Supplementary Materials for

Expansion of Bond Dissociation Prediction with Machine Learning to Medicinally and Environmentally Relevant Chemical Space

Shree Sowndarya S. V[†], Yeonjoon Kim[†], Seonah Kim[†], Peter C. St. John[‡], Robert S. Paton[†]

[†]Department of Chemistry, Colorado State University, Fort Collins, CO, 80523, USA

[‡]Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401, USA

E-mail: seonah.kim@colostate.edu; pstjohn@nvidia.com; robert.paton@colostate.edu

TABLE OF CONTENTS

1. Development of the *BDE-db2* dataset

A total of 244,850 Density Functional Theory (DFT) calculations were performed for both the (closed-shell) parent molecules and homolytically-dissociated (open-shell) radicals. All structures were fully optimized at the (U)M06-2X/def2-TZVP[1, 2] level of theory with Gaussian 16.[3] Starting from the SMILES representation (H-capped for open-shell species) we identify the lowest energy conformer from RDKit[4, 5] and use this geometry as the starting point for DFT optimization and to obtain the energetics and thermochemistry of open- and closed-shell species (Fig. S1A)). The forcefield utilized is MMFF94s, the number of embedded conformers is dependent on the number of rotatable bonds (n) where the mid of 100, 3^n, 1000 is used with the following setting for EmbedMultipleConfs (pruneRmsThresh=0.2, randomSeed=1, useExpTorsionAnglePrefs=True, useBasicKnowledge=True). Following this workflow, 219,834 calculations terminated normally while 25,016 encountered some form of automatically detected error – most frequently the presence of an imaginary frequency or a failure to determine a suitable initial 3D conformer from the SMILES embedding. Before additional data quality checks, these newly added calculations describe 38,277 new small molecules and 199,209 unique BDE values. Augmenting our original efforts[6, 7] with these values results in a total of 509,740 DFT computed molecular enthalpies and 531,244 unique homolytic BDE values.

A linear model was used to detect outliers in computed absolute enthalpy values by regressing against the element counts (including explicit Hs) as independent variables. The linear model uses DFT calculated absolute enthalpy values (in Hartree) as target values. The Inner Quartile Range (IQR) was tabulated for the residuals, and calculations that were more than three times the IQR from the upper quartile were removed. This technique finds molecules with exceptionally high enthalpies for their given molecular composition, typically implying convergence to a particularly unstable conformation. Of the 509,740 enthalpy calculations, 4,309 were removed as enthalpy outliers with this method. The scatter plot for detection of outliers with abnormally large contributions from ΔZPE is shown in (Fig. S1B). The resulting B*DE-db2*, dataset is shared openly and hosted on GitHub at https://github.com/patonlab/BDE-db2. The dataset is available in the folder titled Dataset/bde-db2.

All additional data for studies involving test set 2 and test set 3 can be found in the following GitHub location https://github.com/patonlab/BDE-db2/Datasets

