Prediction of homolytic bond dissociation enthalpies for organic molecules at near chemical accuracy with sub-second computational cost

Peter C. St. John[1,*], Yanfei Guan[2,†], Yeonjoon Kim[1], Seonah Kim[1,*], Robert S. Paton[2,3*]

[1] Biosciences Center, National Renewable Energy Laboratory, 15103 Denver West Parkway, Golden, Colorado 80401, United States
[2] Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, USA
[3] Chemical Research Laboratory, University of Oxford, Mansfield Road, Oxford OX1 3TA, UK
[†] Current address: Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
* *Email*: peter.stjohn@nrel.gov; seonah.kim@nrel.gov; robert.paton@colostate.edu

Abstract

Bond dissociation enthalpies (BDEs) of organic molecules play a fundamental role in determining chemical reactivity. However, BDE computations at sufficiently high levels of quantum mechanical (QM) theory require substantial computing resources. We have therefore developed **A** machine-**L**earning derived, **F**ast, **A**ccurate **B**ond dissociation **E**nthalpy **T**ool (ALFABET), capable of accurately predicting BDEs for organic molecules in a fraction of a second. Automated density functional theory (DFT) calculations at the M06-2X/def2-TZVP level of theory were performed for 42,577 small organic molecules, resulting in 290,664 BDEs. A graph neural network was trained on a subset of these results, achieving a mean absolute error of 0.58 kcal/mol for BDEs of unseen molecules. An interface is available online at https://ml.nrel.gov/bde. The model is further demonstrated on two applications: first, we rapidly and accurately predict major sites of hydrogen abstraction in metabolism of drug-like molecules, and second, we determine the dominant molecular fragmentation pathways during soot formation.

Introduction

Nearly all chemical reactions of organic compounds involve the breaking and formation of covalent bonds. Unsurprisingly, bond energies feature as an essential ingredient in many predictive models of chemical reactivity. Homolytic bond dissociation enthalpies (BDEs) are defined by the enthalpy change for the gas-phase reaction at 298K:

$$A - B \rightarrow A\cdot + B\cdot$$

The cumulative difference between BDE values of all bonds broken and formed in a chemical reaction thus provides an estimate of the overall reaction enthalpy.[1] BDE values are thermodynamic quantities but they are also used widely to predict kinetics. For example, BDE values are used to predict relative reaction rates using well-established Evans-Polanyi-type correlations with bond strengths in radical hydrogen atom abstractions.[2] BDEs also provide insight into thermodynamically accessible reaction mechanisms for a given compound, and their calculation is often the first step in characterizing dominant pathways in combustion,[3] polymer synthesis[4] and thermal stability,[5,6] lignin depolymerization,[7] drug metabolism,[8-10] explosives,[11] organic synthesis planning[12,13] and other applications to energy-related materials.[14]

The accurate measurement and calculation of BDEs underlies numerous applications in organic chemistry. Experimental measurement of BDEs for polyatomic molecules are difficult, but a variety of techniques exist[15] with a typical uncertainty of $\pm$1-2 kcal/mol.[16] Calculation of BDEs with *ab initio* quantum chemistry methods is possible, however the choice of method is known to greatly affect the resulting computational accuracy.[17] Despite this, density functional theory (DFT) computations using M06-2X and M05-2X functionals have been shown to achieve accuracies

comparable to the uncertainties of the underlying experimental measurements.[18] As a result, quantum mechanical (QM) methods play an integral role in calculating radical enthalpies and proposing reaction mechanisms. However, even relatively efficient QM methods such as DFT scale exponentially with basis set size, often taking hours or days to obtain a single BDE value. This conventional workflow requires the geometry of a reactant and its radical products to be optimized and the Hessian of each species evaluated. For flexible compounds this process must be repeated for several alternative conformations. The integration of BDE calculations in molecular design efforts, including quantitative structure-property relationship (QSPR) models, has thus been limited by these computational demands, and the use of BDE calculations for the screening of thousands or millions of candidate structures remains impractical. In this manuscript we describe a new computational workflow that overcomes these limitations.

The rise of machine learning (ML) in quantum chemistry has led to the development of highly-accurate empirical models[19] that have accelerated traditionally difficult QM calculations for predicting enthalpy,[20] optoelectronic properties,[21] and forces.[22] In particular, the rise of graph neural networks (GNNs)[23] in modeling chemical properties has enabled 'end-to-end' learning on molecular structure: a ML strategy where traditional *feature engineering* is replaced by *feature learning* from a graph-based molecular representation. [19] These approaches have led to best-in-class prediction accuracies on a range of applications, especially as the amount of available training data grows.[24,25] An open question in molecular machine learning is whether optimized 3D coordinates are required as inputs to the ML algorithm to reach optimal accuracies. For enthalpy prediction on the QM9 dataset, consisting of all small molecules satisfying known valence rules, 3D-coordinates appear to lead to superior prediction performance.[20] However, a recent study has shown that for some molecules and properties, 3D coordinates did not necessarily lead to improved results over more simple representations of 2D connectivity and atom types (i.e., SMILES[26] notation).[21] Additionally, while precise, absolute QM-derived atomization energies are often inaccurate by up to a full Hartree for common molecules (627 kcal/mol). [27] Direct prediction of reaction energies may therefore be more reliable when compared to experimental values.

For the prediction of BDEs, a previous study leveraged >12,000 DFT calculations and an associative neural network to achieve a mean absolute error (MAE) of 3.4 kcal/mol for unseen bonds relative to DFT results.[28] This model is based on fixed molecular descriptors calculated for each target bond, and thus does not allow the model to learn more detailed descriptions of each bond as more molecular structures and data is added. B3LYP values were used to train this model, however, this functional poorly captures the enthalpies of radical reactions.[29] In our own benchmarking studies this level of theory has an average error 2 kcal/mol larger than other DFT methods against experimental BDE values (see below, Figure 1A). Other existing work has used neural networks to predict the contribution of each bond to the overall atomization energy of closed-shell molecules without explicitly calculating radical enthalpies.[30] While this technique reproduces general trends in overall bond strength, quantitative comparison with experimental BDEs results in MAEs of approximately 10 kcal/mol. More generally, the use of atomization energies as a benchmark for ML algorithms does not guarantee accuracy in predicting more chemically-relevant reaction energies.[31,32] The development of an accurate ML pipeline to quickly estimate BDEs, with acceptable accuracy compared to experimental values, thus remains a challenge.

In this study, we develop **A** machine-**L**earning derived, **F**ast, **A**ccurate **B**ond dissociation **E**nthalpy **T**ool (**ALFABET**) to predict homolytic BDEs at close to chemical accuracy with sub-second computational cost. To accomplish this, we first benchmark several quantum chemistry methods on a database of experimentally measured BDEs,[33] finding that the M06-2X/def2-TZVP level of theory has the optimal tradeoff between empirical accuracy and computational efficiency. A

database of 42,577 closed-shell compounds with nine or fewer heavy atoms and consisting only of C, H, O, and N atoms was then curated from PubChem.[34] Each single bond in the database that was not present in a ring was cleaved to yield two open-shell radicals. DFT enthalpy calculations were then performed on all open and closed-shell molecules to yield 290,664 unique BDEs, representing over 80 days of total CPU time. We then trained a graph neural network on a subset of these results, achieving a MAE of 0.58 kcal/mol when predicting BDEs for unseen closed-shell molecules (compared with DFT results). When compared against experimental values for large molecules not included in the training set, the ML method adds only 1 kcal/mol to the MAE of the DFT approach, while completing in less than a second (compared with over a day per molecule for DFT). The utility of the developed prediction tool was subsequently demonstrated on two separate applications where fast, accurate prediction of the weakest bond in a molecule is required. First, the model was used to rapidly and accurately predict the site C-H oxidation by cytochrome P450 metabolism in large, drug-like molecules. The model replicates the results of much more expensive DFT calculations with an MAE of 1.14 kcal/mol, and 95% of metabolic sites occur at bonds within 2 kcal/mol of the weakest bond in the molecule. Second, the model was used to predict the dominant radicals formed during combustion of fuel molecules, and the identities of these radicals were used as features for a QSPR model of soot formation pathways. These applications demonstrate the broad applicability of the developed tool and demonstrate that bond strength prediction for organic molecules can be reliably performed using fast ML techniques.

