scientific reports

OPEN

Machine learning enabling prediction of the bond dissociation enthalpy of hypervalent iodine from SMILES

Masaya Nakajima✉ & Tetsuhiro Nemoto✉

Machine learning to create models on the basis of big data enables predictions from new input data. Many tasks formerly performed by humans can now be achieved by machine learning algorithms in various fields, including scientific areas. Hypervalent iodine compounds (HVIs) have long been applied as useful reactive molecules. The bond dissociation enthalpy (BDE) value is an important indicator of reactivity and stability. Experimentally measuring the BDE value of HVIs is difficult, however, and the value has been estimated by quantum calculations, especially density functional theory (DFT) calculations. Although DFT calculations can access the BDE value with high accuracy, the process is highly time-consuming. Thus, we aimed to reduce the time for predicting the BDE by applying machine learning. We calculated the BDE of more than 1000 HVIs using DFT calculations, and performed machine learning. Converting SMILES strings to Avalon fingerprints and learning using a traditional Elastic Net made it possible to predict the BDE value with high accuracy. Furthermore, an applicability domain search revealed that the learning model could accurately predict the BDE even for uncovered inputs that were not completely included in the training data.

Organic chemistry enables the synthesis of various molecules by continuously breaking and forming molecular bonds. Bond dissociation enthalpy (BDE) is an indicator of the strength of a chemical bond and is an essential consideration in the design of chemical reactions and reactive molecules. Heat energy or light energy in adequate quantities can be used to break the chemical bond homolytically. Therefore, BDE is a commonly estimated on the basis of thermal measurement[1], kinetics[2], and electricity[3,4]. In recent years, advances in the development of computers, quantum chemistry, and density functional theory (DFT) calculations, have provided remarkably more accurate methodologies for predicting BDE[5–7]. In silico methods can be used to estimate the BDE, even for pinpoint chemical bonds of complicated molecules and imaginary molecules, enabling the design of reactive molecules and estimating the stability of functional molecules before they are synthesised. Even with advanced computer technology, however, the calculation costs of the DFT method remain enormous. The calculation time exponentially increases by the total number of electrons in a molecule. Therefore, obtaining the BDE values of hundreds or thousands of molecules at once by quantum computations remains challenging.

Hypervalent iodine (HVI), which bears over eight valence electrons on iodine, is a reactive molecule used as an oxidant or an alkylating agent in organic synthesis[8–15]. Heterolytic or homolytic cleavage of a weak, three-center four-electron (3c–4e) bond of HVI progresses the chemical reaction. Therefore, the BDE of the 3c–4e bond of HVI is an essential parameter that has been calculated by the DFT method on demand[16–19]. We previously reported the BDE value of 3c–4e bonds in various HVIs on the basis of DFT calculations[19]. We first determined the optimal functional and basis sets for reproducing the 3c–4e bond in silico and calculated a BDE value of 206 HVIs. While this database is helpful for chemists, it is still necessary to calculate the BDE for HVIs that are not available in the database.

Machine learning is currently attracting attention worldwide, and analysing and learning from a population using statistical methods enables immediate prediction of the results from new inputs. In the field of organic chemistry, machine learning is applied for predicting synthetic pathways and reactivity, and optimising reaction conditions[20–26]. Yu and co-workers reported the prediction model of BDE of carbonyl groups with machine learning in 2020[27]. Their important model accurately predicts the BDE value of the C=O bond on the basis of the bond length and bond angle of the relevant site as inputs. Three-dimensional molecular information is required
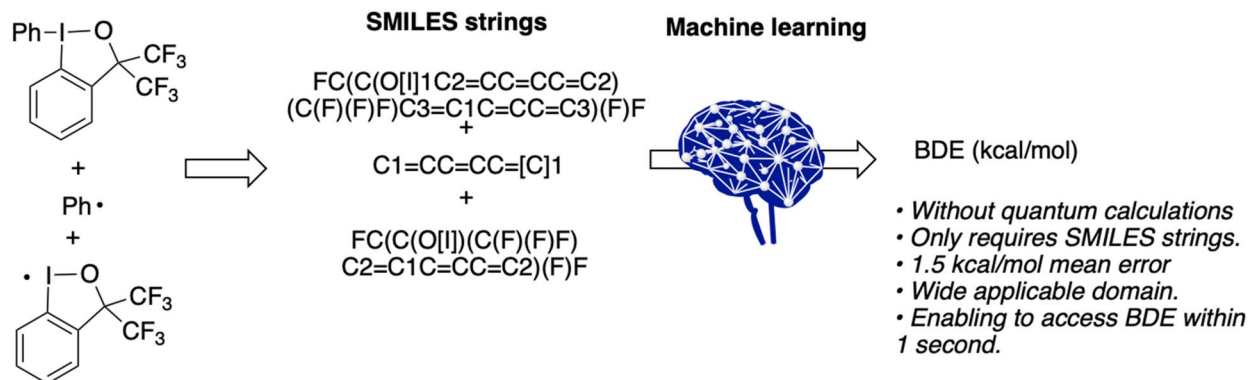
Graduate School of Pharmaceutical Sciences, Chiba University, Chiba, Japan. ✉email: m.nakajima@chiba-u.jp; tnemoto@faculty.chiba-u.jp
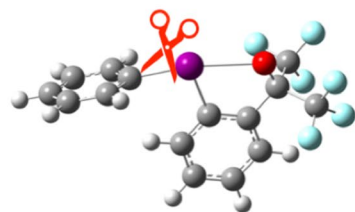
Figure 1. BDE calculations of HVIs using the DFT method (previous work) and machine learning (this work).