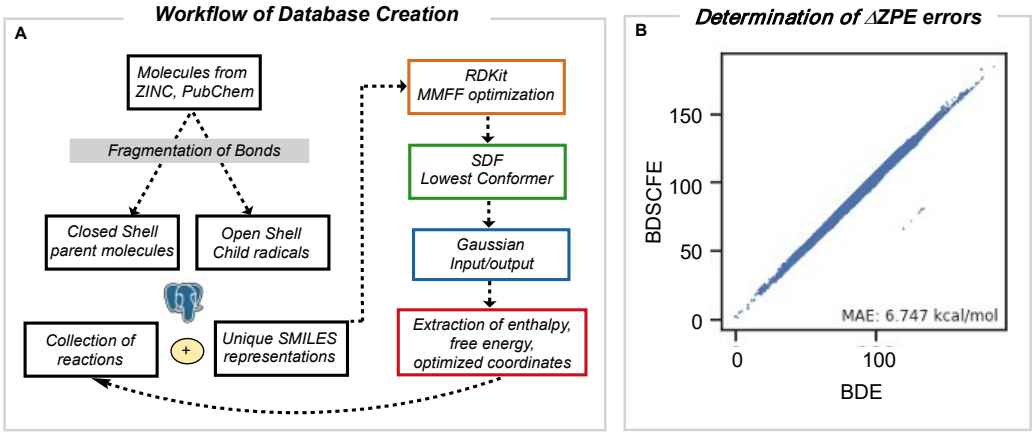

Figure S1. (A) Workflow for the construction of an updated bond dissociation enthalpy database containing halogenated species. (B) Plot of dissociation energy (BDSCFE) vs enthalpy (BDE) exposes errors due to large and unphysical changes in ZPE for a given dissociation, for which the reaction data is removed.