Results

Evaluation of QM methods for calculating homolytic BDEs

In order to ensure that the resulting ML method closely reproduced experimentally-determined BDEs, we performed a benchmark study of common DFT and *ab initio* methods. Computed gas-phase BDE values include unscaled vibrational zero-point energies and thermal corrections to the enthalpy at 298K and 1 atm, using optimized geometries obtained following a conformational search (see below). For a set of 368 experimentally measured BDEs from the *iBond* database,[33] combinations of three different DFT functionals (B3LYP-D3,[35,36] ωB97XD,[37] and M06-2X[38]) and two basis sets (6-31G(d) and def2-TZVP) were compared to DLPNO-CCSD(T)/cc-PVTZ calculations (Figure 1A). As expected, the CCSD(T) calculations took the longest to perform and were the most accurate. Of the DFT methods, the choice of basis set appeared to have the greatest impact on accuracy, with the M06-2X/def2-TZVP combination coming very close to CCSD(T) accuracy. MAEs of the three density functionals followed the order of B3LYP-D3>ωB97XD>M062X for both basis sets. This is consistent with previous benchmarks against the stabilization energy of 43 radical species calculated using CCSD(T)/CBS.[31,39,40] The observed MAE of top performing methods approaches the underlying uncertainty in the experimental measurements.

Conformer sampling was performed using the RDKit library,[41] using the MMFF94s forcefield.[42] Between 100 and 1000 conformers were generated for each molecule, depending on the number of rotatable bonds. The lowest-energy conformer identified by forcefield calculations was then used as an initial guess for subsequent geometry optimization at the higher level of theory. For radicals, initial structures were generated by temporarily replacing the radical with a bonded H atom during force field optimizations. The enthalpy of formation of this first conformer was denoted $\Delta H_{f,0}$. As a reordering of conformational energies often occurs upon reoptimizing MM geometries with a higher level of theory, we analyzed the typical error introduced by only optimizing the MM global minimum energy conformer at the higher level of theory. By optimizing additional higher-energy (i.e., local minima) MM conformers we can calculate the difference between our initial

enthalpy estimation, $\Delta H_{f,0}$, and the Boltzmann-weighted enthalpy of the entire conformer ensemble, $\langle \Delta H_f \rangle$. The difference between these quantities is plotted in Figure 1B, indicating that the median error introduced by only optimizing a single conformer (versus an ensemble of over 100) is only approximately 0.5 kcal/mol, while requiring 1/100th the computational resources. We therefore proceeded with database construction using M06-2X/def2-TZVP to optimize only the most stable MM conformer.

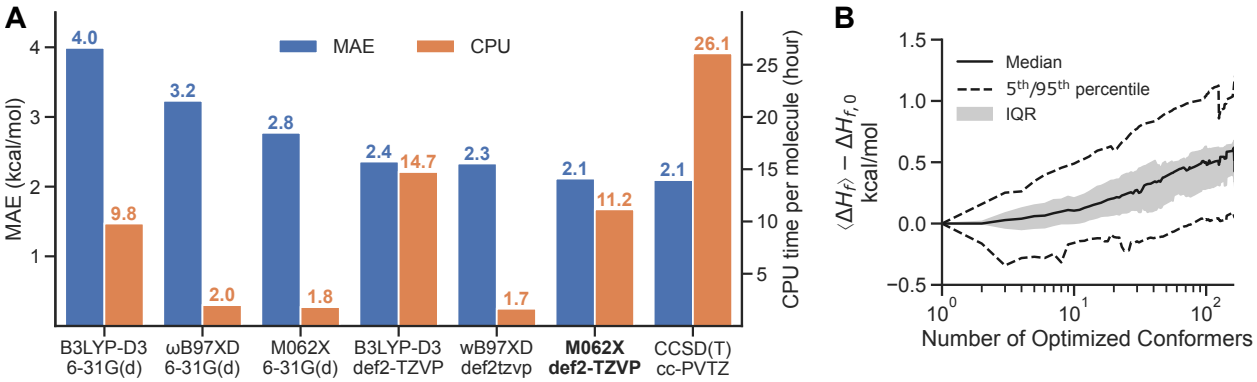

Figure 1: Benchmark study of DFT methods. (A) Trade-off between accuracy (left axis, blue) and computational cost (right axis, orange) for a selection of common QM methods. M062X/def2tzvp was selected for subsequent calculations. MAE and CPU time were averaged over 368 different bonds. (B) Effect of conformer sampling. Molecules were optimized with MMFF94, and the lowest-energy conformers were used to initialize DFT calculations. The plot shows the difference between the Boltzmann average enthalpy for the entire ensemble and the DFT-calculated enthalpy of the first conformer as a function of the number of optimizations performed. Exhaustive conformer sampling only changes the median resulting enthalpies by <0.5 kcal/mol, with a relatively narrow inner quartile range (IQR).

Construction of a machine-learning compatible BDE database

We next developed a large database of BDE values, **BDE-db**, on which to train ALFABET. To maximize the variety of bond strengths for a minimum computational effort, we limited the initial database construction to molecules with 9 or fewer heavy atoms. Additionally, smaller molecules reduce the risk of the geometry optimization finding a local energy minimum substantially higher than the true global minimum.

Construction of BDE-db began with 42,557 "parent" $C_xH_yO_zN_m$ molecules taken from PubChem (Figure 2A). Each single, non-cyclic bond in these molecules was then cleaved to generate two "child" radicals which were also added to the database. Canonicalized SMILES strings with specified configuration at stereogenic centers were used to represent these molecules and remove duplicates (Figure 2B). "Child" radicals were frequently the product of multiple BDE reactions, reducing the number of DFT calculations required. However, this use of the SMILES language presents some complications for database construction. Specifically, bond cleavage occurring within an enantiotopic or diastereotopic group (that are not differentiated by SMILES) forms radicals with a new and unspecified stereocenter in relation to the parent molecule. The creation of new diastereomeric relationships in the products gives rise to non-equivalent BDE values dependent upon the choice of relative configuration. Dissociations resulting in a new stereocenter were omitted from the database.

DFT calculations were then performed for the parent molecules and unique child radicals. A variety of convergence checks were performed to ensure the DFT optimization converged to a

stable structure, including checks for imaginary frequencies and ensuring that the molecule did not further decompose into disconnected molecules (e.g., radical fragmentation of an alkoxyacyl radical into an alkyl radical by loss of CO_2) or suffer an intramolecular rearrangement (e.g., by a [1,*n*]-H shift). Approximately 10% of attempted DFT calculations were discarded, primarily due to imaginary frequencies. A total of 249,374 successful calculations were used to build the BDE-db. These calculations resulted in 484,907 total calculated BDEs, of which 290,664 were unique (methane has only one "unique" BDE value). These numbers highlight the efficiency gains achieved through calculating a large database in parallel and re-using calculation results for child radicals, as typically three QM calculations are required per one BDE.

Development of a graph neural network for predicting BDE

A graph neural network (GNN) was developed to predict BDE directly from molecular structure. GNNs in the past have been used to predict the enthalpy of molecules from their optimized 3D structure, with MAEs close to 0.3 kcal/mol.[22] The application of this technique for the proposed target would require optimized 3D structures of both the parent molecule and child radicals, and prediction errors would likely compound when summing together three separate predictions. We instead sought to develop a model that only required the 2D structure (i.e., SMILES string) of the parent molecule as input. SMILES strings were converted to a graph representation using RDKit (with atoms as nodes and bonds as edges). Each bond in the molecule was represented by two directional edges, pointing in reverse directions between the two bonded atoms.

