Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry

Citation: Li, Yi-Pei et al. "Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry." Journal of Physical Chemistry, 123, 10 (March 2019) 2142-2152 © 2019 American Chemical Society

As Published: http://dx.doi.org/10.1021/acs.jpca.8b10789

Publisher: American Chemical Society (ACS)

Persistent URL: https://hdl.handle.net/1721.1/123874

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry

Yi-Pei Li[§]*, Kehang Han*[§]*, Colin A. Grambow, and William H. Green*[*]

Department of Chemical Engineering, Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, Massachusetts 02139

Submitted to
The Journal of Physical Chemistry A
Nov 7, 2018
Revised version submitted Jan 24, 2019.

§: Both authors contributed equally to this work.

***Corresponding author:** William H. Green: whgreen@mit.edu

Abstract:

Because collecting precise and accurate chemistry data is often challenging, chemistry datasets usually only span a small region of chemical space, which limits the performance and the scope of applicability of data-driven models. To address this issue, we integrated an active learning machine with automatic *ab initio* calculations to form a self-evolving model that can continuously adapt to new species appointed by the users. In the present work, we demonstrate the self-evolving concept by modeling the formation enthalpies of stable closed-shell polycyclic species calculated at the B3LYP/6-31G(2df,p) level of theory. By combining a molecular graph convolutional neural network with a dropout training strategy, the model we developed can predict DFT enthalpies for a broad range of polycyclic species and assess the quality of each predicted value. For the species which the current model is uncertain about, the automatic *ab initio* calculations provide additional training data to improve the performance of the model. For a test set composed of 2,858 cyclic and polycyclic hydrocarbons and oxygenates, the enthalpies predicted by the model agree with the reference DFT values with a root-mean-square error of 2.62 kcal/mol. We found that a model originally trained on hydrocarbons and oxygenates can broaden its prediction coverage to nitrogen-containing species via an active learning process, suggesting that the continuous learning strategy is not only able to improve the model accuracy but is also capable of expanding the predictive capacity of a model to unseen species domains.

Introduction:

Although recent advances in the field of *ab initio* quantum chemistry methods have facilitated quantitative understanding of challenging chemical problems[1–5] and accurate calculations of molecular thermochemistry,[6–16] large-scale theoretical studies are often still limited, at least initially, to the use of empirical methods to rapidly screen out unimportant species, so that only the important species are the subject of CPU-time intensive quantum chemistry calculations. Among the empirical methods developed in the past decades, the Benson group additivity scheme[17] is one of the quickest and most convenient methods to determine thermodynamic properties of molecules without requiring 3D molecular structures. It has achieved great success for accurate prediction of thermochemistry of simple molecules and has been adopted in modeling software for on-the-fly prediction of thermodynamic parameters.[18] However, the performance and the scope of applicability of an empirical model are often limited by the coverage of the chemistry dataset employed when the model was developed. For example, because the additivity scheme was designed for simple organic species, it only uses the properties of individual chemical groups independently to calculate the composite property for a molecule, and thus the contribution of the overall molecular structure to the property is usually not taken into account.[17,19,20] This problem manifests itself for strained structures and can cause significant errors for polycyclic molecules with fused rings.[21] Therefore, application of the additivity method is often restricted to simple chemical systems without the presence of polycyclic species.

It is not an easy task to develop correction schemes for the additivity method to accurately estimate thermochemistry of polycyclic species. The major challenge is that each ring cluster structure has its own specific ring correction and it is impossible to prepare a list of corrections for all polycyclic structures because the number of possible fused ring clusters is exceedingly large. Han et al. developed two algorithms, similarity match and bicyclic decomposition, to ameliorate this problem.[21] Ring strain corrections of small cyclic structures were calculated using *ab initio* methods and organized into a

functional group tree that can find similar matches for any new small cyclics in the similarity match approach. The bicyclic decomposition algorithm can estimate large polycyclic ring strain corrections by decomposing them into smaller ones and adding up the contributions from the fragments. By combining the bicyclic decomposition method with the similarity match approach, one can easily predict ring strains of highly complex polycyclic clusters using the pre-calculated ring strains of simple one-ring and two-ring clusters. This correction scheme successfully reduced the heat of formation error of polycyclic species calculated with the group additivity method from over 60 kcal/mol to about 5 kcal/mol.[21]

However, since the underlying assumption of the bicyclic decomposition scheme is that the contributions of bicyclic ring strain are independent and additive, which is not always accurate, one needs to add more terms to describe the interactions between the decomposed bicyclics to further reduce the error.[21] This is a tedious task because the number of corrections grows exponentially as higher order terms are included and one has to collect more thermodynamic data to determine the values of these higher order corrections. In addition, some applications of the additivity scheme would require considering heteroatomic polycyclic species, which further increases the complexity of the model because the presence of heteroatoms will not only require defining additional groups but also new sets of ring strain corrections. Therefore, even though the group additivity approach is very effective for estimating molecular thermochemistry of simple organic molecules, its prediction accuracy rapidly drops as the species of interest become more and more complicated.

The inherent drawback of the additivity scheme is that the accuracy of the model is dependent on the groups (and ring corrections) chosen by humans. If the list of groups does not cover all the important features of a molecule, e.g., heteroatoms or polycyclic structures, the additivity scheme is unlikely to perform well because the model does not have all the relevant information needed to make good predictions. To address this problem, He et al. developed an automatic and adaptive distance-based group contribution method (DBGC) to avoid manual selection of groups.[22] In DBGC, the intramolecular

interaction between two groups is described by an exponential decay function of the number of bonds between the interacting units. This method performs well for the cases where group interactions cannot be ignored, e.g., highly branched large hydrocarbons. However, DBGC does not explicitly take into account the contribution of global molecular structure, which is important for accurate prediction of thermochemistry of polycyclic species. Since the descriptors of additivity methods were defined to describe a specific chemistry dataset, the molecular structural information that can be perceived by the model is limited by the chemistry dataset considered when the model was developed. Therefore, in addition to collecting new chemistry data, one often needs to redesign the architecture of the model, i.e., define new groups, interactions, or corrections to improve the performance and expand the scope of applicability of the additivity scheme. This process is labor-intensive and time-consuming, and very challenging for non-experts.

To resolve this problem, we adopt a machine learning model that is capable of directly learning structural information that is useful for thermochemistry predictions.[23,24] Over the past ten years, several methods have been proposed including coulomb matrices,[25] symmetry function transformations,[26–28] extended-connectivity fingerprints (ECFPs),[29] and molecular graph convolutions,[30,31] to transform molecules into a fixed-length representation that is suitable for conventional machine learning algorithms. In this work, we adopt the molecular graph convolution method, which has been shown to perform well on a broad range of applications.[32] Therefore, unlike the additivity scheme that requires human input to define chemical groups, the model presented in this work simply uses the "2-D" connectivity structure of a molecule as the input and automatically extracts useful features from the structure to predict molecular properties.

As pointed out by Simm and Reiher,[33] to ensure accurate parameterization of a data-driven model, the training set needs to be representative for the system of interest. However, since it is difficult to include every relevant molecular structure of a chemical system under consideration, one should consider

a continuous refinement scheme, in which new data are constantly added to the training set when necessary.[33,34] Following this line of thought, we developed an ensemble approach to measure the quality of predicted thermochemistry by combining the idea of bootstrap sampling[35] with dropout training in neural networks.[36] Bootstrapping is a sampling method that has been shown to be reliable for determining systematic errors and estimating uncertainties.[37] This feature enables identification of species for which the thermochemical properties are potentially associated with significant uncertainty as judged by the current model, and thus allows for strategic collection of new chemistry data to enhance the performance of the model. Providing new training data for uncertain samples is known as an active learning strategy,[38] and has been shown to be very effective for improving the accuracies of machine learning potential energy surfaces through the application of Gaussian processes[34] or ensemble approaches.[39,40] By combining the active learning scheme with automatic *ab initio* calculations, the machine learning model can effectively identify the species for which the predictions might be inaccurate and automatically derive thermodynamic data by initiating first principles calculations to improve the performance of the model. In the present work, we demonstrate the "self-evolving" feature by modeling the formation enthalpies of polycyclic species derived from density functional theory (DFT).

