Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach

Citation: Grambow, Colin A. et al. "Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach." Journal of Physical Chemistry A 123, 27 (June 2019): 5826-5835 © 2019 American Chemical Society

As Published: http://dx.doi.org/10.1021/acs.jpca.9b04195

Publisher: American Chemical Society (ACS)

Persistent URL: https://hdl.handle.net/1721.1/123828

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

Accurate Thermochemistry with Small Datasets: A Bond Additivity Correction and Transfer Learning Approach

Colin A. Grambow, Yi-Pei Li, and William H. Green*

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

E-mail: whgreen@mit.edu

Phone: +1-617-253-4580

Abstract

Machine learning provides promising new methods for accurate yet rapid prediction of molecular properties, including thermochemistry, which is an integral component of many computer simulations, particularly automated reaction mechanism generation. Often, very large datasets with tens of thousands of molecules are required for training the models, but most datasets of experimental or high-accuracy quantum mechanical quality are much smaller. To overcome these limitations, we calculate new high-level datasets and derive bond additivity corrections to significantly improve enthalpies of formation. We adopt a transfer learning technique to train neural network models that achieve good performance even with a relatively small set of high-accuracy data. The training data for the entropy model is carefully selected so that important conformational effects are captured. The resulting models are generally applicable thermochemistry predictors for organic compounds with oxygen and nitrogen heteroatoms that approach experimental and coupled cluster accuracy while only requiring molecular graph inputs. Due to their versatility and the ease of adding new training data, they are poised to replace conventional estimation methods for thermochemical parameters in reaction mechanism generation. Since high-accuracy data is often sparse, similar transfer learning approaches are expected to be useful for estimating many other molecular properties.

Introduction

Rapid and accurate estimation of molecular properties is a vital component of many chemistry and materials science applications.[1,2] In particular, automated reaction mechanism generation requires estimates of chemical kinetic rates and molecular thermochemistry, which is typically used to calculate reverse reaction rates from the relationship between the Gibbs free energy change of a reaction and the equilibrium constant.[3,4] The temperature-dependent Gibbs free energy of reaction can be computed from the enthalpies of formation, entropies, and heat capacities if one has means for accurately predicting those molecular properties.

In an ideal world, thermochemical properties for all relevant chemical species would be obtained from experiments or high-quality electronic structure calculations. Realistically, the cost associated with obtaining data for each species in this manner would be tremendous because the process of generating a large reaction mechanism may involve more than a million distinct species (including species in reactions later determined to be numerically negligible). An alternative method proposed several decades ago and still in use today is the group additivity method, which decomposes each molecule into groups and sums up the thermochemical contributions from each group. The group values were originally derived from a regression on experimental data,[5] but today most group values are derived from quantum chemistry calculations.[6–8] Group additivity can be applied very rapidly to large numbers of molecules and can provide highly accurate results for some classes of molecules. For example, the thermochemistry of hydrocarbons without cycles is predicted particularly well,[9] but more exotic species, especially heteroatom-containing and fused cyclic compounds, are ill-suited for the group additivity method. Careful collection of new data, manual selection of new groups, and a renewed fitting procedure are required every time incompatible species are encountered, although there have been some efforts towards automatic group selection for a limited set of molecules.[10]

Alternatives are provided by the much more flexible frameworks arising in the field of machine learning. A host of new machine learning methods, especially deep learning methods,

have become available for not just the classical areas of computer vision and natural language processing[11] but also for chemical property predictions.[12–18] Machine learning methods can easily be applied to different chemical domains and training a model that is useful across a broad range of chemistry does not require significant manual engineering of features. The downside is that most methods require very large molecular datasets for training, which are usually only available at low levels of theory.[19] In addition, machine learning models most often treat molecules as rigid structures or graphs, even though effects due to different conformers, especially for entropy, are important in reality.[20]

To overcome the limitation of dataset size, a common technique in machine learning called transfer learning can be employed, in which knowledge learned by training in one domain is transferred to a second domain.[21] In this context, the first domain is a large quantity of low-level density functional theory (DFT) calculations and the second domain is a much smaller collection of thermochemical data from experiments and high-quality quantum mechanical calculations. The information gained from training on a wide array of chemistry greatly enhances the ability to learn from the limited amount of high-level data. Transfer learning and a related technique, Δ-machine learning, have already been successfully employed for energy predictions of molecular geometries,[22,23] and the benefit of transfer learning has been explored across many molecular datasets.[24]

Because high-accuracy data are scarce, our first goal is to construct an enthalpy of formation dataset composed of high-quality explicitly correlated coupled cluster calculations. We also constructed entropy and heat capacity datasets using high-quality DFT calculations. Moreover, we wish to further improve the quality of the enthalpy data by deriving bond additivity corrections (BACs), which are a simple method to correct for systematic errors in energy calculations of electronic structure methods.[25] After supplementing the new datasets with experimental data, the second goal is to train transfer learning models that leverage both existing large low-quality datasets and the newly created, but much smaller, high-quality datasets to obtain models that yield predictions of high accuracy. Furthermore,

we aim to create an entropy prediction model capable of accounting for conformational effects by carefully selecting its training data. The ultimate goal is to use the models as part of the Reaction Mechanism Generator (RMG) software[3] and to replace its group additivity scheme.

Computational Details

Transfer Learning

Transfer learning is a frequently used technique in the machine learning community in which knowledge learned by a model on some task is applied to a different task. Often, a lot of data is available for simpler prediction tasks while only limited data exists in related domains of interest. In the context of molecular property prediction, obtaining large amounts of training data using low-level DFT calculations is a straightforward task, but compiling large sets of wave function-based calculations is associated with significantly higher cost. A transfer learning model in this realm is initially trained on low-level DFT calculations and subsequently refined using the limited high-accuracy data.

A schematic of the complete model used here is shown in Figure 1. The molecular representation and the foundation for the model are based on models used in the studies by Li et al.[26] and by Coley et al.,[18] both of which are based on so-called graph convolutions.[27,28] We refer the interested reader to the descriptions provided in those papers and will limit ourselves here to a concise explanation with a more in-depth treatment of the transfer learning module. Molecules are represented as labeled undirected graphs $\mathcal{M} = (\mathcal{A}, \mathcal{B})$, which are ordered pairs of vertices $\mathcal{A}$ corresponding to the atoms and edges $\mathcal{B}$ corresponding to the bonds. A bond is then given by the unordered pair of atoms $\{x, y\} \in \mathcal{B}$ with $x, y \in \mathcal{A}$. Each atom $a \in \mathcal{A}$ is associated with an atom feature vector $\mathbf{f}_a$ which aggregates the following descriptors: atomic number, the number of non-hydrogen neighbors (heavy atoms), the number of hydrogen neighbors, and ring membership. Similarly, each bond $\{a, y\} \in \mathcal{B}$ is associated with a bond

