries for combustion and pyrolysis. It can also handle up to two ring aromatic chemistries under pyrolysis conditions. In the 300–650 K range, rate predictions tend to be less accurate due to higher sensitivity to the errors in estimates of activation barriers and enthalpies. The lower accuracy makes prediction of product yields and overall reaction rates much less robust, although successful examples exist.<sup>76,77</sup> Most small nitrogen and sulfur functional groups are

covered in the database. RMG can handle their small-molecule chemistries quite well, including some nitrogen–sulfur interactions. However, for more complex heteromolecules, there are gaps in the rate rules and thermodynamic groups.

RMG's methods for estimating solvation energies for small neutral molecules have been well tested against the experimental data and are particularly good for common solvents whose properties are well known. RMG's ability to predict solvent effects on rate coefficients, transport properties, and so forth has not been very extensively tested yet. At present, RMG cannot handle ions.

There are some successful examples of RMG building satisfactory kinetic models for reactions over metal catalysts, implying that the RMG estimates are accurate enough for those cases.<sup>78</sup> However, only a few systems have been studied so far, and it will be some time before RMG can reliably estimate the parameters needed to accurately model catalysis over all metals of interest.

**How Can One Efficiently Use the RMG Database?** The easiest way to access information from the RMG database is the RMG website, located at <https://rmg.mit.edu/>. The web pages provide a graphic user interface (GUI) for browsing through the database entries and searching for specific properties of a molecule or reaction of interest, displaying data in both numerical and corresponding graphical form. The “Molecule Search” tool enables users to search the database for all thermochemistry and transport sources for a given species, as well as exploring the resonance structures thereof.<sup>75</sup> The “Solvation Tool” does the same for solvent and solvation parameters for given solute solvent pairs. The “Kinetic Search” tool allows users to search the database for reactions based on just reactants or reactants and products. It is worth noting that in many cases, in addition to the available library data, the website will return the rate coefficient estimates from both rate rules and group additivity. The latter are generated using a reaction group additivity scheme based on the reaction family kinetic tree and training reactions. The group additivity scheme for estimating reaction rate coefficients (not the one for thermodynamic properties) is no longer maintained as it was found to be appreciably less accurate than rate rules.<sup>20</sup>

When estimates for many species or reactions are required, it becomes much more practical to directly interface with the RMG software. The RMG Python package can be easily installed from Anaconda and can be used to generate bulk or on-the-fly estimates from the RMG database.

## CONCLUSIONS

We have presented the RMG database for chemical property prediction. The RMG database is an ever-expanding database designed for estimating properties important to chemical kinetics. Over the years, it has been expanded and curated to accurately predict properties for many chemical kinetic mechanisms.<sup>28,30,32,42,79,80</sup> Currently, the database is a very solid foundation for building gas-phase mechanisms involving H/C/O/N/S/F/Cl/Br species. Property prediction in the liquid and catalyst phases is state of the art and advancing rapidly. This suite includes estimators for thermochemistry including solvent and adsorption corrections, kinetics for 87 different classes of reactions, solvent viscosity, and species diffusivity in both gas and liquid phases. Every addition to the RMG database is tested in an automated workflow to ensure that estimator performance is maintained or improved as the RMG database is expanded. While there are many databases



available today, the RMG database's focus on chemical mechanism generation makes it uniquely suitable for providing properties for the construction of chemical kinetic mechanisms.

## ■ DATA AND SOFTWARE AVAILABILITY

Most of the data presented is easily accessible through graphical user interfaces provided through the RMG website at <https://rmg.mit.edu>. The database in its entirety is available on GitHub at <https://github.com/ReactionMechanismGenerator/RMG-database>. The software used to access and manipulate the database is also available on Github in a separate repository at <https://github.com/ReactionMechanismGenerator/RMG-Py>. These two packages can be easily installed using Anaconda from the "rmg" channel.