Methods:

Thermochemistry estimator. To ensure effective molecular feature extraction for the species of interest, we implemented a convolutional neural network that operates directly on graphs of arbitrary size and shape following the procedures proposed previously.[23,24,31] As shown in Scheme 1, an input molecule of the molecular convolutional neural network is represented by a molecular matrix $A^{r=0}$ composed of atomic feature vectors. These feature vectors contain information of the local chemical environment of an atom, including atom types, hybridization types, valence structure, and the number of rings of each size containing the atom (see Text S1 and Table S1 in the Supporting Information for the

complete list of features we used). The convolution process gradually merges information of distant atoms by combining feature vectors of connected atoms and the corresponding chemical bond information, which includes bond orders and aromaticity indicators, to generate the molecular matrices with a larger radius ($A^{r=1}$ and $A^{r=2}$) and the corresponding molecular fingerprints. We note that the inclusion of bond order features does not significantly improve model accuracy, and hence such features could be removed to avoid complications when multiple resonance structures are possible. The molecular fingerprints are then combined and passed to a standard neural network with one hidden layer to predict the property of the input molecule. An expanded discussion of the model and an example of input molecular matrix can be found in the work of Coley et al.[24] Interested readers are referred to the detailed description of the convolutional neural network algorithm documented in Section I of the Supporting Information of Coley et al.[24] and the implementation of the model in our Github repository.[41] The major differences between our implementation and previous molecular graph convolution models[23,24,31] are the inclusion of more detailed ring information in the initial molecular matrix (as described in Text S1) and the incorporation of dropout masks to generate ensemble predictions for measuring prediction quality, which will be discussed in detail in the following subsection.

Ensemble predictions. The neural network literature contains a large amount of work on uncertainty estimation based on parametric Bayesian inference.[42,43] However, we estimate uncertainty differently, using a non-parametric ensemble approach motivated by bootstrap sampling.[35] The bootstrap principle is to approximate a population distribution by a sample distribution. In its most common form, bootstrap generates k sets of samples $D_0 \ldots D_k$ from a given data set D by resampling uniformly with replacement.[35] Each bootstrap data set D_i is expected to have a fraction of the unique samples of D and the rest being duplicates. If the original data set is a good approximation of the population of interest, one can derive the sampling distribution of a particular statistic from the collection of its values arising from the k data sets generated by bootstrapping.[35] Similarly, one can train a committee of k models using

the bootstrap data sets and derive ensemble predictions, which is known as bagging or bootstrap aggregating.[44] Since the diversity of the predictions reflects the quality of the models, one can use the standard deviation of the predictions (ensemble spread) to estimate the potential benefits of obtaining new training data to improve the model prediction for a given query.

In this work, the ensemble models were implemented using dropout training with neural networks. That is, instead of building multiple models, we trained one single neural network with multiple dropout masks. Recent work of Gal and Ghahramani shows that one can approximate Bayesian inference using dropout training in neural networks.[43] However, unlike the framework developed by Gal and Ghahramani and the standard dropout procedure in which the mask is generated randomly during each iteration of training, we randomly generated a set of masks before training and saved them along with the weights of the network as part of the model. Since applying dropout masks removes non-output units from a fully-connected network,[36] a standard neural net with *k* dropout masks can be viewed as a collection of *k* sub-networks that share weights. For each training step, one of the sub-networks was randomly selected and optimized with one example (mini-batch of size one). Therefore, each of the sub-networks is expected to see some duplicated examples and only a fraction of the training data just as training ensemble models with bootstrap data sets. The ensemble prediction and ensemble spread were derived by averaging and calculating the standard deviation of the sub-network outputs. Our code for training the ensemble model with a graph convolutional neural network is available online.[41]

Reference Data Sets. The enthalpies of 29,474 cyclic and polycyclic hydrocarbons and oxygenates were used to train and test the model. The ratio between the numbers of training and testing examples was 4:1. A data set consisting of 39,981 nitrogen-containing cyclic and polycyclic molecules was used to examine the self-learning process, the details of which can be found in the subsection entitled Active Learning Process. All of the data were calculated at the B3LYP/6-31G(2df,p) level of theory using the rigid rotor-harmonic oscillator approximation (RRHO) and were extracted from the work of

Ramakrishnan et al[45] unless noted otherwise (see Text S2 in Supporting Information for detailed descriptions of how the enthalpies were calculated).

Since the reference data were calculated at the B3LYP/6-31G(2df,p) level of theory, the enthalpies predicted by the models are the DFT values instead of the (unknown) true enthalpies. The errors reported below only reflect the performance of the neural network in modeling DFT results and should not be interpreted as the accuracy of predicting true enthalpies, because the DFT calculations are themselves associated with errors.[33,46–48] For the same reason, the spreads of ensemble predictions represent the expected model departures from the reference DFT values instead of the true enthalpies. A detailed discussion of the errors associated with the B3LYP/6-31G(2df,p) data and their influence on the performance of the machine learning model can be found the subsection entitled Comparisons with High Level of Theory and Group Additivity method.

Generation of New Data. We also designed and implemented an automatic quantum mechanical calculation package to continuously provide training data for future improvements of the thermochemistry estimator. Similar to previous work such as Chemoton[49] and PACT[50], this package uses RDKit[51] to generate initial 3-D geometries for given molecules and calls the quantum chemistry software Q-Chem[52] to conduct geometry optimization, frequency, and single point calculations at various levels of theory. The automatically generated quantum chemistry data were calculated at the B3LYP/6-31G(2df,p) level of theory using the RRHO approximation in order to be consistent with the data extracted from the work of Ramakrishnan et al.[45] Q-Chem's default settings for convergence tolerances were used for all the calculations. The package automatically parses the output files of quantum mechanical calculations and stores the processed data in a non-relational database under the framework of MongoDB.[53] Features such as communicating with the system scheduler to monitor job status and analyzing convergence failure are implemented in the package to handle large scale quantum mechanical calculations. Advanced functionalities for accurate thermochemistry calculations such as sampling

conformers and rotors are work in progress and are not included in the present work. The prototype of this package can be found in our GitHub repository.[54]

Results and Discussion:

Enthalpy Predictions. As shown in Fig. 1 and Fig. 2, we divided the 5,892 test molecules into six groups based on the constituent elements and the complexity of the ring structures. The simplest test species are small cyclic hydrocarbons containing single or double rings. Just like linear or branched hydrocarbons, the enthalpy of a cyclic hydrocarbon molecule is mainly determined by very basic molecular features, e.g., the types of bonds and the size of the ring in the molecule. Since the convolutional neural network is well capable of extracting these basic features,[24] the enthalpies of the cyclic hydrocarbons predicted by the model agree well with the reference DFT values (Fig. 1a and 2a). Incorporating oxygen atoms in the molecules slightly increases the diversity in chemical bonds. However, it does not affect the accuracy of prediction since the enthalpies of oxygenated cyclic species can also be well described by the model (Fig. 1b and 2b). As listed in Table 1, there is no significant difference between the RMSEs of the cyclic hydrocarbons and the cyclic oxygenates (2.15 and 1.93 kcal/mol), suggesting that the convolutional neural network is capable of adapting to more complex chemical units.

Further examination of the model was carried out for larger linear polycyclics and fused polycyclics containing more than two rings. The former refers to those polycyclic species that have no atoms residing in more than two rings; and the latter refers to those polycyclic species that have at least one atom shared by three or more rings. Because the ring strain of a polycyclic molecule is very sensitive to the configuration of the ring structure, polycyclic ring strain is often not additive, i.e., it does not equal the summation of the expected strain energies of individual rings.[21] Therefore, the convolutional neural network has to extract non-local geometry features to correctly predict the enthalpies of polycyclic species. This is possible with our architecture because we specify the number of rings of each size

containing a given atom. Fig. 1 and Fig. 2 show that most of the enthalpies of polycyclic molecules predicted by the model agree with the DFT reference values, with the best RMSE of 2.91 kcal/mol obtained for oxygenated large linear polycyclics (Table 1), suggesting that the convolutional neural network is capable of encoding non-local geometry information into molecular fingerprints. However, since the number of possible polycyclic structures is extremely large (especially when considering heteroatoms), the training samples encompass only a small fraction of all possible polycyclic configurations. A few outliers exist for the four polycyclic test sets considered here because roughly one fourth of the test molecules include polycyclic cores that are not present in the training set (Table 2), which highlights the necessity for a self-evolving strategy to ensure a model that is accurate for general use, since it is impossible to prepare a comprehensive training set covering the entire chemical space.

Although the enthalpies predicted by the model in general agree with the reference DFT values, they have not yet achieved the level of chemical accuracy, which is in thermochemistry understood as a 95% confidence limit of ± 1 kcal/mol.[55] As listed in Table 1, the 95% confidence interval of the overall test set is about 5 kcal/mol, suggesting that the current convolutional neural network model would not be able to achieve chemical accuracy in predicting actual enthalpy values even if the training data were exact. Since the inherent error associated with the DFT training data also affects the performance of the model on predicting actual enthalpies (as discussed in details in the subsection below), the present machine learning model should be viewed as a low-cost alternative to DFT for rapid screening of unimportant species in large-scale theoretical studies. For the key species in a chemical system, it is still recommended to resort to high-level quantum chemistry methods to derive the best possible enthalpy values.

