Perspective on Automated Predictive Kinetics using Estimates derived from Large Datasets

William H. Green, MIT Dept. of Chemical Engineering, Cambridge, MA 02139

Abstract:

A longstanding project of the chemical kinetics community is to predict reaction rates and the behavior of reacting systems, even for systems where there are no experimental data. Many important reacting systems (atmosphere, combustion, pyrolysis, partial oxidations) involve a large number of reactions occurring simultaneously, and reaction intermediates that have never been observed, making this goal even more challenging. Improvements in our ability to compute rate coefficients and other important parameters accurately from first principles, and improvements in automated kinetic modeling software, have partially overcome many challenges. Indeed, in some cases quite complicated kinetic models have been constructed which accurately predicted the results of independent experiments. However, the process of constructing the models, and deciding which reactions to measure or compute *ab initio*, relies on accurate estimates (and indeed most of the numerical rate parameters in most large kinetic models are estimates.) Machine-learned models trained on large datasets can improve the accuracy of these estimates, and allow a better integration of quantum chemistry and experimental data. The need for continued development of shared (perhaps open-source) software and databases, and some directions for improvement, are highlighted. As we model more complicated systems, many of the weaknesses of the traditional ways of doing chemical kinetic modeling, and of testing kinetic models, have been exposed, identifying several challenges for future research by the community.

Key Words: Modeling, Machine-Learning, Automation, Datasets

Introduction

Historically, chemical processes have been developed using iterative experimentation: (1) discover a new reaction, (2) fit it into a laboratory process including separations/purifications and usually several other reactions to accomplish the overall transformation of interest, (3) then conduct several iterations of scale-up and optimization of reaction conditions, culminating in a pilot plant campaign. A kinetic model is fitted to the data measured in the pilot plant, and that model is used to make the predictions

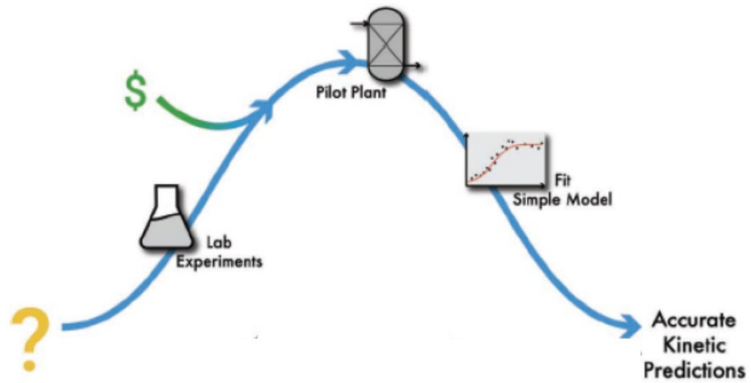

Fig. 1. Schematic of conventional chemical process development workflow.

needed to design the full industrial scale unit needed to commercialize the process, Fig. 1. The full-scale unit is then modestly tweaked to fine-tune its performance.

This conventional procedure works, but has several drawbacks. First, it is expensive: the pilot plant experiments typically cost several million dollars, and this high cost significantly discourages innovations that involve introducing new chemical reactions. Since the pilot plant experiments are expensive, only a limited number are conducted, and so the final process model can only involve a small number of adjustable parameters tuned to match the limited data. Typically this is done by using a reduced simplistic model that only contains the major species, rather than a realistic kinetic model that includes the reactive intermediates. Also, this procedure locks in on particular process design and reaction conditions, since it is usually not practical to consider a wide range of possible processes or reaction conditions experimentally. So in the end, the process that is modeled and built to full industrial scale is typically a local optimum in the limited region of process space tested, not the best possible process. And the simplistic model we have for that process, while usually accurate for interpolation between the reaction conditions of the pilot plan experiments, is usually unreliable in extrapolation. Often the simplistic model cannot represent minor byproducts, or any of several phenomena that can cause upsets in a large reactor.

Could we do things differently? Is there some way to accurately predict the process kinetics without relying on pilot plant experiments? Could we predict the efficiency and rates of several different processes to achieve the same goals on the computer, to more efficiently hone in on the best process? Could we build a more accurate model, that includes the minor by-products and the physics underlying different upsets?

There are at least two distinct issues when trying to make these predictions: (1) Predicting which species and reactions are important in the process, and (2) Predicting the numerical values of all the rate coefficients, equilibrium constants, and other parameters needed for kinetic simulations.

This paper provides my current perspective on these two challenging issues, and how they both depend on our ability to accurately **estimate** rate coefficients and related quantities. In this paper I particularly focus on recent attempts to use machine-learning methods to provide the numerical estimates, which potentially could change how estimation is done. A large number of very talented researchers are currently working on these problems, and many more have worked on them in the past. There are volumes devoted to parts of this problem [1,2] and many reviews of different aspects of the problem.[3,4,5,6]. Here I do not attempt a comprehensive review of this big research area, but instead focus on some specific aspects of these big problems that I am most familiar with, mostly due to recent work done by my collaborators.

Which Species and Reactions are Important?

The first question, “which reactions are occurring, making which species?”, is important in many fields of chemistry, not just kinetics. For example, if one wants to synthesize a molecule, one needs to be able to predict what the major products from combining a set of reactants, reagents, and catalysts at certain reaction conditions will be.[7] But the question is particularly important in kinetics, since to build a kinetic model we need to predict not just the major products, but also the important reactive intermediates and reactions, and the important minor byproducts (e.g. those that are toxic, or which

cannot be allowed to contaminate the final product). And ideally we would like each reaction to be elementary, so we can predict how its rate depends on concentrations.

The definition of 'important' depends on the needs of the modeler: which species can be safely ignored? What is the modeler's error tolerance? It also depends on the reaction conditions: ethanol changes from an inert solvent into a major reactant at high temperatures (e.g. in the process for making ethylene from biomass). Often there is a core set of species and reactions that are currently considered to be important, and then a much larger set of other species and reactions which could conceivably be important, as illustrated in Fig. 2 for the case of propene pyrolysis. In Fig. 2 the black species and solid-arrow reactions are currently thought to be important at the reaction conditions of interest, while it is currently unknown if the dashed reactions and gray species are important. And of course there are thousands of other reactions and species that could conceivably be important, but which are not currently under consideration, these are not shown in the figure. A systematic method is needed, both to propose candidate reactions and species, and to decide which are actually important.

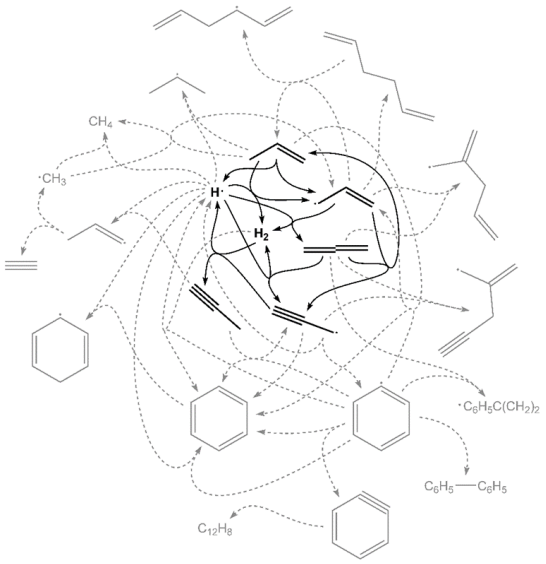

Fig. 2 Some species and reactions that likely (dark) or may (gray) be important in pyrolysis of propene.

