How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 13253–13259
International Edition: doi.org/10.1002/anie.202000959
German Edition: doi.org/10.1002/ange.202000959

Machine Learning

Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning

*Xin Li, Shuo-Qing Zhang, Li-Cheng Xu, and Xin Hong**

Abstract: *Radical C–H bond functionalization provides a versatile approach for elaborating heterocyclic compounds. The synthetic design of this transformation relies heavily on the knowledge of regioselectivity, while a quantified and efficient regioselectivity prediction approach is still elusive. Herein, we report the feasibility of using a machine learning model to predict the transition state barrier from the computed properties of isolated reactants. This enables rapid and reliable regioselectivity prediction for radical C–H bond functionalization of heterocycles. The Random Forest model with physical organic features achieved 94.2 % site accuracy and 89.9 % selectivity accuracy in the out-of-sample test set. The prediction performance was further validated by comparing the machine learning results with additional substituents, heteroarene scaffolds and experimental observations. This work revealed that the combination of mechanism-based computational statistics and machine learning model can serve as a useful strategy for selectivity prediction of organic transformations.*

Introduction

Arene C–H functionalization has received wide interest from academia and industry due to the ubiquitous presence of aromatic rings in biologically active compounds and functional materials.[1] The classic Minisci reaction provides a versatile approach for elaborating arenes through radical-based C–H functionalization, but the application is hindered by the general perception of unpredictable regioselectivity in radical additions.[2] In 2011, Baran and co-workers discovered that alkylsulfinate salts can serve as alkyl radical precursors, which realized a series of powerful radical functionalization methods of aromatic rings in mild conditions[3,4] (Figure 1 A).

The success of arene C–H functionalization using alkylsulfinate salts revealed a key fact, that the innate reactivity of aromatic heterocycles can lead to exceptional control of regioselectivity in certain cases.[4a–c,e,g] Based on the effects from innate reactivity of heterocycle, π-conjugating substituents and electronic properties of radicals, Baran, Blackmond and co-workers developed an empirical guideline for regioselectivity prediction of radical C–H functionalization of heterocycles[5] (Figure 1 B). This empirical approach provides a useful qualitative prediction for the selected heterocycles, while a general and quantified regioselectivity prediction is still challenging, especially considering the wide array of heterocycles, substitution patterns, and sterically and electronically distinctive radicals.

Recently, performance predictions of organic transformation using machine learning (ML) and experimental statistics have seen significant progress.[6] Doyle and co-workers revealed the potential of ML in the navigation of reaction optimization[6e] and prediction of synthetic reaction.[6c] Grzybowski et al. discovered the importance of physically relevant descriptors in the ML prediction of reaction selectivity[6a,g] and molecular thermodynamic properties,[6k] as well as the translation of chemical knowledge in retrosynthetic planning.[7] Denmark and co-workers developed the strategy of universal training set in the predictive modeling of asymmetric catalysis.[6o] Jensen et al. achieved the predictions of reaction outcome and condition of organic transformations using neural network models.[6b,d,h] In addition, Sigman and co-workers established the powerful multidimensional statistical analysis in synthetically relevant catalyst designs.[8]

The above breakthroughs inspired us to establish the desired regioselectivity prediction of radical C–H functionalization of heterocycles using a ML strategy. As alternative to the experimental statistics-based ML predictions, we

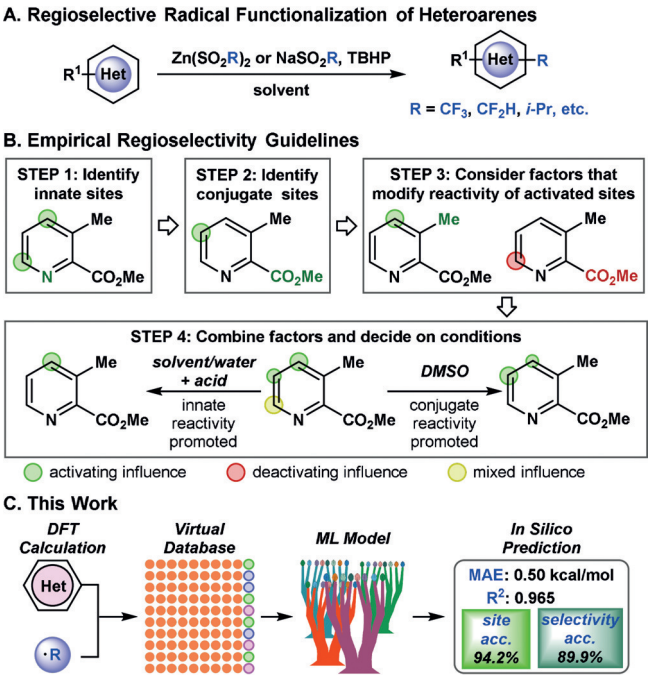

Figure 1. Radical C–H functionalization of heterocycles and strategies of regioselectivity prediction.

[*] X. Li, Dr. S. Q. Zhang, L. C. Xu, Prof. Dr. X. Hong
Department of Chemistry, Zhejiang University
38 Zheda Road, Hangzhou, 310027 (China)
E-mail: hxchem@zju.edu.cn

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202000959.

envisioned that the ML model could connect the transition state barrier with computed properties of intermediates or reactants. This would enable the ML prediction of reactivity and selectivity using mechanism-based computational statistics. Herein we report the development of ML regression models that can achieve regioselectivity prediction of radical C–H functionalization of heterocycles with no experimental input (Figure 1C). The trained Random Forest model with a small number of DFT-computed physical organic features achieved 94.2% site accuracy and 89.9% selectivity accuracy in the out-of-sample test set. Further comparisons with additional substituents, heteroarene scaffolds and experimental results validated this in silico prediction approach.