for the input data, however, and thus time-consuming DFT calculations are inescapable. We considered that an ideal and highly useful method of BDE prediction for chemists should not require quantum computations to prepare the input data. Therefore, we decided to use only structural formula information, such as SMILES strings, to predict the BDE value of HVIs by machine learning (Fig. 1).
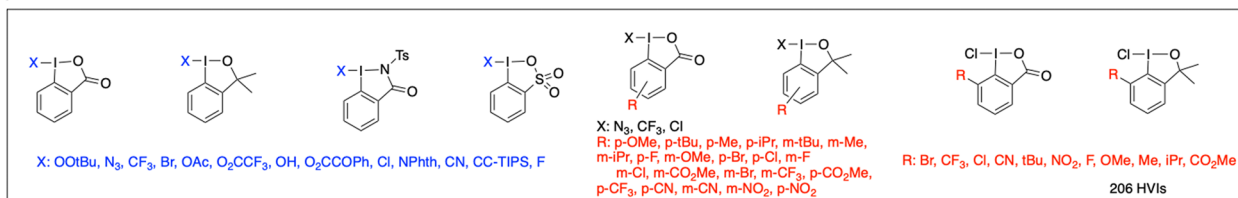
Methods

We first performed DFT calculations to increase the sizes of the data set populations. The DFT calculations were performed using Gaussian16 with MN15[28] functional and SDD[29,30] (for I and Se) and cc-pvTZ[31] (for the others) basis sets[19]. Structure optimizations were carried out with an ultrafine grid at 298.15 K in gas phase. Harmonic vibrational frequencies were computed at the same level of theory to confirm no imaginary vibration was observed for the optimised structure. BDE was calculated from the enthalpy (H) of each species at 298 K according to the following formula:

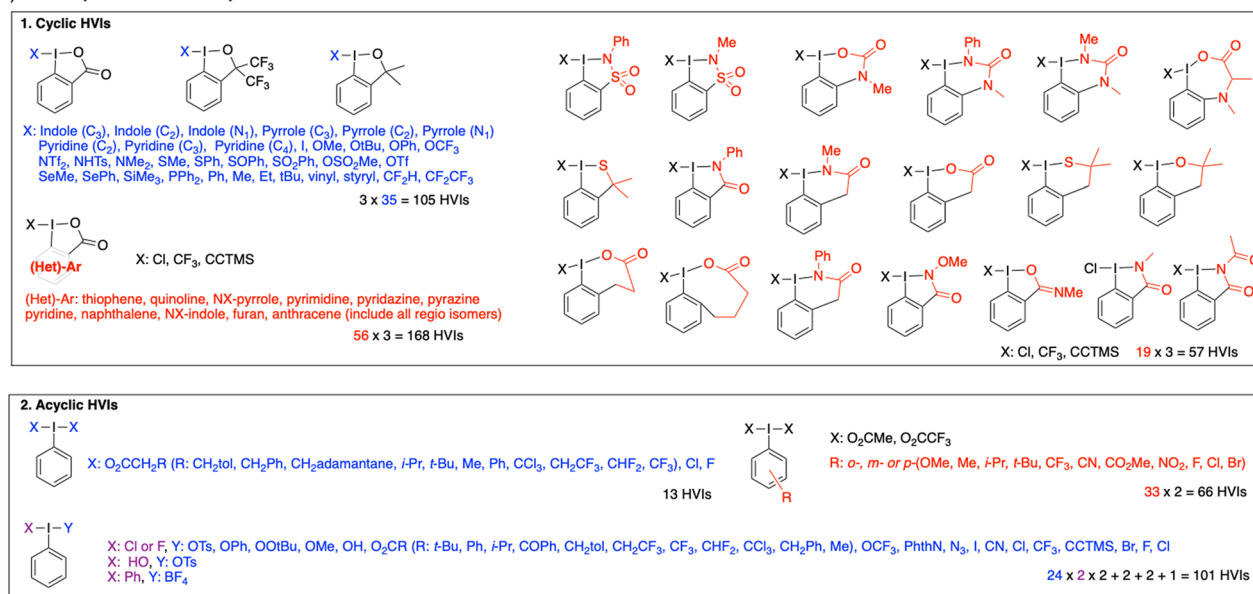$$BDE = y = H_{radicalA}^{298} + H_{radicalB}^{298} - H_{AB}^{298}$$