Measure of Prediction Quality. The standard deviation of the ensemble predictions lie within the range of 0 to 3.75 kcal/mol for most test species, which is in line with the test set mean absolute error (MAE) of 1.74 kcal/mol as shown in Fig. 3a. The ensemble spread can be viewed as a descriptor of the

error distribution to which the prediction belongs. As shown in Fig. 3b, there is a clear positive correlation between the ensemble spread and the standard deviation of the error distribution; therefore, if the spread of ensemble prediction is small, the error distribution that the prediction belongs to should be narrow, which means there is a low probability for that prediction to have a large error. Moreover, because the ensemble spreads correlate with the standard deviations of the error distributions, one can divide the actual errors by the associated ensemble spreads to obtain a "standardized" error distribution as shown in Fig. 4. Although the standardized error distribution is not strictly normal (more weight in the tails compared to a normal distribution), Fig. 4 shows that assuming a normal distribution based on the ensemble spread is not too unreasonable. For instance, if the ensemble spread is 1 kcal/mol, there is a <5% chance that the error in the prediction is larger than 3 kcal/mol.

The underlying assumption of bootstrapping is that population statistics can be obtained from sample data by resampling the data set,[35] which is valid if the data set constitutes a good representation of the entire population. However, chemical space is extremely large so collecting a data set that represents the entirety of chemical space well is extraordinarily difficult. A practical issue is that users might extrapolate the model to an ill-represented molecule domain where both the model prediction and the prediction quality measured by ensemble spread are unreliable. To illustrate this point, we computed model statistics using a model trained using only $C_xH_yO_z$ molecules but a test set composed of 9,995 nitrogen-containing species and found the MAE to be 19.6 kcal/mol, which is much higher than the MAEs of the test hydrocarbons and oxygenates listed in Table 1. The standard deviations of ensemble predictions do not reflect the correct magnitude of the error since the ensemble spreads of the vast majority of test species are lower than 5 kcal/mol as shown in Fig. 5. This is not surprising because the model was only trained on molecules composed of C, H, and O atoms so it does not have any information about the strength of a chemical bond involving nitrogen, and the ensemble model used to measure the prediction qualities is also completely unaware of nitrogen.

However, even though there is no way to accurately quantify the error in the prediction of a molecule coming from an unseen domain, one can still use the spreads of ensemble prediction to identify the "foreign molecules" (at least to some extent). For example, as shown in Fig. 5, about 1.6% of the test hydrocarbons and oxygenates have ensemble spreads higher than 3 kcal/mol; however, roughly 10% of the samples in the nitrogen-containing test set exceed this level. Therefore, if one sets 3 kcal/mol to be the threshold of high-ensemble-spread species, the probability of categorizing a nitrogen-containing species as a high-ensemble-spread sample is about six times higher than that for a hydrocarbon or oxygenate. Although this ratio varies with the choice of cutoff value, for determining which predictions need to be refined, ensemble-spread-based selection should be more (or at least equally) effective compared to random selection since the percentage of nitrogen-containing examples above an level of ensemble spread is always greater than (or equal to) that of hydrocarbons and oxygenates.

Figure 5 also demonstrates that once the model is trained on a few samples of the new nitrogen-containing species, the model starts to recognize the new types of molecules, and thus begins to measure the quality of the predictions more accurately. If one adds 100 nitrogen-containing molecules to the training data, the percentage of the test nitrogen-containing molecules that exceed the 3 kcal/mol ensemble spread level increases to 35%, suggesting that the capacity of the model to identify the foreign molecules (nitrogen-containing species in this case) is significantly improved. Therefore, even though the underlying assumption of bootstrapping does not hold for species from an unseen domain, the ensemble scheme can still be used to select the points that need to be calibrated in an active learning scheme. Of course, including foreign species in the training set will not only facilitate better measurement of prediction quality, but will also improve the performance of the model for these molecules. One example of how the prediction accuracy and the spread of ensemble predictions evolve with the number of training data points will be discussed in the following subsection.

Active Learning Process. To demonstrate how the thermochemistry estimator adapts to a new

type of species, we prepared a dataset composed of 39,981 nitrogen-containing molecules to represent samples from an unseen molecular domain and examined how a model originally trained on hydrocarbons and oxygenates expands its predictive capacity to such unseen species using the active learning scheme. The active learning process starts with calculating the standard deviations of ensemble predictions of the nitrogen-containing molecules and then incorporating the high-spread samples in the training data to update the model. Since in practice there are costs associated with obtaining new training data, the cutoff between high and low ensemble spreads is a parameter that needs to be chosen to balance the costs and the requirements of accuracy in predictions. For this demonstration, the cutoff was set to 3 kcal/mol to mimic a setting that balances training efficiency and efforts for deriving new data. As listed in Table 3, 4,365 out of the 39,981 nitrogen-containing molecules were classified as high-spread species using Model 1, a model that was only trained on hydrocarbons and oxygenates. This is consistent with what we observed in Fig. 5 where roughly 10% of the estimated uncertainties of nitrogen-containing species calculated by a model that has only seen hydrocarbons and oxygenates exceed 3 kcal/mol.

Incorporating the high-spread species identified by Model 1 into the training set can improve the quality of estimated uncertainties. As listed in Table 3, the model generated by the second round of training (Model 2) identifies 7,296 high-spread samples because the prediction qualities of nitrogen-containing species are now estimated more accurately. One might expect more high-spread species to be found after the third round of training (Model 3). However, the number of high-spread molecules identified by Model 3 is significantly lower than in the previous two models because the accuracy of prediction has been improved for the nitrogen-containing species in the training process. As shown in Fig. 6, most of the low-spread species predicted by Model 3 do indeed have small errors (< 6 kcal/mol). Moreover, the test set MAE has decreased from 19.6 to 2.79 kcal/mol during this training process (Table 3), demonstrating that the model has successfully expanded its predictive capacity from hydrocarbons and oxygenates to nitrogen-containing molecules.

Automatic Generation of New Data. The active learning scheme combined with the convolutional neural network, which generalizes molecular feature extraction, provides a convenient framework for the development of a thermochemistry estimator that can be continuously and automatically improved. However, the remaining issue that has to be resolved is the procurement of new training data. Conventionally, collecting data for a machine learning model is a data mining problem. Lots of efforts have been made to develop tools that can be used to extract information published in literature.[56–61] However, the data available in the chemical literature is finite; for many molecules there are no data in the literature. To obtain new data, recent studies have tried to develop robotic platforms to coordinate many chemical experiments and generate data in real time.[62,63] Following the same philosophy, we use automatic quantum mechanical calculations for data generation in combination with the machine learning model to develop a self-evolving thermochemistry estimator as shown in Fig. 7. The self-evolving model is composed of three major components: a thermochemistry central database, a machine learning engine, and an automatic quantum mechanics calculator. The central database is responsible for hosting the information of all species with thermochemistry data as well as molecules without thermochemistry data but of potential interest submitted by users. The machine learning engine, which is the ensemble convolutional neural network discussed above, is responsible for predicting thermochemistry and identifying which of the species without data should be computed using quantum chemistry, based on the ensemble-spread analysis. The thermochemistry data generated by the automatic *ab initio* calculations are sent to the central database where they will serve as additional training examples for the next update of the machine learning engine. This effort enables the thermochemistry estimator to be improved automatically and continuously without the need for human involvement. Therefore, unlike the conventional group additivity approach, users of this model can apply it to species beyond the original training set without worrying about having to refine the model manually.

To examine the self-evolving scheme in real-life applications, we connected the thermochemistry

central database to the Reaction Mechanism Generator (RMG) software package.[18] RMG is a reaction modeling software that reacts a given set of species in all possible ways based on a set of reaction templates, estimates the reaction rates and thermodynamic properties of the reacting species, and simulates the time evolution of a batch reactor at the given reaction conditions. As a user of the thermochemistry estimator, RMG automatically sends molecules with high ensemble spread to the central database. For this work, RMG submitted many highly unsaturated polycyclics that are potentially important in soot formation chemistry, most of which are not well-represented in the original training data set of Ramakrishnan et al.[45] Therefore, the MAE of the new species as predicted by the base model (Model 1) is 24.4 kcal/mol as listed in Table 4. However, with a small amount of additional data generated by the automatic quantum chemistry calculations (~500 data points), MAE is reduced by a factor of two. As the automatic quantum chemistry calculations continue to run, making more training data available, it is expected that the error will continue to drop to values similar to those in Table 1. This exercise again demonstrates the importance of a continuously improving chemistry model since it is impossible to prepare a comprehensive training set covering all the species of potential interest.

Comparisons with High Level of Theory and Group Additivity Method. As discussed above, the reference enthalpies used in this work were calculated at the B3LYP/6-31G(2df,p) level of theory with the RRHO approximation, which are themselves associated with errors. Previous studies have shown that B3LYP energies are often associated with significant errors, primarily due to the absence of long-range dispersion interaction.[33,46–48] Moreover, the RRHO model ignores the effect of rotors and floppy motions, which is known to affect the accuracy of thermochemistry calculations.[16] The RRHO model also fails to describe many features of vibrational spectroscopy for high frequency modes so that zero point energies derived from harmonic frequencies are often down-scaled.[64,65] These errors are inevitably inherited by the models and hence influence the prediction accuracy.