feature vector $\mathbf{f}_{ay}$ only containing information about ring membership. Conventional models might use the bond order and aromaticity indicators as additional features, but these were not included here because the model was found to perform equally well without. This also removes the requirement of selecting specific resonance structures to train on. The ring membership descriptor counts how many rings of each size an atom or a bond is part of. Effectively, this encodes a kind of simplified representation of global molecular structure in the feature vector. All other features only describe the local neighborhood around an atom and the atom itself. Alternatively, or in addition to, a global attention mechanism[29] could be added on top of the graph convolutions to incorporate distal information. The base model is composed of a graph convolutional neural network that converts the molecular representation described by the set of all $\mathbf{f}_a$ and $\mathbf{f}_{ay}$ to a fixed-length molecular feature vector (fingerprint), which is then passed through a final hidden layer before the output layer. The output vector has a single element for the enthalpy of formation model and for the entropy model, and seven elements for the heat capacity model in order to predict heat capacities at seven different temperatures simultaneously. The graph convolution essentially takes each $\mathbf{f}_a$ and $\mathbf{f}_{ay}$ and passes them through neural network layers to incorporate nearest neighbor features into new feature vectors for each atom. This process is repeated for a total of three iterations, thus incorporating information up to a depth of three into the feature vector for each atom. Subsequently, the resulting atom feature vectors are combined and sparsified using a *softmax* activation function to yield the molecular fingerprint.

The base models are trained on B3LYP/6-31G(2df,p) data. The transfer learning models are separate models trained on CCSD(T)-F12/cc-pVDZ-F12//B3LYP/6-31G(2df,p) data with bond additivity corrections for enthalpy of formation and on ωB97X-D3/def2-TZVP for entropy and heat capacities. The quantum mechanical data for the transfer learning models is combined with experimental data. 5% of the available training data for each model was used as validation datasets for early stopping and separate test sets are used to measure model performance. A more detailed description of the datasets will follow in a later

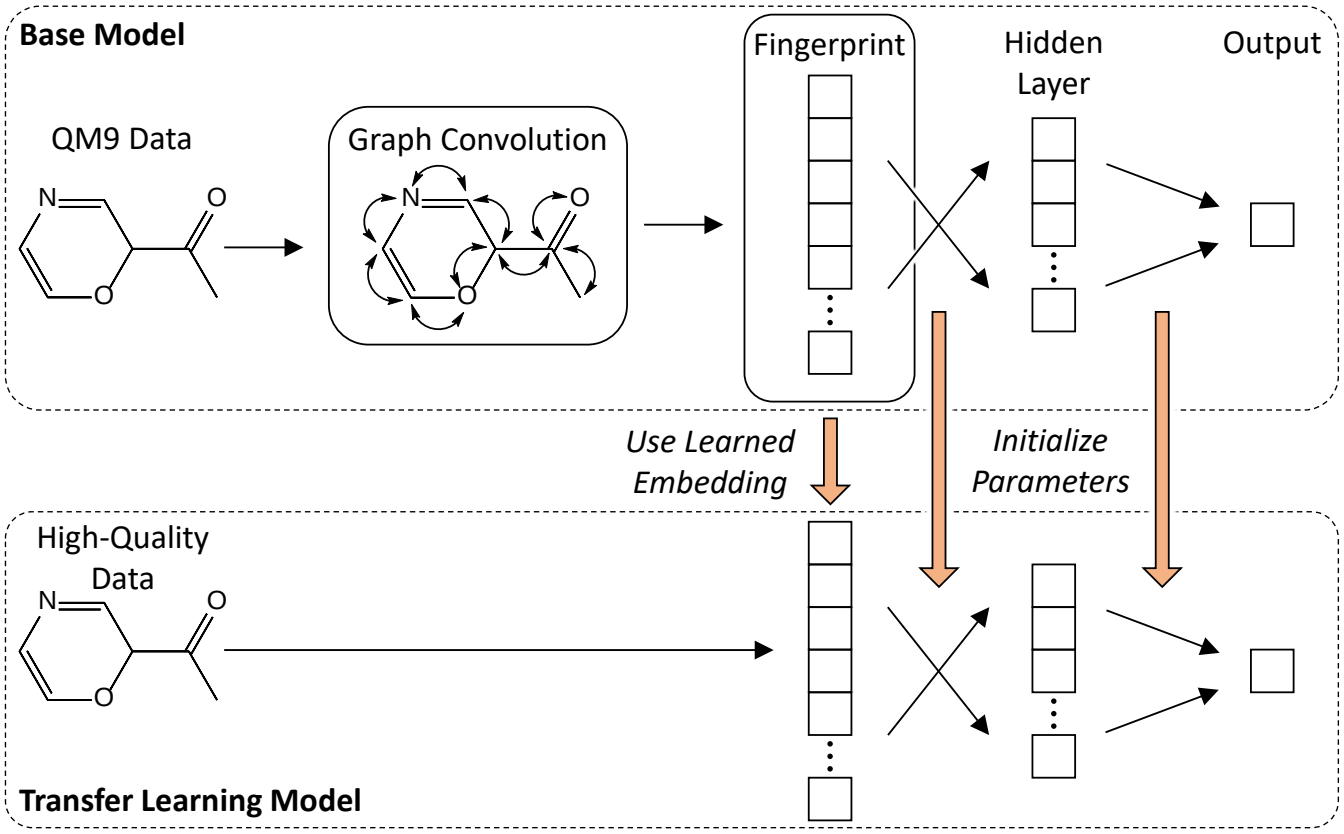


Figure 1: Transfer learning model architecture using a base model to learn a molecular embedding and neural network parameter initialization.

section. The transfer learning models do not retrain the graph convolutions and instead use the learned fingerprint embedding from the base models directly. Additional knowledge is transferred from the base models to the transfer learning models by initializing their weights using the fully trained weights from the base models. A detailed description of the model training and hyperparameters is given in Section S1 in the Supporting Information.