In addition to the BDE data of 206 HVIs, which we reported previously, we newly calculated 510 HVIs by DFT calculations (Fig. 2). Various combinations of iodine-containing backbones and leaving groups were calculated to increase the diversity of the data sets. A total of 330 cyclic HVIs were calculated: 105 molecules with 35 types of leaving groups and three common HVI skeletons for cyclic HVIs, and 225 molecules with 75 types of cyclic HVI skeletons and 3 common leaving groups. In addition, 167 acyclic HVIs were calculated: 13 types of symmetric HVIs, 101 types of asymmetric HVIs, and 66 types of HVIs with 33 HVI skeletons and 2 leaving groups. The 716 types of HVIs were randomly divided into 75% training and 25% test data sets. The training data set was first subjected to a grid search by k-partition cross-validation in each machine learning iterative process to optimise the hyperparameters (see supplementary information for details). For machine learning, three types of structural formulas were converted to SMILES: HVI (neutral), leaving group (radical), and HVI skeleton (radical). Then, fingerprints were generated using an RDkit (version 2019.09.3)[32]: Morgan[33] (Circular, r = 2, 3 or 4), Topological (RDKFingerprint)[32], MACCS[34], and Avalon[35]. In each fingerprint, learning from the training data set was performed with optimised hyperparameters using Elastic Net (EN)[36], support vector (SVR)[37], Neural Network (NN)[38], Random Forest (RF)[39], and LightGBM (LGBM)[40]. The accuracy of the BDE prediction was evaluated by comparison with the test data set. Mean absolute error (MAE) and coefficient of determination (R^2) were used to evaluate the prediction accuracy of the BDE.

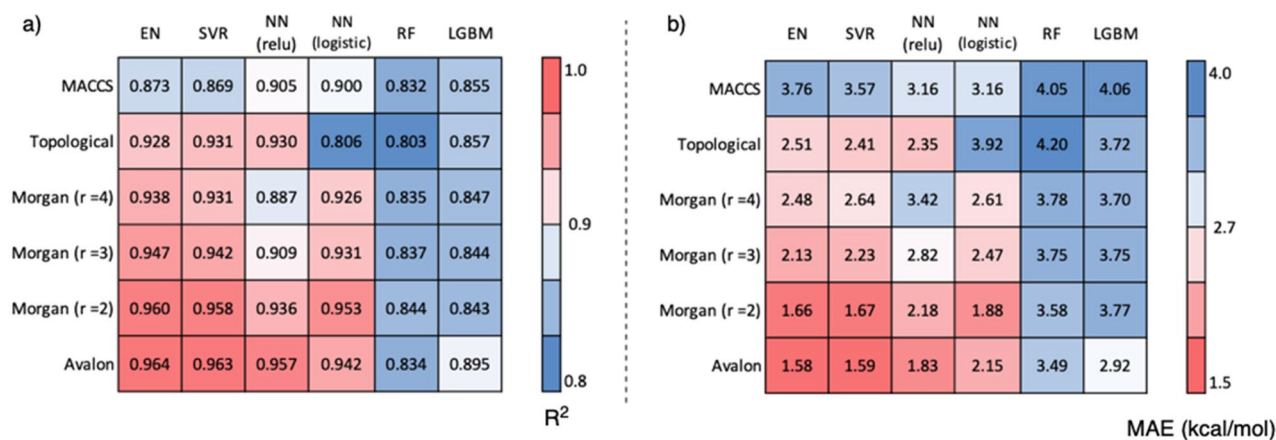$$MAE = \frac{1}{n} \sum_{i=1}^{n} |yi_{DFT} - yi_{ML}|$$

Figure 2. Structures of 716 HVIs for the training and test data sets: **(a)** previously calculated cyclic HVIs, **(b)** additionally calculated 510 HVIs.

Figure 3. Heat map of accuracy of the prediction by various machine learning algorithm and fingerprints.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (yi_{DFT} - yi_{ML})^2}{\sum_{i=1}^{n} (yi_{DFT} - yi_{average\_DFT})^2}$$

The training and testing were performed 10 times (random state = 0–9), and accuracy was evaluated by the average.