## ■ AUTHOR INFORMATION

### Corresponding Author

**William H. Green** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0003-2603-9694](https://orcid.org/0000-0003-2603-9694); Email: [whgreen@mit.edu](mailto:whgreen@mit.edu)

### Authors

**Matthew S. Johnson** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0002-4624-2852](https://orcid.org/0000-0002-4624-2852)

**Xiaorui Dong** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0003-0144-5696](https://orcid.org/0000-0003-0144-5696)

**Alon Grinberg Dana** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; The Wolfson Department of Chemical Engineering, Grand Technion Energy Program (GTEP), Technion—Israel Institute of Technology, Haifa 3200003, Israel; [orcid.org/0000-0001-7545-8719](https://orcid.org/0000-0001-7545-8719)

**Yunsie Chung** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0002-3097-010X](https://orcid.org/0000-0002-3097-010X)

**David Farina, Jr.** – Department of Chemical Engineering, Northeastern University, Boston, Massachusetts 02115, United States; [orcid.org/0000-0002-8670-6121](https://orcid.org/0000-0002-8670-6121)

**Ryan J. Gillis** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0001-9261-6811](https://orcid.org/0000-0001-9261-6811)

**Mengjie Liu** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0003-2414-1986](https://orcid.org/0000-0003-2414-1986)

**Nathan W. Yee** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

**Katrin Blondal** – School of Engineering, Brown University, Providence, Rhode Island 02912, United States; [orcid.org/0000-0002-0964-8589](https://orcid.org/0000-0002-0964-8589)

**Emily Mazeau** – Department of Chemical Engineering, Northeastern University, Boston, Massachusetts 02115, United States; [orcid.org/0000-0001-8844-9563](https://orcid.org/0000-0001-8844-9563)

**Colin A. Grambow** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

**A. Mark Payne** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0002-0180-1796](https://orcid.org/0000-0002-0180-1796)

**Kevin A. Spiekermann** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0002-9484-9253](https://orcid.org/0000-0002-9484-9253)

**Hao-Wei Pang** – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [orcid.org/0000-0001-9381-7500](https://orcid.org/0000-0001-9381-7500)

**C. Franklin Goldsmith** – School of Engineering, Brown University, Providence, Rhode Island 02912, United States; [orcid.org/0000-0002-2212-0172](https://orcid.org/0000-0002-2212-0172)

**Richard H. West** – Department of Chemical Engineering, Northeastern University, Boston, Massachusetts 02115, United States; [orcid.org/0000-0003-3861-6030](https://orcid.org/0000-0003-3861-6030)

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jcim.2c00965>

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

Funding from the Gas Phase Chemical Physics Program of the US Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences (under award number DESC0014901), is appreciated. The work was also supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, through the Exascale Catalytic Chemistry (ECC) Project as part of the Computational Chemical Sciences Program, and by the National Science Foundation under grant no. 1751720. A.G.D. was supported by The George J. Elbaum Scholarship in Engineering, The Ed Satell Foundation, and The Zuckerman STEM Leadership Program. H.P. was supported by the Think Global Education Trust Scholarship.

## ■ REFERENCES

- (1) Manion, J. A.; Huie, R. E.; Levin, R. D.; Burgess, D. R. B., Jr.; Orkin, V. L.; Tsang, W.; McGivern, W. S.; Hudgens, J. W.; Knyazev, V. D.; Atkinson, D. B.; Chai, E.; Tereza, A. M.; Lin, C.-Y.; Allison, T. C.; Mallard, W. G.; Westley, F.; Herron, J. T.; Hampson, R. F.; Frizzell, D. H. *NIST Chemical Kinetics Database*, NIST Standard Reference Database. <https://kinetics.nist.gov/kinetics/welcome.jsp>. Accessed: October 2022.
- (2) Ruscic, B.; Pinzon, R. E.; Von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoy, D.; Wagner, A. F. Active Thermochemical Tables: Thermochemistry for the 21st century. *J. Phys.: Conf. Ser.* **2005**, 16, 078.
- (3) PrImE Kinetics. [https://kinetics.nist.gov/RealFuels/maccr/maccr2011/MACCCR\\_2011\\_Frenklach.pdf](https://kinetics.nist.gov/RealFuels/maccr/maccr2011/MACCCR_2011_Frenklach.pdf).
- (4) ChemKED Database. <https://github.com/pr-omethe-us/ChemKED-database>. Accessed: October 2022.
- (5) Weber, B. W.; Niemeyer, K. E. ChemKED: A Human- and Machine-Readable Data Standard for Chemical Kinetics Experiments. *Int. J. Chem. Kinet.* **2018**, 50, 135–148.
- (6) CloudFlame. <https://cloudflame.kaust.edu.sa/>. Accessed: October 2022.



