Research Article

www.molinf.com

molecular informatics

doi.org/10.1002/minf.202200193

Machine Learning to Predict Homolytic Dissociation Energies of C–H Bonds: Calibration of DFT-based Models with Experimental Data

Wanli Li,[a] Yue Luan,[a] Qingyou Zhang,*[a] and Joao Aires-de-Sousa*[b]

Abstract: Random Forest (RF) QSPR models were developed with a data set of homolytic bond dissociation energies (BDE) previously calculated by B3LYP/6-311++G(d,p)//DFTB for 2263 sp3C–H covalent bonds. The best set of attributes consisted in 114 descriptors of the carbon atom (counts of atom types in 5 spheres around the kernel atom and ring descriptors). The optimized model predicted the DFT-calculated BDE of an independent test set of 224 bonds with MAE = 2.86 kcal/mol. A new data set of 409 bonds from the iBonD database (http://ibond.nankai.edu.cn) was predicted by the RF with a modest MAE (5.36 kcal/mol) but a relatively high R^2 (0.75) against experimental energies. A prediction scheme was explored that corrects the RF prediction with the average deviation observed for the k nearest neighbours (KNN) in an additional memory of experimental data. The corrected predictions achieved MAE = 2.22 kcal/mol for an independent test set of 145 bonds and the corresponding experimental bond energies.

Keywords: density functional calculations · bond energy · machine learning · learning transfer · quantitative structure-property relationship

1 Introduction

The homolytic dissociation energy of C–H bonds is a key parameter playing a decisive role in the assessment of chemical reactivity, e.g., predicting the major possible metabolic sites of xenobiotics,[1] metabolic stability,[2] autooxidation of drugs,[3] anti-oxidant activity,[4] or reaction pathways of pollutants in the atmosphere.[5]

The experimental determination of bond dissociation energies (BDE) for polyatomic molecules is difficult and has a typical uncertainty of ca. 1–2 kcal/mol.[6] Furthermore, predictions from the molecular structural formula are required for compounds not yet prepared or for virtual screening.

Theoretical calculations by quantum chemistry methods can provide accurate estimations of BDEs, but they are too computationally demanding for datasets with millions of compounds. Several methodologies were explored and compared.[7–9] Recently, St. John et al.[6] reported a benchmark study of three DFT functionals (B3LYP-D3, ωB97XD, and M06-2X) and two basis sets (6-31G(d) and def2-TZVP) using 368 experimental BDEs; the M06-2X/def2-TZVP combination performed best and achieved a mean absolute error (MAE) of 2.1 kcal/mol, which approaches the underlying uncertainty in the experimental measurements.

Chemoinformatics quantitative structure-property relationships (QSPR) can predict BDEs very rapidly and achieve high accuracies if appropriate data sets are available for training the models. QSPR studies were reported for specific types of compounds that relied on small data sets of experimental data, and some of them incorporated quantum chemistry descriptors. Cherkasov et al.[10] developed an additive empirical relationship with a data set of 79 molecules to predict the BDE of C–H bonds within 3.75 kcal/mol in molecules where resonance contributions and captodative stabilization are insignificant. Stanger[11] observed a correlation coefficient of 0.951 for a second order polynomial fit between the hybridization (calculated at B3LYP/6-311G*) and the intrinsic C–H BDEs for 17 alkyl C–H bonds. Przybylak and Cronin[12] investigated the C–H bonds at the α-position of 43 ethers and reported a high correlation ($R^2 = 0.852$) in a linear regression analysis of the BDEs versus spin distributions (calculated at B3LYP/6-311G**); correlations were higher within different subcategories of ethers based on structural features. Feng et al.[13] constructed a QSPR equation for the homolytic C–H BDE of strained hydrocarbons, from the hybridization associated

[a] *W. Li, Y. Luan, Q. Zhang*
Henan Engineering Research Center of Industrial Circulating Water Treatment, Henan Joint International Research Laboratory of Environmental Pollution Control Materials, Henan University, Kaifeng, 475004, P.R. China
E-mail: qingyou@vip.henu.edu.cn
[b] *J. Aires-de-Sousa*
LAQV and REQUIMTE, Chemistry Department, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
E-mail: jas@fct.unl.pt

with the C–H bond of the parent compound, the hybridization of the carbon radical, and the extent of spin delocalization of the radical for 89 bonds (calculated at UB3LYP/6-311++G(2df,p)//UB3LYP/6-31G(d)); the model exhibited a correlation coefficient of 0.927 and standard deviation of 2.9 kcal/mol. Zhao et al.[2] proposed an equation with 15 independent parameters associated with structural features that were optimized to predict benzylic C–H BDEs of 303 diverse heterocyclic compounds calculated with the DFT B3LYP functional; a standard deviation of 1.1 kcal/mol was observed for the whole set.

