ACES Asian Chemical Editorial Society

doi.org/10.1002/asia.202300011

CHEMISTRY AN ASIAN JOURNAL Research Article

www.chemasianj.org

Special Collection

Exploring Spectrum-based Molecular Descriptors for Reaction Performance Prediction

Miao-Jiong Tang,[a] Li-Cheng Xu,[a] Shuo-Qing Zhang,*[a] and Xin Hong*[a, b, c]

Abstract: Despite the availability and accuracy of modern spectroscopic characterization, the utilization of spectral information in chemical machine learning is still primitive. Here, we report an optical character recognition-based automatic process to utilize spectral information as molecular descriptors, which directly transforms experimental spectrum images to readable vectors. We demonstrate its machine learning application in the reaction yield dataset of Pd-catalyzed Buchwald-Hartwig cross-coupling with aryl halides. In addition, we also show that the predicted spectrum can serve as an alternative encoding source to support the model training.

Introduction

The rapid development of machine learning (ML) applications has reignited synthetic chemists' interest in prediction for reaction performance,[1–4] which is one of the classic questions in physical organic chemistry dating back to linear free energy relationship[5,6] (LFER). For synthetic chemistry, predicting the performance of unknown reactions based on prior knowledge can provide strong support for the optimization of experimental process and reducing the trial and error costs; data-driven approach has been serving as a powerful strategy in this regard.[7–12] Existing models of reaction performance prediction[10] rely on the molecular descriptors to digitalize the molecular structure (Figure 1). For example, SMILES strings and One-Hot encoding[13] are widely applied and preferentially used in distinguishing compounds in a chemical space. Physical organic parameters (such as HOMO, LUMO, charge and dipole, etc.) have also been used to represent molecule in chemical machine learning.[14–17] In addition, molecule fingerprints[18–21] (FP) and atom-centered symmetry functions[22] (ACSF) are typical methods for substructure encoding.

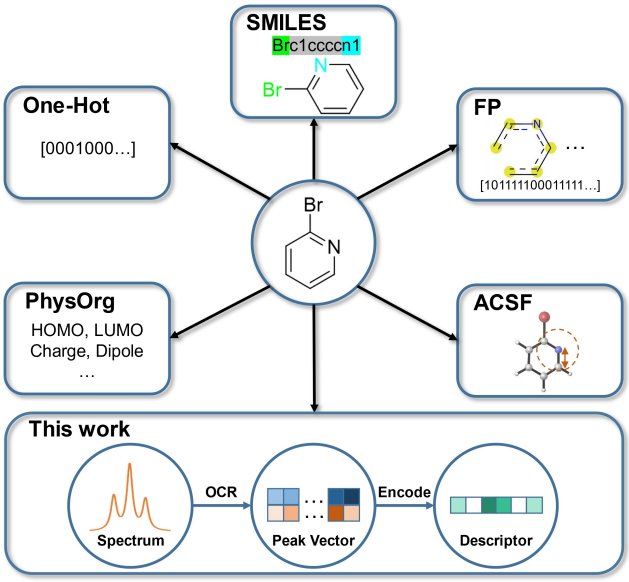


Figure 1. Typical molecular descriptors applied in modern machine learning predictions of reaction performance and this study of spectrum-based molecular descriptors.

In addition to the above molecular descriptors, compound spectrum is an observable representation of the physicochemical properties of a molecule at the microscopic level, which can be easily accessed with an amazing accuracy using modern spectroscopic technologies.[23–26] Current ML applications of the spectral information in synthetic outcome prediction typically use the experimental or computed chemical shifts of certain sites as local descriptors.[27–29] Although this approach is straightforward and intuitive with considerable interpretability, it requires strong domain knowledge of the molecular structure and the structure-activity relationship; in addition, this usage of spectral information is incomplete, which discards the information such as shape and absorption intensity of the spectral peaks. In addition, Roberta et al.[30] developed alternative approach called comparative spectra analysis (CoSA), which calculates the sums of the intensities of NMR spectral peaks within bins as descriptors, with success in the task of predicting

[a] M.-J. Tang, L.-C. Xu, Dr. S.-Q. Zhang, Dr. X. Hong
Center of Chemistry for Frontier Technologies
Department of Chemistry
State Key Laboratory of Clean Energy Utilization
Zhejiang University,
Hangzhou 310027 (P. R. China)
E-mail: hxchem@zju.edu.cn
angellasty@zju.edu.cn

[b] Dr. X. Hong
Beijing National Laboratory for Molecular Sciences
Zhongguancun North First Street NO. 2,
Beijing 100190 (P. R. China)

[c] Dr. X. Hong
Key Laboratory of Precise Synthesis of Functional Molecules
of Zhejiang Province, School of Science
Westlake University
18 Shilongshan Road, Hangzhou 310024,
Zhejiang Province (P. R. China)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/asia.202300011

Special Collection This manuscript is part of a joint special collection on Mechanisms and Selectivities of Organic Reactions – In Celebration of Prof. Kendall N. Houk's 80th birthday.

progesterone activity. Willighagen et al.[31] had also used the ^{1}H and ^{13}C NMR predicted by the ACD predictor directly as QSPR descriptors of molecules to predict the physical properties of molecules.