To examine the performance of the model on predicting true enthalpies, we calculated the

enthalpies of 98 randomly selected cyclic and polycyclic molecules at the CCSD(T)-F12/cc-pVTZ-F12//B3LYP/6-31G(2df,p) level of theory using Molpro.[66–68] Though the geometry optimizations and frequency calculations were still carried out at the B3LYP/6-31G(2df,p) level of theory with RRHO approximations, the zero point energies and the frequencies were scaled by a factor of 0.965 to partially include anharmonic effects.[69] As listed in Table 5, the enthalpies predicted by the neural network model agrees with CCSD(T)-F12/cc-pVTZ-F12 enthalpy values with an RMSE of 3.35 kcal/mol, which is slightly larger than that of the values of B3LYP/6-31G(2df,p) (2.77 kcal/mol). We also benchmarked the group additivity method implemented in the RMG software package,[18] which includes the polycyclic ring strain corrections based on a bicyclic decomposition scheme developed by Han and others.[21] As listed in Table 5, the RMSE of the group additivity method is about 11 kcal/mol, which is roughly three times higher than that of the neural network model. This finding supports the argument that the contributions of bicyclic ring strain are not always independent and additive so that it is important to include higher order corrections or to use a nonlinear approach such as the convolutional neural network model to better describe the ring strain of polycyclic species.

Conclusions:

Although machine learning algorithms have tremendous potential for enhancing chemical simulations, building a reliable molecule-based model is not an easy task because generating accurate chemistry data is often expensive and time-consuming. Since most chemistry datasets only cover a small region of chemical space, data scarcity is often the primary factor limiting the accuracy and the scope of applicability of a data-driven model. To address this issue, we integrated an active learning model with automatic *ab initio* calculations to create a self-evolving machine that can continuously adapt to new species of interest. We implemented a molecular graph convolutional neural network to generalize feature extraction for a broad range of species of potential interest. To enable active learning,

a non-parametric ensemble approach was developed by combining the ideas of bootstrap sampling and dropout training in neural networks to estimate prediction qualities. Therefore, the model can easily identify molecules for which predictions need to be refined based on the spread of ensemble predictions. Our self-evolving machine automatically launches *ab initio* calculations to obtain accurate properties for such molecules to improve the machine learning model.

We examined the idea of a self-evolving machine by modeling the formation enthalpies of polycyclic species. Because the ring strain of a polycyclic molecule is often not additive and very sensitive to details of the fused ring structure, it is difficult to estimate the enthalpy of a polycyclic molecule using local chemical features, such as the types of bonds around each atom. Convolutional fingerprint and convolution methods based on local and near-neighbor properties are unable to give accurate predictions. However, we demonstrate here that if ring information is included as an atom property, the convolutional neural network approach works well. We found that for a test set composed of 5,892 cyclic and polycyclic molecules, the enthalpies determined from our neural network model agree to within an RMSE of 2.62 kcal/mol with reference values. This suggests potential application of molecular graph convolutions including ring information to challenging chemistry tasks that require detailed descriptions of molecular structures.

The reliability of the dropout ensemble approach on measuring prediction quality depends on the validity of the assumption of bootstrap sampling. For molecules drawn from the same domain as the training species, the spread of ensemble predictions can be interpreted as a descriptor for the probability distribution of the prediction error. On the other hand, for examples drawn from a completely unseen domain, there is no rigorous way to correctly quantify the error in a prediction. However, we found that with a properly chosen cutoff criterion, the ensemble-model-based active learning approach is still more effective than random selection for new types of molecules. We demonstrated that an active learning scheme can broaden the applicability of a model originally trained on hydrocarbons and oxygenates to

nitrogen-containing species, suggesting that the self-evolving strategy presented in this work is not only able to improve prediction accuracy but is also capable of expanding the scope of applicability to completely unseen species domains. Here, we demonstrated the effectiveness of combining machine learning with automatic quantum chemistry for predicting enthalpies of formation and the uncertainties in those predictions. We expect this approach will also be effective for automatically constructing models for predicting many other molecular properties which can be computed using quantum chemistry.

The DFT data used here as reference values were calculated at the B3LYP/6-31G(2df,p) level of theory and are themselves associated with significant uncertainty. As a result, the fidelity of the model is affected by the inaccuracies in the DFT data. However, despite the uncertainty in training data, the convolutional neural network model outperforms the group additivity method on enthalpy predictions for polycyclic species (Table 5) due to a better description of the configuration of the ring structure. To overcome the limitation of data accuracy, future work will involve incorporating the best available experimental results, employing more accurate quantum chemistry methods to generate high-quality data, and improving the model accuracy using a transfer learning approach. Another useful extension would be to broaden prediction coverage to radicals and other open-shell species, which are not currently considered in the model.

Supporting Information Available

Additional information as noted in the text, including descriptions of input feature vectors, formulas for enthalpy calculations, training and testing data, computed quantum chemistry properties and optimized molecular geometries. This information is available free of charge via the Internet at http://pubs.acs.org/.

Acknowledgement:

This work was supported by the MIT Consortium for Machine Learning for Pharmaceutical Discovery and Synthesis, and by the DARPA Make-It program under contract ARO W911NF-16-2-0023. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

References:

(1) Bell, A. T.; Head-Gordon, M. Quantum Mechanical Modeling of Catalytic Processes. *Annu. Rev. Chem. Biomol. Eng.* **2011**, *2* (1), 453–477. https://doi.org/10.1146/annurev-chembioeng-061010-114108.

(2) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. – Eur. J.* **2012**, *18* (32), 9955–9964. https://doi.org/10.1002/chem.201200497.

(3) Li, Y.-P.; Head-Gordon, M.; Bell, A. T. Analysis of the Reaction Mechanism and Catalytic Activity of Metal-Substituted Beta Zeolite for the Isomerization of Glucose to Fructose. *ACS Catal.* **2014**, *4* (5), 1537–1545. https://doi.org/10.1021/cs401054f.

(4) Li, Y.-P.; Head-Gordon, M.; Bell, A. T. Computational Study of P-Xylene Synthesis from Ethylene and 2,5-Dimethylfuran Catalyzed by H-BEA. *J. Phys. Chem. C* **2014**, *118* (38), 22090–22095. https://doi.org/10.1021/jp506664c.

(5) Li, Y.-P.; Gomes, J.; Mallikarjun Sharada, S.; Bell, A. T.; Head-Gordon, M. Improved Force-Field Parameters for QM/MM Simulations of the Energies of Adsorption for Molecules in Zeolites and a Free Rotor Correction to the Rigid Rotor Harmonic Oscillator Model for Adsorption Enthalpies. *J. Phys. Chem. C* **2015**, *119* (4), 1840–1850. https://doi.org/10.1021/jp509921r.

(6) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab Initio Thermochemistry. *J. Chem. Phys.* **2004**, *121* (23), 11599–11613. https://doi.org/10.1063/1.1811608.

(7) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-KJ/Mol Predictions. *J. Chem. Phys.* **2006**, *125* (14), 144108. https://doi.org/10.1063/1.2348881.

(8) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. W3 Theory: Robust Computational Thermochemistry in the KJ/Mol Accuracy Range. *J. Chem. Phys.* **2004**, *120* (9), 4129–4141. https://doi.org/10.1063/1.1638736.

(9) Martin, J. M. L.; de Oliveira, G. Towards Standard Methods for Benchmark Quality Ab Initio Thermochemistry—W1 and W2 Theory. *J. Chem. Phys.* **1999**, *111* (5), 1843–1856. https://doi.org/10.1063/1.479454.

(10) Feller, D.; Peterson, K. A.; Dixon, D. A. Further Benchmarks of a Composite, Convergent, Statistically Calibrated Coupled-Cluster-Based Approach for Thermochemical and Spectroscopic Studies. *Mol. Phys.* **2012**, *110* (19–20), 2381–2399. https://doi.org/10.1080/00268976.2012.684897.

(11) Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical Accuracy in Ab Initio Thermochemistry and Spectroscopy: Current Strategies and Future Challenges. *Theor. Chem. Acc.* **2012**, *131* (1), 1079. https://doi.org/10.1007/s00214-011-1079-5.

(12) Feller, D.; Peterson, K. A. An Expanded Calibration Study of the Explicitly Correlated CCSD(T)-F12b Method Using Large Basis Set Standard CCSD(T) Atomization Energies. *J. Chem. Phys.* **2013**, *139* (8), 084110. https://doi.org/10.1063/1.4819125.

(13) Császár, A. G.; Allen, W. D.; Schaefer, H. F. In Pursuit of the Ab Initio Limit for Conformational Energy Prototypes. *J. Chem. Phys.* **1998**, *108* (23), 9751–9764. https://doi.org/10.1063/1.476449.

(14) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126* (8), 084108. https://doi.org/10.1063/1.2436888.