Results and discussion

As a result of the Grid search, we used both the "relu" and "logistic" evaluation functions for NN (see supplementary information for the detailed grid search results). The Avalon fingerprint, which features various factors such as atoms, bonds, and ring information, enables highly accurate prediction with an $R^2 = 0.964$ (Fig. 3a) and

MAE = 1.58 kcal/mol (Fig. 3b) by EN, which was the best score. SVR and NNs also gave high scores. In the Morgan fingerprint, which considers each atom's neighbourhood, the increasing number of recognised atoms gave a lower accuracy, and r = 2 (recognising first and second neighbour atoms) with the EN method giving the highest accuracy, similar to Avalon. The Topological fingerprint, which considers atoms and bond types, gave a high R^2 of 0.931 and a small MAE of 2.41 kcal/mol using the SVR method; however, it was inferior to the Avalon and Morgan fingerprints. The MACCS fingerprint, which counts 166 specific substructures, yielded the worst results. Although it gave an R^2 of 0.905 and an MAE of 3.16 kcal/mol by the NN (relu) method, the errors were small and acceptable. EN and SVR tended to give good results except for the MACCS fingerprint; on the other hand, RF and LGBM, which are decision-tree learning models, predicted BDE with low accuracy in all fingerprints.

Next, we investigated the applicability domain (AD) of these machine learning models[41]. Verifying the AD of the learning model is essential for examining the overfitting of training and the applicable range of uncovered inputs. For the AD search, the BDE of 561 HVIs was newly calculated by DFT calculations and classified into four groups: group A in which the leaving group and the HVI skeleton were individually included in the training data, group B in which the leaving group was included and the HVI skeleton was not included in the training data, group C in which the leaving group was not included and HVI skeleton was included in the training data, and group D in which neither the leaving group nor the HVI skeleton was included in the training data (Fig. 4). All HVIs shown in Fig. 2 were used as training data, and learning by the decision tree, which was an inappropriate learning model, was excluded.

The investigation of AD with group A (Fig. 5aA,bA) demonstrated that the Avalon fingerprint maintained high accuracy, that is, R^2 = 0.932 and MAE = 2.47 kcal/mol with the EN method (Fig. 6A). SVR and NN_r also gave R^2 = 0.920, 0.920 and MAE = 2.70, 2.89 kcal/mol, respectively. The Morgan (r = 2) fingerprint had a slightly lower accuracy with R^2 = 0.911 and MAE = 2.79 kcal/mol with the EN method. On the other hand, in the Topological and MACCS fingerprints, the R^2 value was lower than 0.7 and the minimum MAE was 5.44 kcal/mol (Topological, EN), indicating a significant decrease in accuracy from the test data in Fig. 3. Therefore, overfitting of the training data occurred in the Topological and MACCS fingerprint. With the molecules of group B (Fig. 5aB,bB), which contains new HVI skeletons, the accuracy was slightly decreased but the R^2 value of the Avalon (Fig. 6B) and Morgan (r = 2) fingerprints maintained a high accuracy of 0.880 and 0.863, respectively. In group C (Fig. 5aC,bC), which contains new leaving groups, the Avalon fingerprint could still predict with adequate accuracy with R^2 = 0.828 with the EN method (Fig. 6C). The Morgan (r = 2) fingerprint predicted the BDE value with R^2 = 0.532 and MAE = 8.00 kcal/mol, which are much lower than the values in groups A and B, indicating that prediction with the uncovered leaving groups was not applicable. We considered that because HVI skeletons contain R–I–R' bonds, the Morgan fingerprint could well recognise the pattern of the structure; however, the leaving groups were difficult to learn accurately because of the divergent structures. Finally, we verified the AD of group D (Fig. 5aD,bD), a completely new data set, and revealed that the Avalon fingerprint predicted the BDE value with R^2 = 0.759 and MAE = 5.97 kcal/mol (Fig. 6D). Because the Avalon fingerprint considers a larger variety of features and/or generates the fingerprint with a larger number of bits than MACCS, topological or Morgan, it was possible to appropriately evaluate the similarity of molecules and predict uncovered data with higher accuracy than other fingerprints.

We finally compared the computation time of the DFT method and machine learning method to calculate the BDE value of 561 HVIs of group A-D. In our computational environment, the DFT method required 4272 days (time converted to per core), i.e., 12 years; on the other hand, machine learning completed the 561 predictions from SMILES strings within 3 s, an overwhelming difference in speed.