Several procedures have been proposed to decide which species and reactions should be included in a kinetic simulation. One approach is to choose important reaction sequences by analogy to kinetic models for other systems. For example, if one has a model for the pre-ignition chemistry of one alkane, one can reasonably predict the important reactions in the pre-ignition chemistry of other alkanes [8]. This procedure of modeling-by-analogy is used by many expert kineticists as they begin to model a new system, and has been codified in some software packages that automatically generate kinetic models. [9,10] Another idea is to assume low conversion of the reactant, so its reactions will dominate, and subsequent reactions of the initial products will be less important. This approach is used in software packages such as NetGen [11], and is a good approach for modeling systems where the initial build-up of small concentrations of by-products is harmful, e.g. shelf-stability of drugs.[12]

Yet another concept is to systematically generate every possible reaction from the dark species in Fig. 2, and then numerically test all the gray reactions, pruning away reactions which are calculated to be too slow to be important (as judged by some error tolerance). This "rate-based" approach is used in the Reaction Mechanism Generator (RMG).[13,14,15] The rate-based approach has the advantage of being more general than methods that assume knowledge of analogous system, or low conversion, and it has been impressively successful in a handful of cases.[16,17,18,19,20] However, it has the disadvantage of being slower and more memory-intensive than those approaches, since it requires calculations of a very large number of reaction rates (for all the gray "edge" reactions in Fig. 2 as well as the "core" reactions shown in black). These rate calculations require calculations of the concentrations of each species (e.g. by solving the simulation with the current version of the reaction mechanism). So special work has been done to try to improve the computational efficiency of the rate-based approach [21,22], but even with those improvements it can be challenging to build a reaction network that achieves a user's error tolerance.[23]

Challenges with Species & Reaction Selection

Even if one can successfully complete all the computations, cases are known which confuse the rate-based algorithm, leading it to mis-classify some important species and reactions as unimportant, or vice-versa. More complicated species-selection algorithms have been proposed to try to overcome some of these flaws [24,25], but there are still open questions and a need for better species-selection algorithms.

The species-selection problem is particularly challenging for high-temperature low-pressure systems, where many of the reactions involve "well-skipping", so there is more than one transition state between the reactant(s) and the major products of a reaction.[26] To deal with this issue, RMG includes an extension of the rate-based algorithm for chemically-activated reactions that was originally invented by Matheu [27], then improved by Allen and Goldsmith [28], and recently further improved by Johnson and Grinberg Dana [29]. The numerical/computational challenges can be severe for systems with a large number of important isomer wells; often there are so many possible isomers that it is difficult for a human to correctly enumerate them all, much less judge which are important at particular reaction conditions. Several research groups have been working on aspects of this problem; particularly notable is the automated Potential Energy Surface (PES) search software package KinBot developed by Zador and co-workers [30].

Even in the high-pressure limit, it is not hard to identify important systems where there are so many possible reactions and reactive intermediates that it would be difficult-to-impossible to consider them all. A key practical limitation is that today's differential equation solvers run into numerical difficulties for systems with more than about 10,000 stiff non-linear differential equations (i.e. more than about 10,000 distinct radicals), due to round-off issues solving equations involving the large ill-conditioned Jacobian matrices. But some common mixtures, e.g. diesel fuel, contain more than 10,000 distinct stable molecules, and form more than 100,000 different radicals when they burn. As another example, pyrolysis of biomass or waste polymers involves thousands of distinct radicals from H-abstractions of every C-H bond in the macromolecules, and each of these can react to form several different oligomers and product radicals. As the reactions proceed, multiple sites on the polymer are functionalized, leading to a combinatorial explosion of isomers. Similarly challenging are systems where polymerization or soot formation is occurring, since a detailed model would need to include multiple reactive intermediates of almost every size between the starting material and the macromolecules being formed.

There are a variety of other issues related to dealing with such large simulations, beyond today's numerical solver problems. Not least is the challenge of constructing a complete set of all these important species and reactions (and to identify and debug any omissions). Another major challenge is coming up with accurate kinetic parameters for so many reactions, as discussed in the rest of this paper. Because of these difficulties, some researchers have built reduced models, using techniques such as Structure-Oriented Lumping [31] or the Fragment Method [32] to partially overcome the combinatorics. Further work is needed to better understand the errors introduced by these model-reduction approximations, and to automate these methods to make them easy to use.

Computing k's and K_{eq}'s from First Principles

During most of the 20th century, it was difficult-to-impossible to accurately compute most gas-phase high-pressure-limit rate coefficients $k_\infty(T)$ from first principles, and even more difficult to compute the effects of pressure or solvent on the rates. However, the development of density-functional theory and other quantum chemistry methods, coupled with the rapid advances in computer hardware, have now made DFT calculations routine, and . Rather sophisticated quantum chemistry calculations are now feasible for many reactions [4,5], and they are often quicker and easier than experimental measurements of similar accuracy. This game-changing advance is the reason that it is conceivable that predictions validated by inexpensive laboratory experiments, or in some cases even pure predictions, might be able to replace the conventional approach based on extensive pilot-plant experimentation shown in Fig. 1, while this was impossible in the 20th century.

What accuracy do we need? Usually errors in computed or estimated energies dominate, because they appear in the exponents in the expressions for $k_\infty(T)$ and $K_{eq}(T)$. Since $K_{eq}=\exp(-\Delta G_{rxn}/RT)$, if the computed value of the energy contains an error δG

$$\Delta G_{rxn,computed} = \Delta G_{rxn,true} + \delta G$$

$$K_{eq,computed} = K_{eq,true} \exp(\delta G/RT)$$

So if $|\delta G|>RT$, that error will cause $K_{eq,computed}$ will be off by a factor of $e \approx 2.7$x or more from the true value. Similarly, if the computed reaction barrier differs from the true barrier height by δE,

$$k_{computed} = k_{true} \exp(\delta E/RT)$$

So if $|\delta E| > RT$, that error in the computed barrier energy will cause the computed rate coefficient k to be off by a factor of 2.7x or more.

Can quantum chemistry achieve the needed accuracy? The deviations between high-accuracy experimental values of enthalpies of formation of small molecules and radicals from the Active Thermochemical Tables [33] and computations using various theoretical methods are shown in Fig. 3. Each tick on the x-axes is 10 kcal/mole. All the calculations shown use atom energy corrections (AEC) to bring the numbers from the quantum chemistry calculations onto the conventional elements-in-their-standard-state heat of formation scale. The mean absolute errors (MAE) and root mean square errors

(RMSE) are computed 3 ways: (1) before applying fitted Bond-Additivity Corrections (BAC), (2) after BAC but showing training errors (no test set), and (3) after BAC using leave-one-out cross-validation. The cross-validation method gives the best estimate of the errors in predicting enthalpies of formation of most molecules. The "Before BAC" errors provide estimates of expected errors in predictions for systems with bond types absent from the training data.

The popular low-level DFT method B3LYP-D3BJ, green bars in top figure, mis-predicts many molecules' enthalpies by more than 10 kcal/mole, much larger than RT, corresponding to errors of many orders of magnitude in K_{eq}'s and k's. Fitted bond-additivity corrections (BAC), using the form proposed by Anantharaman and Melius [84], help significantly (purple bars) but even after these tweaks are there are still many errors ~3 kcal/mole or larger. The intermediate-level method CBS-QB3, middle, has been very popular in the combustion/pyrolysis community over the past two decades, since with BAC (purple histogram) it predicts most enthalpies within 2 kcal/mole, without requiring too many computational resources for molecules containing up to about 12 carbon atoms.[34] Note RT > 2 kcal/mole if T>1000 K. The highest-level method shown in Fig. 3, DLPNO-CCSD(T)-F12b/cc-pCVTZ-F12//ωB97M-V/def2-TZVPD, bottom, is so accurate that it does not require BAC corrections to predict almost all of the enthalpies within 2 kcal/mole, green bars.[35] If BAC are applied to this method, most of the enthalpy errors drop to less than 1 kcal/mole, corresponding to about a factor of 3 uncertainty in K_{eq}'s at 500 K. For further discussion of methodology and similar tests of other quantum chemistry methods see [82]. Several research groups have automated quantum calculations of thermochemistry to automatically update the thermochemical parameters for hundreds of species in kinetic models.[88,89]

Note that errors in energies are more tolerable at high temperatures, where RT is larger, than at low temperatures. At present we often can predict k's and K_{eq}'s fairly accurately at combustion temperatures, because we have methods that can predict energies with errors $<RT_{combustion}$. But predicting accurate k's and K_{eq}'s at room temperature is difficult, because most methods cannot compute energies with errors $<RT_{room}$. Much of the most interesting low-temperature chemistry occurs in solvents, making it even more challenging to compute the energies to high accuracy. COSMO-RS [36] calculations based on DFT charge distributions predict the solvent corrections to the thermochemistry for many small neutral molecules in normal solvents with errors of about 1 kcal/mole [37,38,39], an error that adds to the often larger errors in the quantum chemistry prediction of the gas-phase enthalpies of formation. Errors in both types of calculations need to be reduced to enable accurate predictions of kinetics at room temperature.