(15) Keçeli, M.; Elliott, S. N.; Li, Y.-P.; Johnson, M. S.; Cavallotti, C.; Georgievskii, Y.; Green, W. H.; Pelucchi, M.; Wozniak, J. M.; Jasper, A. W.; et al. Automated Computational

Thermochemistry for Butane Oxidation: A Prelude to Predictive Automated Combustion Kinetics. *Proc. Combust. Inst.* **2018**. https://doi.org/10.1016/j.proci.2018.07.113.
(16) Li, Y.-P.; Bell, A. T.; Head-Gordon, M. Thermodynamics of Anharmonic Systems: Uncoupled Mode Approximations for Molecules. *J. Chem. Theory Comput.* **2016**, *12* (6), 2861–2870. https://doi.org/10.1021/acs.jctc.5b01177.
(17) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69* (3), 279–324. https://doi.org/10.1021/cr60259a002.
(18) 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. https://doi.org/10.1016/j.cpc.2016.02.013.
(19) Ritter, E. R.; Bozzelli, J. W. THERM: Thermodynamic Property Estimation for Gas Phase Radicals and Molecules. *Int. J. Chem. Kinet.* **1991**, *23* (9), 767–778. https://doi.org/10.1002/kin.550230903.
(20) Lay, T. H.; Yamada, T.; Tsai, P.-L.; Bozzelli, J. W. Thermodynamic Parameters and Group Additivity Ring Corrections for Three- to Six-Membered Oxygen Heterocyclic Hydrocarbons. *J. Phys. Chem. A* **1997**, *101* (13), 2471–2477. https://doi.org/10.1021/jp9629497.
(21) 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* (4), 294–303. https://doi.org/10.1002/kin.21158.
(22) He, T.; Li, S.; Chi, Y.; Zhang, H.-B.; Wang, Z.; Yang, B.; He, X.; You, X. An Adaptive Distance-Based Group Contribution Method for Thermodynamic Property Prediction. *Phys. Chem. Chem. Phys.* **2016**, *18* (34), 23822–23830. https://doi.org/10.1039/C6CP02929A.
(23) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2224–2232.
(24) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57* (8), 1757–1772. https://doi.org/10.1021/acs.jcim.6b00601.
(25) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301.
(26) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203. https://doi.org/10.1039/C6SC05720A.
(27) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401. https://doi.org/10.1103/PhysRevLett.98.146401.
(28) Khorshidi, A.; Peterson, A. A. Amp: A Modular Approach to Machine Learning in Atomistic Simulations. *Comput. Phys. Commun.* **2016**, *207*, 310–324. https://doi.org/10.1016/j.cpc.2016.05.010.
(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(30) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv:1509.09292* **2015**.
(31) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions:

Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30* (8), 595–608. https://doi.org/10.1007/s10822-016-9938-8.
(32) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.
(33) Simm, G. N.; Reiher, M. Systematic Error Estimation for Chemical Reaction Energies. *J. Chem. Theory Comput.* **2016**, *12* (6), 2762–2773. https://doi.org/10.1021/acs.jctc.6b00318.
(34) Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14* (10), 5238–5248. https://doi.org/10.1021/acs.jctc.8b00504.
(35) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; CRC Press, 1994.
(36) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* **2014**, *15* (1), 1929–1958.
(37) Proppe, J.; Reiher, M. Reliable Estimation of Prediction Uncertainty for Physicochemical Property Models. *J. Chem. Theory Comput.* **2017**, *13* (7), 3297–3317. https://doi.org/10.1021/acs.jctc.7b00235.
(38) Settles, B. Active Learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2012**, *6* (1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018.
(39) A. Peterson, A.; Christensen, R.; Khorshidi, A. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19* (18), 10978–10985. https://doi.org/10.1039/C7CP00375G.
(40) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241733. https://doi.org/10.1063/1.5023802.
(41) *DataDrivenEstimator: A Package of Data Driven Estimators for Thermochemistry and Kinetics*; https://github.com/ReactionMechanismGenerator/DataDrivenEstimator, 2018.
(42) Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Networks. *arXiv:1505.05424* **2015**.
(43) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142* **2015**.
(44) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140. https://doi.org/10.1007/BF00058655.
(45) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. von. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022. https://doi.org/10.1038/sdata.2014.22.
(46) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112* (1), 289–320. https://doi.org/10.1021/cr200107z.
(47) Mardirossian, N.; Head-Gordon, M. ΩB97X-V: A 10-Parameter, Range-Separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-Fittest Strategy. *Phys. Chem. Chem. Phys.* **2014**, *16* (21), 9904–9924. https://doi.org/10.1039/C3CP54374A.
(48) Proppe, J.; Husch, T.; Simm, G. N.; Reiher, M. Uncertainty Quantification for Quantum Chemical Models of Complex Reaction Networks. *Faraday Discuss.* **2017**, *195* (0), 497–520. https://doi.org/10.1039/C6FD00144K.
(49) Simm, G. N.; Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **2017**, *13* (12), 6108–6119. https://doi.org/10.1021/acs.jctc.7b00945.
(50) Keçeli, M.; Elliott, S. N.; Li, Y.-P.; Johnson, M. S.; Cavallotti, C.; Georgievskii, Y.; Green, W. H.; Pelucchi, M.; Wozniak, J. M.; Jasper, A. W.; et al. Automated Computational

Thermochemistry for Butane Oxidation: A Prelude to Predictive Automated Combustion Kinetics. *Proc. Combust. Inst.* **2018**. https://doi.org/10.1016/j.proci.2018.07.113.
(51) *Rdkit: The Official Sources for the RDKit Library*; https://github.com/rdkit/rdkit, 2018.
(52) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113* (2), 184–215. https://doi.org/10.1080/00268976.2014.952696.
(53) *MongoDB: Open Source Document Database*; https://www.mongodb.com/, 2018.
(54) *AutoQM: A Software for Large Scale Quantum Mechanic Calculations*; https://github.com/ReactionMechanismGenerator/autoQM, 2018.
(55) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *Int. J. Quantum Chem.* **2014**, *114* (17), 1097–1101. https://doi.org/10.1002/qua.24605.
(56) Tchoua, R. B.; Chard, K.; Audus, D.; Qin, J.; de Pablo, J.; Foster, I. A Hybrid Human-Computer Approach to the Extraction of Scientific Facts from the Literature. *Procedia Comput. Sci.* **2016**, *80*, 386–397. https://doi.org/10.1016/j.procs.2016.05.338.
(57) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.
(58) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *J. Cheminformatics* **2011**, *3*, 17. https://doi.org/10.1186/1758-2946-3-17.
(59) Rocktäschel, T.; Weidlich, M.; Leser, U. ChemSpot: A Hybrid System for Chemical Named Entity Recognition. *Bioinforma. Oxf. Engl.* **2012**, *28* (12), 1633–1640. https://doi.org/10.1093/bioinformatics/bts183.
(60) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29* (21), 9436–9444. https://doi.org/10.1021/acs.chemmater.7b03500.
(61) Lewinski, N. A.; McInnes, B. T. Using Natural Language Processing Techniques to Inform Research on Nanotechnology. *Beilstein J. Nanotechnol.* **2015**, *6* (1), 1439–1449. https://doi.org/10.3762/bjnano.6.149.
(62) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *figshare* **2018**. https://doi.org/10.26434/chemrxiv.5953606.v1.
(63) Caramelli, D.; Salley, D.; Henson, A.; Camarasa, G. A.; Sharabi, S.; Keenan, G.; Cronin, L. Networking Chemical Robots Using Twitter for #RealTimeChem. *figshare* **2018**. https://doi.org/10.26434/chemrxiv.5952655.v1.
(64) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree−Fock, Møller−Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100* (41), 16502–16513. https://doi.org/10.1021/jp960976r.
(65) Andersson, M. P.; Uvdal, P. New Scale Factors for Harmonic Vibrational Frequencies Using the B3LYP Density Functional Method with the Triple-ζ Basis Set 6-311+G(d,P). *J. Phys. Chem. A* **2005**, *109* (12), 2937–2941. https://doi.org/10.1021/jp045733a.
(66) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: A General-Purpose Quantum Chemistry Program Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (2), 242–253. https://doi.org/10.1002/wcms.82.
(67) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; and others. *MOLPRO*,

Version 2015.1, a Package of Ab Initio Programs; molpro, 2015.
(68) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127* (22), 221106. https://doi.org/10.1063/1.2817618.
(69) National Institute of Standards and Technology (NIST), Precomputed vibrational scaling factors https://cccbdb.nist.gov/vibscalejust.asp (accessed Dec 16, 2018).

Table 1. Statistical errors of enthalpies shown in Fig. 1.