Conclusions

We constructed a BDE prediction model for HVIs from SMILES strings using machine learning, which does not require quantum computations for input data. Avalon fingerprint generation and Elastic Net machine learning made it possible to predict BDE with high accuracy and an MAE of 1.58 kcal/mol. This model exhibited a high applicable range that can be predicted with an MAE of 5.97 kcal/mol, even for completely uncovered inputs. With this model, it is possible to access the predicted value of BDE for HVIs at a remarkable speed compared with modern quantum calculations. We anticipate that machine learning will be carried out by many organic chemists to facilitate the molecular design and reaction design of HVI.
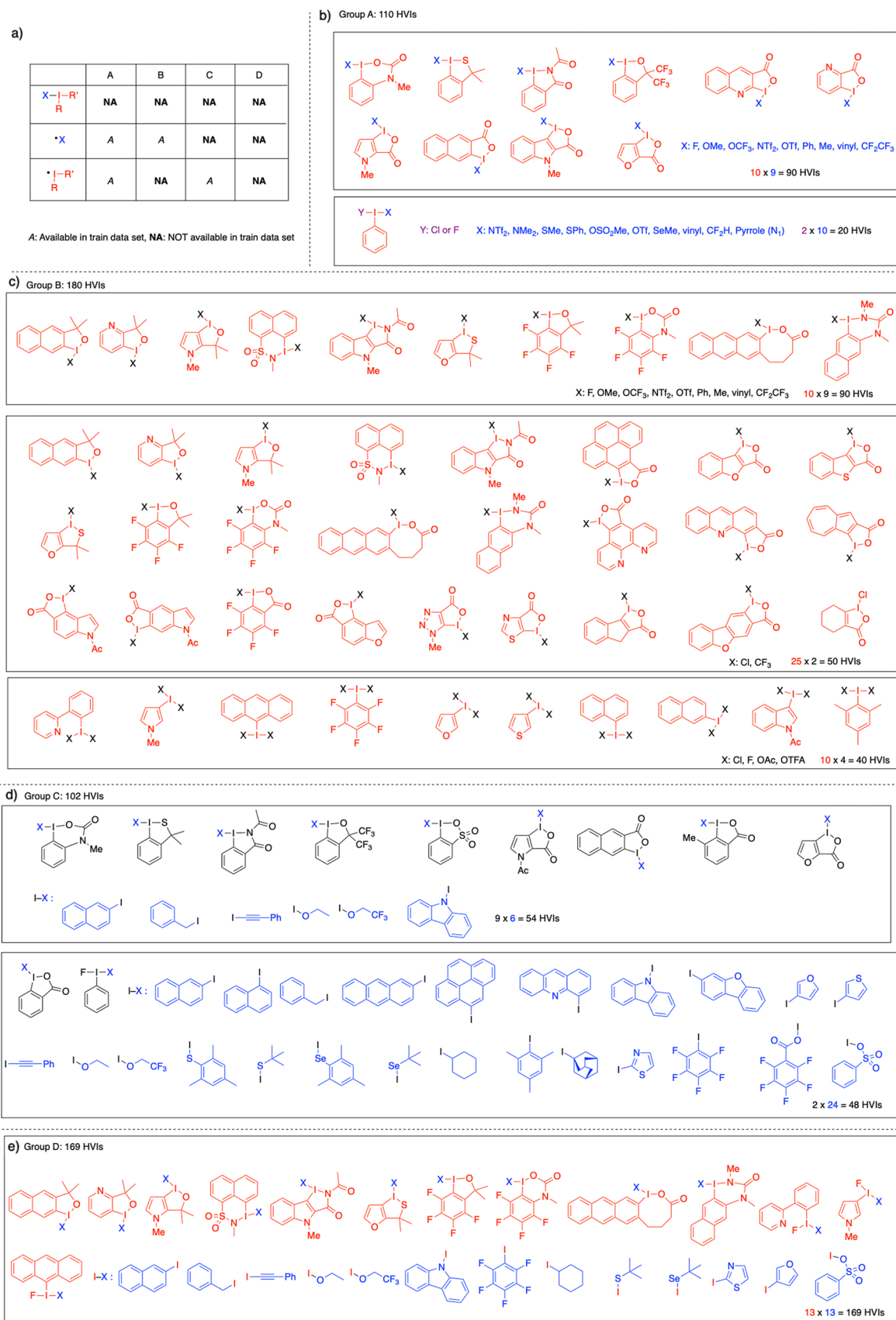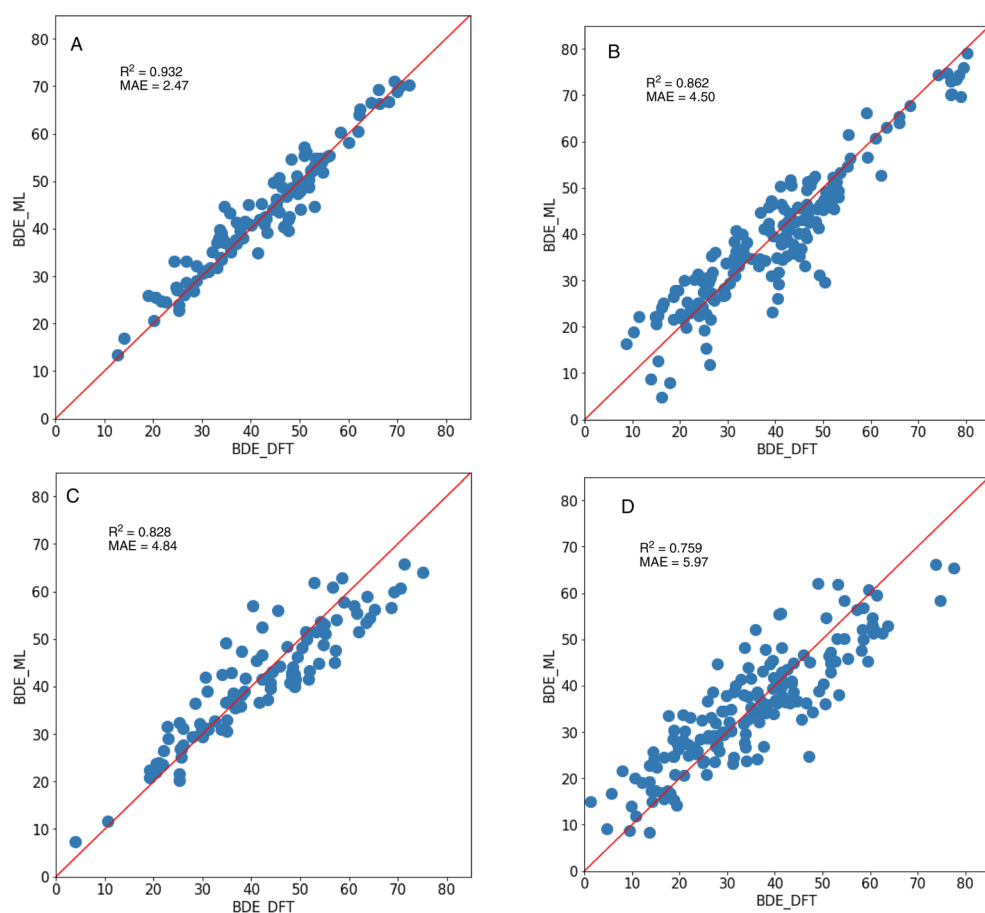
Figure 4. Structures of newly calculated 561 HVIs for AD: (**a**) table for the explanation of group A–D, (**b**) 110 HVIs in group A, (**c**) 180 HVIs in group B, (**d**) 102 HVIs in group C, (**e**) 169 HVIs in group D.
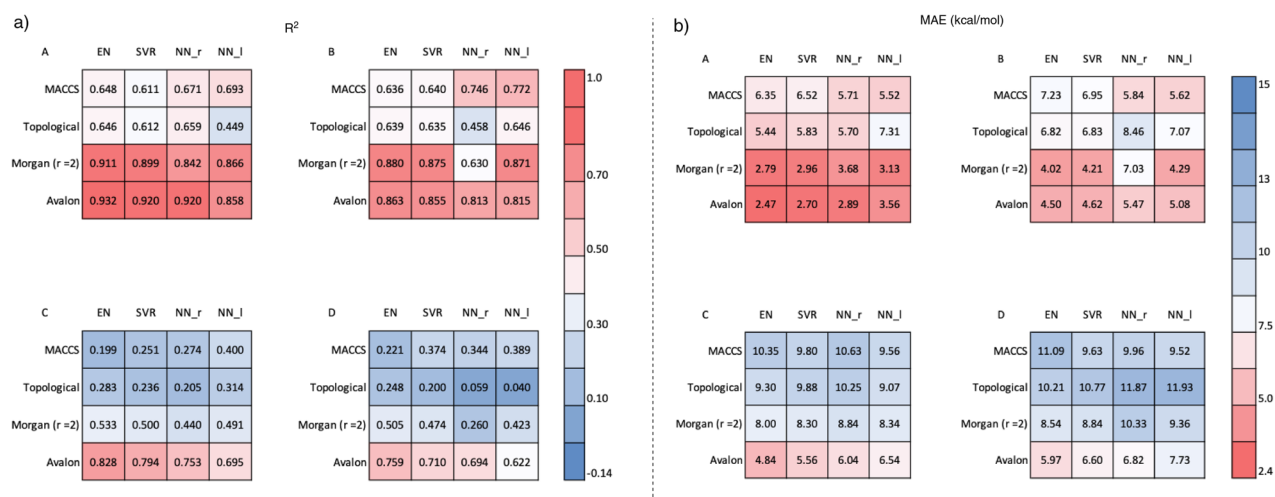
Figure 5. (a) Heat maps of the result (R^2) of AD. (b) Heat maps of the result (MAE) of AD.