In a different approach, large databases with tens/hundreds of thousands of BDEs were calculated by quantum chemistry methods and were used to train machine learning methods. In 2013 one of the authors reported a data set of >12,000 BDEs calculated by B3LYP/6-311++G(d,p)//DFTB and machine learning models trained with 2D structural descriptors to predict the BDE; a MAE of 3.35 kcal/mol and $R^2 = 0.953$ was observed for an independent test set of 887 bonds covering a range of 17.67–202.30 kcal/mol.[14] The data set included neutral molecules and bonds with atoms of elements C, H, O, N, or S.

St John et al.[6] developed a database of 290,664 unique covalent non-cyclic bonds and their BDEs calculated at the M06-2X/def2-TZVP level of theory. The bonds were from 42,557 parent neutral molecules of general formula $C_xH_yO_zN_m$ with 10 or fewer heavy atoms, taken from the PubChem Compound database. A graph neural network trained on a subset of these data achieved a MAE of 0.58 kcal/mol (vs DFT) for 6948 unique BDEs of unseen molecules. Predictions for a set of molecules larger than 10 heavy atoms that were not a part of the training database yielded a MAE of 3.4 kcal/mol against experimental values.

Wen et al.[15] constructed a dataset of over 64,312 unique homolytic and heterolytic bond dissociations of neutral and charged (-1 and +1) molecules including organic and inorganic species (with elements C, H, O, F or Li), closed-shell and radical molecules, and molecules coordinated with metal ions. A chemically inspired graph neural network (BonDNet) was trained to predict BDEs; it maps the difference between the molecular representations of the reactants and products to the BDE of the corresponding reaction. A MAE of 0.51 kcal/mol was reported for a test set of unseen data. The BDE was calculated as the Gibbs free energy of dissociation at the ωB97X-V4 level of theory with the def2-TZVPPD basis set.

Here we report the development of a local model for BDEs of sp3C–H bonds with data previously published.[14] The bonds were represented by atomic descriptors of the carbon atom – counts of atom types in spheres around the atom, and sizes of rings incorporating the atom. The Random Forest algorithm was used for machine learning. The model was validated with a test set and was also challenged with a data set of experimental BDEs. Furthermore, a prediction scheme was explored that corrects the RF prediction with the average deviation observed for the k nearest neighbors (KNN) in an additional memory of experimental data. In this way we investigated the possibility of transferring the knowledge acquired with DFT data at a certain level of theory to an improved prediction of experimental BDEs. Learning transfer in ML approaches is of high interest to overcome the limitations of quantum chemistry calculation of properties and the more limited access to experimental data.[16,17]

2 Methodology

2.1 Data Sets

2.1.1 Data Set of DFT-calculated Bond Dissociation Energies

Data were retrieved for covalent bonds between hydrogen and sp3 carbon atoms from a more general data set consisting of molecular structures and homolytic bond dissociation energies (BDE) for a training set of 4242 compounds and a test set of 100 compounds.[14] The BDE had been calculated from the energies of the molecule and the energies of the two fragments formed by breaking the bond. The geometries of the molecule and the two fragments were optimized with DFTB. Single point energies were calculated for the molecule and for the fragments at B3LYP level in conjunction with the 6–311++G** basis set. The zero point energy correction and vibrational entropic terms were not included.

A subset of 4837 sp3C–H bonds in the training set and 510 sp3C–H bonds in the test set were retrieved. The equivalent C–H bonds were identified according to the uniqueness of the carbon atom by using a highly discriminating atomic index – the aATID index[18] as one of a series of highly selective topological indices[19,20] previously suggested by one of the authors. After excluding equivalent C–H bonds, we obtained 2039 sp3C–H bonds in the training set and 224 sp3C–H bonds in the test set.

2.1.2 Data Set of Experimental Bond Dissociation Energies

Data were retrieved from the Internet Bond-energy Databank (pKa and BDE) – iBonD (http://ibond.nankai.edu.cn) and consisted of 419 molecules with experimental BDE (homolytic bond dissociation enthalpy at 298 K) for one sp3C–H bond in each molecule. Among them, one bond was the same as one of the bonds in the whole DFT data set (based on the aATID index). In addition, one molecule contained an arsenic atom and eight molecules contained silicon atoms. These ten molecules were removed, and the remaining 409 molecules were used as the experimental data set. If more than one BDE value exist for a bond in the data set, the average bond energy was calculated as the final experimental BDE of the bond; the mean absolute deviation from the average was 1.264 kcal/mol for cases

with two values and 1.338 kcal/mol for cases with more than two values.