	MSE (kcal/mol)[a]		MAE (kcal/mol)[a]		RMSE (kcal/mol)[a]		$u_{95\%}$ (kcal/mol)[a]	
	C_xH_y	$C_xH_yO_z$	C_xH_y	$C_xH_yO_z$	C_xH_y	$C_xH_yO_z$	C_xH_y	$C_xH_yO_z$
Small Cyclics	-0.15	-0.09	1.40	1.41	2.15	1.93	4.29	3.85
Large Linear Polycyclics	-0.93	-0.45	3.65	2.08	5.57	2.91	11.05	5.75
Large Fused Polycyclics	-0.05	-0.56	2.73	2.77	3.75	4.08	7.51	8.10
Overall	-0.21		1.74		2.62		5.22	

[a] MSE: mean signed error, MAE: mean absolute error, RMSE: root mean square error, $u_{95\%}$: 95% confidence limit (two standard deviations of the error distribution)[55].

Table 2. Percentage of test molecules containing ring core structures that are not present in the training set.

	Hydrocarbon	Oxygenates
Small Cyclics	3.3%	1.8%
Large Linear Polycyclics	27.7%	21.8%
Large Fused Polycyclics	27.8%	28.2%

Table 3. Active learning of a thermochemistry estimator.

		Model 1	Model 2	Model 3
Number of training samples	Hydrocarbons and oxygenates	23,582	23,582	23,582
	Nitrogen-containing species	0	4,365	11,661
Number of low and high prediction spread samples[a]	High spread	4,365	7,296	1,201
	Low spread	35,616	28,320	27,119
Statistical errors (kcal/mol)[b]	MSE[c]	-16.84	1.02	0.39
	MAE[c]	19.60	4.53	2.79
	RMSE[c]	25.35	6.03	3.77
	$u_{95\%}$[c]	37.88	11.89	7.51

[a] Uncertainties of the molecules in a dataset composed of 39,981 nitrogen-containing cyclic and polycyclic molecules. Samples with high ensemble prediction spreads detected by a model are removed from the dataset and added to the training set to upgrade the model. The cutoff between low and high ensemble spread is 3 kcal/mol.
[b] Model performance on an independent test set composed of 9,995 nitrogen-containing species
[c] MSE: mean signed error, MAE: mean absolute error, RMSE: root mean square error, $u_{95\%}$: 95% confidence limit (two standard deviations of the error distribution)[55].

Table 4. Statistical errors of polycyclic species generated by RMG.

	Training Samples	MSE (kcal/mol)[c]	MAE (kcal/mol)[c]	RMSE (kcal/mol)[c]	$u_{95\%}$ (kcal/mol)[c]
Base Model[a]	Data set of Ramakrishnan et al.[b]	-21.50	24.35	32.25	48.24
Improved Model	Data set of Ramakrishnan et al. and 501 additional data points generated by automatic *ab initio* calculations	-6.05	12.14	18.04	34.10

[a] Model 1 in Table 3.

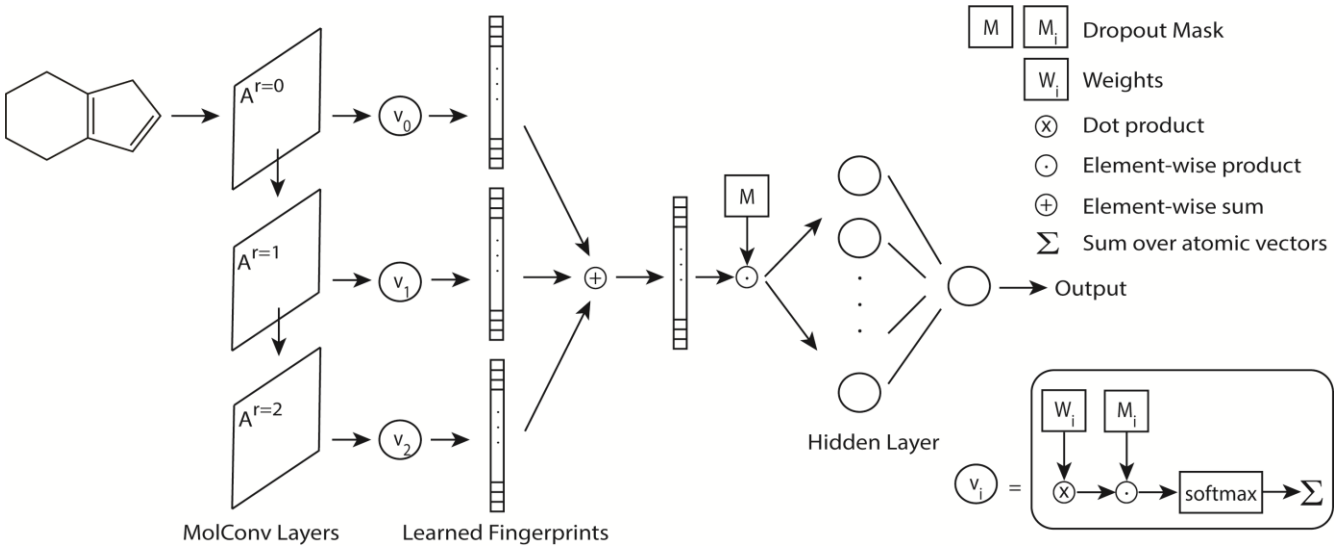
[b] 23,582 cyclic and polycyclic hydrocarbons and oxygenates.

[c] Test set is composed of 157 molecules randomly selected from RMG generated polycyclic species. MSE: mean signed error, MAE: mean absolute error, RMSE: root mean square error, $u_{95\%}$: 95% confidence limit (two standard deviations of the error distribution)[55].

Table 5. Statistical errors of enthalpy calculations.[a]

	MSE (kcal/mol)[b]	MAE (kcal/mol)[b]	RMSE (kcal/mol)[b]	$u_{95\%}$ (kcal/mol)[b]
B3LYP/6-31G(2df,p)[c]	0.29	2.21	2.77	5.55
Conv Neural Network[d]	0.18	2.61	3.35	6.72
Group Additivity[e]	-2.16	6.63	10.83	21.33

[a] Benchmark data are the enthalpies of 98 randomly selected cyclic and polycyclic molecules calculated at the CCSD(T)-F12/cc-pVTZ-F12 level of theory
[b] MSE: mean signed error, MAE: mean absolute error, RMSE: root mean square error, $u_{95\%}$: 95% confidence limit (two standard deviations of the error distribution)[55].
[c] Data extracted from the work of Ramakrishnan et al[45]
[d] Model 1 in Table 3.
[e] The group additivity method implemented in the RMG software package[18]



Scheme 1. Architecture of the molecular convolutional neural network. The input molecular matrix ($A^{r=0}$) is composed of a list of atomic feature vectors and the molecular matrices with larger radius ($A^{r=1}$ and $A^{r=2}$) are derived by combining feature vectors of connected atoms and the corresponding bond information. The molecular matrices are passed through learned mappings and then are summed over atoms to get the one-dimensional molecular fingerprints. Detailed descriptions of the convolution procedure can be found in previous work.[23,31]