- (7) Goteng, G. L.; Nettyam, N.; Sarathy, S. M. CloudFlame: Cyberinfrastructure for combustion research. *Proceedings - 2013 International Conference on Information Science and Cloud Computing Companion*, 2014; pp 294–299 (ISCC-C 2013).
- (8) Turányi, T.; Zsély, I. G.; Varga, T.; Olm, C.; Nagy, T.; Pálvölgyi, R.; Valkó, E.; Busai, Á. ReSpecTh: a joint reaction kinetics, spectroscopy, and thermochemistry information system. In *Proceedings of the 7th European Combustion Meeting*; Budapest, Hungary, 2015, 30, 1–5.
- (9) Smith, D. G. A.; Altarawy, D.; Burns, L. A.; Welborn, M.; Naden, L. N.; Ward, L.; Ellis, S.; Pritchard, B. P.; Crawford, T. D. The MolSSI QCArchive project: An open-source platform to compute, organize, and share quantum chemistry data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *11*, 1491.
- (10) Benson, S. W.; Cruickshank, D. M.; Golden, G. R.; Haugen, R.; O'Neal, F. R.; Rodgers, A. S.; Shaw, H. E.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279–324.
- (11) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. Revised group additivity parameters for the enthalpies of formation of oxygen-containing organic compounds. *J. Phys. Chem.* **1973**, *77*, 1687–1691.
- (12) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M. F.; Marin, G. B. Genesys: Kinetic model construction using chemo-informatics. *Chem. Eng. J.* **2012**, *207–208*, 526–538.
- (13) Ritter, E. R.; Bozzelli, J. W. THERM: Thermodynamic property estimation for gas phase radicals and molecules. *Int. J. Chem. Kinet.* **1991**, *23*, 767–778.
- (14) Li, Y. P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.
- (15) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.
- (16) Han, K. Enabling Automatic Generation of Accurate Kinetic Models for Complicated Chemical Systems, Ph.D. thesis; Massachusetts Institute of Technology, 2018.
- (17) Sirumalla, S. K. Graph neural networks and high throughput quantum chemistry workflows for detailed kinetic modeling, Ph.D. thesis; Northeastern University Boston, Massachusetts, 2021.
- (18) Van de Vijver, R.; Sabbe, M. K.; Reyniers, M. F.; Van Geem, K. M.; Marin, G. B. Ab initio derived group additivity model for intramolecular hydrogen abstraction reactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 10877–10894.
- (19) West, R. H.; Allen, J. W.; Green, W. H. Automatic Reaction Mechanism Generation with Group Additive Kinetics. In *Proceedings of the AIChE 2011 Annual Meeting*; American Institute of Chemical Engineers: Minneapolis, MN, 2011; October 16–21. <https://doi.org/10.1002/chin.201236258>.
- (20) Allen, J. W. *Predictive Chemical Kinetics: Enabling Automatic Mechanism Generation and Evaluation*. Doctoral dissertation, Massachusetts Institute of Technology, 2013, 59–64.
- (21) Johnson, M. S.; Green, W. H. A Machine Learning Based Approach to Reaction Rate Estimation. *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-c98gc.
- (22) Heid, E.; Goldman, S.; Sankaranarayanan, K.; Coley, C. W.; Flamm, C.; Green, W. H. EHreact: Extended Hasse Diagrams for the Extraction and Scoring of Enzymatic Reaction Templates. *J. Chem. Inf. Model.* **2021**, *61*, 4949–4961.
- (23) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.
- (24) Wong, B. M.; Matheu, D. M.; Green, W. H. Temperature and molecular size dependence of the high-pressure limit. *J. Phys. Chem. A* **2003**, *107*, 6206–6211.
- (25) Allen, J. W.; Goldsmith, C. F.; Green, W. H. Automatic estimation of pressure-dependent rate coefficients. *Phys. Chem. Chem. Phys.* **2012**, *14*, 1131–1155.