Transition state theory calculations for $k_\infty(T)$ are similar to calculations of $K_{eq}(T)$. However, transition states more difficult to compute accurately than stable molecules: it is more difficult to provide guesses at transition state geometries that will successfully converge to the desired saddle point, and there are often low-lying electronic states which can cause convergence difficulties and degrade the accuracy of common approximations in the electronic structure calculations. At a simpler level, we do not have good empirical bond additivity corrections for transition states, so at best we might expect errors in barrier heights similar the green histograms in Fig. 3. The implication is that if we want $|\delta E| < RT$ to keep our computed k's within about a factor of 3 of the true values, we need to use rather expensive quantum chemistry methods (at least CBS-QB3, maybe CCSD(T)-F12).

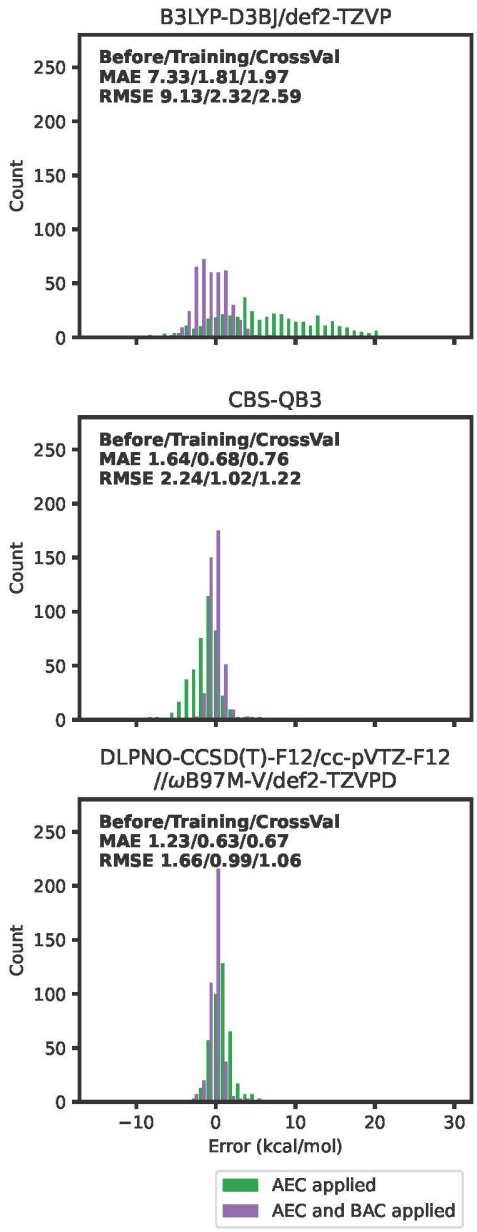

Fig.3 Histograms of Errors between Quantum Chemistry calculations of enthalpies of formation and experimental values from Active Thermochemical Tables (ATcT). Green: errors after applying only atom energy corrections (AEC). Purple: errors after also applying fitted bond-additivity corrections (BAC). Top: Density Functional Theory (B3LYP-D3BJ/def2-TZVP) has RMSE > 9 kcal/mole before BAC. Middle: CBS-QB3, a composite method based on single-determinant wavefunctions, has RMSE close to 1 kcal/mole after BAC. Bottom: CCSD(T)-F12b, a high-level method involving explicit correlation between each pair of electrons, has RMSE < 2 kcal/mole even without BAC. See Ref. [82]

If one uses high-accuracy quantum chemistry methods, the computed rates using transition state theory often closely match experiments.[4,5] One example is shown in Fig. 4, for the reaction phenyl + ethene.[40-45] The effect of a 1 kcal/mole uncertainty in a reaction barrier height is also illustrated in Fig. 4: at high temperature it does not matter much, but at room temperature it leads to large uncertainty in k(T).

Calculations of gas-phase reaction rates at low pressures and high temperatures requires dealing with many extra challenges and complexities, but there has been excellent progress in dealing with them.[46,47,48] Figs. 3 and 4 are some of the many published examples that demonstrate we can compute enthalpies (and so K_{eq}'s) and gas-phase k(T)'s fairly accurately from first-principles.

Recent work suggests that the kinetic solvent effects on small-molecule reactions (of neutral molecules in normal solvents) can also be computed accurately at low cost if the corresponding gas-phase barriers can be predicted accurately [49], though one might need to search for a different conformer of the reactant(s) and transition state in each solvent.[50]

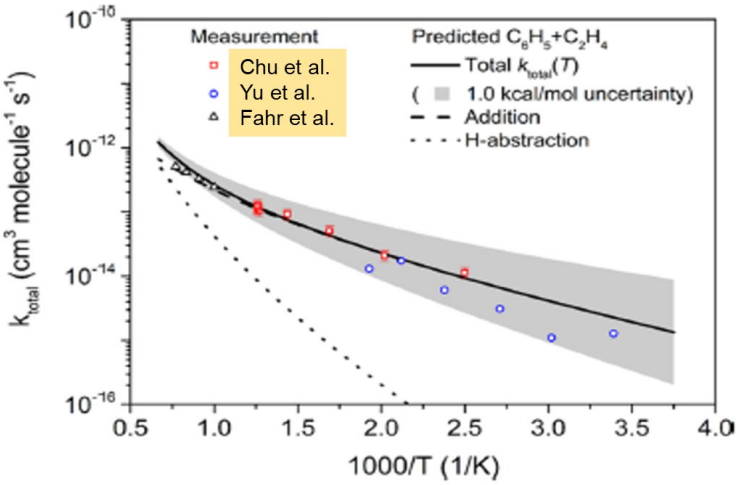

Fig. 4 Predicted and experimental k(T)'s for C_6H_5 + C_2H_4. The experimental data come from Refs. [41-45]. The gray band reflects 1 kcal/mole uncertainty in the computed (G2M(RCC5)//B3LYP) barrier height reported in [40]. This small uncertainty in energy corresponds to a large range of k's (note the y-axis scale spans 6 orders of magnitude).

Challenges in first-principles calculations of kinetic parameters

However, the scope of reactions we can compute accurately is disappointingly narrow. First, accurate first-principles kinetics is mostly confined today to light elements, both for reasons of computational cost, and because the transition metals and lanthanides have low-lying electronic states which reduce the accuracy of single-determinant methods such as CCSD(T) that work so well for organic molecules. The heavy elements introduce additional challenges (relativistic effects, spin-orbit coupling). Supercomputers are advancing rapidly, and many researchers are working to improve quantum

chemistry methods for transition metals, so hopefully this major limitation will be resolved in the coming decade.