While all of the models were trained with a mean squared error loss function, a more intuitive metric for assessing results is the mean absolute error (MAE), which will be reported throughout this paper. In addition, 95% confidence intervals (CI) calculated as twice the root-mean-square error (RMSE), which are commonly reported in thermodynamic tables,[30] are listed as well. The trained models are available in the Supporting Information and can be easily used in conjunction with the *DataDrivenEstimator* package available on GitHub.[31]

Thermochemistry Calculations

Electronic structure calculations were performed at a variety of levels of theory. The B3LYP/6-31G(2df,p) level of theory is used for low-level geometry optimizations and frequency calculations (used for calculating low-level enthalpies of formation, entropies, and heat capacities). A scale factor of 0.965 is applied to the computed harmonic frequencies used to compute the zero-point energy (ZPE).[32] High-level geometry optimizations and frequency calculations were performed at the ωB97X-D3/def2-TZVP level of theory corrected by a scale factor of 0.975 (used for calculating entropies and heat capacities at the higher level of theory).[33] High-level energies were calculated at the CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) level of theory (used for calculating high-level enthalpies of formation). The double-zeta basis set was selected in order to allow for a large number of coupled cluster calculations while maintaining reasonable accuracy (the accuracy after fitting bond additivity corrections is shown in a later section). The geometries for some molecules selected for CCSD(T)-F12 calculations were taken directly from the published QM9 dataset.[19] They were not reoptimized and we did not attempt to confirm that they are the lowest-lying conformers. For new

geometry calculations, the lowest energy conformer was selected based on a conformer search using the RDKit[34] and the MMFF94 force field. For enthalpy calculations at 298 K, contributions from other conformers can mostly be neglected.[20] On the other hand, entropy is more strongly affected by conformational variations, so we only calculated molecules without rotatable bonds for the entropy models and for molecules with rotatable bonds we captured conformational effects implicitly by using experimental data (a more detailed description of the datasets is available in a later section). All DFT calculations made use of the Q-Chem 5.1 electronic structure code[35] and the coupled cluster calculations used Molpro 2015.[36–38]

Standard ideal gas statistical thermodynamic models were used to compute rigid rotor harmonic oscillator (RRHO) partition functions. Enthalpies and entropies were calculated at 298 K and heat capacities were calculated at seven different temperatures—300 K, 400 K, 500 K, 600 K, 800 K, 1000 K and 1500 K. Symmetry contributions are not included in the entropies because RMG incorporates them automatically during mechanism generation. Therefore, the partition function for entropy was not divided by the external symmetry number. In fact, training a machine learning model to predict entropies that are symmetry-corrected is a more difficult task because the model has to implicitly learn symmetry numbers, which instead could easily be applied after training a model that does not include symmetry. Of course, correct computational determination of symmetry numbers, whether by estimating point groups from three-dimensional molecular geometries or from a computation based on a molecular graph representation, is a complex task in itself already discussed in the literature[39–41] and is outside the scope of this study.

Calculation of the enthalpy of formation follows the atomization energy approach detailed by Curtiss et al.[42] Atomization energies obtained from ab initio calculations are often not very accurate, because atoms and standard-state forms of some elements (e.g., graphite, $O_2(^3\Sigma_g^-)$) have significantly different electronic states than the closed-shell organic molecules studied here. To improve the accuracy of the formation enthalpy, it is common to use bond additivity corrections (BACs), which are empirical corrections to molecular energies and

enthalpies of formation that use a few fitted parameters to correct for systematic errors in electronic structure calculations for some bond types. Fitting the parameters to a set of relatively few (tens or hundreds) low-uncertainty experimental data can significantly improve the error of calculations relative to experimental data and generalizes well beyond molecules in the reference dataset because the corrections are specific to atoms and bonds rather than the molecule as a whole. We use BACs as described by Anantharaman and Melius,[25] which involves fitting three parameters per atom type where one parameter is an atomic correction. Calculation details are given in Section S2 in the Supporting Information.

Datasets

Training effective machine learning models is to a certain extent an exercise in dataset selection. As such, we are using an array of representative datasets from literature, proprietary sources, and some created by ourselves. Many of the electronic structure calculations and geometries are either taken directly from the popular QM9 dataset[19] with up to nine non-hydrogen atoms, or subsets of molecules are selected from the set of all molecules in QM9 in order to be calculated at a different level of theory. QM9 properties, which include the results of energy and harmonic vibrational frequency calculations, are available at the B3LYP/6-31G(2df,p) level of theory. Because we are only interested in HCNO-containing molecules and because diffuse functions were not included in the basis set, we removed all fluorine-containing molecules. We also removed the set of molecules that failed the consistency check described in the original publication, which involves converting force field, semi-empirical, and density functional theory (DFT) geometries to InChI strings and verifying that they are identical.[19] Other than the high-accuracy data for fitting bond additivity corrections, experimental data were obtained from a version of the NIST-TRC database,[43] henceforth simply referred to as NIST data. All coupled cluster and DFT calculations done by us are available in the Supporting Information. An overview of the datasets is given in Table 1.

We only considered species with an even number of electrons, i.e., no doublet radicals.

Table 1: Enthalpy of formation ($\Delta_f H^\circ_{298}$), entropy (S°_{298}), and heat capacity (C_p) datasets.

Dataset	Name	Property	Level(s)	Size
1	`bac_fit`	$\Delta_f H^\circ_{298}$	CC[a], expt.	147
2	`bac_test`	$\Delta_f H^\circ_{298}$	CC[a], expt.	412
3	`base`	$\Delta_f H^\circ_{298}$, S°_{298}, C_p	DFT-low[b]	∼130k
4	`tf_h_1`	$\Delta_f H^\circ_{298}$	CC[a]	∼10k
5	`tf_h_2`	$\Delta_f H^\circ_{298}$	expt.	∼3k
6	`tf_s`	S°_{298}	DFT-high[c]+expt.	∼3k
7	`tf_c`	C_p	DFT-high[c]+expt.	∼2k
8	`tf_h_test`[d]	$\Delta_f H^\circ_{298}$	GA[e], DFT-low[b], CC[a]+expt.	∼1.2k
9	`tf_s_test`[d]	S°_{298}	GA[e], DFT-low[b], DFT-high[c]+expt.	∼0.3k
10	`tf_c_test`[d]	C_p	GA[e], DFT-low[b], DFT-high[c]+expt.	∼0.2k

[a]CCSD(T)-F12/cc-pVDZ-F12//B3LYP/6-31G(2df,p) + BAC; [b]B3LYP/6-31G(2df,p); [c]ωB97X-D3/def2-TZVP; [d]Contain the same molecules; [e]Group additivity.

For fitting the bond additivity corrections, we selected a dataset of highly accurate experimental enthalpies of formation (`bac_fit`) and calculated the corresponding coupled cluster enthalpies of formation. The uncertainty in each experimental enthalpy value in the `bac_fit` dataset is at most 0.5 kcal mol^{-1}, but the majority are significantly lower. In thermodynamic tables, uncertainty is typically provided as 95% confidence intervals, which approximately correspond to two standard deviations to the left and to the right of the mean.[30] For the most part, the uncertainty in these data adhere to that standard, but we were not able to verify the uncertainty quantification used in some of the sources. This set of 147 enthalpy values spans diverse chemical species of both small and large size involving most permutations of bonds between HCNO atoms. It is obtained from a variety of sources[44–47] including the Active Thermochemical Tables[30,48] and is available in the Supporting Information. We selected an additional set of 412 molecules from the NIST data for testing the fitted BACs (`bac_test`). The test set molecules are selected to have a more varied set of molecules; in particular `bac_test` includes somewhat larger molecules and potentially more complex electronic structure effects. Unlike `bac_fit`, the uncertainties of `bac_test` are not readily

available, so there is an assumption that the experimental data are reasonably well known.