Figure 6. Correlation plots of the prediction by EN with Avalon fingerprint.

Data availability

Computational details including the results of grid search, geometry and energy of HVIs by DFT, and the list of SMILES and the value of BDE_{DFT} are provided in Supplementary Information.

References

1. Szwarc, M. The estimation of bond-dissociation energies by pyrolyric methods. *Chem. Rev. (Washington, DC, US)* **47**, 75–173. https://doi.org/10.1021/cr60146a002 (1950).
2. Kerr, J. A. Bond dissociation energies by kinetic methods. *Chem. Rev.* **66**, 465–500 (1966).
3. Fu, Y. *et al.* Quantum-chemical predictions of redox potentials of organic anions in dimethyl sulfoxide and reevaluation of bond dissociation enthalpies measured by the electrochemical methods. *J. Phys. Chem. A* **110**, 5874–5886. https://doi.org/10.1021/jp055682x (2006).
4. Okajima, M. *et al.* Generation of diarylcarbenium ion pools via electrochemical C–H bond dissociation. *Bull. Chem. Soc. Jpn.* **82**, 594–599. https://doi.org/10.1246/bcsj.82.594 (2009).
5. 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.* **43**, 2005–2013. https://doi.org/10.1021/ci034033k (2003).
6. Yao, X.-Q., Hou, X.-J., Jiao, H., Xiang, H.-W. & Li, Y.-W. Accurate calculations of bond dissociation enthalpies with density functional methods. *J. Phys. Chem. A* **107**, 9991–9996. https://doi.org/10.1021/jp0361125 (2003).
7. Kim, S. *et al.* Computational study of bond dissociation enthalpies for a large range of native and modified lignins. *J. Phys. Chem. Lett.* **2**, 2846–2852. https://doi.org/10.1021/jz201182w (2011).
8. Kita, Y., Tohma, H., Kikuchi, K., Inagaki, M. & Yakura, T. Hypervalent iodine oxidation of *N*-acyltyramines: Synthesis of quinol ethers, spirohexadienones, and hexahydroindol-6-ones. *J. Org. Chem.* **56**, 435–438. https://doi.org/10.1021/jo00001a082 (1991).
9. Kita, Y. *et al.* Hypervalent iodine-induced nucleophilic substitution of para-substituted phenol ethers. Generation of cation radicals as reactive intermediates. *J. Am. Chem. Soc.* **116**, 3684–3691. https://doi.org/10.1021/ja00088a003 (1994).
10. Zhdankin, V. V. *et al.* Preparation, X-ray crystal structure, and chemistry of stable azidoiodinanes—Derivatives of benziodoxole. *J. Am. Chem. Soc.* **118**, 5192–5197. https://doi.org/10.1021/ja954119x (1996).
11. Kieltsch, I., Eisenberger, P. & Togni, A. Mild electrophilic trifluoromethylation of carbon- and sulfur-centered nucleophiles by a hypervalent iodine(III)-CF3 reagent. *Angew. Chem. Int. Ed.* **46**, 754–757. https://doi.org/10.1002/anie.200603497 (2007).
12. Phipps, R. J. & Gaunt, M. J. A meta-selective copper-catalyzed C-H bond arylation. *Science (Washington, DC, US)* **323**, 1593–1597. https://doi.org/10.1126/science.1169975 (2009).
13. Brand, J. P. & Waser, J. Direct alkynylation of thiophenes: Cooperative activation of TIPS-EBX with gold and Broensted acids. *Angew. Chem. Int. Ed.* **49**, 7304–7307. https://doi.org/10.1002/anie.201003179 (2010).
14. Matsumoto, K., Nakajima, M. & Nemoto, T. Visible light-induced direct S0 → Tn transition of benzophenone promotes C(sp3)-H alkynylation of ethers and amides. *J. Org. Chem.* **85**, 11802–11811. https://doi.org/10.1021/acs.joc.0c01573 (2020).
15. Nakajima, M. *et al.* A direct S0→Tn transition in the photoreaction of heavy-atom-containing molecules. *Angew. Chem. Int. Ed.* **59**, 6847–6852. https://doi.org/10.1002/anie.201915181 (2020).
16. Konnick, M. M. *et al.* Selective CH functionalization of methane, ethane, and propane by a perfluoroarene iodine(III) complex. *Angew. Chem. Int. Ed.* **53**, 10490–10494. https://doi.org/10.1002/anie.201406185 (2014).
17. Li, M., Wang, Y., Xue, X.-S. & Cheng, J.-P. A systematic assessment of trifluoromethyl radical donor abilities of electrophilic trifluoromethylating reagents. *Asian J. Org. Chem.* **6**, 235–240. https://doi.org/10.1002/ajoc.201600539 (2017).
18. Yang, J.-D., Li, M. & Xue, X.-S. Computational I(III)-X BDEs for benziodoxol(on)e-based hypervalent iodine reagents: Implications for their functional group transfer abilities. *Chin. J. Chem.* **37**, 359–363. https://doi.org/10.1002/cjoc.201800549 (2019).