Results and Discussion

The designed workflow for our ML approach is described in Figure 2A. We first developed a series of Python scripts to automatically achieve the generation, collection and analysis of the DFT-computed statistics. This large quantity of data was subject to data preprocessing, which normalized the features to a standard distribution. Subsequent performance evaluation selected the desired ML model from the candidate algorithms. The complexity of the selected model was further reduced through a feature selection process, which led to the final ML model for application purpose. Feature ranking of the final ML model allowed the interpretation of the controlling factors of selectivity prediction.

The sample space is critical for the generalization ability of the ML model.[9] A wide array of arene scaffolds, substitution patterns and radicals were carefully selected to establish the computational statistics (Figure 2B). The selected arene scaffolds cover general five- and six-membered ring heterocycles in organic chemistry, including both neutral and protonated states. In addition to the parent aromatic compounds, mono-substitution with cyano or methoxyl substituent at all possible positions were initially considered. These two substituents allow the ML model to map the variation of electronic properties of arenes. Further tests of the generalization ability on substituent patterns identified the necessity to include CF_3 and *t*Bu substituents in model training (see below). The total number of computed aromatic compounds in the initial model training was 109, leading to 262 distinctive positions for radical C–H functionalization. For the chemical space of radical, we included the radicals that can be generated from commercially available alkylsulfinate salts[4a–e,h,10] as well as a number of additional cases that were present in related experimental studies.[4f,5,11] The designed sample space presented a collection of 3406 radical C–H functionalization reactions, and 5174 regioisomeric competitions.

To train the ML model, DFT calculations[12] were performed to provide the computational statistics. A previous

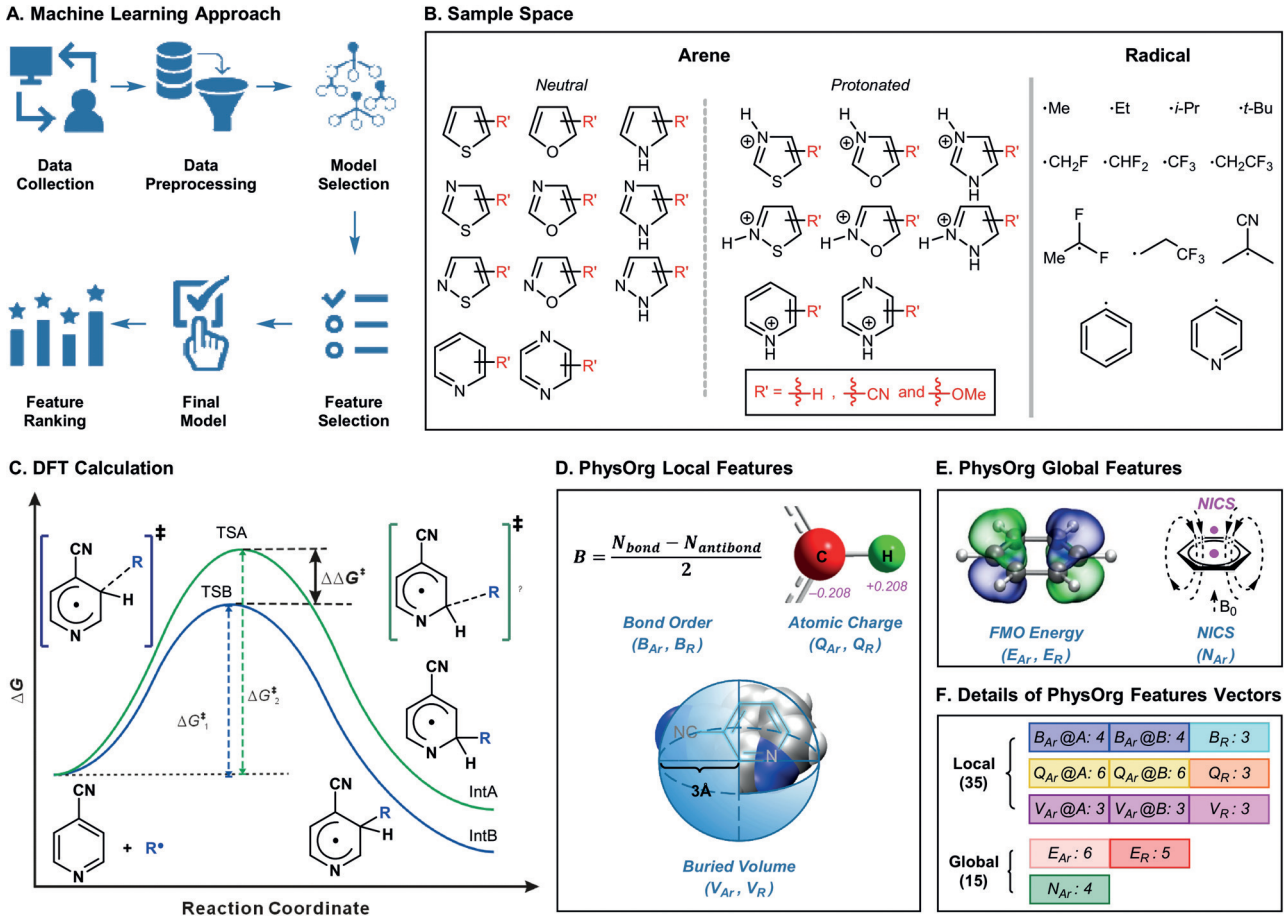

Figure 2. Key information of designed machine learning approach for regioselectivity prediction.