As described earlier, enthalpies of formation, entropies, and heat capacities are first trained on a large dataset of low-quality data (`base`) and then on a smaller dataset of high-quality data. The low-quality data (`base`) are taken as the roughly 129 000 HCNO molecules in the QM9 set filtering out those with identified inconsistencies, supplemented by 1700 molecules from the NIST-TRC, `bac_fit`, and `bac_test` datasets recalculated at the B3LYP/6-31G(2df,p) level of theory to match the level of QM9 (`base`). The additional molecules correspond to those that do not overlap with the species already present in QM9.

There are three different transfer learning models: an enthalpy of formation model, an entropy model, and a heat capacity model. For each of these models, the training datasets are composed of both calculated and experimental data. The experimental data contain many molecules that are significantly larger than those in QM9 with up to 42 non-hydrogen atoms.

For the enthalpy of formation model, the high-quality training data is a combination of the 147 experimental data points for fitting BACs (`bac_fit`), experimental data for about 2700 NIST molecules (`tf_h_2`), and a selection of approximately 9800 explicitly correlated coupled cluster calculations (CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) + BAC) corresponding to molecules sampled at random from QM9 (`tf_h_1`). Note that non-random selections can improve model performance,[26,49] but such an active selection scheme is not the focus of the present study.

The entropy model is trained on 2300 NIST data and 900 ωB97X-D3/def2-TZVP DFT calculations (`tf_s`). The NIST entropy data are of mixed accuracy, with some data from direct experimental measurements but much of it from indirect methods and extrapolations. Internal and external symmetry number contributions are calculated using RMG for each NIST molecule and are removed from the experimental entropy because the goal is to train a model that can be used in RMG, which adds its computed symmetry contributions in during a reaction mechanism simulation. The 900 DFT calculations correspond to molecules

randomly selected from QM9 with the constraint of being exclusively composed of cyclic or polycyclic cores without rotatable bonds.

The heat capacity model is trained on 1100 NIST data points and the same 900 ωB97X-D3/def2-TZVP DFT molecules (`tf_c`). The experimental heat capacities are from a mix of direct and indirect methods, with considerable variance in error bars. 5% of the training data available for each model were reserved as a held-out validation dataset to stop the training before overfitting.

Lastly, we selected molecules for test sets (`tf_h_test`, `tf_s_test`, `tf_c_test`) that are not present in any of the training datasets. For each property, the selection consists of roughly 10% of all molecules with available high-accuracy data. Because all molecules for which high-accuracy data are available are also present in the low-accuracy training data, each molecule in the test sets has both high- and low-accuracy properties. For example, none of the molecules in `tf_s_test` are in `tf_s` or `base`.

Results and Discussion

Bond Additivity Corrections

As outlined previously, we calculated enthalpies of formation using the atomization energy method and then added corrections. Here, we compare the accuracy of the calculated values with and without fitted bond additivity corrections (BACs). The level of theory for the single point energy calculations is CCSD(T)-F12a/cc-pVDZ-F12, which will also be referred to as F12 for simplicity. BACs are fitted to minimize the difference between F12 enthalpies of formation calculated from quantum chemistry and the corresponding experimental data for the `bac_fit` dataset in Table 1.

Overall, the fitting procedure reduced the average error across `bac_fit` very significantly. Before adding BACs, the MAE between the F12 enthalpies of formation and the high-accuracy experimental data was $8.98\,\mathrm{kcal\,mol^{-1}}$ and the RMSE was $10.45\,\mathrm{kcal\,mol^{-1}}$. This

very large error is surprising considering it has previously been shown that chemical accuracy ($1\,\mathrm{kcal\,mol^{-1}}$) is possible with a double-zeta basis for certain molecular reaction energies.[50] Of course, there is a large error-cancelling effect between reactant and products for both atoms and bonds when calculating reaction energies that does not occur for enthalpies of formation calculated from atomization energies because the electronic structure of molecules and atoms is very different. Figure 2 shows that the error across `bac_fit` increases roughly linearly with increasing numbers of heavy atoms in a molecule. Such a trend indicates there is a large systematic error in the uncorrected values that scales with the number of heavy atoms. Knizia et al. report an MAE of $1.86\,\mathrm{kcal\,mol^{-1}}$ for CCSD(T)-F12a with a double-zeta basis set for atomization energies,[50] which is significantly smaller than our error. However, they are benchmarking against conventional CCSD(T)/CBS instead of experimental data and 47 out of the 49 molecules in their benchmark dataset only have one or two non-hydrogen (heavy) atoms each. As shown in Figure 2, The errors for molecules in the `bac_fit` dataset containing one or two heavy atoms are in line with those reported by Knizia et al. on their test set. We are not aware of any discussion regarding abnormal enthalpies of formation with double-zeta CCSD(T)-F12a in the literature, potentially because most studies that employ explicitly correlated coupled cluster methods use triple-zeta and larger basis sets, which are prohibitively expensive for the present study.

After adding fitted BACs, the MAE computed across `bac_fit` is reduced to only $0.70\,\mathrm{kcal\,mol^{-1}}$ and the RMSE becomes $1.18\,\mathrm{kcal\,mol^{-1}}$ (95% CI: $2.36\,\mathrm{kcal\,mol^{-1}}$). Furthermore, the systematic error in Figure 2 has been removed. In fitting BACs, the atomic energies $E_{0,i}$ in Equation (S3) in the Supporting Information have effectively been redefined by adding corrections for each element (given by the α_i values in Equation (S5) in the Supporting Information). As hypothesized, the derived BACs generalize well, which is demonstrated by a reduction in MAE from $13.90\,\mathrm{kcal\,mol^{-1}}$ to $0.98\,\mathrm{kcal\,mol^{-1}}$ for the test set `bac_test` (RMSE decreases from $14.74\,\mathrm{kcal\,mol^{-1}}$ to $1.31\,\mathrm{kcal\,mol^{-1}}$), demonstrating that double-zeta calculations can yield good results if corrected with BACs. The uncertainties in the experimental enthalpy

of formation values for the test set are not known and are likely higher than the very accurate data in `bac_fit`, which may be part of the reason for the slightly larger error across `bac_test`. Fitting parameters for specific bond types instead of using three parameters per atom type (using Equation (S4) instead of Equation (S5) in the Supporting Information) would lead to an almost identical reduction in error (MAE: 0.67 kcal mol^{-1}, RMSE: 1.16 kcal mol^{-1} for `bac_fit`) but may be sensitive to the resonance structure used for each molecule. Additionally, with that approach the atom corrections would be absorbed as part of the bond corrections instead of being treated separately.