Recent years have witnessed a fast development of the usage of spectral information in chemical machine learning. The recognition of protein structures by machine learning methods using spectral information[32] as well as spectrum prediction from protein structure[33,34] have been developed and applied to protein molecular dynamics[35] and real-time interaction simulations.[36] There are increasing attempts to apply data-driven ML methods to automate the linking of the spectrum of molecules to structures.[37–39] Recent work of Luo, Jiang and co-workers also implemented spectral information to improve the graph model's capability for retrosynthesis planning.[40] Despite these exciting advances, there currently lacks the ML workflow to automatically process the experimental spectrum images, which limits the application of spectrum-based molecular descriptors in reaction performance prediction. In this work, we report a procedure that automatically process the spectrum image to ML-readable vectors. The effectiveness of the spectrum-based descriptors is demonstrated in the reaction yield prediction task of Pd-catalyzed Buchwald-Hartwig reaction with aryl halides.[41] We further show that it is possible to directly use software-predicted spectrum as an alternative data and knowledge source to support the model training.

Results and Discussion

Considering that existing molecular databases (SciFinder®,[42] NIST,[43] etc.) usually store molecular spectrum data in the form of image, we first developed a tool that automatically process the spectrum images to ML-readable vectors. Figure 2 elaborates the designed generation procedure of spectrum descriptor for formatted spectrum images using ^{1}H NMR of 2-bromopyridine as an example. IR and mass spectra are also processable with details given in Figure S3. For these common types of compound spectrum, we applied the optical character recognition (OCR) technology to process the images, which can obtain the chemical shifts and intensities in an automatic fashion. This peak information is stored as the normalized spectral peak vector. To unify the dimensionality of the spectrum descriptors, we projected the peak vectors onto a one-dimensional grid with a customizable resolutions and range of chemical shifts, which produced the final spectrum descriptor. For the studied spectra in this work, we set the sampling ranges of 1.5–10 ppm and 40–210 ppm for ^{1}H NMR and ^{13}C NMR respectively. The resolutions were determined to be 1 ppm and 5 ppm after hyperparameter optimization (Table S4). In addition to image-type spectrums, the scatter plot-type spectrums stored in JCAMP[44,45] format can also be converted into spectral peak vectors with a simple pre-processing. The entire image processing generally requires about 700 milliseconds for typical organic compounds (about 650 milliseconds for the 2-bromopyridine example in Figure 2), which can support the large-scale virtual screening.

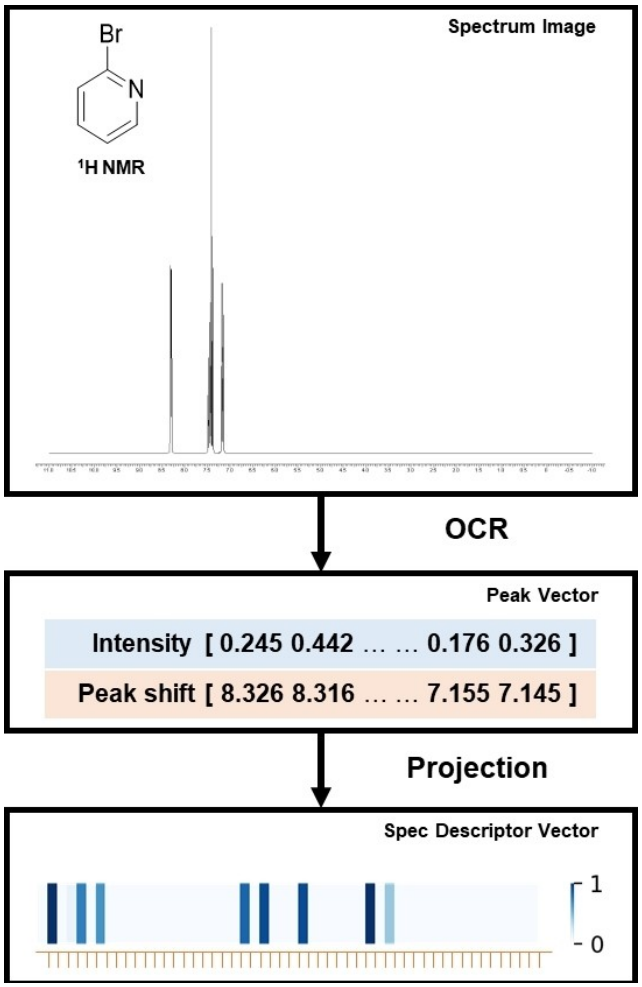
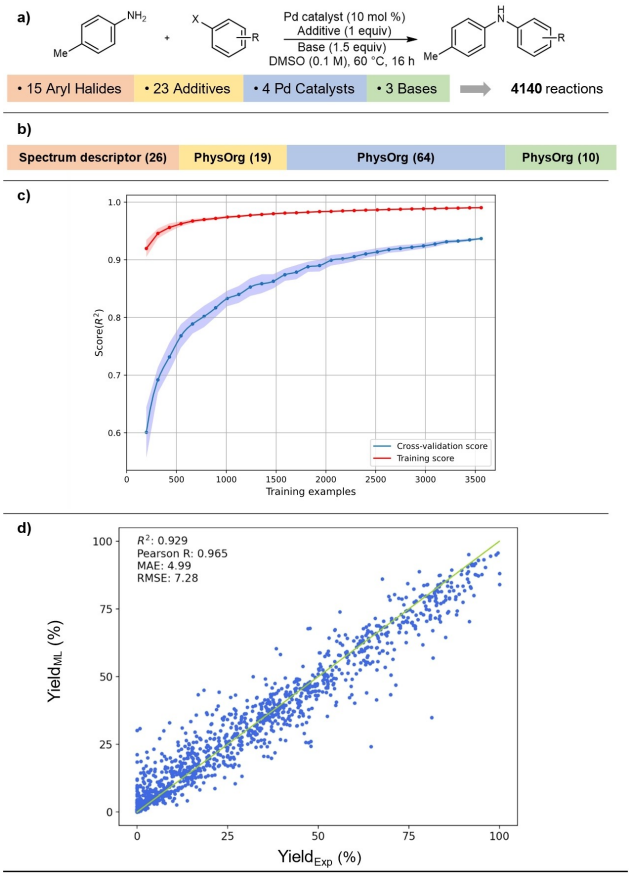