The data set was randomly partitioned into a training set composed of 327 bonds and a test set of 82 bonds.

2.2 Bond Descriptors

2.2.1 Atomic Type (AT) Descriptors

As all the bonds in this study have a hydrogen atom in common, the bonds were represented by descriptors of the carbon atom - counts of atom types in spheres around the atom. The atom types were described in a previous work,[21] and are defined in terms of the element, number of attached hydrogen atoms, number of attached non-hydrogen atoms and aromaticity – 34 atom types. The definition of atom types based on elements and number of neighbors does not require information concerning bond orders, thus avoiding differences in descriptors caused by alternative representations of the same structure (e.g., mesomers). Aromaticity was detected with the CXCALC tool from the JChem package v. 6.1.3, 2013, ChemAxon (http://www.chemaxon.com).

The AT descriptors of the bond were derived from counts of atom types in several spheres around the carbon atom (called kernel atom for its descriptors). In this case, 34*n* descriptors were generated for each atom. Here *n* is the number of layers around the atom (the layer of the kernel atom is the first layer). The default number of layers is 5. The descriptors are developed in terms of topological connectivity spheres and no 3D coordinates are involved.

2.2.2 Modified Atom Type (MAT) Descriptors

No ring information was introduced into the 34 atom types, except aromaticity indirectly implying ring information. However, the number and size of the rings encompassing the carbon atom of the bond were suggested as important factors affecting the BDE. Thus, the AT descriptors were extended by adding the sizes of rings encompassing the kernel atom – modified atom type descriptors (MAT). Additional 20 descriptors were appended to the AT descriptors to encode the sizes of rings from 3 to 22 members – each descriptor is the number of corresponding rings. Two examples are illustrated in Figure 1. In total, 34*n* + 20 MAT descriptors were generated with an in-house program written in Java. Here, *n* is the number of layers, and the default number of layers is 5.

2.2.3 Modified Distance Descriptors (MD)

Molecular MD descriptors were previously suggested and applied to the prediction of HOMO and LUMO energies by the authors.[22] Herein, atomic MD descriptors were obtained by extracting the corresponding atomic part of the molecular descriptors. These descriptors were designed exclusively based on the molecular connectivity and making no use of bond orders and atomic formal charges. Modified distances (MD) descriptors were implemented that count the pairs of atoms in a molecule at specific "modified distances" defined in terms of the radius of the atoms and electronegativity of neighbors. The descriptors consist in the counts of pairs of atoms within specific intervals of modified distances. Herein the default parameters are 1010 intervals, a resolution of 0.017, interatomic distances up to 4 bonds, and a distance factor of 4. As with MAT descriptors, the MD descriptors of the sp3 carbon atom was used to represent the sp3C–H bond.

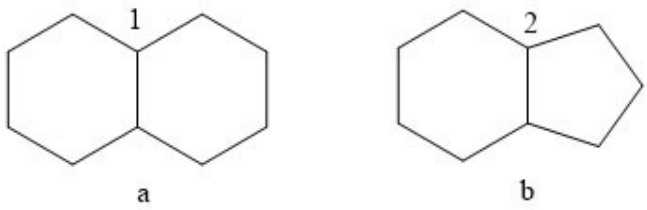

Figure 1. Ring descriptors in MAT descriptors: atom 1 belongs to two six-member rings and one ten-member ring, thus, the fourth descriptor (representing six-member ring) is 2 and the eighth descriptor (representing ten-member ring) is 1; atom 2 belongs to a five-, a six- and a nine-member ring, hence the third, fourth and seventh descriptors are 1.

2.3 Machine Learning (ML) Algorithm

Random Forests (RF)[23] were employed to predict the BDE from bond descriptors. They are ensembles of unpruned regression trees created with bootstrap samples of the training data. The models were first assessed with the embedded prediction error for the objects left out in the bootstrap procedure (out-of-bag estimation, OOB). The definition of distance between objects (required for the *k*-nearest neighbors calibration of predictions) was based on the number of trees in the ensemble that assign the objects to the same terminal node (divided by the total number of trees). Therefore, such a comparison relies on the descriptors that were chosen by the RF to build the model. Here, RFs were grown with the R program[24] version 3.6.1, using the randomForest library.[25] The number of trees in the forest was set to 1,000, and the remaining parameters were set to default.