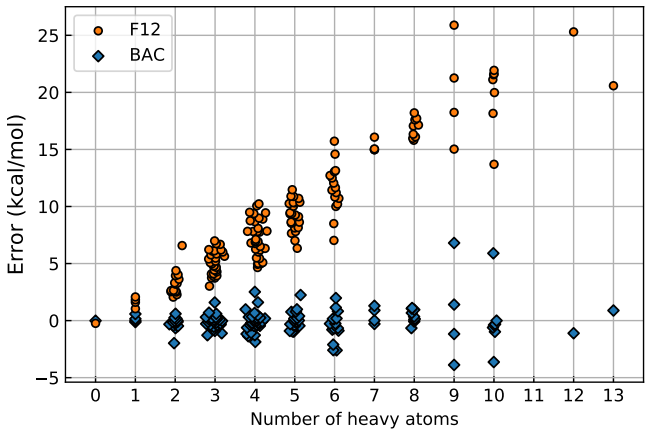


Figure 2: Enthalpy of formation errors vs. the number of heavy atoms in each molecule before (F12) and after (BAC) fitting bond additivity corrections on dataset `bac_fit`.

The distribution of errors with and without BACs is shown in Figure 3. Each error is computed as the subtraction of the experimental value from the calculated value. The systematic error observed in Figure 2 manifests itself as a very wide range of errors much greater in magnitude than after fitting BACs. The majority of enthalpies of formation computed from atomization energies are in error by more than 5 kcal mol^{-1}. Including BACs leads to a tight distribution centered at zero with all but two molecules in error by less than 5 kcal mol^{-1}. The highest error of 6.80 kcal mol^{-1} corresponds to phenyl isocyanate. An error of such a large magnitude will cause issues in reaction mechanism generation, but the likelihood of such errors is small. Using a triple-zeta basis would most likely reduce the probability of large errors even further, but computational restrictions necessitated the use

of a double-zeta basis here.

As before, the BAC procedure generalizes well to the test set `bac_test` as shown in the second panel of Figure 3. After fitting BACs, none of the molecules in `bac_test` are in error by more than $5\,\mathrm{kcal\,mol^{-1}}$.

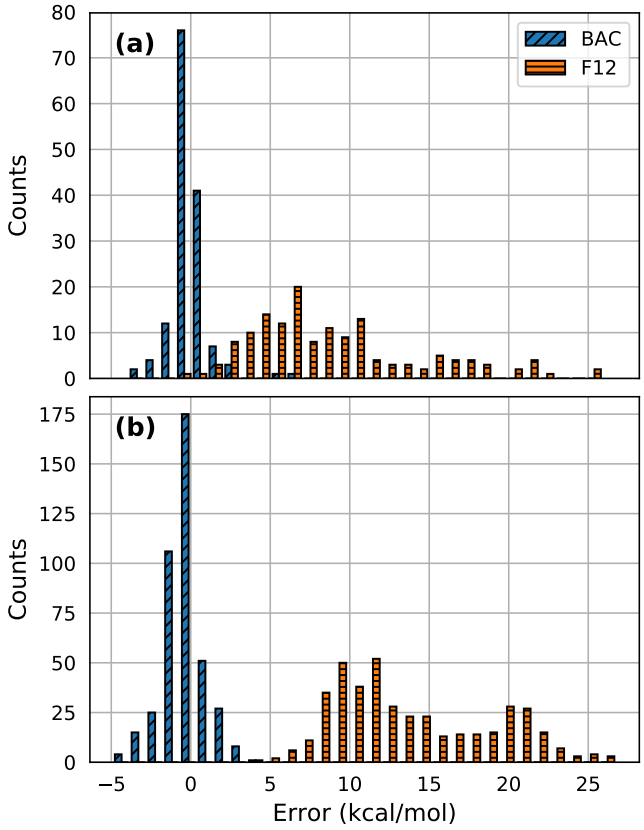


Figure 3: Distribution of enthalpy of formation errors relative to experiment (calculated minus experimental value) in the `bac_fit` dataset (a) and the `bac_test` dataset (b) before (F12) and after (BAC) fitting corrections.

Transfer Learning

First, the three base models (Figure 1), one for enthalpy of formation, one for entropy, and one for heat capacities, were trained on the `base` dataset (Table 1). In order to train the enthalpy of formation transfer learning model, BACs were applied to all high-level CCSD(T)-F12/cc-pVDZ-F12 data to form datasets `tf_h_1` and `tf_h_test`. The training data for the transfer learning enthalpy of formation model is the combination of the coupled cluster

(`tf_h_1`) and experimental data (`tf_h_2`). Similarly, the training data for entropy and heat capacity are both a combination of high-level DFT and experimental data (`tf_s` and `tf_c`, respectively). The transfer learning models used the mapping that converts a molecular graph to a fixed-length vector learned during training of the base models. The remaining neural network parameters were initialized using the corresponding weights in the base models. For all models, the molecules in the test datasets, `tf_h_test`, `tf_s_test`, and `tf_c_test`, are identical and their properties are available at both the low and high level so that performance can be measured both in terms of precision and accuracy. In this context, model *precision* is measured by how well the model predictions match the values at the level of theory of the training data:

$$\text{precision} = \frac{1}{N}\sum_{i=1}^{N}\left|p_i^{\text{model}} - p_i^*\right| \tag{1}$$

where p_i^* is the value of the property at the same level of theory as the data used to train the model. Equation (1) corresponds to MAE and the equation for RMSE is analogous. Model *accuracy* is measured by how well the model predictions match the true values of the property, which are approximated using experimental data or coupled cluster data:

$$\text{accuracy} = \frac{1}{N}\sum_{i=1}^{N}\left|p_i^{\text{model}} - \hat{p}_i\right| \tag{2}$$

where $\hat{p}_i$ corresponds to the "true" value of the property. Naturally, high accuracy is the most desirable property of a machine learning model for property prediction. Additionally, RMG was used to calculate group additivity estimates of the thermochemical properties for the test set molecules to enable comparison to current RMG predictions.[3,8]

The *accuracies* are shown in Table 2 in terms of mean absolute error (MAE), root-mean-square error (RMSE), and 95% confidence interval (CI). For all three properties, the predictions afforded by the transfer learning model are clearly better than those of the base model and especially those of group additivity. Therefore, the transfer learning model is more

suitable for simulations in RMG. Because of the molecules available in the QM9 database, the test set contains many complex structures, such as fused and bridged polycyclic compounds with several heteroatoms. These types of molecules are especially difficult to model with group additivity because contributions to thermochemistry are not solely additive across the groups present in the molecule but are strongly influenced by less local contributions like ring strain. Even though the group additivity scheme implemented in RMG has sophisticated ring strain corrections,[8] it lacks the ability to model many such molecules. If the test set were only composed of linear hydrocarbons, it would be very likely that group additivity would outperform the transfer learning model since group additivity was trained to even higher-accuracy data than most of the training data used here. For more complex RMG simulations involving fused cyclic molecules, the transfer learning model is a better choice.