Second, while we can compute the thermochemistry and reactions of ions and of electronically-excited states (e.g. photochemistry), the accuracy is significantly degraded compared to the reactions of ground-state neutrals, particularly in condensed phase. For example, the solvation energy of ions in water is often mis-predicted by about 5 kcal/mole [90], which is much too inaccurate to use in kinetic modeling.

Third, computational cost is a serious challenge as we move from small molecules (e.g. one aromatic ring) to the medium-sized molecules that are so important in organic, pharmaceutical, and biological chemistry. The accurate methods are based on coupled-cluster calculations (e.g. CCSD(T)), where the cost of computing the energy at a single molecular geometry scales roughly as the number of electrons to the 7^{th} power. So for example computing the energy of a dimer (or a monomer + monomer Transition State) is about $2^7 = 128$ times slower than computing its monomer. But as molecules get larger they also have more conformations, and so more geometries that need to be computed. Often many of the conformers correspond to the gauche-gauche'-trans arrangements around single bonds; there are about $3^{Nrotors}$ of those conformers. So if a monomer has 6 rotatable single bonds, it will have about $3^6 = 729$ conformers, and its dimer with 12 rotatable single bonds will have about $3^{12} = 531{,}441$ conformers, i.e. about 729 times as many conformers as the monomer. If one were to attempt to compute the energies of each conformer at the same high (e.g. CCSD(T)) level the CPU time to compute the dimer (or monomer + monomer TS) would be about 128*729 = 93,312 times longer than computing the monomer at the same level of theory. So although it may be practical to very accurately compute the properties of the monomer, it might be infeasible to compute the dimer or the monomer + monomer TS. There has been a lot of excellent work to find methods with near CCSD(T) accuracy but better scaling, but more work is needed on methods for handling conformers/rotors accurately and efficiently, particularly conformers of transition states.

When trying to decide which species belongs in a kinetic model and which can be safely omitted, Fig. 2, it would be very helpful to have reliable numbers for the rate coefficients of each potential reaction. However, in many systems there are known to be hundreds or thousands of important species, and most species can react rapidly with all the radicals and some of the stable molecules, often in multiple ways. So today one would like to have reliable k's for perhaps a million reactions, and in the future even more. Probably in coming decades, as both computational chemistry methods and computer hardware improve, it might become possible to compute a million reactions of medium size molecules accurately from first principles, but today it is not practical. So today we instead need to rely on *estimates*, as discussed next.

How to Estimate Kinetic Parameters?

In the 20^{th} century (and even earlier) many methods were proposed for estimating rate coefficients or thermochemical parameters (which provide K_{eq}'s and so the relationship between the k's for a forward reaction and its reverse) were proposed. These included the Arrhenius equation (k = f(T)) [51], the Lindemann-Hinshelwood equation (k=f(P)) [52], the Hammett equations [53] for how both rate coefficients and equilibrium constants depend on substituents on an aromatic ring, the Evans-Polanyi equation ($k = f(\Delta H_{rxn})$) [54], Benson's group contribution method for thermochemistry (and so K_{eq}) and estimation methods for Arrhenius A factors [55], and several other correlations. Most of these

correlations could be cast so that the handful of adjustable parameters could be unambiguously determined by linear least-squares fitting to a small number of data (though sometimes even these linear least-squares calculations ran into ill-conditioning problems). Several of these estimation methods involve first identifying which functional groups are in the molecules, or which bonds are changing during the reaction, and then using the appropriate linear correlation(s), shown schematically in Fig. 5. In modern parlance, that first step converts a molecule or reaction into a "fingerprint" vector, and the second step converts the vector into the estimate of the quantity of interest (e.g log(k) or log(K_{eq})).

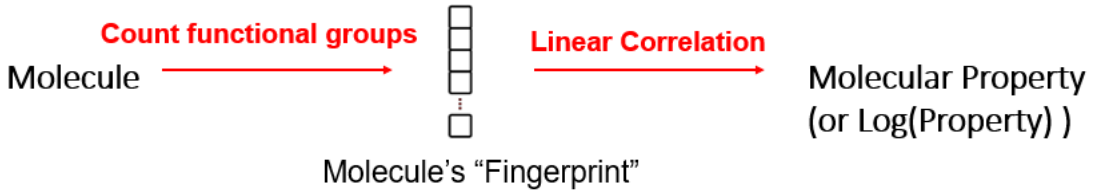

Fig. 5. Common 20th century estimation strategy for molecular or reaction properties. For example, one entry in the fingerprint could be the number of a certain functional group in the molecule.

As reliable quantum chemistry data became available, these were initially used to augment the experimental data, and fitted using similar linear correlations [56,57,58]. However in the last decade it has become practical to compute an enormous number of molecules and reactions using quantum chemistry [37,59,60,61], so much data that it can make sense to use much more complicated nonlinear correlations, a.k.a. "machine-learning (ML)", Fig. 6.

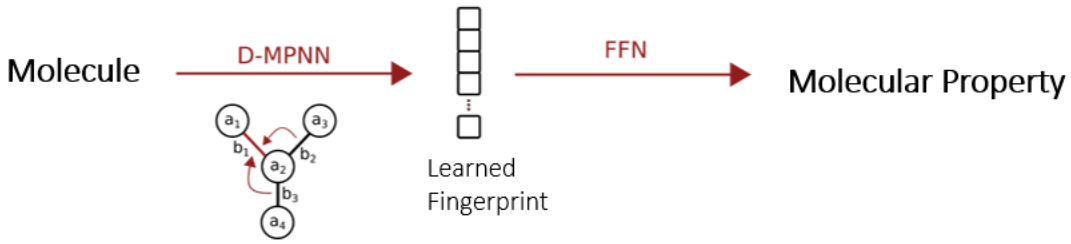

Fig. 6. The "machine-learning" non-linear correlation used by Chemprop and similar programs. The procedure "learns" which set of attributes is best to include in the fingerprint, using a type of graph convolution called a directed message passing neural net (D-MPNN). It simultaneously tunes a feed-forward neural net (FFN) to predict the property of interest from that learned fingerprint.

Fig. 6 shows one of several different non-linear many-parameter fitting forms that have been used to correlate molecules or reactions with their properties. The form shown in Fig. 6 is used in Chemprop [62], a very popular open-source "ML for Chemistry" software package (more than 100,000 downloads per month in early 2024). The Chemprop-type approach is very effective if one has a large amount of data (e.g. on >1,000 molecules or reactions). If one has a smaller data set, e.g. data on about 100 molecules, other correlation methods with fewer adjustable parameters may be more satisfactory, e.g. the Reaction Mechanism Generator software package uses trees to classify reactions [63,64]. A key advantage of having a large dataset is that one can reserve a large "test set" as well as a "validation set" and still have a large amount of data remaining to train the estimator. The validation dataset is used to

detect when the model is starting to overfit the training data, and so when one should stop the training procedure. The reserved test set then provides a realistic assessment of the accuracy of the estimates, albeit on molecules or reactions similar to those in the test set. Helpful tips are available [65] on how best to do this fitting procedure and estimate the uncertainties in the predictions, with warnings about common pitfalls.

Some properties depend on more than one molecule, e.g. solvation energy of a solute (Molecule 1) in a solvent (Molecule 2); that situation can be modeled as shown in Fig. 7. Slightly more complicated variations are required for reactions to maintain atom-mapping through the transition state [66] or for properties where Molecule 1 and Molecule 2 play the same roles, so (Molecule 1, Molecule 2) should give exactly the same property value as (Molecule 2, Molecule 1).[62] For predicting the properties of mixtures, one must also input the mole fractions of each molecule in the mixture.