2. Breakdown of types of bond dissociation in the *BDE-db2* dataset

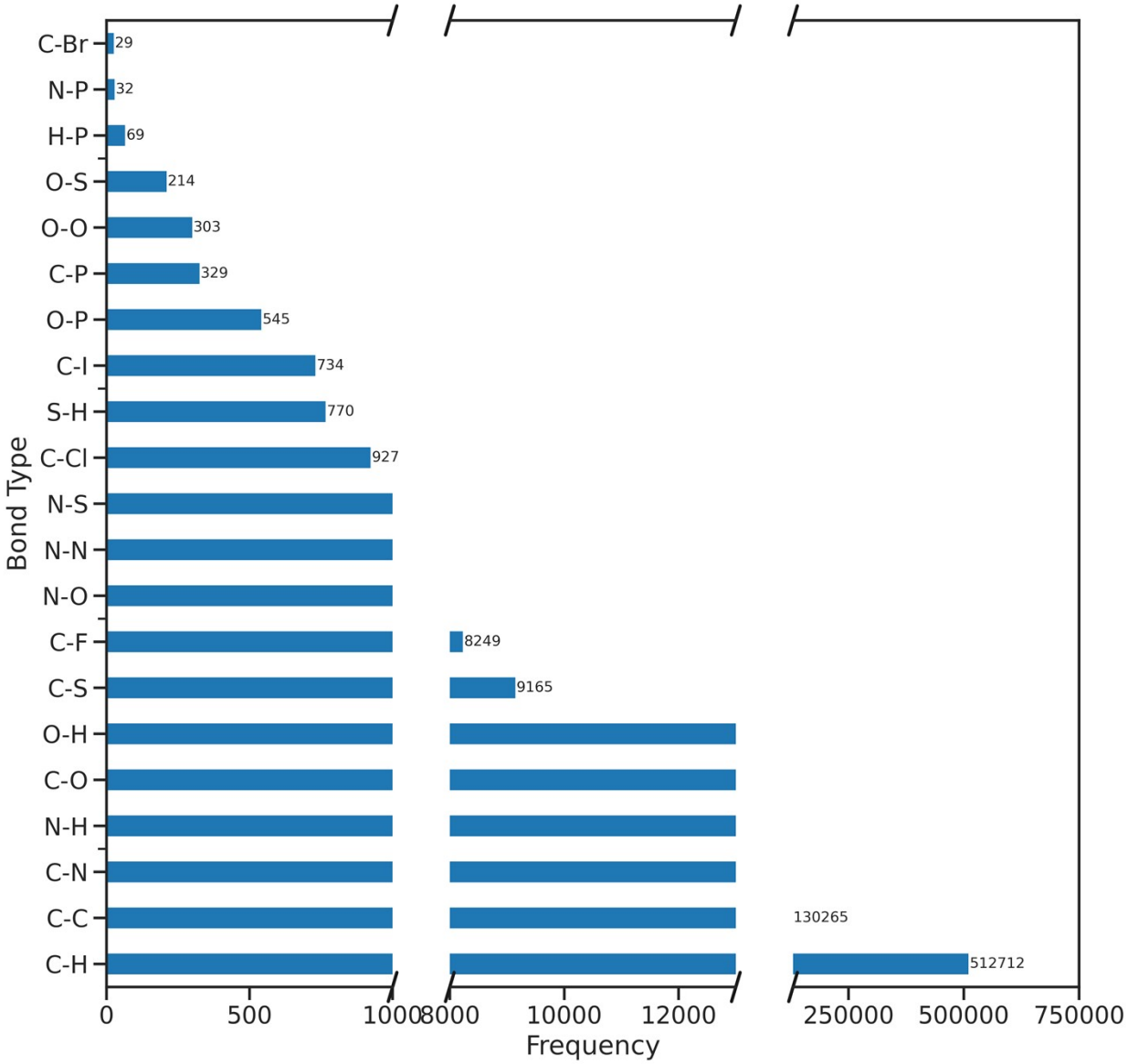

Figure S2. Frequency counts of all dissociation bond-types greater than 25 present in *BDE-db2*.