Table 2: Test set (`tf_h_test`, `tf_s_test`, `tf_c_test`) *accuracies* of enthalpy of formation ($\Delta_f H^\circ_{298}$), entropy (S°_{298}), and heat capacity (C_p) predictions for the transfer learning models (TF), the base models, and group additivity (GA). $\Delta_f H^\circ_{298}$ in $\mathrm{kcal\,mol^{-1}}$ and S°_{298}/C_p in $\mathrm{cal\,mol^{-1}\,K^{-1}}$.

$\Delta_f H^\circ_{298}$	MAE	RMSE	95% CI
TF	1.78	2.80	5.60
Base	4.76	6.47	12.94
GA	9.99	16.17	32.35
S°_{298}	MAE	RMSE	95% CI
TF	0.80	1.16	2.31
Base	9.74	13.67	27.34
GA	11.25	18.51	37.02
C_p[a]	MAE	RMSE	95% CI
TF	0.74	1.21	2.41
Base	2.48	3.32	6.63
GA	3.41	5.44	10.88

[a]Average across 7 temperatures.

The *precisions* for the base models are shown in Table 3. The precisions for the transfer learning models are identical to their accuracies, so they are the values in Table 2. Except for the MAE corresponding to enthalpy of formation, the precisions of the base models are

significantly worse than those of the transfer learning models, which is surprising given the amount of data the base models were trained on. However, this effect is compounded by the fact that many of the molecules in the test set are drawn from the data with experimentally available properties, which are proportionally underrepresented in the training data for the base models compared to molecules drawn from QM9. For example, the MAE calculated across the *validation* sets used for early stopping (which are randomly drawn from the training data) is 1.56 kcal mol^{-1} for the base model and 1.42 kcal mol^{-1} for the transfer learning model, which are similar in magnitude. Similarly, for the validation datasets the MAE is 0.85 cal mol^{-1} K^{-1} and 0.76 cal mol^{-1} K^{-1} for the entropy base and transfer learning models, respectively; and 0.60 cal mol^{-1} K^{-1} and 0.71 cal mol^{-1} K^{-1} for the heat capacity base and transfer learning models, respectively. Regardless, this indicates that less than a tenth of all available molecules have to be calculated at the high level or obtained from experiment in order to train a model that reaches the same precision as a low-level model trained on all available molecules.

Table 3: Test set (`tf_h_test`, `tf_s_test`, `tf_c_test`) *precisions* of enthalpy of formation ($\Delta_f H^\circ_{298}$), entropy (S°_{298}), and heat capacity (C_p) predictions for the base models only. $\Delta_f H^\circ_{298}$ in kcal mol^{-1} **and** S°_{298}/C_p **in** cal mol^{-1} K^{-1}**.**

Property	MAE	RMSE	95% CI
$\Delta_f H^\circ_{298}$	1.69	3.51	7.03
S°_{298}	1.50	2.23	4.46
C_p [a]	2.34	5.85	11.70

[a] Average across 7 temperatures.

As mentioned in previous sections, entropy is strongly affected by conformational effects.[20] Experimental data naturally includes all the important conformers and internal rotors. However, the quantum chemistry calculations used here are for a single conformer, and do not include corrections for internal rotation. Table 4 shows that by combining experimental and quantum chemistry calculations into a single training set affords predictions of nearly identical quality for molecules with and without internal rotors. Because the training data for the base model is composed exclusively of static electronic structure calculations,

its accuracy and precision for molecules with internal rotors is lowered significantly. The training data for the heat capacity model were chosen in the same manner as for the entropy model and the results for splitting the test data into molecules with and without internal rotors are similar and shown in Section S3 of the Supporting Information.

Table 4: Test set (`tf_s_test`) *accuracies* and *precisions* of entropy (S°_{298}) predictions in $\mathrm{cal\,mol^{-1}\,K^{-1}}$ for the transfer learning model (TF) and the base model split by molecules with and without internal rotors.

Accuracy	MAE	RMSE	95% CI
TF (no rotors)	0.72	1.10	2.19
TF (with rotors)	0.85	1.19	2.37
Base (no rotors)	1.44	1.75	3.51
Base (with rotors)	14.12	16.85	33.70
Precision	**MAE**	**RMSE**	**95% CI**
Base (no rotors)	0.96	1.33	2.65
Base (with rotors)	1.78	2.58	5.16

Parity plots and frequency distributions of the errors for the different transfer learning models are shown in Figure 4. The error distributions show that while most molecules are predicted well by the transfer learning model, several predictions are poor, albeit more accurate than the base model and group additivity on average. In theory, prediction quality can be improved by providing more information in the form of input atom and bond featurization, for example, by incorporating molecular geometry, but molecular representation in RMG is inherently graph-based and lacks geometrical information. Furthermore, thermochemistry may be strongly affected by different molecular geometries. Rapid estimation of molecular geometries may be possible with distance geometry based three-dimensional embedding and subsequent force field optimization as is available in cheminformatics packages like the RDKit,[34] but exhaustive conformer searches for the selection of the lowest energy conformation may be prohibitively expensive for large molecules in RMG and it is not clear when distance geometry-based approaches might fail.

Obtaining nearly 10 000 high-level data points for the enthalpy of formation model, as

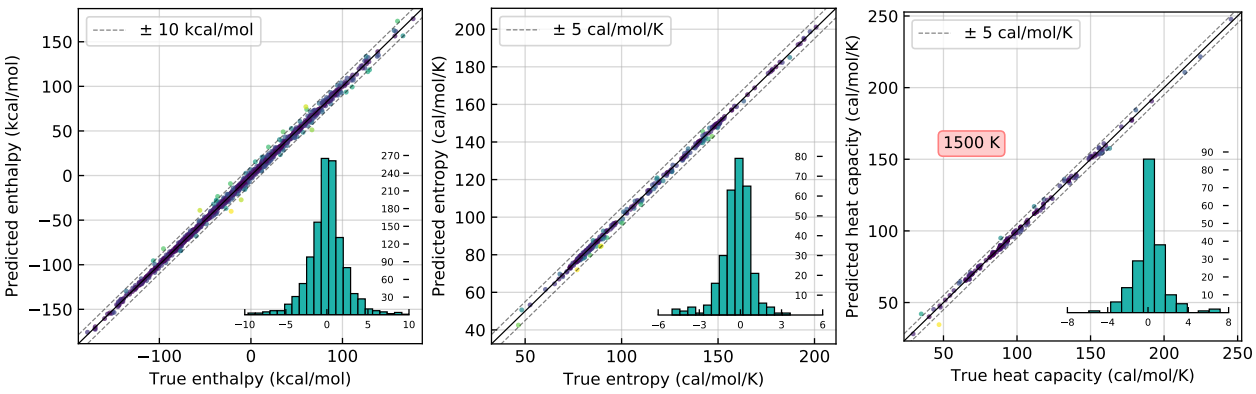


Figure 4: Parity plots of the experimental and high-level electronic structure calculations ("true"), and the values predicted by the transfer learning models for the test sets `tf_h_test`, `tf_s_test`, and `tf_c_test`. Dashed lines of 10 kcal mol^{-1} and 5 cal mol^{-1} K^{-1} are shown to guide the eye. Frequency distributions of the signed errors ("predicted"−"true") are superimposed. The heat capacity plot is for the values at 1500 K, which has the largest errors.