Table 1: The regression performances (R^2) of the combinations of molecular features and ML models for regioselectivity prediction.

Local[a]	Global[a]	Ada	BRR	DTree	GB	GPR	KNR	KRR	LR	LSVR	Lasso	NN	RF	Ridge	SVR	XGB
ACSF (48)	--	0.566	0.274	0.911	0.829	0.000	0.184	0.257	0.456	0.240	0.000	0.299	0.949	0.266	0.227	0.939
	BoB (413)	0.606	0.212	0.911	0.840	0.000	0.000	0.128	0.324	0.000	0.000	0.419	0.955	0.227	0.000	0.942
	FP (1358)	0.611	0.200	0.914	0.847	0.000	0.032	0.207	0.182	0.000	0.000	0.609	0.953	0.218	0.121	0.941
SOAP (15876)	--	0.799	0.896	0.928	0.928	0.000	0.366	0.856	0.901	0.815	0.594	0.961	0.964	0.856	0.459	0.966
	BoB (413)	0.799	0.894	0.925	0.928	0.000	0.330	0.863	0.902	0.829	0.594	0.955	0.964	0.866	0.395	0.966
	FP (1358)	0.800	0.893	0.927	0.931	0.000	0.365	0.825	0.902	0.814	0.594	0.967	0.964	0.881	0.477	0.968
PhysOrg local (35)	--	0.704	0.648	0.906	0.874	0.789	0.688	0.629	0.648	0.619	0.000	0.804	0.949	0.633	0.546	0.945
	PhysOrg global (15)	0.702	0.645	0.929	0.885	0.878	0.561	0.625	0.645	0.616	0.000	0.838	0.963	0.631	0.486	0.955

experimental and computational study elucidated that the radical addition step is irreversible and determines the regioselectivity of the overall radical C–H functionalization.[13] This work provides the critical mechanistic basis for our computations. The regioselectivities in the sample space were computed by comparing the free energy barriers of the competing radical additions (Figure 2C).[14]

The training of the ML model applied a library of widely used molecular features, which describe both the local atomic properties in the forming C–C bond of radical addition and the global molecular properties of the arene and the radical. The local features include atom-centered symmetry functions[15] (ACSF) and smooth overlap of atomic positions[16] (SOAP). The global features include bag of bonds[17] (BOB) and molecular fingerprints[18] (FP). In light of Grzybowski's recent discoveries that physically relevant features are useful in the ML prediction of the selectivity in Diels–Alder reactions[6g] and the pK_a of the C–H bond,[6k] we also selected a set of chemical descriptors with physical organic basis (PhysOrg, Figure 2D,E). This collection of 50 features allows the description of both electronic and steric effects based on frontier molecular orbital (FMO) energy, atomic charge, buried volume,[19] NICS value[20] as well as Wiberg bond index[21] (Figure 2F).

We performed a systematic evaluation of the combinations of molecular features and ML models. The regression performances using five-fold cross validation and random splitting are summarized in Table 1. SOAP local feature generally outperformed ACSF local feature in most tested ML models. The addition of global features (BoB or FP) could further improve the regression limitedly. This led to the best combination of SOAP/FP features and XGBoost (XGB)[22] regressor, with a R^2 of 0.968. It is worth noting that the tree-based ML regressors, Decision Tree (DTree),[23] Random Forest (RF)[23] and XGB, generally have better performances than the other ML models, which may be related to the size of the sample space and the nature of the selectivity prediction. In addition to these known features which require a large feature space (Table 1), we found that the usage of PhysOrg feature can achieve competitive performance with a significantly smaller feature space. With only 35 descriptors, the PhysOrg local feature with RF model achieved a nice regression with 0.949 R^2. The addition of PhysOrg global feature further improved R^2 to 0.963. This suggests that the PhysOrg feature is highly condensed in selectivity-determining information, which allows the regioselectivity prediction with only a handful of descriptors.

We next evaluated the key aspects of the promising combinations, XGB model using SOAP/FP features and RF model using PhysOrg features. The SOAP/FP-XGB model only requires the geometric coordinates of the arene and radical reactants, which offers potential for high-throughput virtual screening if geometry from low level of theory is applicable. This motivated us to further examine three levels of geometry in the SOAP/FP-XGB model, MMFF94 force field,[24] PM7 semi-empirical theory[25] and B3LYP[26]/6-311 + G(2d,p). All three SOAP/FP-XGB models provided satisfying regression performances (Figure 3A–C). The site accuracies (the chance to predict the correct reaction site between two positions) and the selectivity accuracies (the chance to correctly determine if the selectivity is high, low or insignificant)[27] are all higher than 90%. The performance of SOAP/FP(MMFF94)-XGB model (Figure 3C) is encouraging because it suggests that the SOAP feature does not necessarily require computationally expensive geometries to achieve the desired ML prediction of reaction performance, which is critical for high-throughput screening purpose.