3. Prediction accuracy of newly added bond types in held-out test set

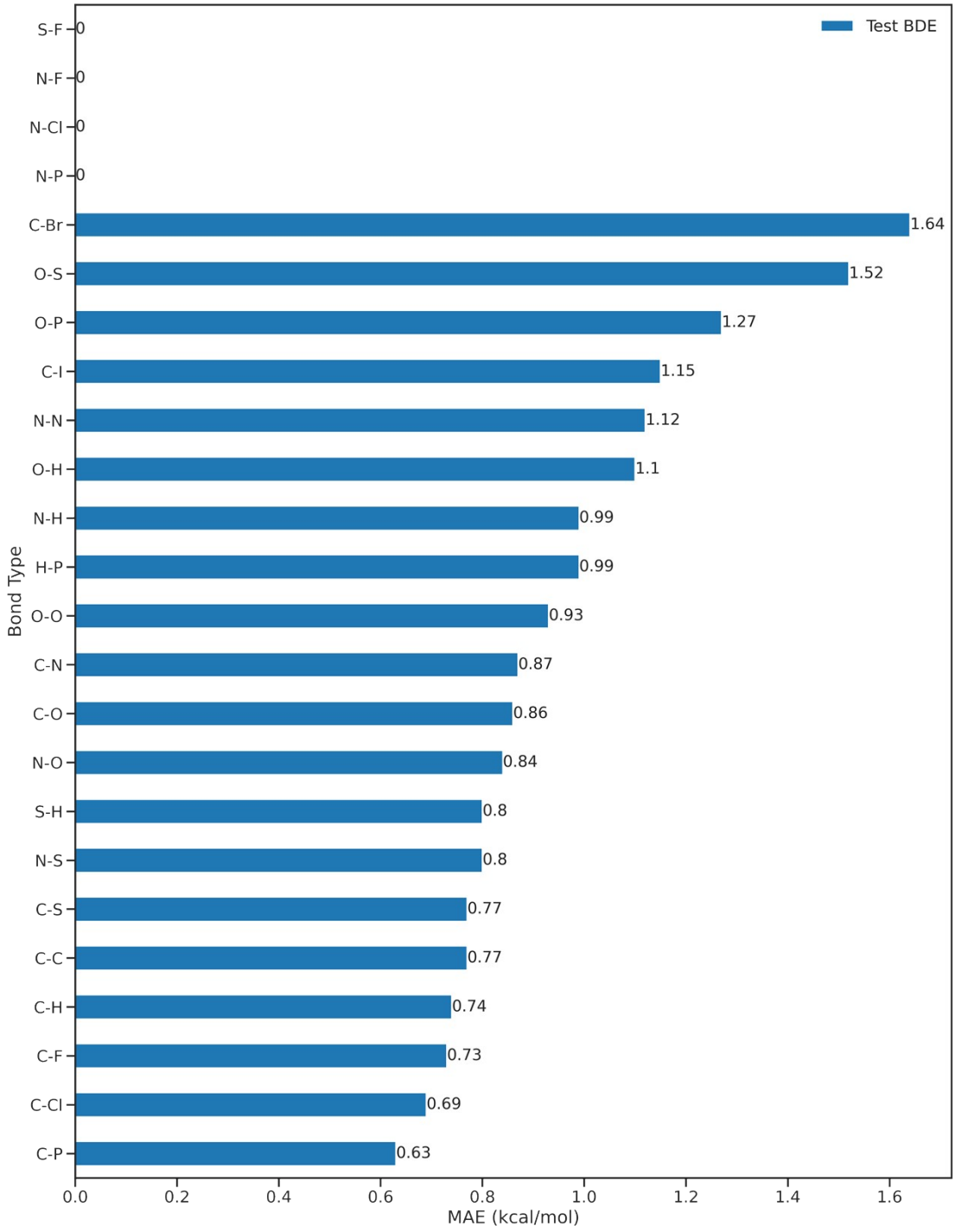

Figure S3. Prediction accuracy relative to DFT oracle for each bond type in held-out set set with Model 1.

4. Model 1: Performance on an external test set of halogenated heterocycles

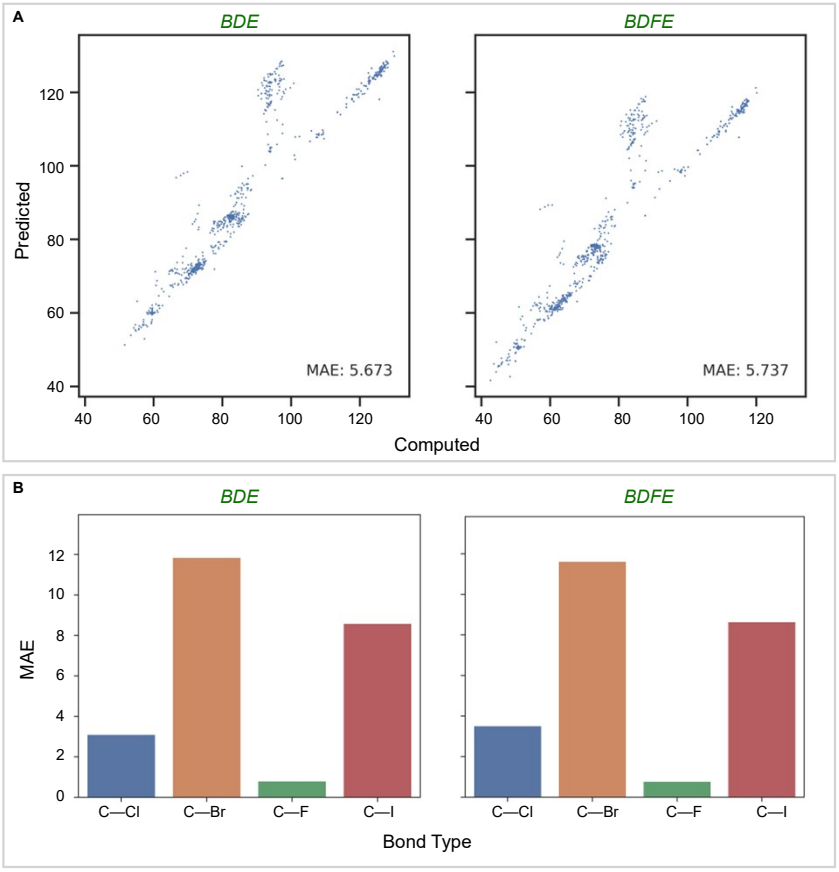

Figure S4. (A) Parity plots for BDE and BDFE prediction (kcal/mol) for aryl halides with Model 1. (B) Spread of errors for aryl halides according to bond type with Model 1.