19. Matsumoto, K., Nakajima, M. & Nemoto, T. Determination of the best functional and basis sets for optimization of the structure of hypervalent iodines and calculation of their first and second bond dissociation enthalpies. *J. Phys. Org. Chem.* https://doi.org/10.1002/poc.3961 (2019).
20. Gao, H. *et al.* Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476. https://doi.org/10.1021/acscentsci.8b00357 (2018).
21. Walker, E. *et al.* Learning to predict reaction conditions: Relationships between solvent, molecular structure, and catalyst. *J. Chem. Inf. Model.* **59**, 3645–3654. https://doi.org/10.1021/acs.jcim.9b00313 (2019).
22. Fu, Z. *et al.* Optimizing chemical reaction conditions using deep learning: A case study for the Suzuki-Miyaura cross-coupling reaction. *Org. Chem. Front.* **7**, 2269–2277. https://doi.org/10.1039/d0qo00544d (2020).
23. Kondo, M. *et al.* Exploration of flow reaction conditions using machine-learning for enantioselective organocatalyzed Rauhut-Currier and [3+2] annulation sequence. *Chem. Commun. (Cambridge, UK)* **56**, 1259–1262. https://doi.org/10.1039/c9cc08526b (2020).
24. Jorner, K., Tomberg, A., Bauer, C., Skold, C. & Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* **5**, 240–255. https://doi.org/10.1038/s41570-021-00260-x (2021).
25. Kim, H. W. *et al.* Reaction condition optimization for non-oxidative conversion of methane using artificial intelligence. *React. Chem. Eng.* **6**, 235–243. https://doi.org/10.1039/d0re00378f (2021).
26. Matsubara, S. Digitization of organic synthesis—How synthetic organic chemists use AI technology. *Chem. Lett.* **50**, 475–481. https://doi.org/10.1246/cl.200802 (2021).
27. Yu, H. *et al.* Using machine learning to predict the dissociation energy of organic carbonyls. *J. Phys. Chem. A* **124**, 3844–3850. https://doi.org/10.1021/acs.jpca.0c01280 (2020).
28. Yu, H. S., He, X., Li, S. L. & Truhlar, D. G. MN15: A Kohn-Sham global-hybrid exchange-correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. *Chem. Sci.* **7**, 5032–5051. https://doi.org/10.1039/c6sc00705h (2016).
29. Dolg, M., Wedig, U., Stoll, H. & Preuss, H. Energy-adjusted ab initio pseudopotentials for the first row transition elements. *J. Chem. Phys.* **86**, 866–872. https://doi.org/10.1063/1.452288 (1987).
30. Andrae, D., Haeussermann, U., Dolg, M., Stoll, H. & Preuss, H. Energy-adjusted ab initio pseudopotentials for the second and third row transition elements. *Theor. Chim. Acta* **77**, 123–141. https://doi.org/10.1007/bf01114537 (1990).
31. Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023. https://doi.org/10.1063/1.456153 (1989).
32. *RDKit: Open-Source Cheminformatics Software.* https://www.rdkit.org/.
33. Morgan, H. L. Generation of a unique machine description for chemical structures—A technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113. https://doi.org/10.1021/c160017a018 (1965).
34. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280. https://doi.org/10.1021/ci010132r (2002).
35. Gedeck, P., Rohde, B. & Bartels, C. QSAR—How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **46**, 1924–1936. https://doi.org/10.1021/ci050413p (2006).
36. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).
37. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
38. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554–2558 (1982).

39. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition*. 278–282 (IEEE, 2021).
40. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 3146–3154 (2017).
41. Tetko, I. V. *et al.* Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **48**, 1733–1746 (2008).

Acknowledgements

Author contributions

M.N. conceived this research and performed all calculations. All authors discussed and co-wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99369-8.

Correspondence and requests for materials should be addressed to M.N. or T.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.