A prediction protocol was devised that uses an additional memory of experimental BDEs to correct (calibrate) the predictions obtained with the RF trained with BDEs calculated by DFT methods. The procedure is schematically represented in Figure 2. The bonds in the additional database are predicted by the RF, these predictions are

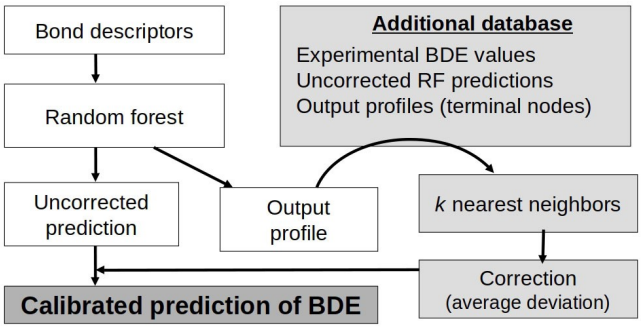

Figure 2. A small database of experimental BDEs is used to calibrate the predictions obtained by a RF trained with DFT data.

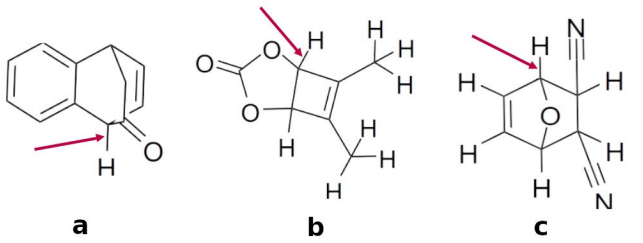

Figure 3. The three main outliers of the training set predicted by the DFT-based RF model (bonds indicated by the arrows). a: DFT = 113.69 kcal/mol, predicted = 78.68 kcal/mol; b: DFT = 49.8 kcal/mol, predicted = 84.42 kcal/mol; c: DFT = 113.62 kcal/mol, predicted = 80.47 kcal/mol.

Table 1. RF prediction of DFT-calculated BDE based on AT and MD descriptors encoding different layers (MAE and RMSE in kcal/mol).

Type of descriptors/number of descriptors/number of layers	OOB - training set (R^2/MAE/RMSE)	Test set (R^2/MAE/RMSE)
AT/98/5	0.765/3.407/5.464	0.813/3.145/5.027
AT/204/6	0.762/3.444/5.502	0.807/3.211/5.078
AT/238/7	0.762/3.446/5.499	0.809/3.236/5.080
MD/652/4	0.673/4.058/6.465	0.768/3.441/5.588
MD/663/5	0.670/4.095/6.502	0.766/3.465/5.633

Table 2. RF prediction of DFT-calculated BDE based on AT descriptors and AT descriptors augmented with ring information – MAT descriptors. MAE and RMSE are in kcal/mol.

Descriptors/number of descriptors	OOB - training set (R^2/MAE/RMSE)	Test set (R^2/MAE/RMSE)	OOB - whole data set (R^2/MAE/RMSE)
AT/98	0.765/3.407/5.464	0.813/3.145/5.027	0.770/3.354/5.403
MAT/114	0.812/3.068/4.906	0.846/2.862/4.567	0.818/3.006/4.822

compared with the experimental values and the deviations are computed, which are used to correct the RF predictions of similar new bonds.

3 Results and Discussion

3.1 Machine Learning Models Trained with DFT Bond Energies

Random forest models were trained to predict the C–H bond energies obtained from DFT calculations. The bonds were represented by the atom type descriptors (AT) and modified distance descriptors (MD) encoding up to 7 layers. Constant descriptors were removed. The number of descriptors and the results for various experiments are in Table 1 including the out-of-bag (OOB) error estimation of the training set and predictions for the test set.

AT descriptors performed better than MD descriptors but increasing the number of layers above 5 did not improve predictions. Training a RF with the whole data set represented by AT descriptors up to 5 layers yielded OOB predictions with $R^2 = 0.770$, MAE = 3.354 kcal/mol and RMSE = 5.403 kcal/mol.

To investigate the shortcoming of AT descriptors, the three predicted bonds with the highest errors were identified – Figure 3. All of them are connected to fused rings, which suggested the inclusion of information concerning rings in augmented AT descriptors.

A new model was trained with the inclusion of 16 additional descriptors encoding the sizes of rings to which the carbon atom of the C–H bond belongs. The results improved significantly, obtaining a MAE of 3.07 kcal/mol in the OOB estimation and 2.86 kcal/mol for the test set (Table 2 and Figure 4). The predictions for the test set were slightly more accurate than those obtained for the same test set with a previously developed global model.[14] The global model predicted the test set with $R^2 = 0.840$, MAE = 2.952 kcal/mol and RMSE = 4.667 kcal/mol.