- (26) Johnson, M. S.; Green, W. H. Examining the Accuracy of Methods for Obtaining Pressure Dependent Rate Coefficients. *Faraday Discuss.* **2022**, DOI: 10.1039/d2fd00040g.
- (27) Class, C. A.; Liu, M.; Vandeputte, A. G.; Green, W. H. Automatic mechanism generation for pyrolysis of di-tert-butyl sulfide. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21651–21658.
- (28) Class, C. A.; Vasilou, A. K.; Kida, Y.; Timko, M. T.; Green, W. H. Detailed kinetic model for hexyl sulfide pyrolysis and its desulfurization by supercritical water. *Phys. Chem. Chem. Phys.* **2019**, *21*, 10311–10324.
- (29) Vandewiele, N. M.; Magoon, G. R.; Van Geem, K. M.; Reyniers, M. F.; Green, W. H.; Marin, G. B. Kinetic Modeling of Jet Propellant-10 Pyrolysis. *Energy Fuels* **2015**, *29*, 413–427.
- (30) Chu, T. C.; Buras, Z. J.; Oßwald, P.; Liu, M.; Goldman, M. J.; Green, W. H. Modeling of aromatics formation in fuel-rich methane oxy-combustion with an automatically generated pressure-dependent mechanism. *Phys. Chem. Chem. Phys.* **2019**, *21*, 813–832.
- (31) 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*, No. e1900051.
- (32) Johnson, M. S.; Nimlos, M. R.; Ninnemann, E.; Laich, A.; Fioroni, G. M.; Kang, D.; Bu, L.; Ranasinghe, D.; Khanniche, S.; Goldsborough, S. S.; Vasu, S. S.; Green, W. H. Oxidation and pyrolysis of methyl propyl ether. *Int. J. Chem. Kinet.* **2021**, *53*, 915–938.
- (33) Sumathi, R.; Green, W. H. Thermodynamic properties of ketenes: Group additivity values from quantum chemical calculations. *J. Phys. Chem. A* **2002**, *106*, 7937–7949.
- (34) Sumathi, R.; Green, W. H. Missing thermochemical groups for large unsaturated hydrocarbons: Contrasting predictions of G2 and CBS-Q. *J. Phys. Chem. A* **2002**, *106*, 11141–11149.
- (35) Sumathi, R.; Green, W. H., Jr. Oxygenate, oxyalkyl and alkoxy carbonyl thermochemistry and rates for hydrogen abstraction from oxygenates. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3402–3417.
- (36) Farina, D. S.; Sirumalla, S. K.; Mazeau, E. J.; West, R. H. Extensive High-Accuracy Thermochemistry and Group Additivity Values for Halocarbon Combustion Modeling. *Ind. Eng. Chem. Res.* **2021**, *60*, 15492–15501.
- (37) 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. *J. Phys. Chem.* **1995**, *99*, 14514–14527.
- (38) 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.
- (39) Lai, L.; Khanniche, S.; Green, W. H. Thermochemistry and Group Additivity Values for Fused Two-Ring Species and Radicals. *J. Phys. Chem. A* **2019**, *123*, 3418–3428.
- (40) 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.
- (41) Ince, A.; Carstensen, H.; Sabbe, M.; Reyniers, M.; Marin, G. B. Group additive modeling of substituent effects in monocyclic aromatic hydrocarbon radicals. *AIChE J.* **2017**, *63*, 2089–2106.
- (42) Blondal, K.; Jelic, J.; Mazeau, E.; Studt, F.; West, R. H.; Goldsmith, C. F. Computer-Generated Kinetics for Coupled Heterogeneous/Homogeneous Systems: A Case Study in Catalytic Combustion of Methane on Platinum. *Ind. Eng. Chem. Res.* **2019**, *58*, 17682–17691.
- (43) Goldsmith, C. F.; West, R. H. Automatic Generation of Microkinetic Mechanisms for Heterogeneous Catalysis. *J. Phys. Chem. C* **2017**, *121*, 9970–9981.
- (44) Reymond, J.-L.; Ruddigkeit, L.; Blum, L.; van Deursen, R. The enumeration of chemical space. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 717–733.
- (45) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.
- (46) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