5. Dataset composition: *BDE-db2* vs. halogenated heterocycle test set

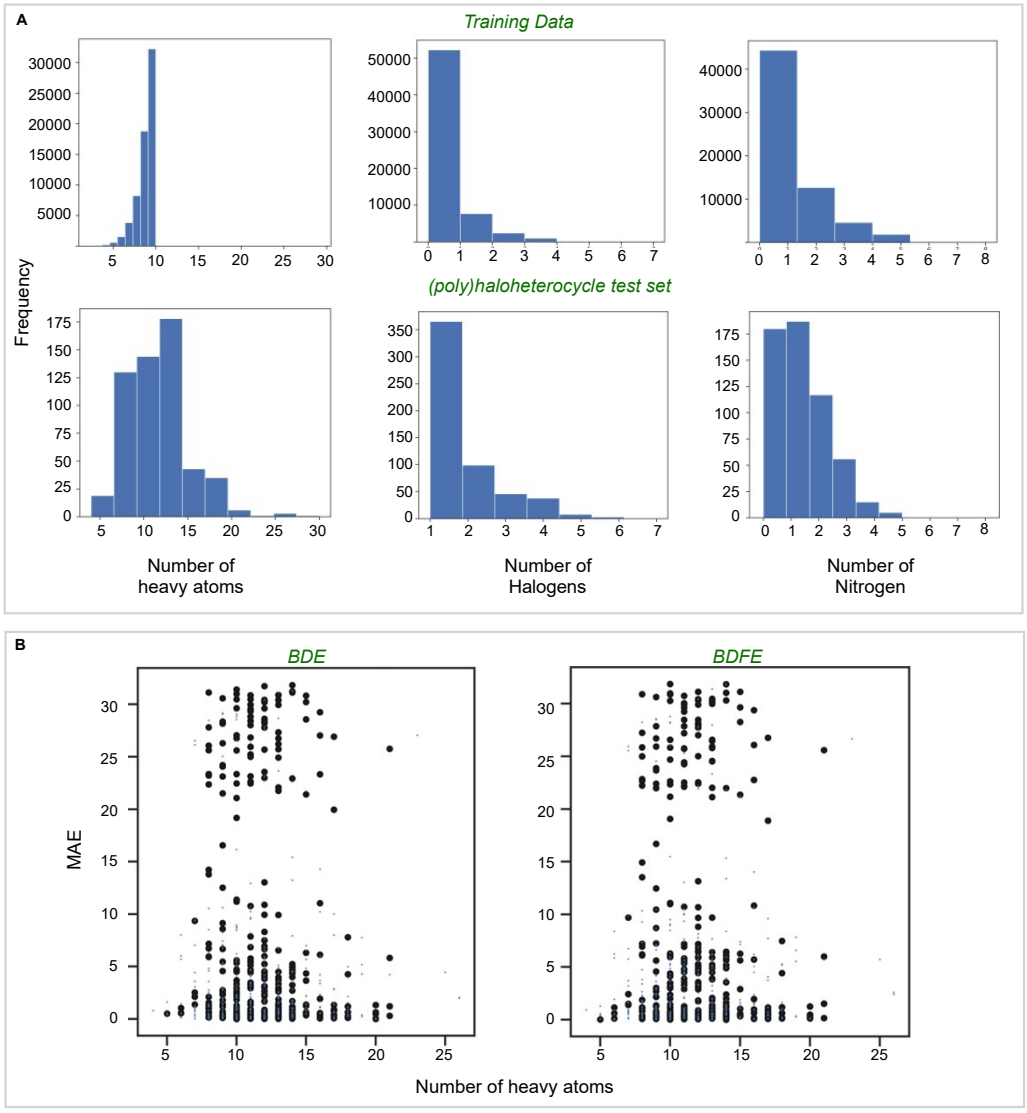

Figure S5. (A) Comparison of the atomic composition of the model training set and aryl halide external test set. (B) The presence of multiple halogens in a molecule is associated with larger prediction errors (black dots represent molecules with more than one halogen).