For the PhysOrg-RF model, feature selection technique was applied to further decrease its complexity by selecting the key features that are most relevant for the regioselectivity prediction. This can improve its generalization ability without jeopardizing the overall accuracy.[28] 32 out of the original 50 features were identified as the best subset using recursive feature elimination with cross-validation (RFECV),[29] which

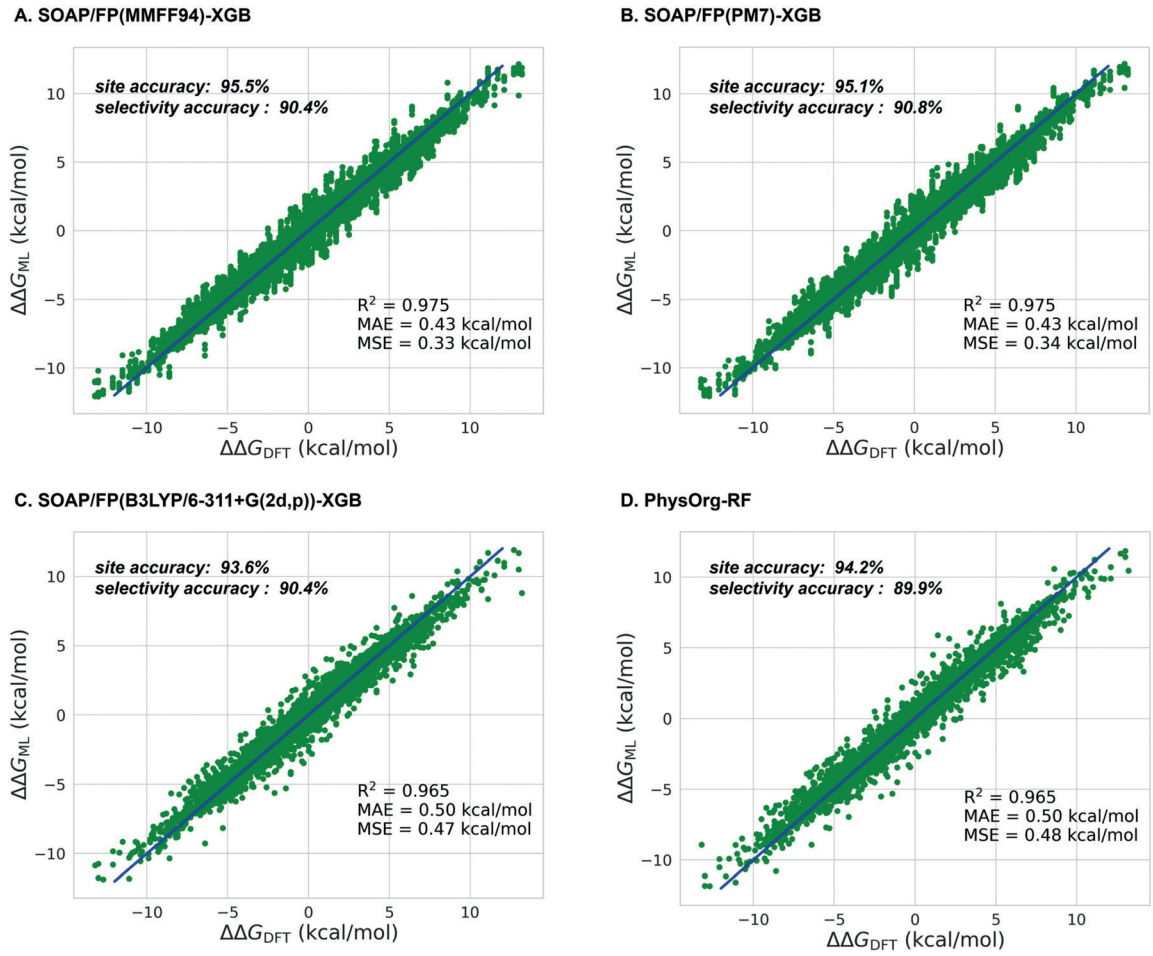

Figure 3. Regression performances of SOAP/FP-XGB models and PhysOrg-RF model.

finalized the RF model setting. The regression performance of the PhysOrg-RF model is shown in Figure 3D, with a site accuracy of 94.2% and a selectivity accuracy of 89.9%. In addition to the high descriptive ability, the physical organic descriptors offered the interpretability of the PhysOrg-RF model. Feature ranking of these descriptors can provide mechanistic insights in the control of regioselectivity (Figure S1 in the Supporting Information), as in the ML prediction of C–N cross coupling reaction by Doyle and coworkers.[6c] The computation of the required features in the PhysOrg-RF model can be easily achieved in any common quantum-chemical software packages, which generally takes minutes in modern PC for arenes and radicals in related transformations.

Comparing with the above ML-training of barrier differences ($\Delta\Delta G$) between the regioisomeric radical additions, the ML-training of absolute barriers (ΔG) and subsequent regioselectivity prediction using the ML-predicted barriers showed worse performance. Using the PhysOrg-RF model as a demonstration, the ML training of ΔG can achieve satisfying performance with R^2 of 0.939 and MAE of 0.79 kcal mol^{-1} (Figure 4A). However, transferring these ML-predicted ΔG to the corresponding $\Delta\Delta G$ showed a noticeable accuracy reduction (Figure 4B vs. Figure 3D). The site accuracy is 87.1%, and the selectivity accuracy is 80.1%. This suggests that the ML prediction of reaction performances would be benefited by customization to the target property (reactivity, chemo-, regio- or enantioselectivity), instead of developing a ML model for barrier regression and applying such model in various related selectivity predictions.