- (47) Grinberg Dana, A.; Johnson, M.; Allen, J.; Sharma, S.; Raman, S.; Liu, M.; Gao, C.; Grambow, C.; Goldman, M.; Ranasinghe, D.; Gillis, R.; Payne, A. M.; Li, Y.-P.; Dames, E.; Buras, Z.; Vandewiele, N.; Yee, N.; Merchant, S.; Buesser, B.; Class, C.; Goldsmith, F.; West, R.; Green, W. Automated Reaction Kinetics and Network Exploration (Arkane): A Statistical Mechanics, Thermodynamics, Transition State Theory, and Master Equation Software, **2022**. DOI: [10.26434/chemrxiv-2022-4klsm](https://doi.org/10.26434/chemrxiv-2022-4klsm)
- (48) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. H. Predicting Infrared Spectra with Message Passing Neural Networks. *J. Chem. Inf. Model.* **2021**, *61*, 2594–2609.
- (49) Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* **1987**, *23*, 243–246.
- (50) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Determination of McGowan Volumes for Ions and Correlation with van der Waals Volumes. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1848–1854.
- (51) Viswanath, D. S.; Ghosh, T. K.; Prasad, D. H.; Dutt, N. V.; Rani, K. Y. *Viscosity of Liquids: Theory, Estimation, Experiment, and Data*; Springer Netherlands, 2007; pp 1–660.
- (52) Bloxham, J. C.; Redd, M. E.; Giles, N. F.; Knotts, T. A.; Wilding, W. V. Proper Use of the DIPPR 801 Database for Creation of Models, Methods, and Processes. *J. Chem. Eng. Data* **2021**, *66*, 3–10.
- (53) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun.* **1987**, *57*, 233–243.
- (54) Joback, K. G. A unified approach to physical property estimation using multivariate statistical techniques, Ph.D. Thesis, Massachusetts Institute of Technology, 1984.
- (55) Devotta, S.; Pendyala, V. R. Modified Joback Group Contribution Method for Normal Boiling Point of Aliphatic Halogenated Compounds. *Ind. Eng. Chem. Res.* **1992**, *31*, 2042–2046.
- (56) Tee, L. S.; Gotoh, S.; Stewart, W. E. Molecular Parameters for Normal Fluids. Lennard-Jones 12-6 Potential. *Ind. Eng. Chem. Fundam.* **1966**, *5*, 356–363.
- (57) Abraham, M. H.; Platts, J. A.; Hersey, A.; Leo, A. J.; Taft, R. W. Correlation and estimation of gas-chloroform and water-chloroform partition coefficients by a linear free energy relationship method. *J. Pharm. Sci.* **1999**, *88*, 670–679.
- (58) Mintz, C.; Clark, M.; Acree, W. E.; Abraham, M. H. Enthalpy of Solvation Correlations for Gaseous Solutes Dissolved in Water and in 1-Octanol Based on the Abraham Model. *J. Chem. Inf. Model.* **2007**, *47*, 115–121.
- (59) Chung, Y.; Gillis, R. J.; Green, W. H. Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data. *AIChE J.* **2020**, *66*, No. e16976.
- (60) Bell, I. H.; Wronski, J.; Quoilin, S.; Lemort, V. Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library coolprop. *Ind. Eng. Chem. Res.* **2014**, *53*, 2498–2508.
- (61) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446.
- (62) Mitnik, N.; Haba, S.; Grinberg Dana, A. Non-physical Species in Chemical Kinetic Models: A Case Study of Diazenyl Hydroxy and Diazenyl Peroxide. *ChemPhysChem* **2022**, DOI: [10.1002/cphc.202200373](https://doi.org/10.1002/cphc.202200373).
- (63) Johnson, R. D., II. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, 2020; <http://cccbdb.nist.gov/>.