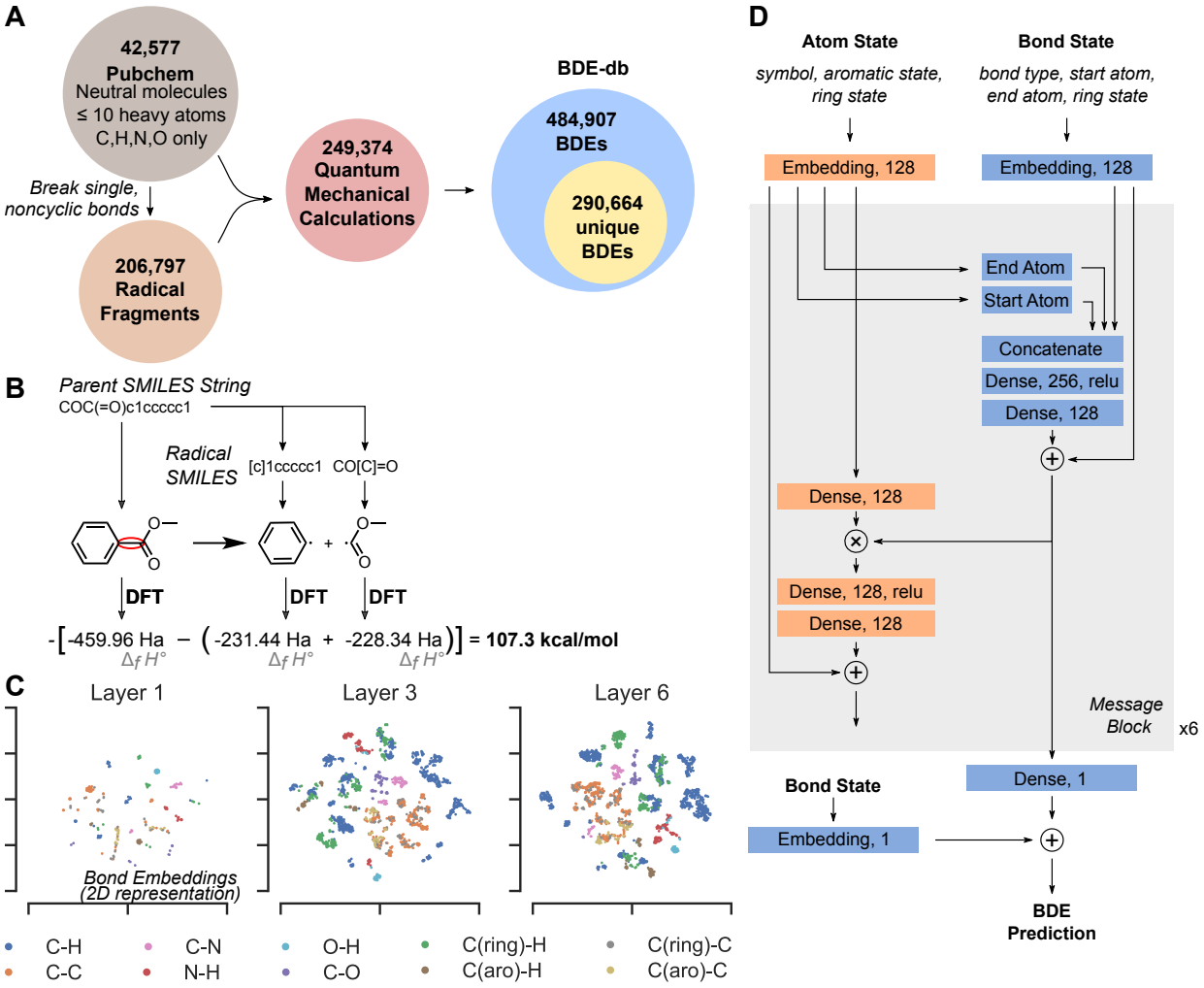

Figure 2: Overview of database construction and GNN structure. (A) Size of key elements of BDE-db. (B) Indexing and calculation of a single BDE reaction. For a given cleaved bond, SMILES strings of the parent molecule and two resulting radicals are passed for DFT optimization. (C) 2D representations of bond embeddings are shown via the t-SNE algorithm after the first, third, and final message passing layers. Initially, bonds of similar classes are clustered close together in embedding space. For deeper layers of the model, representations of the bonds become more detailed as they represent its specific local environment. (D) Structure of the GNN. Atom and bond state vectors are updated through a series of 6 message passing blocks. The final embedding layer is then used to predict the BDE of each bond.

GNNs operate by mixing information between neighboring nodes and edges. By iteratively updating node and edge internal states depending on the internal states of their neighbors, embedding vectors are generated that serve as a finite-dimensional description of each atom or bond's local environment (Figure 2C). For BDE prediction, bond embedding vectors at the final layer are reduced through a linear layer to predict the BDE (predictions from both the forward and backward bond edge are averaged together). The overall network structure was inspired by a model from Jørgensen *et al*.,[43] but with a simplified interaction structure. As only 2D inputs are used, atom and bond vectors are initialized with embedding layers based on a number of properties inferred via RDKit (Figure 2D). In each message passing layer, bond states are first updated with information from neighboring atoms, and atom states are then updated with information from neighboring bonds. Residual connections were used for each message passing layer in order to aid convergence of deeper models.[44] Six message passing layers were used in the final model, as no improvement in accuracy was seen for additional layers. The final model structure contains 1.06M parameters. Bond states from the final message passing layer are reduced to a single BDE prediction by passing them through a linear layer. Following SchNet,[22] these predictions were added to a single mean BDE value for each bond class to generate the final prediction. BDE predictions are therefore generated simultaneously for each bond in the input molecule.

Validation ("dev") and test sets were each constructed from all BDEs associated with 1000 parent molecules. The training set thus consisted of 40,577 unique parent molecules and 276,717 unique BDEs. Performance of the final model was tested against the held-out test set, consisting of 6,948 unique BDEs. The MAE on these bonds was 0.58 kcal/mol, with 95% of predictions falling within 2.25 kcal/mol of their DFT-calculated values (Figure 3A). Since the goal of the method is ultimately to reproduce experimental BDE measurements, the speed and accuracy of the GNN on the iBond database was compared to the DFT method (Figure 3B). For molecules that were a part of the training set, the ML method is able to closely reproduce DFT results with only a slight increase in MAE (2.4 kcal/mol for ML, 2.1 kcal/mol for DFT). However, a more difficult test of the ML approach is for molecules larger than 9 heavy atoms that were not a part of the training database. For these larger molecules, average DFT computation times were over a day per molecule. However, the accuracy of the ML method remained acceptable, adding less than 1 kcal/mol to the MAE of the DFT method (3.4 kcal/mol for ML, 2.5 kcal/mol for DFT). For these molecules, ALFABET was able to predict BDEs for all the bonds in the molecule in under 1ms per molecule.

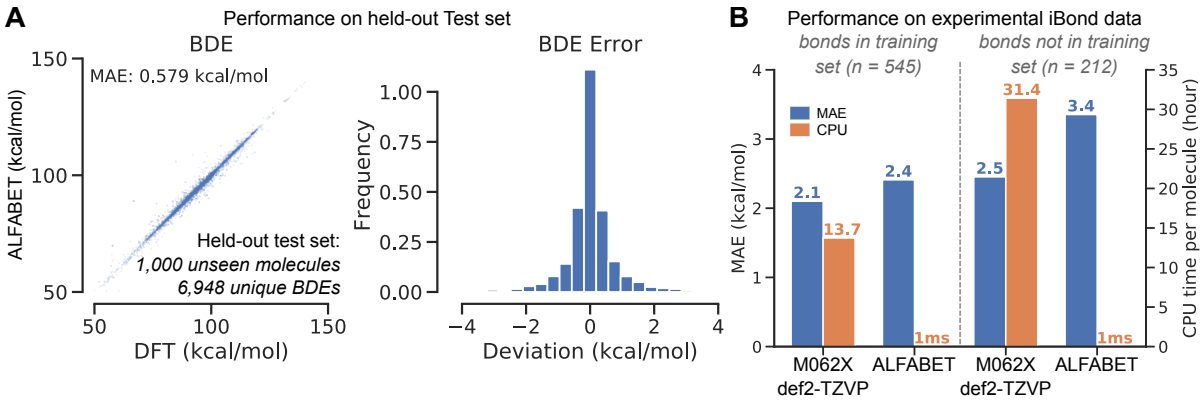

Figure 3: Performance of the ML BDE prediction algorithm. (A) Performance on the held-out, DFT-generated test set. (*left*) Parity plot of ALFABET predictions vs. DFT calculations. (*right*) Histogram of prediction errors. The model achieves an MAE of 0.579 kcal/mol on unseen molecules. (B) Performance of the model on experimentally measured BDEs from the iBond database. Prediction accuracy was quantified separately for bonds inside the training database (left) and those outside it (right). Molecules and bonds outside the training set tended to be much larger, thus resulting in larger DFT error and long DFT computational times.