was done in this study, is already associated with large computational cost. Therefore, it is important to know how much data is really needed to obtain acceptable results. To assess this, we trained different models on various fractions of the approximately 9800 F12 data (`tf_h_1` in Table 1) and tested on all of the remaining data. For example, the smallest training set considered by us is composed of 81 molecules with the test error being reported on the remaining 9724 molecules. The results are shown in Figure 5. Remarkably, the MAE is already smaller than 3 kcal mol^{-1} when only training on 81 molecules. This suggests that only very few data points are required to adapt the information learned during training of the base model to be suitable for predictions in the high-level domain. Predictions of practical importance can already be achieved with less than 1000 high-level training data, which is less than 1% of the low-level training data used in the base model. Interestingly, the lowest error in Figure 5 is smaller than the error in Table 2, even though the experimental molecules were not included here and the usual assumption in machine learning is that more data leads to smaller errors. However, the experimental data tend to be more diverse than the molecules in the `tf_h_1` dataset, at least in terms of molecular size, which renders the learning task somewhat more difficult and may explain this difference.

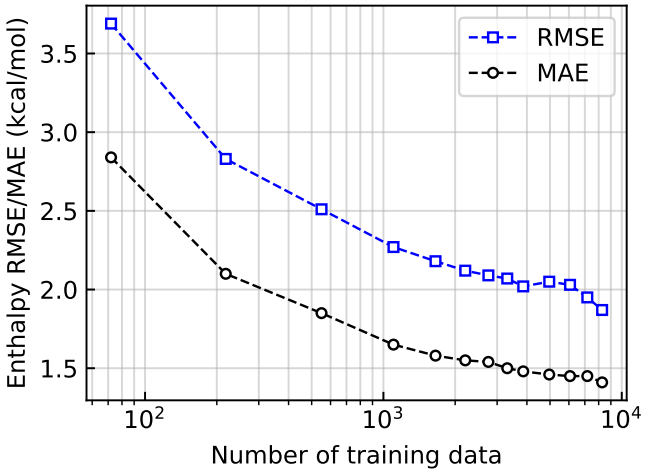


Figure 5: Test error of enthalpy of formation model with varying number of CCSD(T)-F12 training data points. The test error is computed across all molecules in dataset `tf_h_1` that were not trained on.

Conclusions

With the continual development of new methods and the rapid expansion of molecular databases, machine learning is ideally suited for chemical property prediction in automated reaction mechanism generation. Because most methods are agnostic to the type of input molecule, machine learning frameworks are much more flexible than conventional (e.g., group additivity) ones. Nonetheless, the amount of required training data is usually very large and especially difficult to obtain at levels of theory that are of practical importance. To address this issue, we created an extensive dataset of explicitly correlated coupled cluster enthalpies of formation, albeit still much smaller than available low-quality datasets. We fitted bond additivity corrections to reduce the mean absolute error compared to experiment to less than 1 kcal mol^{-1}. We also collected an array of experimental data and calculated additional high-level density functional data for new entropy and heat capacity datasets.

In order to train useful machine learning models with the comparatively small amount of high-quality training data, we employed a transfer learning approach to predict enthalpy of formation, entropy, and heat capacities at several temperatures. Three base models were trained on 130 000 molecules, which were used to initialize parameters in the high-level neu-

ral network models and which provided learned molecular embeddings to convert molecular graphs into appropriate fixed-length vector representations. Subsequent training of the transfer learning models resulted in models capable of thermochemical property prediction with accuracies far exceeding those of the base models and group additivity. By combining an experimental dataset containing molecules with many rotatable bonds with a DFT dataset only composed of rigid molecules, the entropy and heat capacity transfer learning models achieve equally accurate predictions for molecules with and without rotatable bonds. We showed that fewer than 1000 high-level training data points are required to obtain a useful enthalpy of formation model.

Several improvements can be made to both the methods and the data in the future. The larger error of the test sets compared to the validation datasets used for early stopping and the presence of predictions with large errors hint at issues with generalizability to significantly different chemical domains. To combat this issue, the current model design could be improved by incorporating novel ideas from methods that have been shown to generalize well to larger molecules, for example, incorporating 3D geometries into the graph convolution or constructing the convolution in a more atom-wise fashion.[13] The current datasets contain no radicals, but thermochemical predictions of radical species are still possible in RMG by using the hydrogen atom bond increment method (HBI) to predict their properties from their stable counterparts.[51] Alternatively, radicals could be directly included in the training data, thus directly enabling prediction of their properties without the need for HBI. Moreover, the developed models are limited to the realm of organic chemistry and extension to transition metal chemistry is not trivial due to difficulties with generating training data.

We have shown that transfer learning coupled with novel high-quality data is an effective technique to obtain accurate thermochemistry predictions suitable for automated reaction mechanism generation while only requiring small datasets on the order of a few thousand molecules. We expect that further refinement of the methods and data will lead to general-purpose property prediction schemes in the near future.

Supporting Information Available

The following files are available free of charge.

- The datasets used for training and testing the models with molecular geometries and thermochemical properties: B3LYP/6-31G(2df,p) calculations (enthalpies of formation, entropies, and heat capacities), CCSD(T)-F12a/cc-pVDZ-F12 + BAC calculations (enthalpies of formation), ωB97X-D3/def2-TZVP calculations (entropies and heat capacities)
- A list of molecular identifiers for molecules obtained from the NIST-TRC database and corresponding CAS numbers for each dataset
- Detailed description of model training and hyperparameter selection; description of thermochemistry calculations and bond additivity corrections; test set accuracies and precisions of heat capacity models split by molecules with and without internal rotors; description of remaining Supporting Information files; description of data extraction from NIST-TRC
- Trained ML models

Acknowledgement

We gratefully acknowledge financial support from ExxonMobil under grant no. EM09079.

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) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

(2) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.

(3) 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.