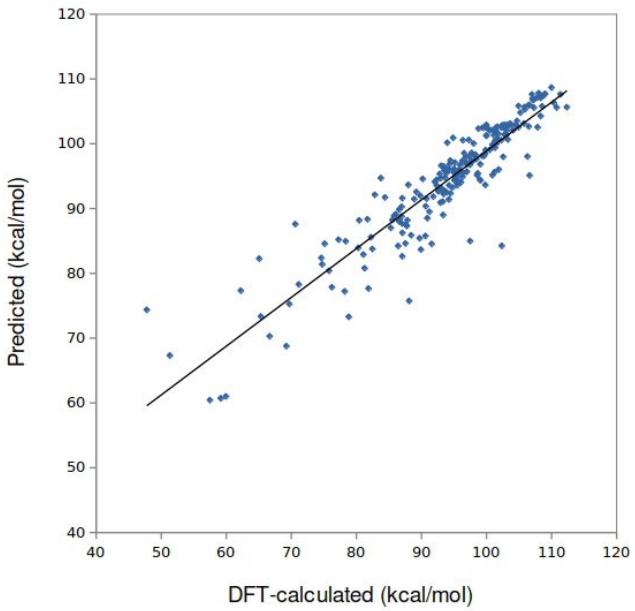

Figure 4. RF predictions of the test set compared with the DFT-calculated bond energies.

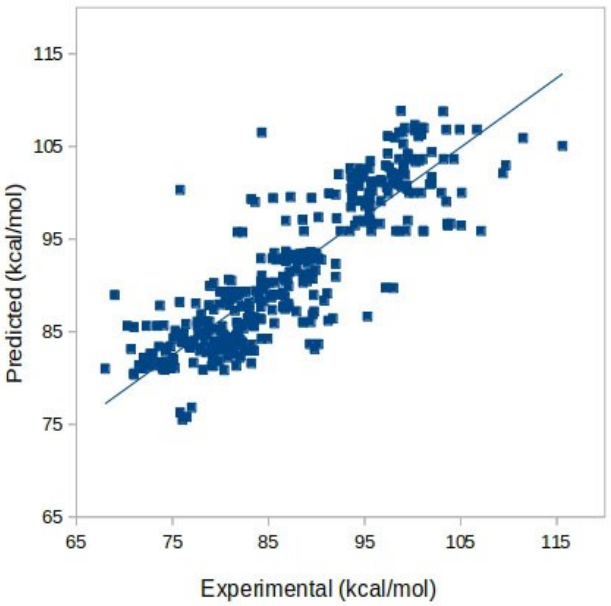

Figure 5. Prediction of 409 BDEs by a RF model trained with DFT-calculated bond energies and comparison with experimental values.

3.2 Prediction of Experimental Bond Energies with DFT-based ML Models

C–H bonds with sp3-hybridized carbon atoms were retrieved from the iBonD database (http://ibond.nankai.edu.cn) with their experimental homolytic bond energies. The 409 bonds in this data set were predicted by the RF model trained with DFT data. A graphical representation of the predictions vs the experimental values is displayed in Figure 5. The mean absolute deviation between the RF predictions and the experimental values was 5.36 kcal/mol, but a relatively high R^2 (0.75, Figure 5) suggested that systematic deviations of the DFT methods employed for the original training data might be a major source of the observed errors. In the next section we report the calibration of RF models (trained with DFT data) with a small data set of experimental data.

3.3 Calibration of DFT-based ML Models with Experimental Bond Energies

The data set of 409 bonds and their experimental energies were used to explore the possibility of approximating the ML predictions to the experimental values based on an additional small database of experimental data. A subset of 327 bonds was used as the additional database and the remaining 82 bonds as a test set.

The 82 bonds of the test set were submitted to the RF, which had been trained to predict the DFT-calculated energies. Predictions were obtained (uncorrected predictions), as well as the RF profile of each bond – the terminal nodes of the forest trees assigned to the bond. The RF predictions and profiles were pre-calculated for the additional database of 327 bonds. These were used to calculate distances between the 82 bonds of the test set and the bonds of the additional database, and to obtain the k-nearest neighbors (KNN) of each test bond. The procedure is illustrated in Figures 2 and 6. The mean deviation between the RF prediction and the experimental value for the KNN was used to correct the RF prediction for each test bond. The results are presented in Table 3 and Figure 7.

Table 3. MAE of calibrated predictions for a test set of 82 BDEs (corrections based on the experimental values of the *k*-nearest neighbors in the additional database, *k* from 1 to 5).

k	1	2	3	4	5
MAE of the test set (kcal/mol)	1.917	1.623	1.519	1.532	1.558

Correction with the most similar bond yielded a MAE of 1.917 kcal/mol, and the best results were observed with the 3 nearest neighbors (1.519 kcal/mol). To investigate the relationship between the similarity of the KNN and the improvement of the predictions, the errors were determined separately for groups of bonds in the test set with specific ranges of RF similarities to the nearest neighbor (> 0.9, 0.9–0.7, 0.7–0.5, 0.5–0.3 and < 0.3) – Table 4.