Analysis of ALFABET prediction outliers

During construction of BDE-db and ALFABET, we conducted error analyses of preliminary data and models to refine the GNN structure and correct common DFT errors. In this section, we present a more extensive analysis of the remaining large prediction errors (>10 kcal/mol) for bonds in the training, validation, and test sets (Figure 4). In evaluating errors in DFT and ML calculations, additional BDE calculations were performed at the composite G4 level of theory to serve as a "ground truth" reference.[45] G4 radical formation enthalpies lie close to experimental values (4.5-6.2 kJ/mol), albeit at an increased computational cost relative to DFT.[39]

ML predictions using deep neural networks have been criticized as being "black-box" in nature. However, in this study we use the bond embedding vectors from the final message passing layer to interpret the ALFABET predictions, generating a quantitative "similarity score" to bonds contained in the training database (see methods). These embeddings are subsequently reduced to a single BDE prediction, and thus neighboring bond BDEs indicate how the GNN interprets the input molecule. We found that significant errors can arise in either DFT reference data or the ALFABET predictions due to several recurring structural motifs. In this section, we present examples of several classes of errors that lead to disagreement between DFT calculations and predicted BDEs.

The loss of stabilizing non-covalent interactions such as intramolecular hydrogen bonds by bond dissociation result in prediction errors (Figure 4A). Relative to the internally H-bonded conformer **1a**, the G4 BDE value is 90.8 kcal/mol. Our DFT reference value was correctly generated using this more stable conformation. However, ALFABET underpredicts this C-H bond strength by 15 kcal/mol – and is much closer to the hypothetical BDE value of 79.0 kcal/mol for the less stable conformer (**1b**) lacking an H-bond. We can attribute this prediction error to a failure to account for this strong H-bond in the parent compound. Inspection of nearest neighbor structures in the training database (including a similar bond for a 7-membered cycloheptanone) confirm this to be the case, since optimized structures for these molecules lacked internal H-bonds and have DFT values in the ~80 kcal/mol range (Figure 5A). For molecules where an intermolecular H-bond is lost or disrupted upon bond cleavage, predictions will tend to underestimate the true BDE value.

Conversely, the development of new stabilizing interactions in radical products result in anomalously low BDE values that are overestimated by ALFABET predictions (Figure 4B). For example, the carboxyl radical formed from ***cis*-3** undergoes ring-closure to form a stabilized radical that results in an anomalously small BDE value of 51.4 kcal/mol. While the DFT value lies close to this, the prediction is an overestimate by more than 40 kcal/mol. However, ***trans*-3**, which differs only by the configuration of the central C=C bond, has a BDE value of 88.0 kcal/mol. Ring-closure cannot occur in this case. The BDE prediction lies close to this value and the failure for *cis*-3 can be attributed to the occurrence of radical cyclization.

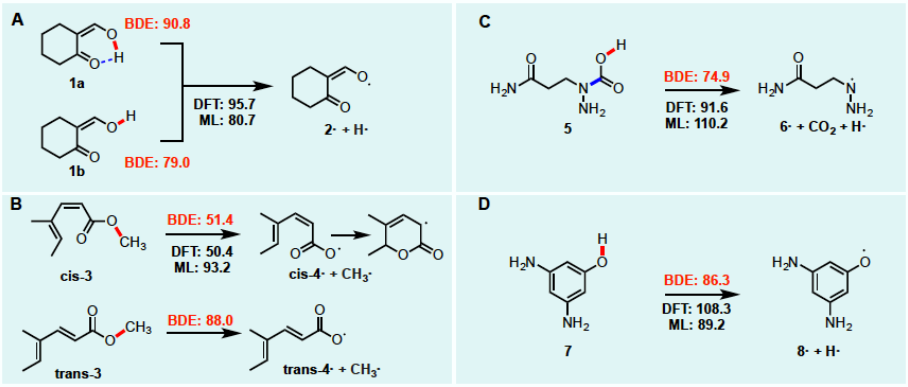

Figure 4: Error analysis of predicted (ML) and DFT-calculated BDE values against ground-truth G4 values (in kcal/mol) for representative molecules with large prediction errors. G4 BDEs are shown in red.

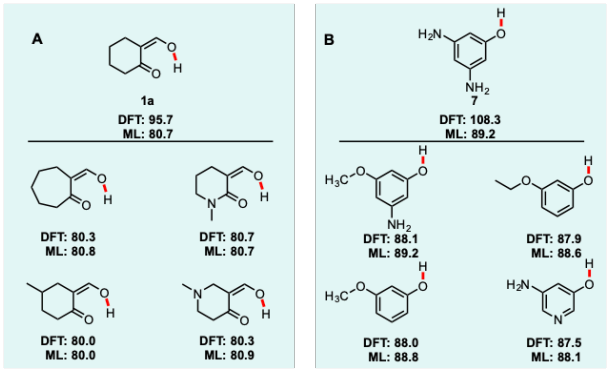

Figure 5: Similar bonds from the BDE-db database for two query bonds (top) from the BDE prediction outliers. (A) Bond from Figure 4A (**1a**). (B) Bond from Figure 4D (**7**).

In constructing the BDE-db database, we omitted reactions where a bond dissociation resulted in an unstable radical that further decomposed into smaller species. While G4 calculations (which use uB3LYP/6-31G(2df,p) geometries) suggest that O-H dissociation of a carbamic acid group (Figure 4C), results in the spontaneous loss of CO_2, M06-2X calculations result in a weakly-bound adduct with a N-C bond length of 1.63 Å. Relative to the G4 value, both DFT and ML predictions in this case are inaccurate.

Another scenario resulting in BDE prediction outliers arises from difficult-to-converge electronic structure calculations for strongly delocalized systems (Figure 4D). The O-H BDE values for phenols **7** is predicted by ALFABET as 89.2 kcal/mol, whereas the reference DFT value is much higher at 108.3 kcal/mol. The G4 value is much closer to the predicted BDE and suggest that in this case, it is the DFT value that is erroneous. Indeed, phenolic O-H bonds of neighboring molecules in the database have similar BDEs to the predicted value and further indicate that the

DFT result is the outlier (Figure 5B). The overestimate by DFT results from the convergence of open-shell structures to an incorrect electronic state. We found this was sensitive to the input structure used for geometry optimization and difficult to filter automatically (calculations are fully converged with a stable wavefunction) without prior knowledge of an expected BDE value.

In general, the most egregious ML-DFT prediction errors arise for conformations or electronic structures atypical with respect to the rest of the training database. Using 3D features as inputs to the ML model might alleviate some of these prediction errors, although this would increase the computational cost of the ML predictions (as 3D coordinates would be required to generate predictions) and the possibility would remain of passing sub-optimal 3D inputs to the ML model and generating correspondingly poor DFT predictions. Additional filtering of DFT results might allow more accurate ALFABET predictions. However, ML prediction methods will likely never be able to appropriately predict the results of medium- to long-range intramolecular interactions without sufficient training examples.