6. Model 2: Performance on *BDE-bd2* and an external test set of halogenated heterocycles

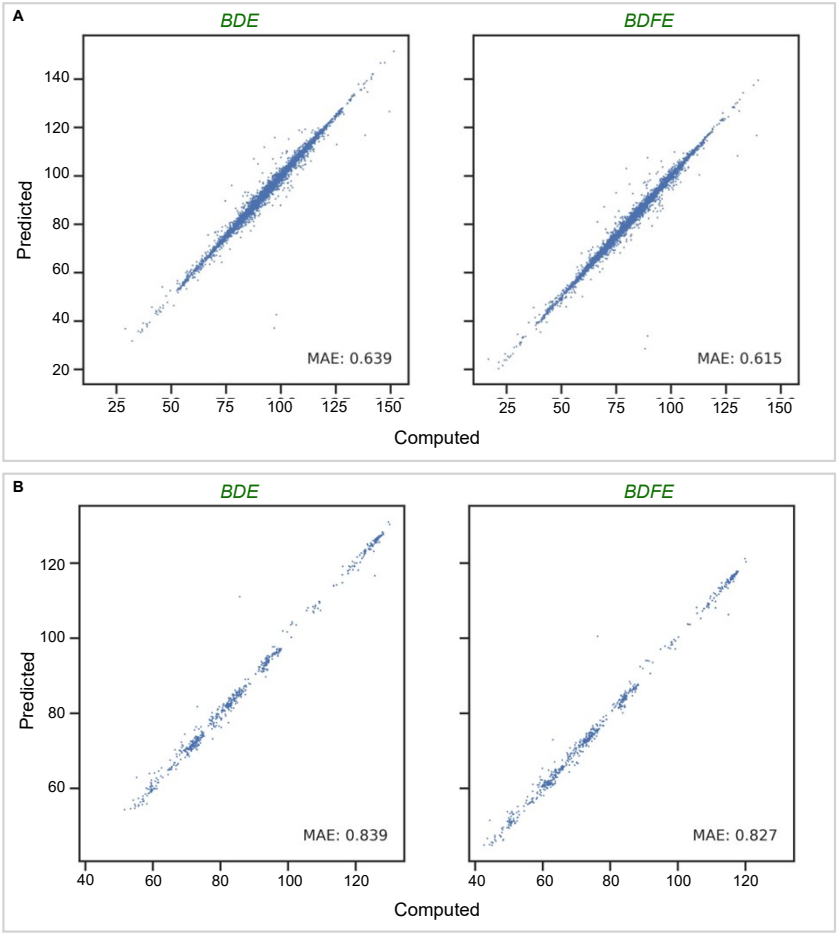

Figure S6. Parity plots for prediction of BDE and BDFE (kcal/mol) with Model 2 for (A) held-out test set and (B) external test set of halogenated heterocycles.

7. Composition of the polyhaloalkyl test set

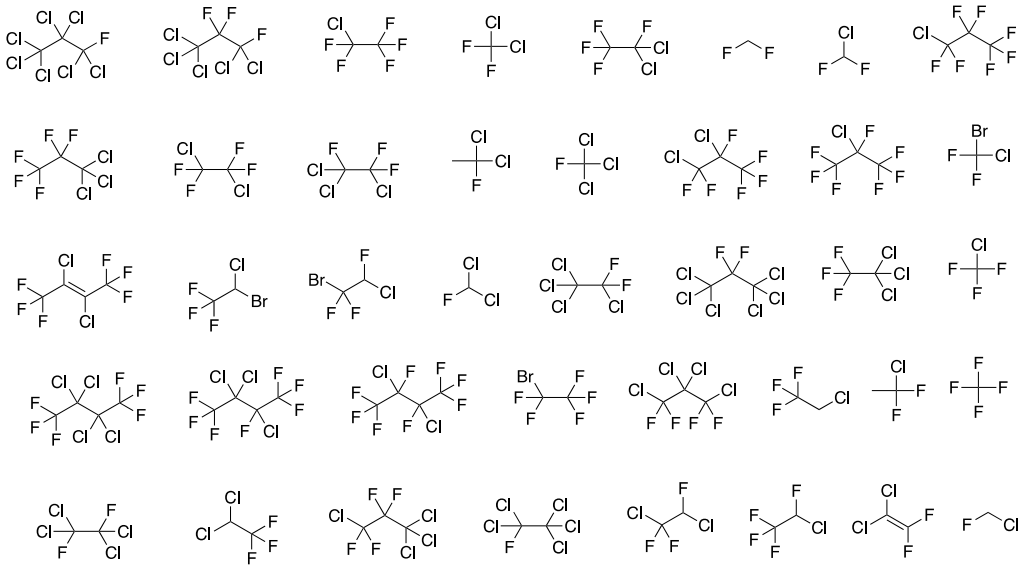

Figure S7. Molecules in the polyhaloalkyl test set.