To further evaluate the generalization ability of the developed PhysOrg-RF and SOAP/FP(MMFF94)-XGB models in the target regioselectivity challenges, especially the "unseen" classes of compounds, we tested these two models in three additional datasets of substituents, arene scaffolds and experimental examples (Figure 5).[30] Figure 5A includes the MAEs of the ML-predicted $\Delta\Delta G$s for ten typical substituents that are compatible in the radical C–H functionalization of arenes,[3–5] comparing with the DFT-calculated regioselectivities of CF_3 radical C–H functoinalization for representative substituted heteroarenes. Three substituents (H, CN, OMe) were present in the original dataset (Figure 2), which have the expected consistent performances. The additional seven substituents showed higher errors of predictions, especially the alarming performances of CH_3, *t*Bu, Cl and CF_3 substituents (Figure 5A). This points out that the original dataset did not thoroughly map the chemical space of substituent variation, which results in limited generalization ability in this dimension. H, OMe, CN are the substituents that are capable to describe the conjugative electronic effect, while the inductive effect (CF_3, Cl) and extreme steric effect (*t*Bu) are

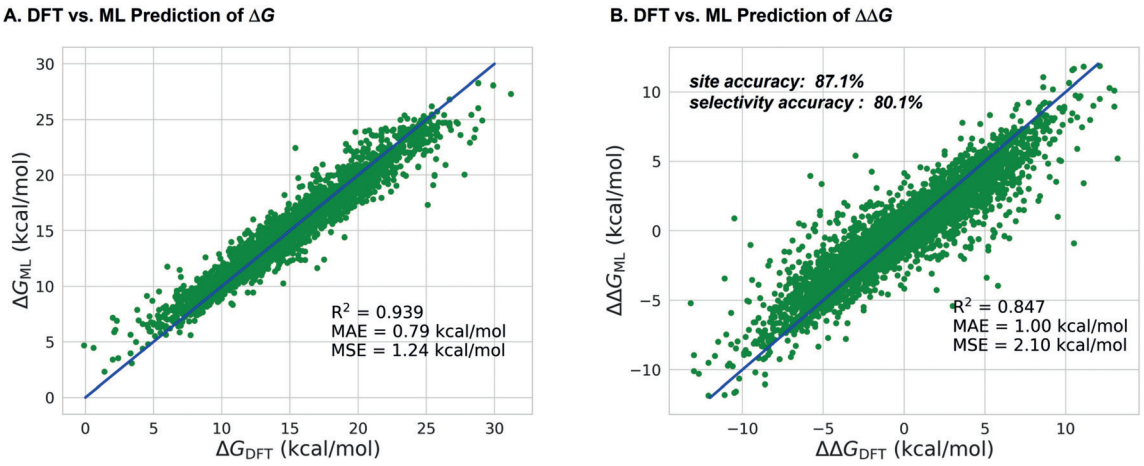

Figure 4. Regression performance of the PhysOrg-RF model in radical addition barriers (A) and the regioselectivity predictions using the ML-predicted barriers (B).