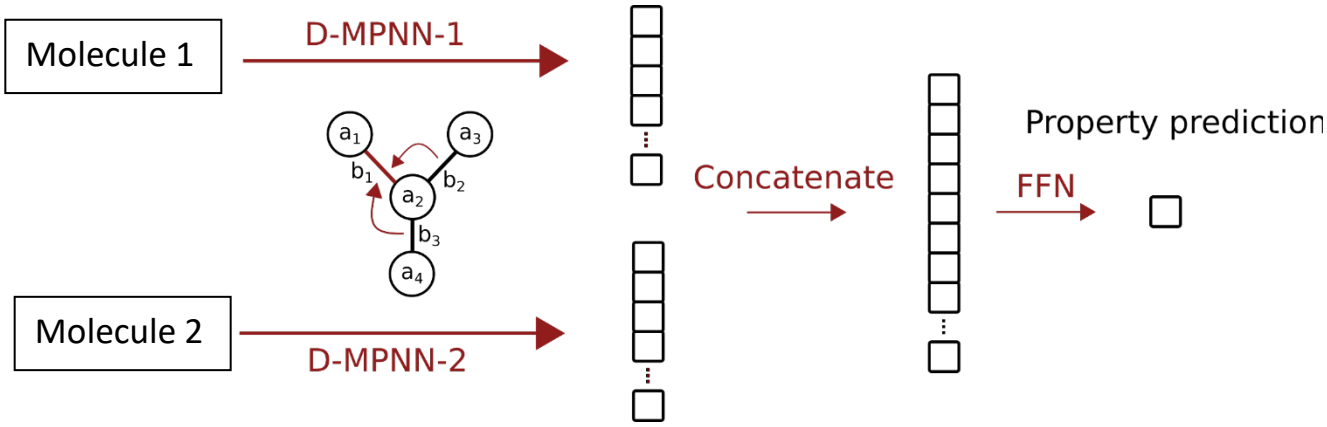

Fig. 7. Properties that depend on more than one molecule can be modeled by concatenating their learned fingerprints.

Many data are needed to "teach" a machine-learning model the rules of chemistry. This is often a problem in chemical kinetics, since we usually have only small experimental datasets, and often the data available is very "clumpy", with many data for certain types of molecule or reactions, and zero data on some other types of molecules or reactions. The need for data is illustrated for solvation free energy in Fig. 8 (orange curve) using calculations and data reported by Vermeire and Green.[37]

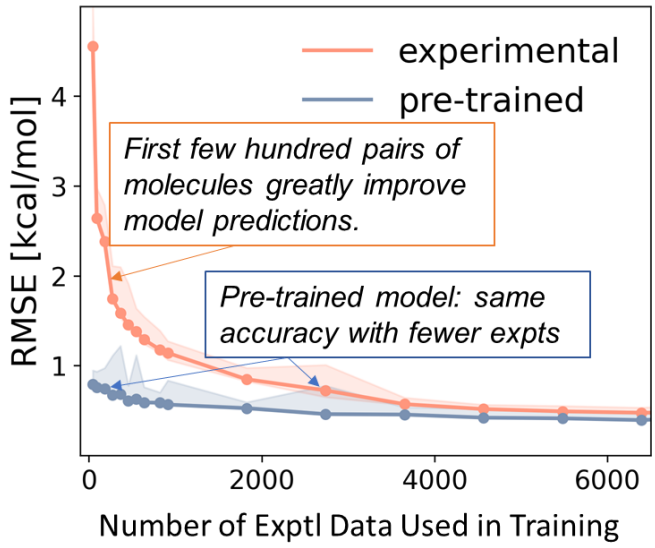

Fig. 8. Achieving good accuracy with a model for solvation free energy based solely on experimental data requires many data. Pre-training a model on a large number of quantum (COSMO-RS) calculations greatly reduces the number of experimental data needed. The Model vs. Expt. root-mean-square error (RMSE) asymptotes at the experimental noise level; in that large-data limit the model can be significantly more accurate than the experiments. See [37].

One way to avoid the need to purchase or synthesize a large number of molecules, and then perform thousands of experiments, is to instead start with a large number of quantum chemistry calculations. Pre-training a solvation model with many quantum chemical (COSMO-RS [36]) calculations, and then calibrating that model with ~200 solute-solvent pairs gives (Fig. 8, blue curve) about the same prediction accuracy as the purely experimental model (orange curve) trained with ~3,000 solute-solvent pairs. This approach has the important advantage that one can easily compute molecules which are not purchasable, hard to synthesize, or short-lived. My group has constructed several large quantum chemistry datasets [37,60,61,67,71,73], and used them and the available experimental data to train estimators for several different quantities important in kinetics: thermochemistry of both stable molecules and radicals, reaction barriers, transition state geometries, gas-phase k's, solvent effects on k's and K_{eq}'s, and IR and UV/vis spectra. [37,38,39,61,63,66,68,69,70,71,72,83]

In recent years, it has become possible to use the computer to automatically discover transition state geometries [30, 74, 75, 76] and so to compute reaction barriers for large numbers of reactions.[60,73,77]. With these large data sets, one can use machine-learning methods to build models for rapidly estimating the barriers for other reactions.[66,72] A particularly interesting example of this approach is Spiekermann et al.'s [72] development of a fast estimator for predicting unimolecular reaction barriers (for neutral singlet PES's involving only a few first row elements) trained by CCSD(T)-F12 data, which achieves an accuracy better than possible with DFT calculations, Fig. 9.

Fast reaction-barrier estimators of this quality would be very helpful for deciding which reactions should be included in a kinetic model, recall Fig. 2. In principle they could be combined with fast estimators for Arrhenius A-factors to provide fast estimates of $k_\infty(T)$, but to our knowledge no one has yet provided a similarly large data set of reliable Arrhenius A-factors that would be needed to train this type of

comprehensive neural-net rate-estimation model. Instead, the best rate-estimates available today are for a limited set of reaction types, often based on tree classifiers, that assign different rate rules to different types of reactions. [63,64,78, 87] Tree classifiers are expected to be more suitable than neural nets if data is sparse, which currently the case for many types of reactions. For a recent example of how these tree classifiers can be constructed automatically see [63]. The tree classifiers provide a relatively easy way to estimate the uncertainty in the estimates [63], compared to the rather complicated and often too-optimistic uncertainty estimates from today's neural-net approaches.[79,80]

It is possible to extend these approaches to liquid phase, at least for cases where all the reactants, products and the solvent are neutral molecules. The kinetic solvent effects on small-molecule reactions can be computed accurately if the corresponding gas-phase barriers can be predicted accurately [49] and rapidly predicted using machine-learning approaches.[61] The kinetic solvent effect estimator was trained using a large dataset of COSMO-RS calculations, but tested using held-out experimental measurements, with an RMSE error in $\Delta\Delta G^{\ddagger}$ < 1 kcal/mole [61], comparable to the accuracy of COSMO-RS calculations at predicting kinetic solvent effects.[49] However, it should be kept in mind that the number of reactions where the kinetic solvent effects on the rate coefficient has been experimentally determined and reported is rather small. Measurements on a broader range of reactions in various solvents and at various temperatures would be helpful to more thoroughly test the accuracy of kinetic solvent effect predictions.

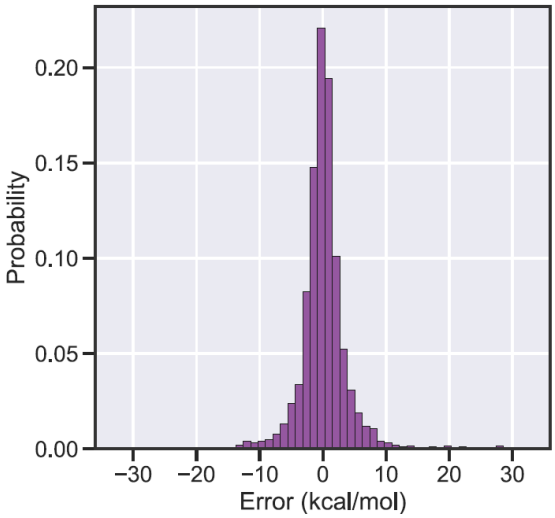

Fig. 9. Histogram of errors between fast estimator of reaction barrier heights and CCSD(T)-F12 calculations on more than a thousand test reactions. The mean absolute error in the fast estimates is ~3 kcal/mole, better than the typical accuracy of DFT calculations of barrier heights. For details on the datasets and model see [72,73].