8. Model 2: Performance on polyhaloalkyl test set

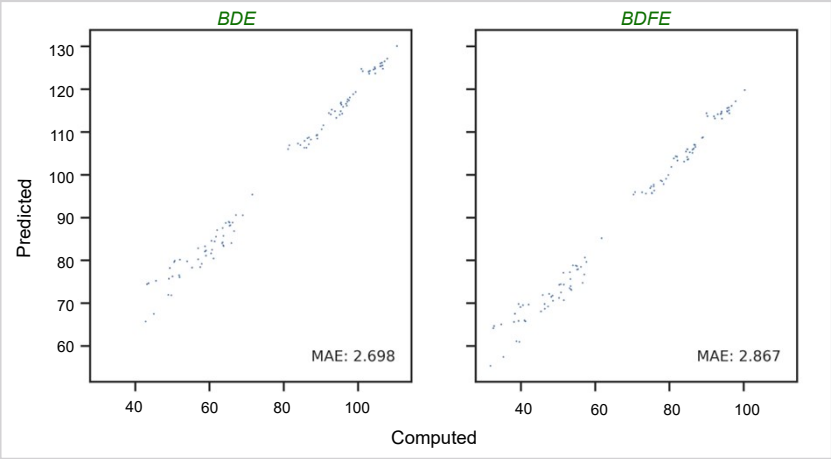

Figure S8. Parity plots for prediction of BDE and BDFE (kcal/mol) for polyhaloalkyl test set with Model 2.

9. Polyhaloalkyl molecules added to training data

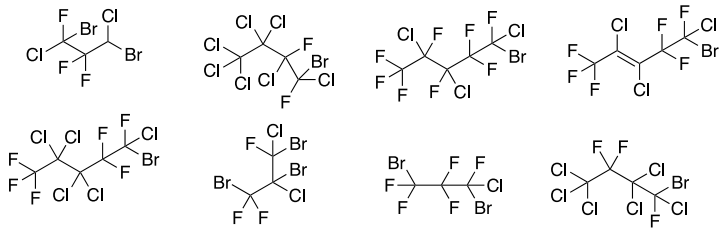

Figure S9. Molecular structures added to train Model 3. N.B. The graph representation used does not distinguish between (*R*)- and (*S*)-stereogenic centers, and so configuration at these centers is not shown.