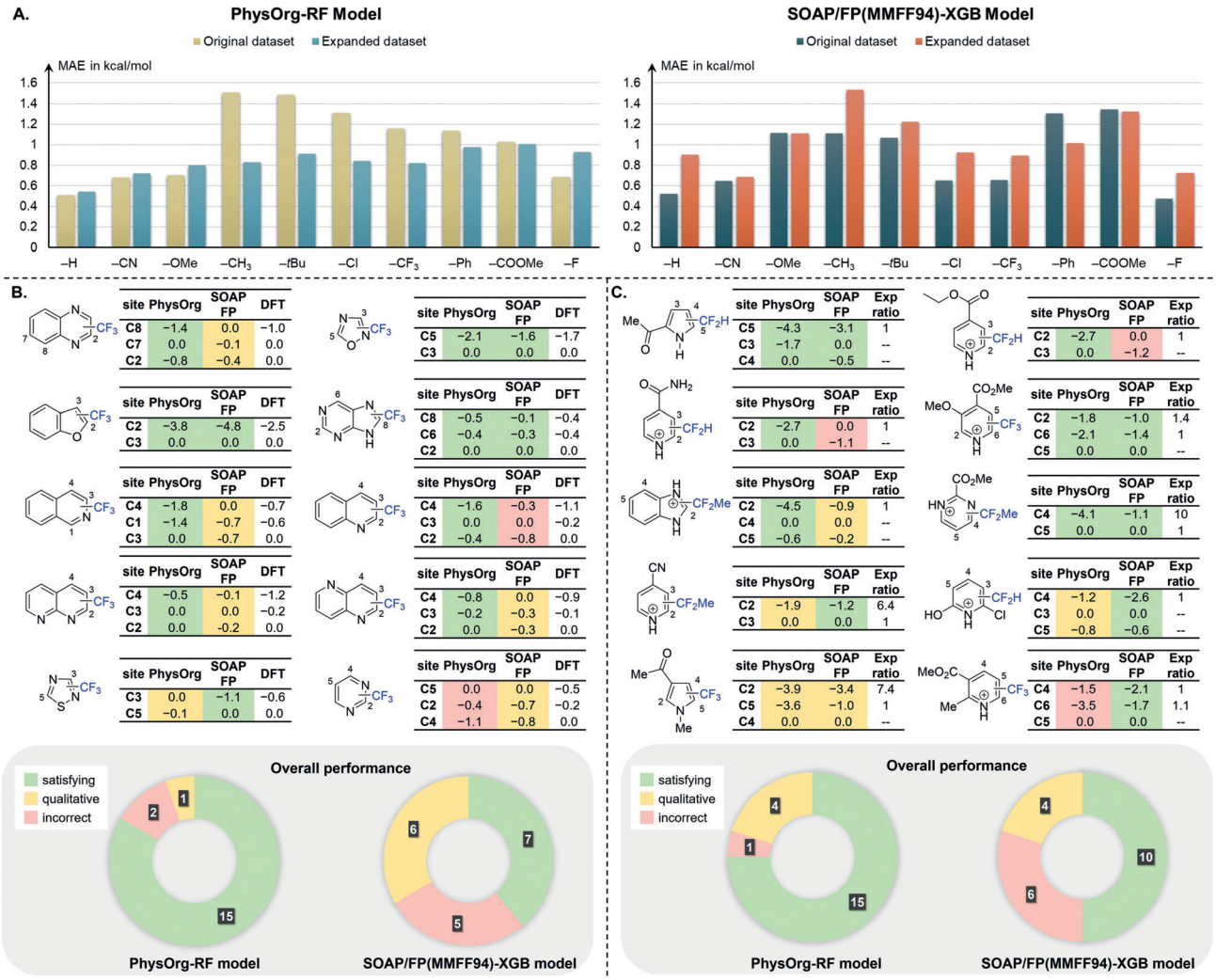

Figure 5. Evaluation of generalization ability of the PhysOrg-RF and SOAP/FP(MMFF94)-XGB models in additional datasets of substituent (A), arene scaffold (B) and experimental examples (C). Selected cases are presented in the evaluation results of scaffold (B) and experimental examples (C).[30]

not properly represented. The issues with CH_3 could due to the lack of alkyl substituents in the original dataset.

Considering these limitations, additional DFT calculations of transition states and reactants for the CF_3- and *t*Bu-

substituted cases were performed to expand the dataset to 9438 regioisomeric competitions.[31] This target-orientated inclusion of new data indeed improved the performance of PhysOrg-RF model, leading to a consistent predictive ability for the tested substituents (Figure 5A). The chemical knowledge embedded in the physical organic descriptors are particularly suitable for such iterative improvement, and user feedbacks of the PhysOrg-RF model will serve as a strong engine to continuously improve its predictive ability. For the SOAP/FP(MMFF94)-XGB model, including new data of CF_3 and *t*Bu substituents did not improve the performance (Figure 5A). This again emphasizes the sharp contrast of feature dimension and nature between the dozens of physical organic descriptors and the thousands of SOAP and FP descriptors.

The evaluations of heteroarene scaffolds and experimental examples also corroborated the generalization ability of the PhysOrg-RF model. For 18 arene scaffolds that were not present in the dataset, satisfying regioselectivity predictions of CF_3 radical functionalization were found for the updated PhysOrg-RF model trained by the expanded dataset (Figure 5B). 15 heteroarenes have the satisfying predictions, and representative examples of these predictions are included in Figure 5B. For 20 experimental examples from Baran's reports,[3–5] the PhysOrg-RF model also provided desired predictions for the majority of cases (Figure 5C). In contrast, quite a few predictions using SOAP/FP(MMFF94)-XGB model were not satisfying (Figure 5B,C). This alerted the limited generalization ability of the developed SOAP/FP-(MMFF94)-XGB model. Unlike the PhysOrg descriptors (32 descriptors after feature selection), the large feature space of SOAP (15876 descriptors) and FP (1358 descriptors) would require a significantly larger training set to match the feature space and support the desired generalization ability. Therefore, the PhysOrg-RF model is the recommended model for the target regioselectivity prediction. The regioselectivity prediction using the computational statistics-based ML strategy and physical organic descriptors can be transferred to additional selectivity predictions of organic transformations, which are currently under investigation in our laboratory.

Conclusion

In summary, rapid and reliable regioselectivity prediction of radical C–H functionalization of heterocycles was realized by a computational statistics-based ML strategy. The combination of physical organic features and random forest algorithm provides a satisfying regression model that connects the DFT-computed properties of isolated reactants with transition state barriers. The random forest model with 32 physical organic descriptors achieved 94.2% site accuracy and 89.9% selectivity accuracy for the out-of-sample test set. The applicability of this ML model was further validated by comparing the predictions with DFT-computed selectivity in additional datasets containing "unseen" substituents and arene scaffolds, as well as experimental results. This work provides a useful tool for regioselectivity prediction in the widely used radical C–H functionalization of heterocycles and reveals the potential of mechanism-based computational statistics as a complimentary data source in performance prediction of organic transformations.

Acknowledgements

Financial support from the NSFC (21702182, 21873081 for X.H.), Fundamental Research Funds for the Central Universities (2019QNA3009, X.H.), Gold Experts Plan of Zhejiang University and China Postdoctoral Science Foundation (2018M640546 for S.Q.Z.) is gratefully acknowledged. Calculations were performed on the high-performance computing system at the Department of Chemistry, Zhejiang University.

Conflict of interest

The authors declare no conflict of interest.