Challenges in Estimating Kinetic Parameters

These approaches based on high-throughput quantum chemical calculations, while sometimes very helpful (see Fig. 8), are not a panacea: they are limited to molecules and properties which can be computed to a reasonable accuracy with quantum chemistry with the computer resources available. In

particular it is usually impractical to compute large numbers of macromolecules, or molecules with complicated electronic structures (e.g. many metal complexes), so the quantum chemistry data sets are biased towards small simple molecules. Even for relatively small simple molecules, it can be impossible to guarantee that one has found all the important (i.e. low energy) conformers.[50] We also do not know how to compute some important properties (e.g. solvation of ions, photochemistry in condensed phase) to the accuracy required for predictive kinetic modeling. And in some cases, such as reaction barriers, we believe we know how to do the quantum chemistry calculations, but it can be challenging to determine the barrier heights by experiment to an accuracy sufficient to really test if the quantum chemistry is accurate. All of these issues lead to important gaps and errors in the training datasets for any rate estimator, and can force the estimator to work in extrapolation (e.g. to estimate reactions of large molecules from training data that only includes small molecules), leading to uncontrolled errors in the estimates.

Quantum chemistry calculations are possible on many more molecules and reactions than are practical to address experimentally – this is a huge advantage over relying solely on experimental data to train an estimator. As a result, several large datasets of reaction barriers and thermochemistry have been computed in recent years [59,60,73,85,86], and several additional large datasets are expected to be published soon. However, there are often systematic gaps in the quantum chemistry datasets even if the dataset ostensibly covers the reaction one is trying to estimate. This can happen because the quantum chemistry fails to converge (or gives wrong values) for molecules or reactions with unusual electronic structures or geometries. So for example the widely used QM9 [59] dataset omits some small organic species with highly strained structures, and Grambow et al. [76] were unable to find satisfactory transition states for about 20% of the reactions they considered despite trying several different approaches. A few percent of Grambow's large set of successfully computed saddle points [60] were later identified to have significant deviations in barrier heights from CCSD(T)-F12 values [73], presumably due to convergence problems or the inability of the DFT method employed to handle the electronic structures for those transition states. Also, many data sets are only computed using DFT or even lower-accuracy methods, not the expensive high-accuracy methods needed for quantitatively-accurate predictions of rates or equilibria. Only a few of the data sets have included the bond-additivity or isodesmic reaction corrections needed to achieve thermochemistry close to experimental values, Fig. 3.[81,82,84] These problems all propagate into errors in a rate or thermochemistry estimator trained using these data. Some of these problems can be hard to uncover, particularly if the estimator is pretty good and so predicts most k(T) accurately. In the worst case, a serious mis-prediction might not become apparent until years later, when new experimental data or a much better quantum chemistry calculation becomes available, demonstrating that the estimate and/or the original quantum chemistry calculation for a particular reaction was incorrect. *Caveat emptor*!

Also, while quantum chemical calculations are often much faster than experiments, it is important to keep in mind that they are not zero-cost to society: someone is paying for the expensive supercomputer hardware and the electricity it is consuming to generate large data sets, and the associated greenhouse gas emissions are damaging the climate. Often, as in Fig. 8, there is limited value in pushing to larger and larger dataset sizes. So it is best practice to build models at the same time one is computing (or measuring) the data, to better assess the value of the next tranche of data relative to the cost of obtaining it.

Conclusion

Predictive kinetic modeling based largely on quantum chemistry has advanced to the state where it can compete with traditional heavily empirical methods of constructing quantitative kinetic models in several subfields of chemistry. However, there is significant room for improving its scope and reliability when the challenges highlighted in this paper are overcome. A key, both today and in the future, is the continued development of methods for computing and estimating k(T) and K_{eq}(T), with better identification of when those calculations or estimates are significantly uncertain. Achieving better estimates relies on access to more and better data both from quantum chemical calculations and from experiments. More open data sharing, and more recognition of the people who do the hard work involved in effectively sharing data and compiling curated data sets (ideally with error bars and provenance information on each number), would help us reach this goal. As the community routinely uses ever larger datasets and kinetic simulations, improved software tools are needed to assist humans in working with and checking the data, the models, and model vs. experiment comparisons, as well as better algorithms to convert the raw power of new computers into easy-to-use tools for predicting the behavior of chemical systems.

Acknowledgements

I gratefully acknowledge Beat Buesser for creating the original version of Fig. 3, and Hao-Wei Pang and Haoyang Wu for doing the calculations and creating the updated version shown in this paper. The work presented here was primarily supported by the US Department of Energy's Office of Basic Energy Sciences's Gas Phase Chemical Physics program through grant DE-SC0014901.

Literature Cited

[1] Mathematical Modeling of Complex Reaction Systems, ed. by Tiziano Faravelli, Flavio Manenti, and Eliseo Ranzi. *Computer-Aided Chemical Engineering* series, volume **45** (2019).

[2] Chemical Engineering Kinetics, ed. by G.B. Marin. *Advances in Chemical Engineering* series, volume **32** (2007).

[3] Tomlin AS, Turanyi T, Pilling MJ. Mathematical Tools for the Construction, Investigation, and Reduction of Combustion Mechanisms. *Comprehensive Chemical Kinetics* 1997; **35**: 293 – 437.

[4] Klippenstein SJ. From theoretical reaction dynamics to chemical modeling of combustion. *Proceedings of the Combustion Institute* 2017; **36:** 77-111.

[5] Klippenstein SJ. Spiers Memorial Lecture: Theory of Unimolecular Reactions. *Faraday Discussion* 2022; **238**: 11-67.

[6] Vereecken L, Aumont B, Barnes I, Bozzelli JW, Gillen MR, Mellouki A, Goldman MA, Green WH, Madronich S, Orlando JJ, Picquet-Varrault B, Rickard AR, Stockwell WR, Wallington TJ, Carter WPL.

Perspective on Mechanism Development and Structure-Activity Relationships for Gas-phase Atmospheric Chemistry. *International Journal of Chemical Kinetics* 2018; **50**: 435-469.

[7] Coley CW, Green WH, Jensen KF. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* 2018; **51**: 1281–1289.

[8] Merchant SS, Goldsmith CF, Vandeputte AG, Burke MP, Klippenstein SJ, Green WH. Understanding low-temperature first-stage ignition delay: Propane. *Combustion & Flame* 2015; **162**: 3658-3673.

[9] Warth V, Battin-Leclerc F, Fournet R, Glaude PA, Come GM, Scacchi G. Computer based generation of reaction mechanisms for gas-phase oxidation. *Comput. Chem.* 2000; **24**: 541-60.

[10] Blurock ES. Reaction: System for Modeling Chemical Reactions. *J. Chem. Inf. Comput. Sci.* 1995; **35**:607–616.

[11] Broadbelt LJ, Stark SM, Klein MT. *Industrial & Engineering Chemistry Research* 1994; **33**: 790-799.

[12] Wu H, Grinberg Dana A, Ranasinghe DS, Pickard FC, Wood GPF, Zelesky T, Sluggett GW, Mustakis J, Green WH. Kinetic Modeling of API Oxidation: 2. Imipramine Stress Testing. *Molecular Pharmaceutics* 2022**; 19**: 1526–1539.