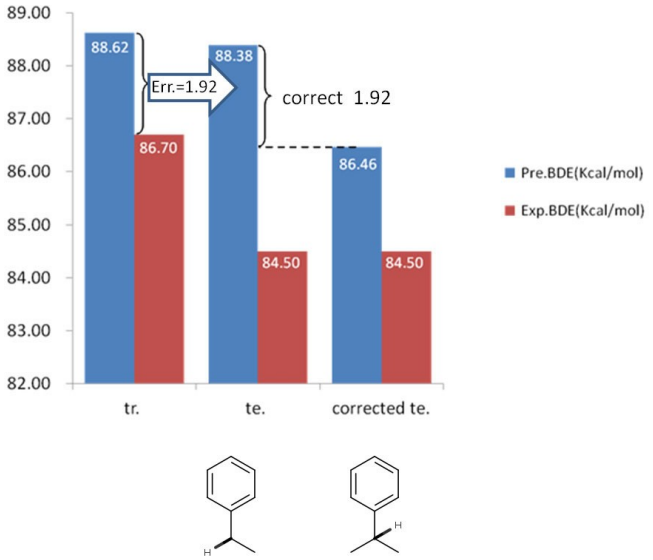

Figure 6. Calibration of a DFT-based RF prediction with experimental data. The prediction of the test set C–H bond of the molecule on the right (te) was corrected by the observed deviation for the most similar C–H bond in the additional database (tr, left molecule). Involved bonds are highlighted in bold.

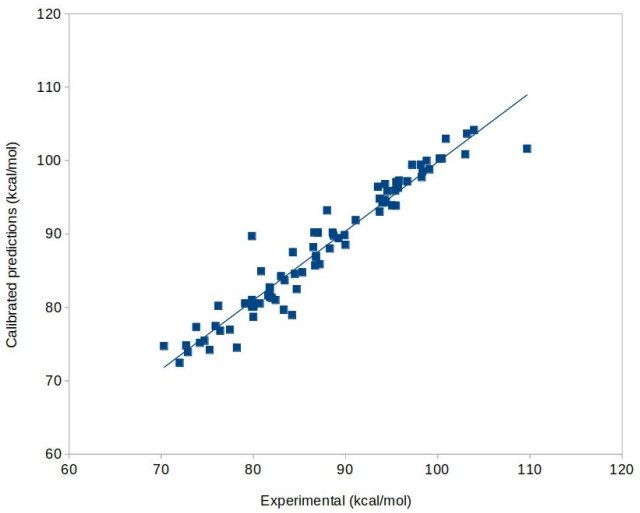

Figure 7. Prediction of 82 BDEs by a DFT-based RF calibrated with experimental data (calibration based on the KNN in the additional database) and comparison with experimental values.

A calibration was also investigated with a linear regression of experimental BDEs against the predictions by the RF trained with the DFT data. This was built with the bonds in the additional database, and the following equation was obtained: calibrated prediction = 1.0026 × RF uncorrected prediction - 4.631 kcal/mol. The MAEs obtained by this alternative calibration for the test set, as well as the uncorrected predictions, are included in Table 4 separately

Table 4. MAE (kcal/mol) of predictions for a test set of 82 BDEs separately determined for bonds with specific ranges of RF similarity to the nearest neighbor in the additional database (predictions obtained by the DFT-based RF with a) calibration based on the experimental values of the KNN, b) calibration based on a default correction obtained from a linear regression and c) uncorrected).

RF similarity	> 0.9	0.9–0.7	0.7–0.5	0.5–0.3	< 0.3
a) RF calibrated with KNN	2.236	1.267	1.989	1.575	3.096
b) RF calibrated with default correction	2.412	3.773	2.874	1.938	3.877
c) Uncorrected RF	5.361	6.186	5.347	4.347	3.999

for bonds with specific ranges of RF similarity to the nearest neighbor in the additional database.

The results suggest that a high similarity to a KNN is not required for significant improvement of the predictions. The larger error for the KNN-based corrections with similarities > 0.9 is partly due to one bond in this range yielding a large error (9.95 kcal/mol). If that bond is excluded, the MAE in the range of similarity > 0.9 becomes 1.6 kcal/mol. The bond and its nearest neighbor are displayed in Figure 8.

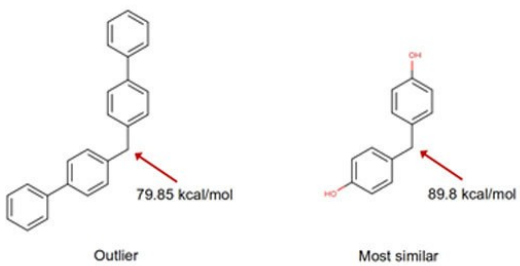

Figure 8. Illustration of an outlier concerning a C–H bond with a KNN with high RF similarity in the additional database.