Keywords: machine learning · mechanism-based computational statistics · radical C–H functionalization · random forest model · regioselectivity prediction

[1] a) X. Chen, K. M. Engle, D.-H. Wang, J.-Q. Yu, *Angew. Chem. Int. Ed.* **2009**, *48*, 5094; *Angew. Chem.* **2009**, *121*, 5196; b) J. Wencel-Delord, T. Dröge, F. Liu, F. Glorius, *Chem. Soc. Rev.* **2011**, *40*, 4740; c) L. Ping, D. S. Chung, J. Bouffard, S.-G. Lee, *Chem. Soc. Rev.* **2017**, *46*, 4299; d) K. Murakami, S. Yamada, T. Kaneda, K. Itami, *Chem. Rev.* **2017**, *117*, 9302; e) R. Shang, L. Ilies, E. Nakamura, *Chem. Rev.* **2017**, *117*, 9086; f) P. Gandeepan, T. Müller, D. Zell, G. Cera, S. Warratz, L. Ackermann, *Chem. Rev.* **2019**, *119*, 2192; g) J. A. Labinger, *Chem. Rev.* **2017**, *117*, 8483; h) P. B. Arockiam, C. Bruneau, P. H. Dixneuf, *Chem. Rev.* **2012**, *112*, 5879; i) T. W. Lyons, M. S. Sanford, *Chem. Rev.* **2010**, *110*, 1147; j) D. A. Colby, R. G. Bergman, J. A. Ellman, *Chem. Rev.* **2010**, *110*, 624.

[2] a) F. Minisci, R. Bernardi, F. Bertini, R. Galli, M. Perchinummo, *Tetrahedron* **1971**, *27*, 3575; b) F. Minisci, F. Fontana, E. Vismara, *J. Heterocycl. Chem.* **1990**, *27*, 79; c) D. C. Harrowven, B. J. Sutton, S. Coulton, *Org. Biomol. Chem.* **2003**, *1*, 4047; d) D. C. Harrowven, B. J. Sutton, in *Prog. Heterocycl. Chem. Vol. 16* (Eds.: G. W. Gribble, J. Joule), Elsevier, Amsterdam, **2005**, pp. 27–53; e) W. R. Bowman, J. M. D. Storey, *Chem. Soc. Rev.* **2007**, *36*, 1803; f) M. A. J. Duncton, *MedChemComm* **2011**, *2*, 1135; g) P. S. Fier, J. F. Hartwig, *J. Am. Chem. Soc.* **2014**, *136*, 10139; h) J. Tauber, D. Imbri, T. Opatz, *Molecules* **2014**, *19*, 16190.

[3] For reviews, see: a) M. Yan, J. C. Lo, J. T. Edwards, P. S. Baran, *J. Am. Chem. Soc.* **2016**, *138*, 12692; b) J. M. Smith, J. A. Dixon, J. N. deGruyter, P. S. Baran, *J. Med. Chem.* **2019**, *62*, 2256.

[4] a) Y. Ji, T. Brueckl, R. D. Baxter, Y. Fujiwara, I. B. Seiple, S. Su, D. G. Blackmond, P. S. Baran, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 14411; b) Y. Fujiwara, J. A. Dixon, F. O'Hara, E. D. Funder, D. D. Dixon, R. A. Rodriguez, R. D. Baxter, B. Herlé, N. Sach, M. R. Collins, Y. Ishihara, P. S. Baran, *Nature* **2012**, *492*, 95; c) Y. Fujiwara, J. A. Dixon, R. A. Rodriguez, R. D. Baxter, D. D. Dixon, M. R. Collins, D. G. Blackmond, P. S. Baran, *J. Am. Chem. Soc.* **2012**, *134*, 1494; d) F. O'Hara, R. D. Baxter, A. G. O'Brien, M. R. Collins, J. A. Dixon, Y. Fujiwara, Y. Ishihara, P. S. Baran, *Nat. Protoc.* **2013**, *8*, 1042; e) Q. Zhou, A. Ruffoni, R.

Gianatassio, Y. Fujiwara, E. Sella, D. Shabat, P. S. Baran, *Angew. Chem. Int. Ed.* **2013**, *52*, 3949; *Angew. Chem.* **2013**, *125*, 4041; f) R. Gianatassio, S. Kawamura, C. L. Eprile, K. Foo, J. Ge, A. C. Burns, M. R. Collins, P. S. Baran, *Angew. Chem. Int. Ed.* **2014**, *53*, 9851; *Angew. Chem.* **2014**, *126*, 10009; g) J. Gui, Q. Zhou, C.-M. Pan, Y. Yabe, A. C. Burns, M. R. Collins, M. A. Ornelas, Y. Ishihara, P. S. Baran, *J. Am. Chem. Soc.* **2014**, *136*, 4853; h) F. O'Hara, A. C. Burns, M. R. Collins, D. Dalvie, M. A. Ornelas, A. D. N. Vaz, Y. Fujiwara, P. S. Baran, *J. Med. Chem.* **2014**, *57*, 1616.

[5] F. O'Hara, D. G. Blackmond, P. S. Baran, *J. Am. Chem. Soc.* **2013**, *135*, 12122.

[6] a) C. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 3582; b) C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434; c) D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186; d) H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2018**, *4*, 1465; e) M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *J. Am. Chem. Soc.* **2018**, *140*, 5004; f) P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, *Chem. Sci.* **2018**, *9*, 6091; g) W. Beker, E. P. Gajewska, T. Badowski, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2019**, *58*, 4515; *Angew. Chem.* **2019**, *131*, 4563; h) C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370; i) I. W. Davies, *Nature* **2019**, *570*, 175; j) A. Filipa de Almeida, R. Moreira, T. Rodrigues, *Nat. Rev. Chem.* **2019**, *3*, 589; k) R. Roszak, W. Beker, K. Molga, B. A. Grzybowski, *J. Am. Chem. Soc.* **2019**, *141*, 17142; l) P. L. Kang, C. Shang, Z.-P. Liu, *J. Am. Chem. Soc.* **2019**, *141*, 20525; m) A. R. Rosales, J. Wahlers, E. Lime, R. E. Meadows, K. W. Leslie, R. Savin, F. Bella, E. Hansen, P. Helquistl, R. H. Munday, O. Wiest, P.-O. Norrby, *Nat. Catal.* **2019**, *2*, 41; n) A. Tomberg, M. J. Johansson, P.-O. Norrby, *J. Org. Chem.* **2019**, *84*, 4695; o) A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, 247.