10. Model 3: Performance on *BDE-db2*, halogenated heterocycle, and polyhaloalkyl test sets.

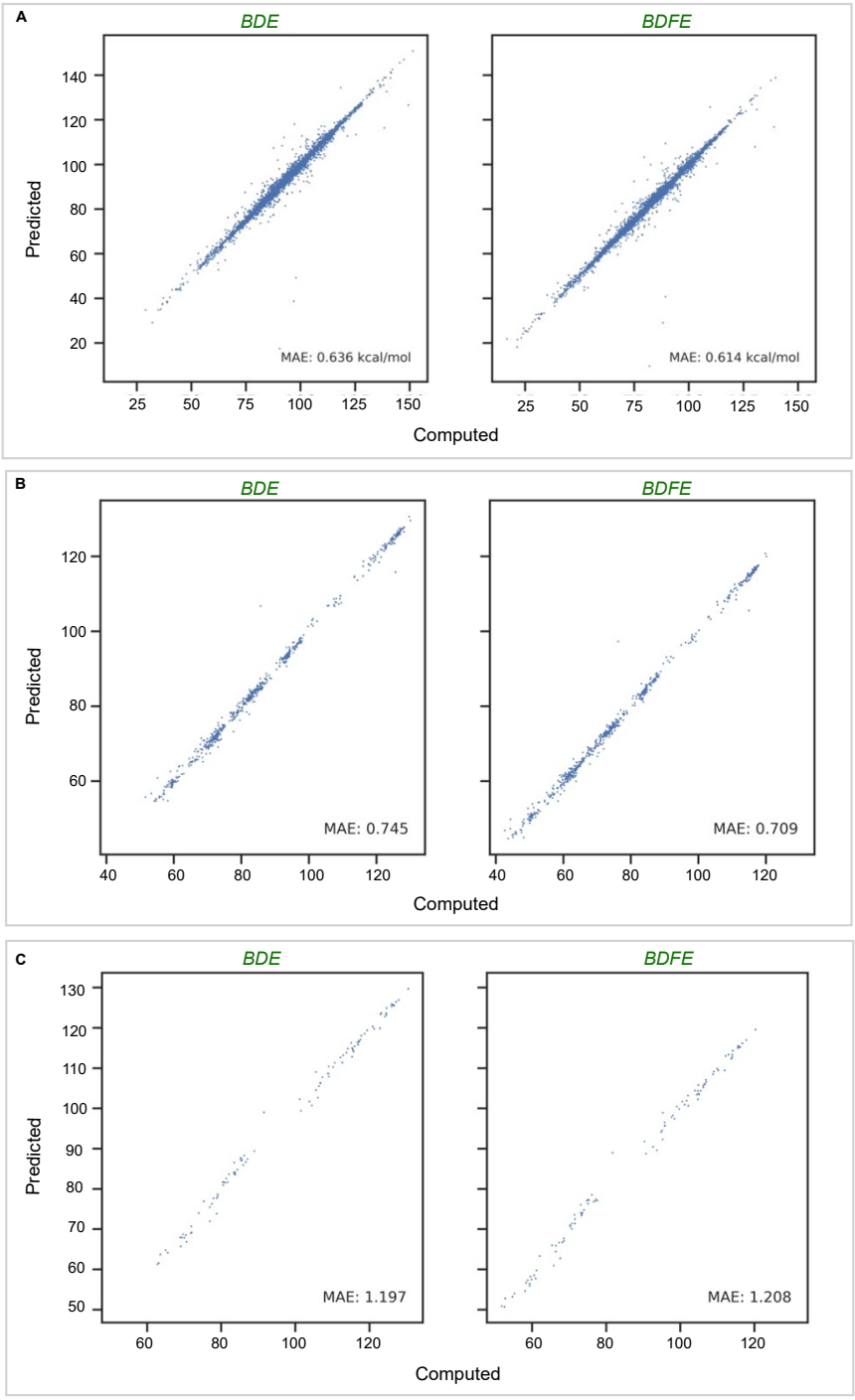

Figure S10. BDE and BDFE predictions (kcal/mol) obtained with an improved model, Model 3: (A) held-out test set, (B) halogenated heterocycles and (C) polyhaloalkyl test set.

11. Additional Details on comparison of traditional cheminformatics features and QM features.

For the models developed with Random Forest the following input and parameters were utilized. The inputs of fingerprints were created using Morgan Fingerprints with a radius of 3 and 512 bits defined around the bond of interest by specifying the atoms involved in the bond. The hyperparameters for modelling with random forest (RandomForestRegressor) was scanned using RandomizedSearchCV. The hyperparameter search include n_estimators = [100,200,300,400], max_features = [1,3,5,7, 'auto'], max_depth = [15, 10, 100, 1000]. Each model on the learning curve is optimized to get the respective hyperparameters for 10 different runs. The best parameters is chosen across the 10 runs to test on the held out test set 1. For graph neural network models with added QM description of bond lengths, the bond lengths were curated for each bond from the respective DFT calculation. The RBFExpansion bond length (dimension of 128) is concatenated with the tokenized embedding of the bond state built from RDKit features. This updated bond state is utilized in the message passing operation in the graph neural networks. The newly developed model with the QM features was test on the held-out test set 1.

References

1. 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*, 1095-1099.
2. 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. Acc.* **2008,** *120*, 215-241.
3. 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.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
4. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Info. Model.* **2015,** *55*, 2562-2574.
5. "RDKit: Open-source cheminformatics. https://www.rdkit.org".
6. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020,** *11*.
7. St. John, P. C.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S., Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci. Data* **2020,** *7*, 244.