The atoms within the five layers of the two sp3C–H bonds are the same. As a result, the descriptors of the two bonds are the same and the similarity is 1. The optimized model involving counts of atom types up to 5 layers of atoms around the kernel bond cannot predict effects arising from structural features further away.

The bonds less similar to the additional database (proximity < 0.3) were less well predicted after calibration (MAE = 3.096 kcal/mol). The results of Table 4 show this is

Table 5. Prediction of dissociation energy of sp3C–H bonds by RF models trained with experimental data.

Descriptors/number of descriptors	OOB - training set (R^2/MAE/RMSE)	Test set (R^2/MAE/RMSE)	OOB - whole data set (R^2/MAE/RMSE)
AT/170	0.8483/2.428/3.612	0.9152/1.714/2.607	0.865/2.224/3.376
MAT/190	0.8658/2.314/3.413	0.921/1.631/2.509	0.8795/2.123/3.200
MD/1010	0.709/3.287/5.060	0.761/2.826/4.447	0.7416/3.034/4.705
MAT + MD/1200	0.8492/2.384/3.610	0.9024/1.788/2.789	0.8629/2.213/3.409

due to the not so good correction applied by the calibration. Other bonds more similar to the additional database were predicted by the uncorrected RF with even higher MAE and the calibration lowered the error much more significantly. The calibration with a default correction based on the linear regression improved the RF predictions but not so much as the KNN calibration.

Prediction of the test set by a simple KNN based on the Euclidean distance of the normalized bond descriptors (instead of RF profiles) and the 327 experimental bond energies yielded MAEs of 2.3–2.5 kcal/mol, which are higher than the RF KNN-calibrated predictions. This shows that the knowledge acquired by the RF was relevant for the best calibration method.

The results of calibrated predictions were compared with the results of the available ALFABET graph neural networks trained with 276,717 BDE calculated by M06-2X/def2-TZVP.[6] We used 145 C–H bonds that are simultaneously in the data set of experimental values and in the test set of the reference work.[14] Now the additional database with experimental data consisted in the remaining 409–145 = 264 bonds. The 145 bonds were predicted by the RF, and the KNN-calibrated predictions achieved a MAE = 2.22 kcal/mol. The ALFABET model predicted the same bonds with a MAE of 3.11 kcal/mol.

3.4 Machine Learning Models Trained with Experimental Bond Energies

RF models were also built exclusively with the training set of 327 bonds and their experimental BDE using different descriptors (AT, MAT, MD, and MAT + MD). The same test set of 82 bonds was used to evaluate the models. The results are shown in Table 5 and Figure 9.

According to the OOB of the training set the best model was obtained with MAT descriptors. The MAE of the test set is 1.63 kcal/mol. The plot of the experimental BDE and predicted BDE for test set are shown in Figure 9. The results show that the accuracy of the DFT-RF calibrated with experimental data could surpass the accuracy of models trained with experimental data.

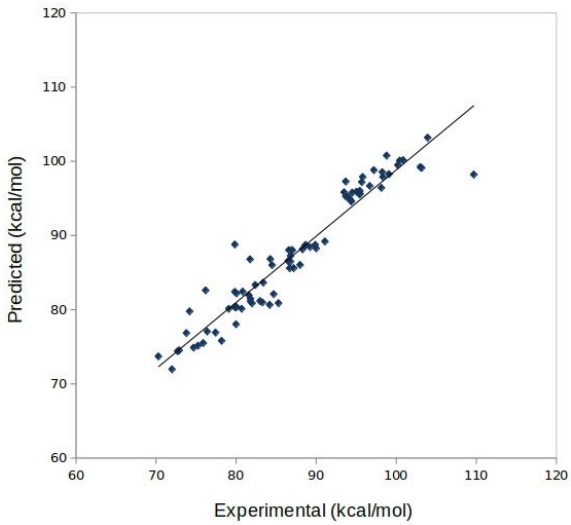

Figure 9. Prediction of 82 BDEs in the test set by a RF model trained with experimental data.

4 Conclusions

Fast QSPR predictions of DFT-calculated Csp3-H BDEs were achieved by RF models with a MAE of 2.9 kcal/mol (vs DFT calculations). The comparison of predicted and experimental values for a data set of 409 bonds yielded a mean absolute deviation of 5.4 kcal/mol and $R^2 = 0.75$. A calibration scheme was devised to approximate the RF predictions to the experimental values using an additional small data set of experimental data. Prediction of experimental values were thus achieved for an independent test set with MAE = 1.52 kcal/mol for 3 KNN.