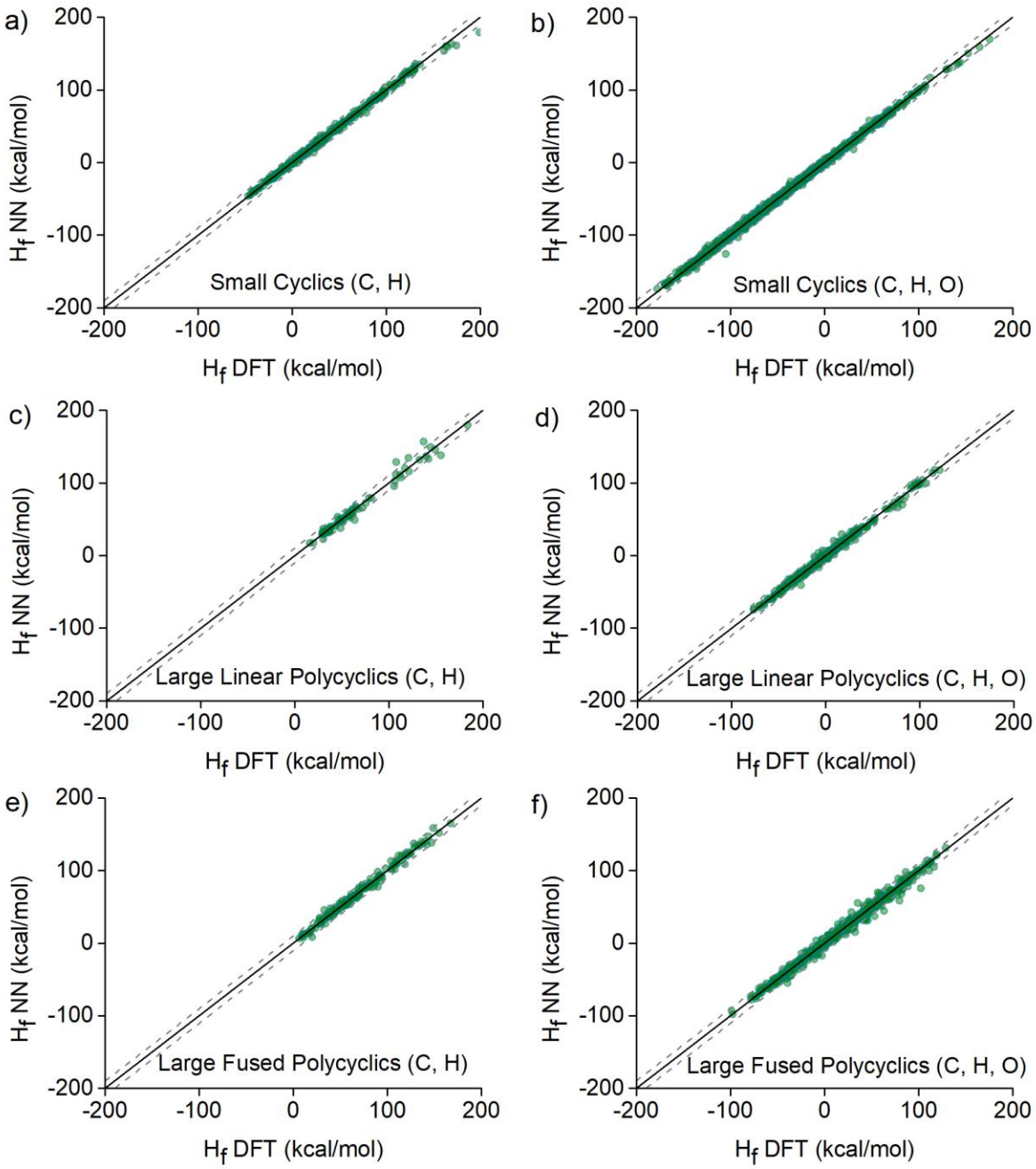


Figure 1. Parity plots of the reference formation enthalpies (DFT) and the values predicted by the convolutional neural network (NN). The left and the right panels are the results of hydrocarbon (C, H) and oxygenated (C, H, O) test samples, respectively. An error bar of $\pm 10\ kcal/mol$ is shown by the dashed lines.

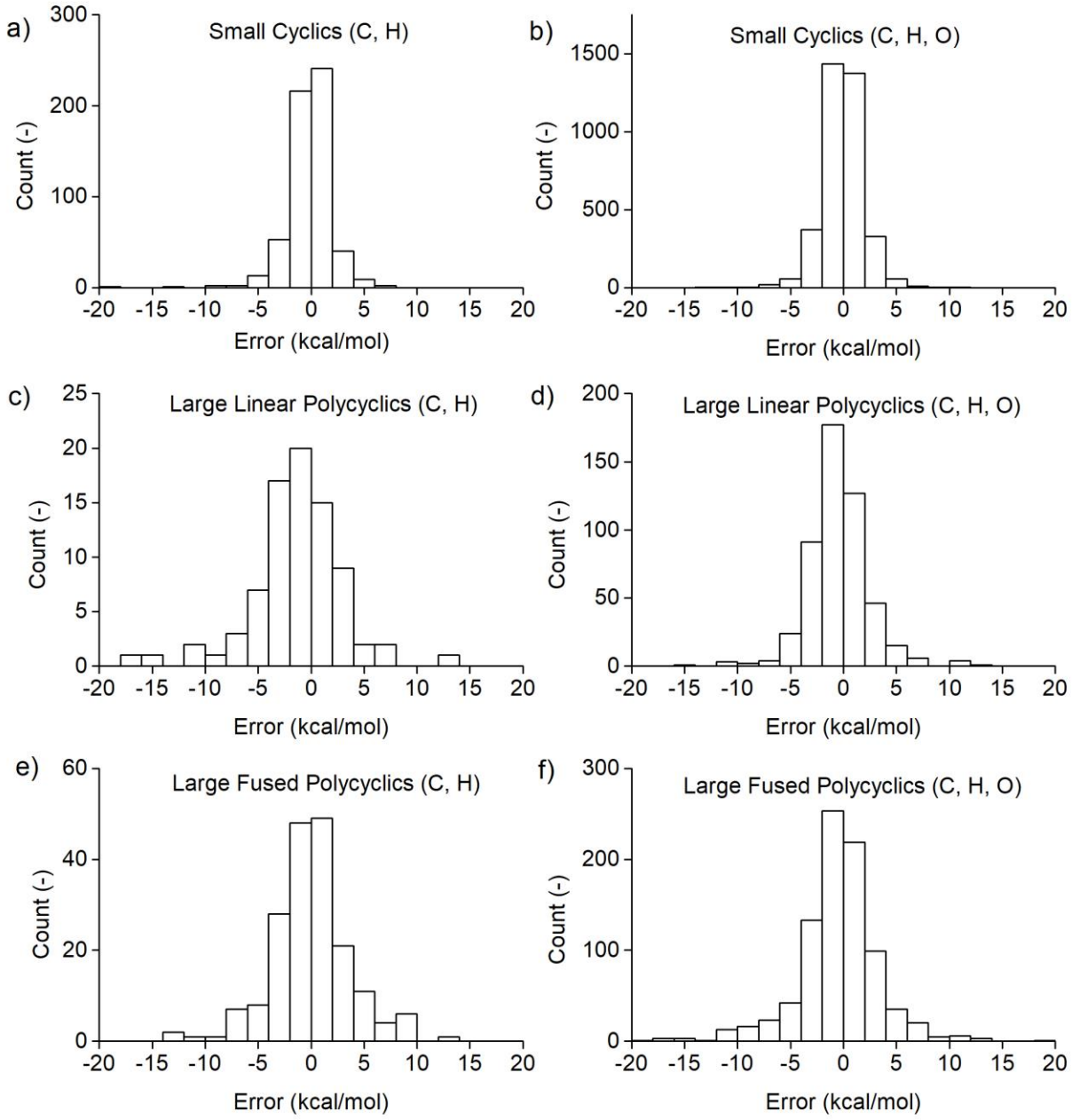


Figure 2. Error distributions of the test sets. The errors are defined as the difference between the formation enthalpies predicted by the convolutional neural network and the reference DFT values. The left and the right panels are the results of hydrocarbon (C, H) and oxygenated (C, H, O) test samples, respectively.

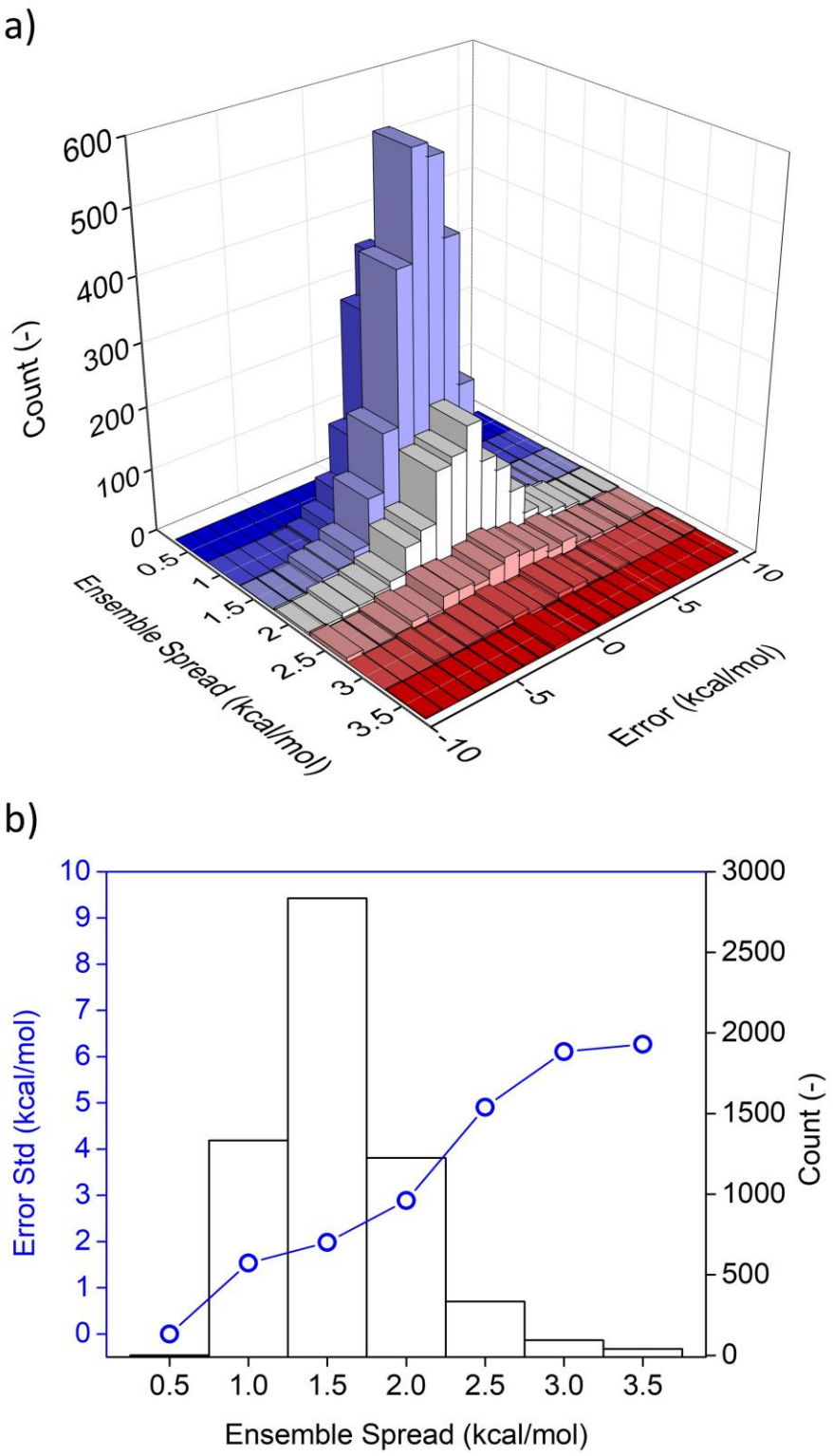


Figure 3. Errors and standard deviations of enthalpies predicted by the ensemble model (ensemble spread). The first panel, (a), shows that the predictions with higher ensemble spreads tend to have a broader error distribution. This observation can be confirmed by the second panel, (b), which shows a positive correlation between the ensemble spreads and the standard deviations of the error distributions.