Application to Bond Dissociation in Large Molecules

We used ALFABET to predict the C-C, C-O, and C-H bonds in methyl linolenate, an unsaturated fatty acid methyl ester found in biodiesel (Figure 6). BDE values of biodiesel molecules are difficult to obtain experimentally and computational estimates are important for characterizing combustion chemistry, particularly the initial stages of pyrolysis. DFT BDE values have been obtained previously for methyl linolenate, in addition to multireference averaged coupled-pair functional (MRACPF2) values, which due to the large molecular size, were estimated using small surrogate models. The presence of C(sp^3)-H, C(sp^2)-H, C(sp^3)-O, C(sp^3)-C(sp^3), and C(sp^3)-C(sp^2) bond types and carbonyl and olefin functional groups provides a good opportunity to test model performance. Pleasingly, our model provides BDE values very close to M08-HX/ma-TZVP (MAE of 0.97 kcal/mol, R2 of 0.987[46]) and MRACPF2/CBS (MAE of 1.99 kcal/mol, R2 of 0.957[42]), across 33 single bonds ranging in strengths by 34 kcal/mol. The BDE values of weaker C-C and C-H bonds □-to the carbonyl and in allylic (and doubly-allylic) positions, along with those of stronger C(sp2)-C and C(sp2)-H bonds are all correctly described. This prediction, taking less than a second to complete, demonstrates the utility and accuracy of ALFABET for BDE prediction of larger, flexible hydrocarbons that are challenging to study by DFT and impossible for *ab initio* methods.

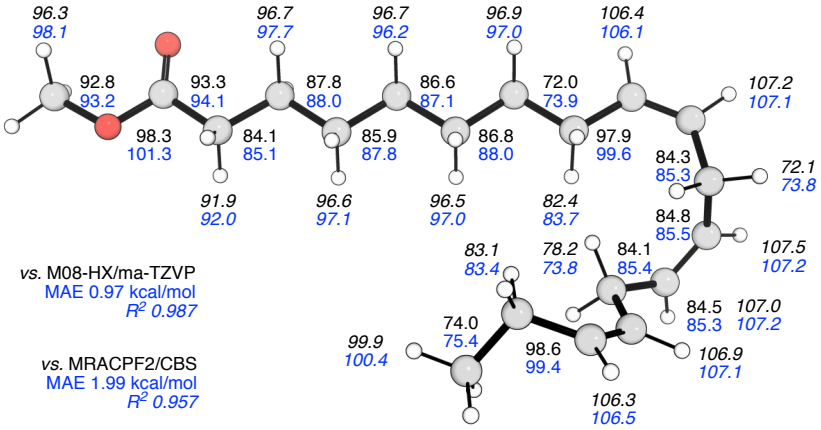

Figure 6: Application of the ML approach to quantitative BDE prediction in a large organic molecule. Comparison of BDE values (kcal/mol) of the C-C, C-O and C-H (italicized) bonds in methyl linolenate. ML values are in blue and M08-HX/ma-TZVP values are in black.

Application to Site of P450 Metabolism Prediction

The main advantage of the proposed method is that, due to its computational speed, it can be used in forward screening applications where DFT calculations would be infeasible. We therefore demonstrate the method's applicability to two design challenges where BDEs play an important role in determining a molecule's suitability. The first application is the pharmaceutical development of drug molecules, where predicting how a compound is likely to be metabolized can reduce failure rates in clinical trials.[47] Many xenobiotics are degraded by the cytochrome P450 enzyme, where the site of metabolism has been shown to correlate with the weakest C-H bond in the molecule.[9]

Calculation of C-H BDEs in drug screening, however, is a computationally expensive task, and we thus determined whether ALFABET demonstrates similar accuracy to a DFT-based calculation approach. We constructed a database of 28 drugs and their site of metabolism by cytochrome P450 from literature sources.[9,48-52] Drugs considered ranged in size from 6 to 32 heavy atoms. DFT calculations were then performed to determine the BDEs of all C-H bonds, and BDEs were also predicted using the developed GNN (Figure 7A).

We then developed a site of metabolism classifier using the calculated BDEs. The weakest bonds in the molecule, within a certain energy tolerance, were predicted as possible targets for cytochrome P450. The accuracy of the classifier, for BDEs derived both from DFT and from ALFABET, were quantified using a receiver operating characteristic (ROC) curve, Figure 7B. This curve plots the true positive rate versus the false positive rate as the classifier tolerance is adjusted. The area under the curve (AUC) of the ROC curve thus represents a quantitative measure of the classifier's performance, ranging from 0.5 (random guessing) to 1.0 (perfect predictions). The AUC for the DFT and ML-based classifiers was 0.86 and 0.87, respectively, indicating that the developed GNN is as accurate as DFT-based methods for predicting the site of metabolism, while requiring a fraction of the computational cost.

To verify that ALFABET predictions are accurate for BDEs of drug molecules much larger than those used to construct the training set, DFT calculations then performed for 82 top-selling drug molecules.[53] These molecules ranged in size between 8 and 34 heavy atoms. Only H-atom BDEs were considered, resulting in 748 unique bonds broken. Despite only being trained on smaller molecules, the GNN successfully predicts the BDEs for much larger species, resulting in a MAE of 1.14 kcal/mol (Figure 7C).

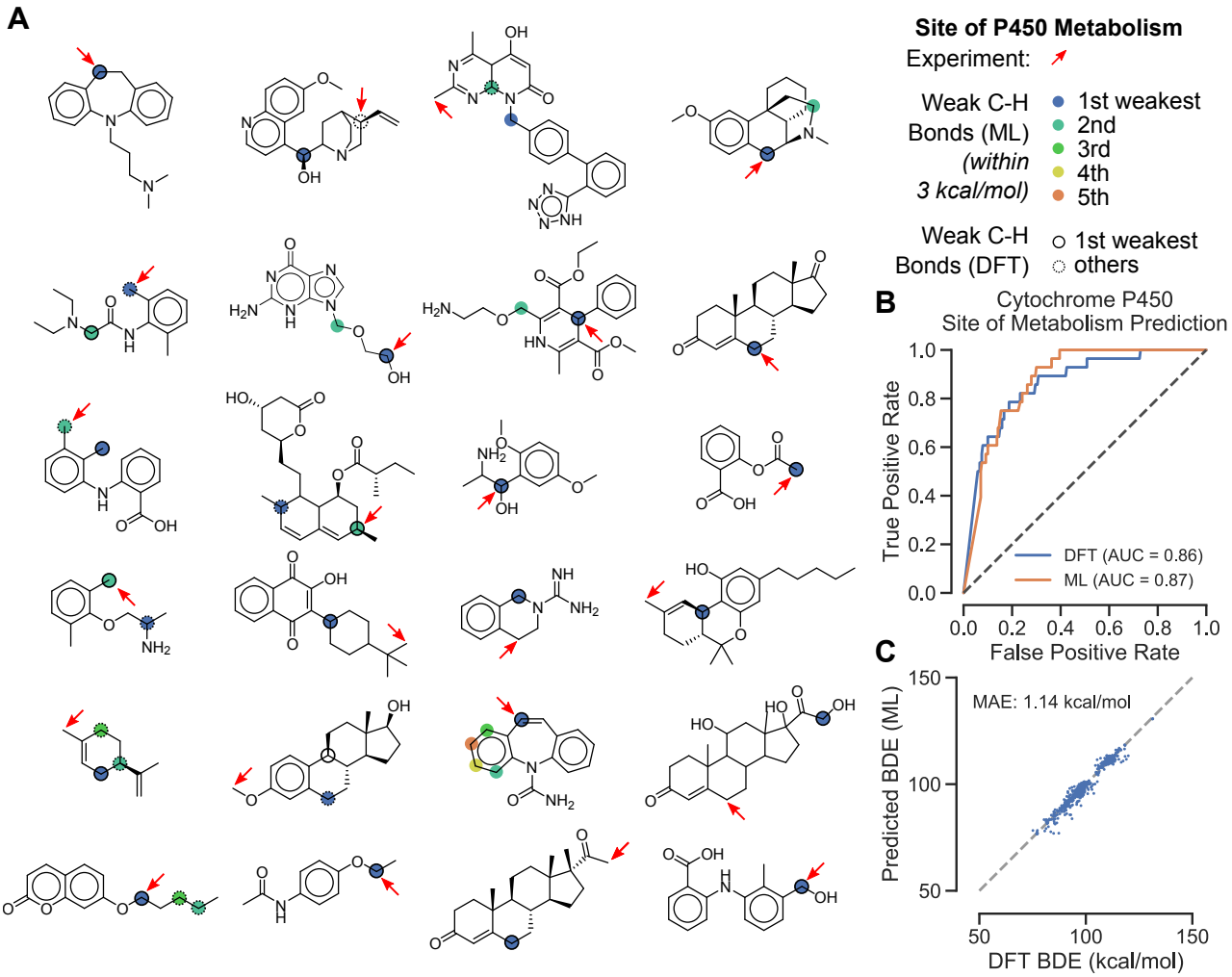

Figure 7: Application of ALFABET to predict site of cytochrome P450 metabolism. (A) Structures of many of drug molecules where the site of metabolism is known. Arrows indicate the experimentally determined breaking bond, while colors and circles indicate weakest bonds determined by ML and DFT, respectively. (B) ROC curve for classifiers that predict the metabolic site through BDEs generated through ML or DFT. Both approaches yield similar performance. (C) Accuracy of the ML method in predicting BDEs for 82 large, drug-like molecules.