[13] Susnow RG, Dean AM, Green WH, Peczak P, Broadbelt LJ. Rate-Based Construction of Kinetic Models for Complex Systems. *Journal of Physical Chemistry A* 1997; **101**: 3731-3740.

[14] Gao CW, Allen JW, Green WH, West RH. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Computer Physics Communications* 2016; **203**: 212-225.

[15] Liu M, Grinberg Dana A, Johnson MS, Goldman MJ, Jocher A, Payne AM, Grambow CA, Han K, Yee NW, Mazeau EJ, Blondal K, West RH, Goldsmith CF, Green WH. RMG 3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model* 2021; **61**: 2686-2696.

[16] Matheu DM, Dean AM, Grenda JM, Green WH. Mechanism Generation with Integrated Pressure-Dependence: A New Model for Methane Pyrolysis. *Journal of Physical Chemistry A* 2003; **107**: 8552-8565.

[17] Hansen N, Merchant SS, Harper MR, Green WH. The Predictive Capability of an Automatically Generated Combustion Chemistry Mechanism: Chemical Structures of Premixed *iso*-Butanol Flames. *Combustion and Flame* 2013; **160**: 2343-2351.

[18] Gudiyella S, Buras ZJ, Chu T-C, Lengyel I, Pannala S, Green WH. A modeling study of high-temperature pyrolysis of natural gas. *Industrial & Engineering Chemistry Research* 2018; **57**: 7404–7420.

[19] Chu T-C, Buras ZJ, Oßwald P, Liu M, Goldman MJ, Green WH. 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.

[20] Zhang P, Yee NW, Filip SV, Hetrick CE, Yang B, Green WH**.** Modeling Study of the anti-knock tendency of substituted phenols as additives: An application of the Reaction Mechanism Generator (RMG). *Physical Chemistry Chemical Physics* 2018; **20**: 10637-10649.

[21] Han K, Green WH, West RH. On-the-Fly Pruning for Rate-Based Reaction Mechanism Generation. *Computers and Chemical Engineering* 2017; **100**: 1-8.

[22] Jocher A, Vandewiele NM, Han K, Liu M, Gao CW, Gillis RJ, Green WH. Scalability strategies for automated reaction mechanism generation. *Computers & Chemical Engineering* 2019; **131**: 106578.

[23] Payne AM, Spiekermann KA, Green WH. Detailed Reaction Mechanism for 350-400 C Pyrolysis of an Alkane, Aromatic, and long-chain Alkylaromatic Mixture. *Energy & Fuels* 2022; **36**: 1635-1646.

[24] Pang H-W, Forsuelo M, Dong X, Hawtof RE, Ranasinghe DS, Green WH. Detailed Multiphase Chemical Kinetic Model for Polymer Fouling in a Distillation Column. *Industrial and Engineering Chemistry Research* 2023; **62**: 14266–14285.

[25] Johnson MS, Pang H-W, Liu M, Green WH. Species Selection for Automatic Chemical Kinetic Mechanism Generation. 2024; DOI:10.26434/chemrxiv-2023-wwrqf-v2

[26] Miller JA, Klippenstein SJ. Master equation methods in gas phase chemical kinetics. *Journal of Physical Chemistry A* 2006; **110**:10528-10544

[27] Matheu DM, Lada TA, Green WH, Dean AM, Grenda JM. Rate-Based Screening of Pressure-Dependent Reaction Networks. *Computer Physics Communications* 2001; **138**: 237-249.

[28] Allen JW, Goldsmith CF, Green WH. Automatic Estimation of Pressure-Dependent Rate Coefficients. *Physical Chemistry Chemical Physics* 2012; **14**: 1131 - 1155.

[29] Johnson MS, Grinberg Dana A, Green WH. A Workflow for Automatic Generation and Efficient Refinement of Pressure Dependent Networks. *Combustion & Flame* 2023; **257**: 112516.

[30] Van de Vijver R, Zador J. KinBot: Automated stationary point search on potential energy surfaces**.** *Computer Physics Communications* 2020; **248**: 106947.

[31] Quann RJ, Jaffe SB. Structure-Oriented Lumping: Describing the Chemistry of Complex Hydrocarbon Mixtures. *Ind. Eng. Chem. Res.* 1992, **31**: 2483–2497.

[32] Han K, Green WH. A Fragment-based Mechanistic Kinetic Modeling Framework for Complex Systems. *Industrial & Engineering Chemistry Research* 2018; **57**: 14022–14030.

[33] Ruscic B, Bross DH. Active Thermochemical Tables (ATcT) Thermochemical Values ver. 1.122r. DOI: 10.17038/CSE/1822363.

[34] Montgomery JA, Frisch MJ, Ochterski JW, Petersson GA. A complete basis set model chemistry VII. Use of the minimum population localization method. *J. Chem. Phys.* 2000; **112**: 6532–6542.

[35] Knizia G, Adler TB, Werner HJ. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* 2009; **130**: 054104.

[36] Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans.* 1993; *2*, 799−805.

[37] Vermeire FH, Green WH. Transfer learning for solvation free energies: from quantum chemistry to experiments. *Chemical Engineering Journal* 2021; **418**: 129307.

[38] Chung Y, Vermeire FH, Wu H, Walker PJ, Abraham MH, Green WH. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *Journal of Chemical Information and Modeling* 2022*;* **62:** 433-446.

[39] Vermeire FH, Chung Y, Green WH. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *Journal of the American Chemical Society* 2022; **144**: 0785–10797.

[40] Tokmakov IV, Lin MC. Combined Quantum Chemical/RRKM-ME Computational Study of the Phenyl + Ethylene, Vinyl + Benzene, and H + Styrene Reactions*. J. Phys. Chem. A* 2004; **108**: 9697−9714.

[41] Chu T-C, Buras ZJ, Eyob B, Smith MC, Liu M, Green WH. Direct Kinetics and Product Measurement of Phenyl Radical + Ethylene. *Journal of Physical Chemistry A* 2020; **124**: 2352-2365.

[42] Yu T, Lin MC. Kinetics of Phenyl Radical Reactions Studied by the Cavity-Ring-Down Method. *J. Am. Chem. Soc.* 1993; **115**: 4371−4372.

[43] Yu T, Lin MC. Kinetics of the Phenyl Radical Reaction with Ethylene: An RRKM Theoretical Analysis of Low and High Temperature Data. *Combust. Flame* 1995; **100**: 169−176.

[44] Fahr A, Stein SE. Reactions of Vinyl and Phenyl Radicals with Ethyne, Ethene and Benzene. *Symp. (Int.) Combust.* 1989; **22**: 1023−1029.

[45] Fahr A, Mallard GW, Stein SE. Reactions of Phenyl Radicals with Ethene, Ethyne, and Benzene. *Symp. (Int.) Combust*. 1988; **21**: 825−831.

[46] Georgievskii Y, Miller JA, Burke MP, Klippenstein SJ. Reformulation and solution of the master equation for multiple-well chemical reactions. *J. Phys. Chem. A* 2013*;* **117**: 12146–12154.

[47] Jasper AW, Pelzer KM, Miller JA, Kamarchik E, Harding LB, Klippenstein SJ. Predictive *a priori* pressure-dependent kinetics*. Science* 2014; **346**(6214): 1212-1215.

[48] Zhang RM, Xu X, Truhlar DG. Low-Pressure Limit of Competitive Unimolecular Reactions. *Journal of the American Chemical Society* 2020; ***142***(37): 16064-16071.

[49] Chung Y, Green WH*.* Computing kinetic solvent effects and liquid phase rate constants using quantum chemistry and COSMO-RS methods. *Journal of Physical Chemistry A* 2023; **127**: 5637-5651.

[50] Pattanaik L, Menon A, Settels V, Spiekermann K, Tan Z, Vermeire FH, Sandfort F, Eiden P, Green WH. ConfSolv: Prediction of solute conformer free energies across a range of solvents. *Journal of Physical Chemistry B* 2023; **127**: 10151-10170.