[7] a) S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904; *Angew. Chem.* **2016**, *128*, 6004; b) B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos, T. Klucznik, *Chem* **2018**, *4*, 390; c) T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. Startek, G. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem* **2018**, *4*, 522; d) K. Molga, P. Dittwald, B. A. Grzybowski, *Chem* **2019**, *5*, 460; e) W. Jaworski, S. Szymkuc, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, B. A. Grzybowski, *Nat. Commun.* **2019**, *10*, 1434; f) K. Molga, E. P. Gajewska, S. Szymkuć, B. A. Grzybowski, *React. Chem. Eng.* **2019**, *4*, 1506; g) E. P. Gajewska, S. Szymkuć, P. Dittwald, M. Startek, O. Popik, J. Mlynarski, B. A. Grzybowski, *Chem* **2020**, *6*, 280; h) T. Badowski, E. P. Gajewska, K. Molga, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2020**, *59*, 725; *Angew. Chem.* **2020**, *132*, 735.

[8] For reviews, see: a) M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292; b) F. D. Toste, M. S. Sigman, S. J. Miller, *Acc. Chem. Res.* **2017**, *50*, 609; c) S. G. Robinson, M. S. Sigman, *Acc. Chem. Res.* **2020**, *53*, 289.

[9] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.

[10] For the commercial availability of the alkylsulfinate salts, see: https://www.sigmaaldrich.com/catalog/product/aldrich/ald00444?lang=en&region=US.

[11] a) I. B. Seiple, S. Su, R. A. Rodriguez, R. Gianatassio, Y. Fujiwara, A. L. Sobel, P. S. Baran, *J. Am. Chem. Soc.* **2010**, *132*, 13194; b) Y. Fujiwara, V. Domingo, I. B. Seiple, R. Gianatassio, M. Del Bel, P. S. Baran, *J. Am. Chem. Soc.* **2011**, *133*, 3292.

[12] All DFT calculations were performed with Gaussian 09 software package. Computational details are included in the Supporting Information.

[13] R. D. Baxter, Y. Liang, X. Hong, T. A. Brown, R. N. Zare, K. N. Houk, P. S. Baran, D. G. Blackmond, *ACS Cent. Sci.* **2015**, *1*, 456.

[14] 3342 transition states of the 3406 radical additions in the sample space were located. This leads to 5190 regioisomeric competitions for the initial ML training.

[15] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.

[16] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754.

[17] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326.

[18] a) J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273; b) D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742.

[19] a) A. Poater, F. Ragone, S. Giudice, C. Costabile, R. Dorta, S. P. Nolan, L. Cavallo, *Organometallics* **2008**, *27*, 2679; b) A. Poater, B. Cosenza, A. Correa, S. Giudice, F. Ragone, V. Scarano, L. Cavallo, *Eur. J. Inorg. Chem.* **2009**, 1759; c) H. Clavier, S. P. Nolan, *Chem. Commun.* **2010**, *46*, 841; d) A. Poater, F. Ragone, R. Mariz, R. Dorta, L. Cavallo, *Chem. Eur. J.* **2010**, *16*, 14348.

[20] Z. Chen, C. S. Wannere, C. Corminboeuf, R. Puchta, P. v. R. Schleyer, *Chem. Rev.* **2005**, *105*, 3842.

[21] a) K. B. Wiberg, *Tetrahedron* **1968**, *24*, 1083; b) I. Mayer, *Chem. Phys. Lett.* **1983**, *97*, 270; c) I. Mayer, *J. Comput. Chem.* **2007**, *28*, 204.

[22] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, Association for Computing Machinery, San Francisco, California, USA, **2016**, pp. 785 – 794.

[23] L. Breiman, *Machine Learning* **2001**, *45*, 5.

[24] a) T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 490; b) T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 520; c) T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 553; d) T. A. Halgren, R. B. Nachbar, *J. Comput. Chem.* **1996**, *17*, 587; e) T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 616.

[25] J. J. P. Stewart, *J. Mol. Model.* **2013**, *19*, 1.

[26] a) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785; b) A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[27] The selectivities are separated into three categories following the general consesus of synthetic chemists: high (larger than 10:1), low (3:1 to 10:1), and insignificant (smaller than 3:1).

[28] a) Y. Saeys, I. Inza, P. Larrañaga, *Bioinformatics* **2007**, *23*, 2507; b) M. Shahlaei, *Chem. Rev.* **2013**, *113*, 8093.

[29] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Machine Learning* **2002**, *46*, 389.

[30] Detailed performances and determinations of the prediction qualitiy of the ML model are included in the Supporting Information.

[31] 6114 transition states of the designed 6214 radical additions in the expanded sample space were located. This leads to 9370 regioisomeric competitions for the training of updated PhysOrg-RF and SOAP/FP(MMFF94)-XGB models.

Manuscript received: January 19, 2020
Revised manuscript received: March 30, 2020
Accepted manuscript online: May 2, 2020
Version of record online: May 26, 2020