(4) Atkins, P.; de Paula, J. *Elements of Physical Chemistry*, 7th ed.; Oxford University Press: New York, 2017.

(5) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.

(6) Sumathi, R.; Green, W. H. Thermodynamic Properties of Ketenes: Group Additivity Values from Quantum Chemical Calculations. *J. Phys. Chem. A* **2002**, *106*, 7937–7949.

(7) Sebbar, N.; Bozzelli, J. W.; Bockhorn, H. Thermochemical Properties, Rotation Barriers, Bond Energies, and Group Additivity for Vinyl, Phenyl, Ethynyl, and Allyl Peroxides. *J. Phys. Chem. A* **2004**, *108*, 8353–8366.

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

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

(10) 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*, 23822–23830.

(11) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.

(12) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**, arXiv:1704.01212. `http://arxiv.org/abs/1704.01212` (accessed January 10, 2019).

(13) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(14) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. **2017**, arXiv:1706.08566. `https://arxiv.org/abs/1706.08566` (accessed January 8, 2019).

(15) Hy, T. S.; Trivedi, S.; Pan, H.; Anderson, B. M.; Kondor, R. Predicting Molecular Properties with Covariant Compositional Networks. *J. Chem. Phys.* **2018**, *148*, 241745.

(16) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(17) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(18) 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*, 1757–1772.

(19) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

(20) Li, Y.-P.; Bell, A. T.; Head-Gordon, M. Thermodynamics of Anharmonic Systems:

Uncoupled Mode Approximations for Molecules. *J. Chem. Theory Comput.* **2016**, *12*, 2861–2870.

(21) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.

(22) S Smith, J.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. Outsmarting Quantum Chemistry Through Transfer Learning. **2018**, ChemRxiv. `https://chemrxiv.org/articles/Outsmarting_Quantum_Chemistry_Through_Transfer_Learning/6744440/1` (accessed September 17, 2018).

(23) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(24) Fare, C.; Turcani, L.; Pyzer-Knapp, E. O. Powerful, Transferable Representations for Molecules Through Intelligent Task Selection in Deep Multitask Networks. **2018**, arXiv:1809.06334. `https://arxiv.org/abs/1809.06334` (accessed January 12, 2019).

(25) Anantharaman, B.; Melius, C. F. Bond Additivity Corrections for G3B3 and G3MP2B3 Quantum Chemistry Methods. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.

(26) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(27) Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. **2015**, arXiv:1509.09292. `http://arxiv.org/abs/1509.09292` (accessed August 16, 2018).

(28) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 595–608.

(29) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. **2014**, arXiv:1409.0473. `http://arxiv.org/abs/1409.0473` (accessed January 9, 2019).

(30) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *Int. J. Quantum Chem. 114*, 1097–1101.

(31) Han, K.; Li, Y.-P.; Grambow, C. A. DataDrivenEstimator: A Package of Data Driven Estimators for Thermochemistry and Kinetics. 2019; `https://github.com/ReactionMechanismGenerator/DataDrivenEstimator`, (accessed January 7, 2019).

(32) National Institute of Standards and Technology (NIST), Precomputed Vibrational Scaling Factors. 2013; `https://cccbdb.nist.gov/vibscalejust.asp`, (accessed October 19, 2018).

(33) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(34) Landrum, G. RDKit: Open-Source Cheminformatics. 2006; `http://www.rdkit.org`, (accessed August 6, 2018).

(35) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; 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*, 184–215.

(36) 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*, 242–253.

(37) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R. et al. *MOLPRO*, Version 2015.1, A Package of Ab Initio Programs. 2015; Molpro: Cardiff, U.K., 2015.

(38) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(39) Ivanov, J.; Schüürmann, G. Simple Algorithms for Determining the Molecular Symmetry. *J. Chem. Inf. Model* **1999**, *39*, 728–737.

(40) Chen, W.; Huang, J.; Gilson, M. K. Identification of Symmetries in Molecules and Complexes. *J. Chem. Inf. Model* **2004**, *44*, 1301–1313.

(41) Vandewiele, N. M.; Van de Vijver, R.; Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B. Symmetry Calculation for Molecules and Transition States. *J. Comput. Chem.* **2015**, *36*, 181–192.

(42) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(43) NIST Thermodynamics Research Center, NIST/TRC Table Database. CD-ROM, 2004.

(44) Cioslowski, J.; Schimeczek, M.; Liu, G.; Stoyanov, V. A Set of Standard Enthalpies of Formation for Benchmarking, Calibration, and Parametrization of Electronic Structure Methods. *J. Chem. Phys.* **2000**, *113*, 9377–9389.

(45) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A.; Frisch, M. J. Calibration and Comparison of the Gaussian-2, Complete Basis Set,

and Density Functional Methods for Computational Thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.

(46) Emel'yanenko, V. N.; Verevkin, S. P.; Varfolomeev, M. A.; Turovtsev, V. V.; Orlov, Y. D. Thermochemical Properties of Formamide Revisited: New Experiment and Quantum Mechanical Calculations. *J. Chem. Eng. Data* **2011**, *56*, 4183–4187.

(47) Månsson, M. Non-Bonded Oxygen-Oxygen Interactions in 2,4,10-Trioxaadamantane and 1,3,5,6,9-Pentoxecane. *Acta Chem. Scand. B* **1974**, *28*, 895–899.

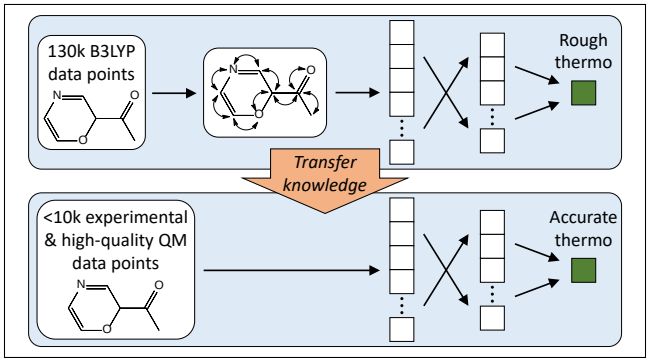
(48) Ruscic, B.; Bross, D. H. Active Thermochemical Tables (ATcT) Values Based on Ver. 1.122d of the Thermochemical Network. 2018; `https://atct.anl.gov`, (accessed October 10, 2018).

(49) 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*, 241733.

(50) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 Methods: Theory and Benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(51) Lay, T. H.; Bozzelli, J. W.; Dean, A. M.; Ritter, E. R. Hydrogen Atom Bond Increments for Calculation of Thermodynamic Properties of Hydrocarbon Radical Species. *J. Phys. Chem.* **1995**, *99*, 14514–14527.

TOC Graphic


130k B3LYP
data points
Rough
thermo
Transfer
knowledge
<10k experimental
& high-quality QM
data points
Accurate
thermo