[51] Arrhenius, SA. Uber die Dissociationswarme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Z. Phys. Chem*. 1889; **4**: 96–116.

[52] Lindemann FA, Arrhenius S, Langmuir I, Dhar NR, Perrin J, Lewis WCMC. Discussion on "the radiation theory of chemical action". *Transactions of the Faraday Society* 1922; **17**: 598.

[53] Hammett, LP. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* 1937; **59**: 96-103.

[54] Evans MG, Polanyi M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *J. Chem. Soc. Faraday Trans* 1936; **32**: 1333.

[55] Benson SW. *Thermochemical Kinetics 2nd ed*. Wiley NY 1976.

[56] Lay TH, Bozzelli JW, Dean AM, Ritter ER. Hydrogen Atom Bond Increments for the Calculation of Thermodynamic Properties of Hydrocarbon Radicals. *Journal of Physical Chemistry* 1995; **99**: 14514-14527.

[57] Sumathi R, Green, WH. *A priori* Rate Constants for Kinetic Modeling. *Theoretical Chemistry Accounts* 2002; **108**: 187-213.

[58] Sumathi R, Green WH. Missing thermochemical groups for large unsaturated hydrocarbons: Contrasting predictions of G2 and CBS-Q. *Journal of Physical Chemistry A* 2002; **106:**11141-11149.

[59] Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 2014; **1**: 140022.

[60] Grambow CA, Pattanaik L, Green WH. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data* 2020; **7**: 137.

[61] Chung Y, Green WH. Machine learning from quantum chemistry to predict experimental kinetic solvent effects**.** *Chemical Science* 2024; **15**: 2410-2424.

[62] Heid E, Greenman K, Chung Y, Li S-C, Graff D, Vermeire FH, Wu H, Green WH, McGill CJ. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* 2024; **64**: 9–17.

[63] Johnson MS, Green WH. A machine-learning based approach to reaction rate estimation. *Reaction Chemistry and Engineering* (2024) in press, available online.

[64] Johnson MS, Dong X, Grinberg Dana A, Chung Y, Farina D, Gillis RJ, Liu M, Yee NW, Blondal K, Mazeau E, Grambow CA, Payne AM, Spiekermann K, Pang H-W, Goldsmith CF, West RH, Green WH. The RMG Database for Molecular Property Prediction. *Journal of Chemical Information and Modeling* 2022; **62**: 4906-4915.

[65] Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing Uncertainty in Machine Learning for Chemistry. *Journal of Chemical Information and Modeling* 2023; **63**: 4012-4049.

[66] Heid E, Green WH. Machine learning of reaction properties via learned representations of the condensed graph of reaction (CGR). *Journal of Chemical Information and Modeling* 2022; **62**: 2101–2110.

[67] Grambow CA, Li Y-P, Green WH. Accurate Thermochemistry with Small Datasets: A Bond Additivity Correction and Transfer Learning Approach. *Journal of Physical Chemistry A* 2019**; 123**: 5826-5835.

[68] Grambow CA, Pattanaik L, Green WH. Deep Learning of Activation Energies. *Journal of Physical Chemistry Letters* 2020; **11**: 2992-2997.

[69] Pattanaik L, Ingraham JB, Grambow CA, Green WH. Generating Transition States with Deep Learning. *Phys. Chem. Chem. Phys.* 2020; **22**: 23618-23626.

[70] McGill CJ, Forsuelo M, Guan Y, Green WH. Message Passing Neural Networks for Infrared Spectra Prediction. *J. Chem. Inf. Model* 2021; **61**: 2594-2609.

[71] Greenman KP, Green WH, Gomez-Bombarelli R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chemical Science* **13**: 1152-1162 (2022)

[72] Spiekermann KA, Pattanaik L, Green WH. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *Journal of Physical Chemistry A* 2022; **126:**3976-3986.

[73] Spiekermann KA, Pattanaik L, Green WH. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data* 2022; **9**: 417.

[74] Maeda S, Taketsugu T, Morokuma K. Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method. *J Comput Chem.* 2014; **35**(2):166-73.

[75] Suleimanov YV, Green WH. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *Journal of Chemical Theory & Computation* 2015; **11**: 4248–4259.

[76] Grambow CA, Jamal A, Li Y-P, Green WH, Zador J, Suleimanov YV. Unexpected Unimolecular Reaction Pathways of a gamma-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *Journal of the American Chemical Society* 2018; **140**: 1035-1048.

[77] Zhao Q, Savoie BM. Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces. *Angewandte Chemie Intl. Ed.* 2022; **61**: e202210693.

[78] Wang K, Dean AM. Rate Rules and Reaction Classes. *Computer-Aided Chemical Engineering* 2019; **45**: 203-257.

[79] Scalia G, Grambow CA, Pernici B, Li Y-PGreen WH. Evaluating Scalable Uncertainty Estimation Methods for DNN-Based Molecular Property Prediction", *Journal of Chemical Information and Modeling* 2020; **60**: 2697–2717.

[80] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty Quantification using Neural Networks for Property Prediction. *Journal of Chemical Information and Modeling* 2020; **60**: 3770-3780.

[81] Buerger P, Akroyd J, Mosbach S, Kraft M. A systematic method to estimate and validate enthalpies of formation using error-cancelling balanced reactions. *Combust. Flame* 2018; **187**: 105-121.

[82] Wu H, Payne AM, Pang H-W, Menon A, Grambow C, Ranasinghe D, Dong X, Grinberg Dana A, Green WH. Towards Accurate Quantum Mechanical Thermochemistry: (1) Extensible Implementation and

Comparison of Bond Additivity Corrections and Isodesmic Reactions. *Journal of Physical Chemistry A* 2024 (accepted).

[83] Pang H-W, Dong X, Johnson MS, Green WH. Subgraph Isomorphic Decision Tree to Predict Radical Thermochemistry with Bounded Uncertainty Estimation. *Journal of Physical Chemistry A* 2024; **128**: 2891–2907.

[84] Anantharaman B, Melius C. Bond additivity corrections for G3B3 and G3MP2B3 quantum chemistry methods. *Journal of Physical Chemistry A* 2005; **109**: 1734-1747.

[85] Zhao Q, Vaddadi SM, Woulfe M, Ogunfowora LA, Garimella SS, Isayev O, Savoie BM. *Scientific Data* 2024; **10**: 145.

[86] Khan D, Benali A, Kim SYH, von Rudorff GF, von Lilienfeld OA. Towards comprehensive coverage of chemical space: Quantum mechanical properties of 836k constitutional and conformational closed shell neutral isomers consisting of HCNOFSiPSClBr. *arXiv*:2405.05961. DOI: 10.48550/arXiv.2405.05961

[87] Li N, Girhe S, Zhang M, Chen B, Zhang Y, Liu S, Pitsch H. A machine learning method to predict rate constants for various reactions in combustion kinetic models. *Combustion & Flame* 2024; **263**: 113375.

[88] Keceli M, Elliott S, Li Y-P, Johnson MS, Cavallotti C, Georgievskii Y, Green WH, Pelucchi M, Wozniak JM, Jasper AW, Klippenstein SJ. Automated Computational Thermochemistry for Butane Oxidation: A Prelude to Predictive Automated Combustion Kinetics. *Proceedings of the Combustion Institute* 2019*;* **37**:363-371.

[89] Pio G, Dong X, Bolzano E, Green WH. Automatically Generated Model for Light Alkene Combustion. *Combustion & Flame* 2022; **241**:112080.

[90] Zheng JW, Green WH. Experimental Compilation and Computation of Hydration Free Energies for Ionic Solutes. *Journal of Physical Chemistry A* 2023; **127**:10268-10281.