Predicting combustion mechanisms from weakest bonds

In addition to metabolite decomposition, BDEs are essential in determining predominant combustion kinetic mechanisms. We next applied ALFABET to construct a mechanistically-inspired model of soot formation during combustion of new fuel chemistries. The yield sooting index (YSI) is an experimental measurement of the amount of soot a substance forms during combustion in a test flame,[54,55] and is an important parameter to consider during selection of potential fuel blendstocks.[56] While methods to predict YSI quickly from molecular structure exist,[55,57] these models do not leverage recent mechanistic understandings of how soot formation proceeds. Specifically, formation and growth of polyaromatic hydrocarbons (PAHs), the main component of particulate matter, is governed by the recombination of radicals formed in the combustion process.

In this study, we use our newly developed ML approach to predict the weakest bond in each of a set of 217 different fuel molecules with measured YSI values. The identities of the two radicals

that form are then used to construct a QSPR model to predict soot formation. Instead of a series of descriptors or functional groups, each molecule was represented by only two parameters: one for each of the two radicals formed during cleaving of the weakest bond. These parameters are shared between molecules that decompose to form identical radicals (Figure 8A). Molecules were chosen such that each radical was the result of at least 2 molecule decompositions.

We performed a leave-one-out cross-validation to determine the ability of the model to predict YSI for unseen molecules. In each cross-validation fold, a single compound was removed from the dataset and a weighted least-squares regression (with data weighted by their experimental uncertainty) was performed on the remainder of the data. Fitted radical weights are then used to predict the YSI of the held-out molecule. The cross-validated predictive accuracy of the new model, based on ALFABET predictions, achieves a weighted least-squares loss less than half that of a recently developed group contribution model on the same dataset (Figure 8B).[55] These results demonstrate that AFLABET predictions can improve forward screening approaches in which bond energy is an important parameter.

We further verified that ALFABET is accurate for larger molecules outside the training set considered in this application. For the 91 molecules with YSI measurements and between 11-20 heavy atoms, DFT calculations were performed to confirm the predicted BDEs. The resulting prediction error was even lower than for the withheld test set predictions (Figure 8C), demonstrating the ability of the model to scale to larger molecules.

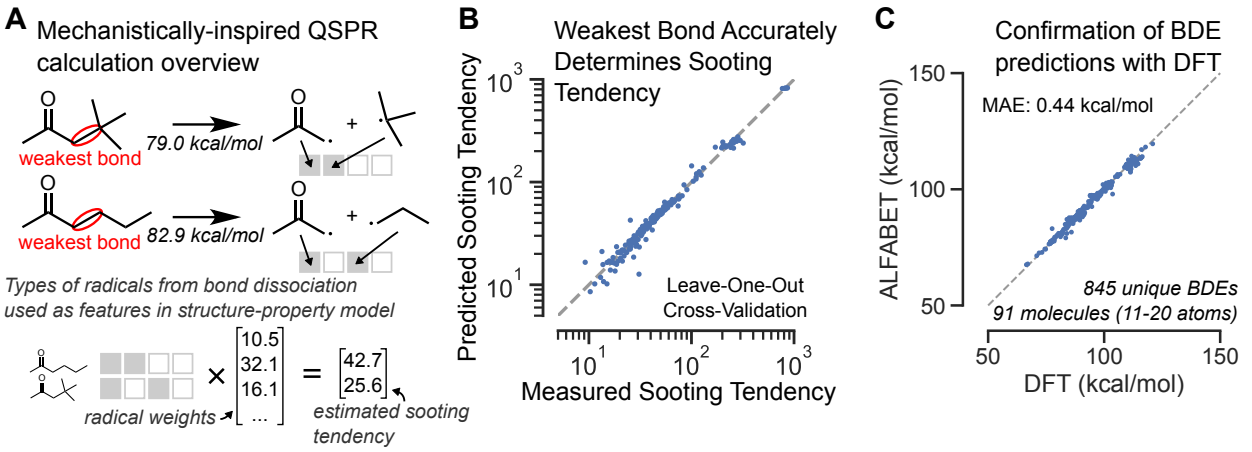

Figure 8: Development of a model for sooting tendency based on fast BDE calculation (A) Overview of the QSPR approach. ALFABET predictions are used to determine a molecules weakest bond, which identifies the radicals used as features in the QSPR. (B) Results of the QSPR model under leave-one-out cross-validation. The model achieves a superior accuracy to a previous group-contribution method. (C) Confirmation of the predictions for molecules larger than those included in the training set.

Methods

Computational details for calculating homolytic BDEs

To sample radical conformations, H atoms were added to radical centers prior to MMFF structure optimization and removed afterwards. MMFF94s performs well in conformational and noncovalent benchmarks involving neutral, closed-shell molecules,[58] however, it was not parametrized for radicals.[42] Unrestricted Kohn-Sham DFT calculations of radicals were carried out with careful consideration of electronic structures because M06-2X showed less accurate results in some aromatic radicals.[59,60] Specifically, spatial and spin symmetry of orbitals were broken by using the

initial guess of mixed HOMO-LUMO with assuming no point-group symmetry of the structure. The stability of "wavefunctions" was also analyzed to confirm that the most stable electronic state had been found.[61] Convergence to the wrong electronic state occurred most frequently for aromatic radicals. Gaussian 16[62] was used for all DFT calculations with a default ultra-fine grid for all numerical integration and for the G4 calculations to analyze outliers. DLPNO-CCSD(T) calculations were carried out with ORCA 4.0 as a single-point energy correction to the B3LYP-D3/6-31G(d) enthalpy using optimized geometries from B3LYP-D3/6-31G(d).[39]

All optimizations were checked for convergence to an energy minimum, which included checking for proper termination flags from Gaussian and ensuring the resulting structure had no imaginary vibrational frequencies. In addition, we verified that the molecule did not decompose into separate molecules during the Gaussian optimization by ensuring that all bond lengths (expected from the Lewis structure) were less than 0.4Å plus the sum of the covalent radii of the participating atoms. Finally, statistical tests on the completed database were used to screen for molecules with abnormally large enthalpies. For a given chemical formula (i.e., elemental composition), a linear model was used to predict overall molecule enthalpy. If residuals from this linear fit were greater than 3 inner quartile ranges from the predicted enthalpy, the molecule was discarded. This step removed a handful of high-energy, hypothetical molecules or ones that converged to unreasonable geometries. The BDE-db dataset has been published in an open-source database available on Figshare.[63]

Graph Neural Network Development

Determining the optimal inputs and structure to the GNN developed in this study was an iterative process in order to find one that yielded the lowest validation error. Nodes and edges were assigned to independent classes depending on a number of features. For nodes, unique classes were assigned based on an atom's symbol, chirality tag, aromaticity, presence in Ring (3, 4, 5, or $\geq$ 6), number of neighbors, and number of neighbor H's. Edge classes were assigned based on the start atom symbol, end atom symbol, and presence of the bond in ring (3, 4, 5, or $\geq$ 6). The edge interaction network and atom state updating layers from Jorgensen *et al*.[43] were simplified by removing layers until losses began to increase, and residual connections were added to the end of each message passing layers while batch normalization layers[64] were added to the beginning of each message passing layer. The number of message passing layers was varied between 2 and 12, with validation losses not decreasing after six layers. Since the number of atoms for molecules in the training set was capped at nine, this allows messages to traverse the entire molecule except in a few select cases.