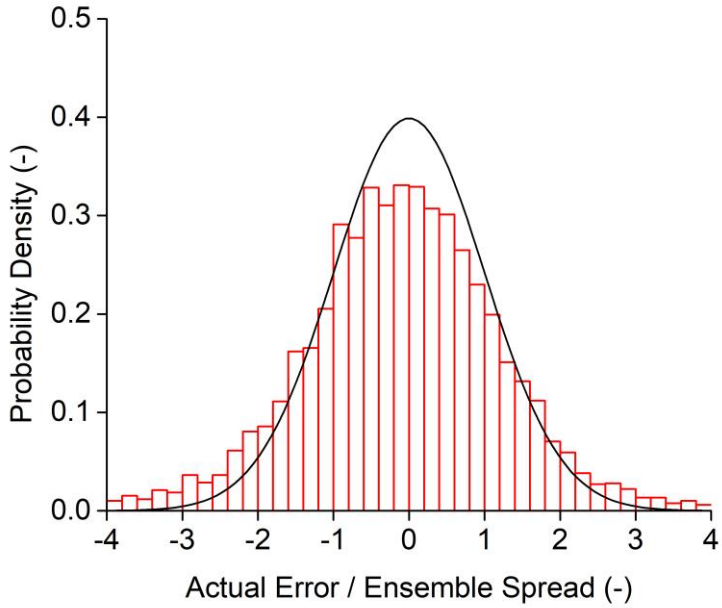


Figure 4. Distribution of standardized error (actual error/ensemble spread). About 59%, 87%, and 96% of the values fall in the range of ± 1, ± 2, and ± 3, respectively. The black curve is a standardized normal distribution for reference.

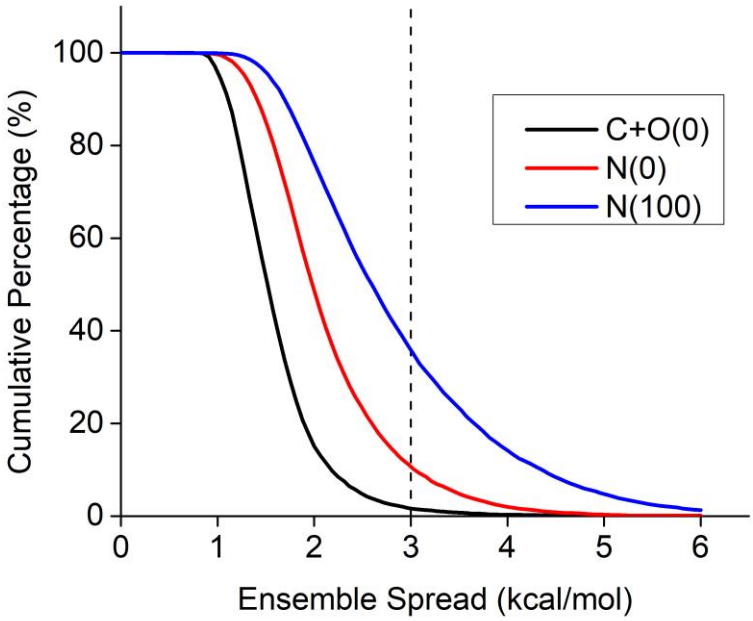


Figure 5. Cumulative percentage of test molecules predicted to lie above a certain level of ensemble spread. The black curve, C+O(0), is the combined result of the hydrocarbon and oxygenate test sets listed in Table 1. The red and blue curves, N(0) and N(100), are results of a test set composed of 9,995 nitrogen-containing species. The numbers in the parentheses correspond to the number of nitrogen-containing species in the training data for each model.

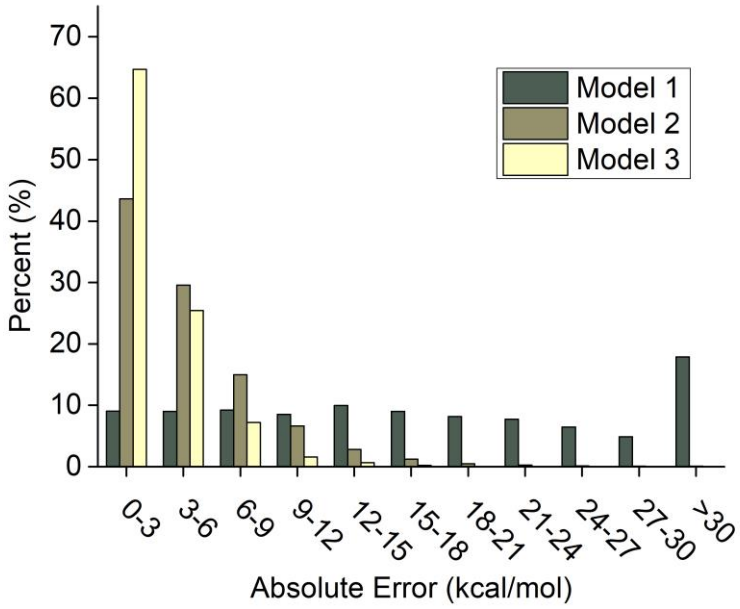


Figure 6. Error distributions of the species estimated by the models to have ensemble spread < 3 kcal/mol. As the model is improved by adding training data, it both becomes more accurate and more reliably identifies which predictions are uncertain.

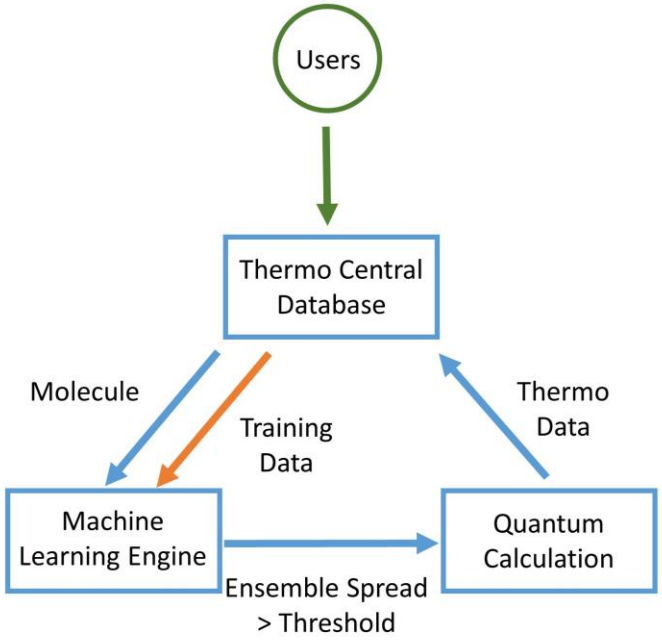
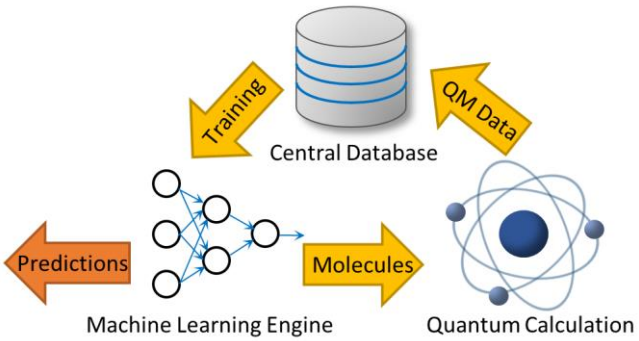


Figure 7. Schematic of the self-evolving thermochemistry estimator. The program spawns quantum calculations for species if identifies as uncertain, then adds the newly computed data to the training set, continuously improving the model.

TOC Graphic


Training
QM Data
Central Database
Predictions
Molecules
Machine Learning Engine
Quantum Calculation