Acknowledgements

This work was supported by the Associate Laboratory for Green Chemistry – LAQV, which is financed by national funds from Fundação para a Ciência e Tecnologia (FCT/MCTES), Portugal, under grant UIDB/50006/2020. JAS thanks David Ponting and co-workers at Lhasa Limited for useful suggestions and discussions. This work was also supported by the National Natural Science Foundation of China [Grant number 21875061, 21975066] and the program for Science

& Technology Innovation Team in Universities of Henan Province [Grant number 19IRTSTHN029].

Conflict of Interest

None declared.

Data Availability Statement

The data that support the findings of this study are available in the Supplementary Material of ref. 14 and from the Internet Bond-energy Databank (pKa and BDE) - iBonD (http://ibond.nankai.edu.cn).

References

[1] K. L. M. Drew, J. Reynisson, *Eur. J. Med. Chem.* **2012**, *56*, 48–55.
[2] S. W. Zhao, L. Liu, Y. Fu, Q. X. Guo, *J. Phys. Org. Chem.* **2005**, *18*, 353—367.
[3] P. Lienard, J. Gavartin, G. Boccardi, M. Meunier, *Pharm. Res.* **2015**, *32*, 300—310.
[4] Q. V. Vo, P. C. Nam, N. M. Thong, N. T. Trung, C. D. Phan, A. Mechler, *ACS Omega* **2019**, *4*, 8935–8942.
[5] W. S. McGivern, A. Derecskei-Kovacs, S. W. North, J. S. Francisco, *J. Phys. Chem. A* **2000**, *104*, 436–442.
[6] P. C. St John, Y. Guan, Y. Kim, S. Kim, R. S. Paton, *Nat. Commun.* **2020**, *11*, 2328.
[7] A. S. Menon, G. P. F. Wood, D. Moran, L. Radom, *J. Phys. Chem. A* **2007**, *111*, 13638–13644.
[8] J. M. Hudzik, J. W. Bozzelli, J. M. Simmie, *J. Phys. Chem. A* **2014**, *118*, 9364–9379.
[9] Y. Feng, L. Liu, J. T. Wang, H. Huang, Q. X. Guo, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2005–2013.
[10] A. Cherkasov, M. Jonsson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1222–1226.
[11] A. Stanger, *Eur. J. Org. Chem.* **2007**, *34*, 5717–5725.
[12] K. R. Przybylak, M. T. D. Cronin, *J. Mol. Struct.* **2010**, *955*, 165–170.
[13] Y. Feng, L. Liu, J. T. Wang, S. W. Zhao, Q. X. Guo, *J. Org. Chem.* **2004**, *69*, 3129–3138.
[14] X. Qu, D. A. R. S. Latino, J. Aires-de-Sousa, *J. Cheminf.* **2013**, *5*, 34–47.
[15] M. Wen, S. M. Blau, E. W. C. Spotte-Smith, S. Dwaraknath, K. A. Persson, *Chem. Sci.* **2021**, *12*, 1858–1868.
[16] C. A. Grambow, Y. P. Li, W. H. Green, *J. Phys. Chem. A* **2019**, *123*, 5826–5835.
[17] H. Han, S. Choi, *J. Phys. Chem. Lett.* **2021**, *12*, 3662–3668.
[18] Y. Luan, X. L. Li, W. L. Li, W. Li, Y. M. Zhou, Q. Y. Zhang, A. M. Pang, Development uniqueness test of highly selective atomic topological indices based on the number of attached hydrogen atoms. Manuscript submitted for publication.
[19] K. Xiao, M. Chen, T. Zhao, Q. Zhang, *Chemom. Intell. Lab. Syst.* **2018**, *178*, 56–64.
[20] T. Wu, M. Chen, K. Xiao, Y. Zhou, Q. Zhang, *Chem. J. Chin. Univ.* **2019**, *40*, 1158–1163.
[21] Q. Zhang, F. Zheng, R. Fartaria, D. A. R. S. Latino, X. Qu, T. Campos, T. Zhao, J. Aires-de-Sousa, *Chemom. Intell. Lab. Syst.* **2014**, *134*, 158–163.
[22] F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2017**, *57*, 11–21.
[23] L. Breiman, *Machine Learning* **2001**, *45*, 5–32.
[24] R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing; http://www.R-project.org: Vienna, Austria, 2015.
[25] A. Liaw, M. Wiener, Classification and Regression by Random Forest, RNews. 2 (**2002**) 18–22.

Received: August 1, 2022
Accepted: September 27, 2022
Published online on October 19, 2022