The loss function optimized the mean absolute error of all BDEs in the molecule, masking bonds for which DFT values were not available. Since edges in the model are directional, each bond has two corresponding edge states. During training, the BDE prediction of each directional edge is separately scored, while at test time the BDE prediction from both edges is averaged. The model was trained for 500 epochs using a batch size of 128 molecules with the ADAM optimizer using a learning rate of 1E-3 and a decay rate of 1E-5.

GNN Implementation

GNN models were implemented using the Python nfp library (https://github.com/nrel/nfp), which provides extensions to the Keras deep learning framework for modeling graph-valued systems. Models were trained using a single Nvidia Tesla V100 GPU for approximately 10-12 hours. Weights for the final trained model and python scripts to generate predictions for new molecules has been made available through a Github repository (https://github.com/NREL/alfabet). Python scripts to train the model and Jupyter notebooks to create the figures in the paper are available at https://github.com/pstjohn/bde_model_methods.

Calculating Neighboring Bonds

Intermediate layers in the GNN could be used to search for similar bonds in the DFT database for a given query bond. Embedding vectors for all bonds with calculated BDE values were generated from the output of the final message passing layer, a 128-dimensional vector. For computational efficiency, these vectors were reduced to a 10-dimensional vector through a principal component analysis (PCA). A nearest-neighbors search was then used to find the 10 closest bonds in the BDE-db database. The scikit-learn library[65] was used to perform the PCA and nearest-neighbors searches.

Conclusions

In this study, we have developed a ML prediction tool to quickly calculate homolytic BDEs for organic molecules containing C, H, O, and N atoms, at an accuracy comparable with state-of-the-art DFT approaches. An interface for the developed prediction tool is available online at https://ml.nrel.gov/bde. Because BDEs are intrinsic properties of covalently bonded molecules, their relative strengths are important parameters in a wide range of chemical studies. We therefore expect our tool to enable high-throughput and accurate development of novel compounds. Beyond the application areas to drug design and combustion pathways considered in this paper, we expect our tool to be useful in understanding polymer thermal stability, lignin depolymerization pathways, explosives, and high-performance energy-related materials. More broadly, this study demonstrates the potential for deep learning techniques to accelerate quantum mechanical investigations where high-throughput computations are possible but time-consuming. Future work will look to expand these approaches to transition state structures.

Acknowledgement

We thank Michael Bartlett for assistance constructing and deploying the BDE prediction website. We also thank Kristin Munch for helpful conversations and assistance setting up the database for managing Gaussian calculations. Computational resources for P. C. St. John, Y. Kim, and S. Kim were provided by the Computational Sciences Center at National Renewable Energy Laboratory. R.S.P. gratefully acknowledges the RMACC Summit supercomputer supported by the National Science Foundation (ACI-1532235 and ACI-1532236), the University of Colorado Boulder and Colorado State University; the Extreme Science and Engineering Discovery Environment (XSEDE) through allocation TG-CHE180056; the support of NVIDIA Corporation for the donation of a Titan Xp GPU. This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy under the Co-Optima imitative. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

(1) Benson, S. *Thermochemical kinetics: methods for the estimation of thermochemical data and rate parameters*; Wiley: New York, 1976.

(2) Gani, T. Z. H.; Kulik, H. J. Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane to Methanol Conversion by Fe IV═O. *ACS Catal.* **2018**, *8* (2), 975–986 DOI: 10.1021/acscatal.7b03597.

(3) Kim, S.; Fioroni, G. M.; Ji-Woong, P.; Robichaud, D. J.; Das, D. D.; John, P. C. S.; Tianfeng, L.; McEnally, C. S.; Pfefferle, L. D.; Paton, R. S.; et al. Experimental and

theoretical insight into the soot tendencies of the methylcyclohexene isomers. *Proceedings of the Combustion Institute* **2018**, 1–8 DOI: 10.1016/j.proci.2018.06.095.
(4) Lin, C. Y.; Marque, S. R. A.; Matyjaszewski, K.; Coote, M. L. Linear-Free Energy Relationships for Modeling Structure–Reactivity Trends in Controlled Radical Polymerization. *Macromolecules* **2011**, *44* (19), 7568–7583 DOI: 10.1021/ma2014996.
(5) Giannetti, E. Thermal stability and bond dissociation energy of fluorinated polymers: A critical evaluation. *Journal of Fluorine Chemistry* **2005**, *126* (4), 623–630 DOI: 10.1016/j.jfluchem.2005.01.008.
(6) Bian, C.; Wang, S.; Liu, Y.; Jing, X. Thermal stability of phenolic resin: new insights based on bond dissociation energy and reactivity of functional groups. *RSC Adv.* **2016**, *6* (60), 55007–55016 DOI: 10.1039/C6RA07597E.
(7) Kim, S.; Chmely, S. C.; Nimlos, M. R.; Bomble, Y. J.; Foust, T. D.; Paton, R. S.; Beckham, G. T. Computational Study of Bond Dissociation Enthalpies for a Large Range of Native and Modified Lignins. *J. Phys. Chem. Lett.* **2011**, *2* (22), 2846–2852 DOI: 10.1021/jz201182w.
(8) Lienard, P.; Gavartin, J.; Boccardi, G.; Meunier, M. Predicting Drug Substances Autoxidation. *Pharm Res* **2014**, *32* (1), 300–310 DOI: 10.1007/s11095-014-1463-7.
(9) Drew, K. L. M.; Reynisson, J. The impact of carbon-hydrogen bond dissociation energies on the prediction of the cytochrome P450 mediated major metabolic site of drug-like compounds. *European Journal of Medicinal Chemistry* **2012**, *56* (C), 48–55 DOI: 10.1016/j.ejmech.2012.08.017.
(10) Zhao, S.-W.; Liu, L.; Fu, Y.; Guo, Q.-X. Assessment of the metabolic stability of the methyl groups in heterocyclic compounds using C-H bond dissociation energies: Effects of diverse aromatic groups on the stability of methyl radicals. *Journal of Physical Organic Chemistry* **2005**, *18* (4), 353–367 DOI: 10.1002/poc.856.
(11) Harris, N. J.; Lammertsma, K. Ab Initio Density Functional Computations of Conformations and Bond Dissociation Energies for Hexahydro-1,3,5-trinitro-1,3,5-triazine. *J. Am. Chem. Soc.* **1997**, *119* (28), 6583–6589 DOI: 10.1021/ja970392i.
(12) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, *33* (6-7), 469–476 DOI: 10.1002/minf.201400052.
(13) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186–190 DOI: 10.1126/science.aar5169.
(14) Wilcox, D. A.; Agarkar, V.; Mukherjee, S.; Boudouris, B. W. Stable Radical Materials for Energy Applications. *Annu. Rev. Chem. Biomol. Eng.* **2018**, *9* (1), 83–103 DOI: 10.1146/annurev-chembioeng-060817-083945.
(15) Blanksby, S. J.; Ellison, G. B. Bond Dissociation Energies of Organic Molecules. *Acc. Chem. Res.* **2003**, *36* (4), 255–263 DOI: 10.1021/ar020230d.
(16) Luo, Y. R. *Comprehensive handbook of chemical bond energies*; 2007.
(17) Feng, Y.; Liu, L.; Wang, J.-T.; Huang, H.; Guo, Q.-X. Assessment of Experimental Bond Dissociation Energies Using Composite ab Initio Methods and Evaluation of the Performances of Density Functional Methods in the Calculation of Bond Dissociation Energies. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2005–2013 DOI: 10.1021/ci034033k.
(18) Zhao, Y.; Truhlar, D. G. How Well Can New-Generation Density Functionals Describe the Energetics of Bond-Dissociation Reactions Producing Radicals? *J. Phys. Chem. A* **2008**, *112* (6), 1095–1099 DOI: 10.1021/jp7109127.
(19) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59* (6), 2545–2559 DOI: 10.1021/acs.jcim.9b00266.
(20) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv.org*. April 4, 2017.