- (64) Ruscic, B.; Bross, D. *Active Thermochemical Tables (ATcT) Values Based on Ver. 1.122d of the Thermochemical*, 2018; <https://atct.anl.gov/Thermochemical%20Data/version%201.122d/index.php>.
- (65) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A., Jr; Frisch, M. J. Calibration and comparison of the Gaussian-2, complete basis set, and density functional methods for computational thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.
- (66) Anantharaman, B.; Melius, C. F. Bond additivity corrections for G3B3 and G3MP2B3 quantum chemistry methods. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.
- (67) Spiekermann, K.; Pattanaik, L.; Green, W. H. *High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions*; Nature Publishing Group, 2022.
- (68) Pedley, J. *Thermochemical Data and Structures of Organic Compounds*; CRC Press, 1994; Vol. 1.
- (69) Burke, M. P.; Chaos, M.; Ju, Y.; Dryer, F. L.; Klippenstein, S. J. Comprehensive H<sub>2</sub>/O<sub>2</sub> kinetic model for high-pressure combustion. *Int. J. Chem. Kinet.* **2012**, *44*, 444–474.
- (70) Hashemi, H.; Christensen, J. M.; Gersen, S.; Levinsky, H.; Klippenstein, S. J.; Glarborg, P. High-pressure oxidation of methane. *Combust. Flame* **2016**, *172*, 349–364.
- (71) Bugler, J.; Somers, K. P.; Simmie, J. M.; Güthe, F.; Curran, H. J. Modeling Nitrogen Species as Pollutants: Thermochemical Influences. *J. Phys. Chem. A* **2016**, *120*, 7192–7197.
- (72) Glarborg, P.; Miller, J. A.; Ruscic, B.; Klippenstein, S. J. Modeling nitrogen chemistry in combustion. *Prog. Energy Combust. Sci.* **2018**, *67*, 31–68.
- (73) Song, Y.; Hashemi, H.; Christensen, J. M.; Zou, C.; Haynes, B. S.; Marshall, P.; Glarborg, P. An Exploratory Flow Reactor Study of H<sub>2</sub>S Oxidation at 30–100 Bar. *Int. J. Chem. Kinet.* **2017**, *49*, 37–52.
- (74) Wood, G. P. F.; Radom, L.; Petersson, G. A.; Barnes, E. C.; Frisch, M. J.; Montgomery, J. A. A restricted-open-shell complete-basis-set model chemistry. *J. Chem. Phys.* **2006**, *125*, 094106.
- (75) Grinberg Dana, A.; Liu, M.; Green, W. H. Automated chemical resonance generation and structure filtration for kinetic modeling. *Int. J. Chem. Kinet.* **2019**, *51*, 760–776.
- (76) Grinberg Dana, A.; Wu, H.; Ranasinghe, D. S.; Pickard, F. C.; Wood, G. P. F.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation: (1) The AIBN/H<sub>2</sub>O/CH<sub>3</sub>OH Radical “Soup”. *Mol. Pharmaceutics* **2021**, *18*, 3037–3049.
- (77) Wu, H.; Grinberg Dana, A.; Ranasinghe, D. S.; Pickard, F. C.; Wood, G. P. F.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation: (2) Imipramine Stress Testing. *Mol. Pharmaceutics* **2022**, *19*, 1526–1539.
- (78) Mazeau, E. J.; Satpute, P.; Blöndal, K.; Goldsmith, C. F.; West, R. H. Automated Mechanism Generation Using Linear Scaling Relationships and Sensitivity Analyses Applied to Catalytic Partial Oxidation of Methane. *ACS Catal.* **2021**, *11*, 7114–7125.
- (79) Class, C. A.; Liu, M.; Vandeputte, A. G.; Green, W. H. Automatic mechanism generation for pyrolysis of di-tert-butyl sulfide. *Phys. Chem. Chem. Phys.* **2016**, *18*. DOI: [10.1039/c6cp02202b](https://doi.org/10.1039/c6cp02202b)
- (80) Dames, E. E.; Rosen, A. S.; Weber, B. W.; Gao, C. W.; Sung, C. J.; Green, W. H. A detailed combined experimental and theoretical study on dimethyl ether/propane blended oxidation. *Combust. Flame* **2016**, *168*, 310–330.