(21) St John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **2019**, *150* (23), 234111 DOI: 10.1063/1.5099132.
(22) 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. *arXiv.org*. June 26, 2017.
(23) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv.org*. June 4, 2018.
(24) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Lilienfeld, von, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13* (11), 5255–5264 DOI: 10.1021/acs.jctc.7b00577.
(25) Feinberg, E. N.; Sheridan, R.; Joshi, E.; Pande, V. S.; Cheng, A. C. Step Change Improvement in ADMET Prediction with PotentialNet Deep Featurization. *arXiv.org*. March 27, 2019.
(26) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36 DOI: 10.1021/ci00057a005.
(27) Hoffmann, R.; Schleyer, P. V. R.; Schaefer, H. F., III. Predicting Molecules-More Realism, Please! *Angew. Chem. Int. Ed.* **2008**, *47* (38), 7164–7167 DOI: 10.1002/anie.200801206.
(28) Qu, X.; Latino, D. A.; Aires-de-Sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *Journal of Cheminformatics* **2013**, *5* (1), 1–13 DOI: 10.1186/1758-2946-5-34.
(29) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. Should Contemporary Density Functional Theory Methods Be Used to Study the Thermodynamics of Radical Reactions? *J. Phys. Chem. A* **2007**, *111* (42), 10754–10768 DOI: 10.1021/jp075837w.
(30) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8* (12), 2689–2694 DOI: 10.1021/acs.jpclett.7b01072.
(31) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13* (14), 6670–19 DOI: 10.1039/c0cp02984j.
(32) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19* (48), 32184–32215 DOI: 10.1039/C7CP04913G.
(33) Internet Bond-energy Databank (pKa and BDE)--iBonD Home Page.
(34) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **2018**, *47* (D1), D1102–D1109 DOI: 10.1093/nar/gky1033.
(35) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652 DOI: 10.1063/1.464913.
(36) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initioparametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104–154120 DOI: 10.1063/1.3382344.

(37) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10* (44), 6615–6616 DOI: 10.1039/b810189b.
(38) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor Chem Account* **2007**, *120* (1-3), 215–241 DOI: 10.1007/s00214-007-0310-x.
(39) Neese, F.; Schwabe, T.; Kossmann, S.; Schirmer, B.; Grimme, S. Assessment of Orbital-Optimized, Spin-Component Scaled Second-Order Many-Body Perturbation Theory for Thermochemistry and Kinetics. *J. Chem. Theory Comput.* **2009**, *5* (11), 3060–3073 DOI: 10.1021/ct9003299.
(40) Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals—Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2010**, *7* (2), 291–309 DOI: 10.1021/ct100466k.
(41) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574 DOI: 10.1021/acs.jcim.5b00654.
(42) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5-6), 490–519 DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.
(43) Jørgensen, P. B.; Jacobsen, K. W.; Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. *arXiv.org*. June 8, 2018.
(44) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv.org*. December 10, 2015.
(45) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126* (8), 084108–084113 DOI: 10.1063/1.2436888.
(46) Li, X.; Xu, X.; You, X.; Truhlar, D. G. Benchmark Calculations for Bond Dissociation Enthalpies of Unsaturated Methyl Esters and the Bond Dissociation Enthalpies of Methyl Linolenate. *J. Phys. Chem. A* **2016**, *120* (23), 4025–4036 DOI: 10.1021/acs.jpca.6b02600.
(47) de Groot, M. J. Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discovery Today* **2006**, *11* (13-14), 601–606 DOI: 10.1016/j.drudis.2006.05.001.
(48) Lienard, P.; Gavartin, J.; Boccardi, G.; Meunier, M. Predicting Drug Substances Autoxidation. *Pharm Res* **2014**, *32* (1), 300–310 DOI: 10.1007/s11095-014-1463-7.
(49) Andersson, T.; Broo, A.; Evertsson, E. Prediction of Drug Candidates' Sensitivity Toward Autoxidation: Computational Estimation of C-H Dissociation Energies of Carbon-Centered Radicals. *Journal of Pharmaceutical Sciences* **2014**, *103* (7), 1949–1955 DOI: 10.1002/jps.23986.
(50) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting Drug Metabolism: A Site of Metabolism Prediction Tool Applied to the Cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46* (12), 2313–2324 DOI: 10.1021/jm021104i.
(51) Kumar, G. N.; Surapaneni, S. Role of drug metabolism in drug discovery and development; John Wiley & Sons, Ltd, 2001; Vol. 21, pp 397–411.
(52) Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **2006**, *34* (90001), D668–D672 DOI: 10.1093/nar/gkj067.
(53) The Top 300 of 2018. *https://clincalc.com/DrugStats/Top300Drugs.aspx*.
(54) McEnally, C. S.; Pfefferle, L. D. Improved sooting tendency measurements for aromatic hydrocarbons and their implications for naphthalene formation pathways. *Combustion and Flame* **2007**, *148* (4), 210–222 DOI: 10.1016/j.combustflame.2006.11.003.

(55) Das, D. D.; St John, P. C.; McEnally, C. S.; Kim, S.; Pfefferle, L. D. Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale. *Combustion and Flame* **2018**, *190*, 349–364 DOI: 10.1016/j.combustflame.2017.12.005.
(56) Huo, X.; Huq, N. A.; Stunkel, J.; Cleveland, N. S.; Starace, A. K.; Settle, A. E.; York, A. M.; Nelson, R. S.; Brandner, D. G.; Fouts, L.; et al. Tailoring diesel bioblendstock from integrated catalytic upgrading of carboxylic acids: a “fuel property first” approach. *Green Chem.* **2019**, *4*, 83–15 DOI: 10.1039/C9GC01820D.
(57) St John, P. C.; Kairys, P.; Das, D. D.; McEnally, C. S.; Pfefferle, L. D.; Robichaud, D. J.; Nimlos, M. R.; Zigler, B. T.; McCormick, R. L.; Foust, T. D.; et al. A Quantitative Model for the Prediction of Sooting Tendency from Molecular Structure. *Energy Fuels* **2017**, *31* (9), 9983–9990 DOI: 10.1021/acs.energyfuels.7b00616.
(58) Paton, R. S.; Goodman, J. M. Hydrogen Bonding and π-Stacking: How Reliable are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *J. Chem. Inf. Model.* **2009**, *49* (4), 944–955 DOI: 10.1021/ci900009f.
(59) Tishchenko, O.; Truhlar, D. G. Benchmark Ab Initio Calculations of the Barrier Height and Transition-State Geometry for Hydrogen Abstraction from a Phenolic Antioxidant by a Peroxy Radical and Its Use to Assess the Performance of Density Functionals. *J. Phys. Chem. Lett.* **2012**, *3* (19), 2834–2839 DOI: 10.1021/jz3011817.
(60) Galano, A.; Muñoz-Rugeles, L.; Alvarez-Idaboy, J. R.; Bao, J. L.; Truhlar, D. G. Hydrogen Abstraction Reactions from Phenolic Compounds by Peroxyl Radicals: Multireference Character and Density Functional Theory Rate Constants. *J. Phys. Chem. A* **2016**, *120* (27), 4634–4642 DOI: 10.1021/acs.jpca.5b07662.
(61) Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree–Fock theory. *J. Chem. Phys.* **1977**, *66* (7), 3045–3050 DOI: 10.1063/1.434318.
(62) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. Gaussian 16 Rev. C.01. **2016**.
(63) John, P. S.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. BDE-db: A collection of 290,664 Homolytic Bond Dissociation Enthalpies for Small Organic Molecules. Figshare November 4, 2019.
(64) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv.org*. February 10, 2015.
(65) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* **2011**, *12*, 2825–2830.