Figure 2. Generation procedure of spectrum descriptor using OCR to access the peak shift and intensity information from spectrum image. With hyperparameter optimization, the sampling range of 1.5-10 ppm and resolution of 1 ppm was applied for the ^{1}H NMR spectra, and the sampling range of 40–210 ppm and resolution of 5 ppm was applied for the ^{13}C NMR spectra.

To demonstrate the effectiveness of the spectrum-based descriptors, we next evaluated its performance in the yield prediction task of Pd-catalyzed Buchwald-Hartwig cross-coupling of aryl halides. This high-quality dataset was generated in Doyle's previous ML study of yield prediction,[41] which is maps the reaction yields of 4140 cross-coupling reactions involving 15 aryl halides, 23 additives, 4 palladium catalysts and 3 bases (Figure 3a). We collected the spectrum images (^{1}H NMR, ^{13}C NMR) of the 15 aryl halides in the dataset from the SciFinder® database. The spectrum images of the rest compounds in the dataset are not all available, so the physical organic descriptors in Doyle's original study were used. The final reaction coding is a 119-dimensional vector with 26 dimensions of spectrum descriptors for aryl halides and 93 dimensions of physical organic parameters for the rest compounds (Figure 3b). Using these encodings, the learning curve of the random forest (RF) model for regression is shown in Figure 3c. This learning curve is built by random extraction of subsets from the total dataset



	R^2	RMSE
Trial 1	0.928	7.37
Trial 2	0.930	7.26
Trial 3	0.932	7.15
Trial 4	0.928	7.34
Trial 5	0.932	7.18
Trial 6	0.929	7.30
Trial 7	0.931	7.22
Trial 8	0.925	7.49
Trial 9	0.930	7.25
Trial 10	0.930	7.27
Average	**0.929**	**7.28**

Figure 3. Application of spectrum descriptor in machine learning yield prediction of Pd-catalyzed Buchwald-Hartwig reaction. a) Overview of dataset b) Composition of the reaction encodings c) Learning curve of the random forest model, performed under ten-fold cross-validation using standard deviation (σ) as an uncertainty measure. d) Regression performances in 10-fold cross-validation.

for cross validation. In each subset, we performed 10 10-fold cross validations to obtain the average mean value as well as the error of prediction. This learning curve suggested that a sample space of over 3000 examples (about 70% of the dataset) is necessary to reach the desired training performance. Therefore, we used RF model with a 70/30 split of training and test data.

Figure 3d shows the regression performances of the RF model with the designed combination of spectrum descriptors and physical organic parameters. The model provides a convincing correlation between the predicted and experimental yields with a coefficient of determination R^2 of 0.929, Pearson correlation coefficient R of 0.965, mean absolute error (MAE) of 4.99% and root mean square error (RMSE) of 7.28%. To remove the influence of random dataset splitting, we further evaluated the model performances with ten trials. While each trial has a different random splitting of the dataset, the model's predictive ability is steady (Figure 3d). The R^2 values of the ten trials varied between 0.925 to 0.931, and the RMSEs are all below 7.50. These ML results are comparable to Doyle's original regression performance under the same RF algorithm, which provide strong support that the spectrum-based descriptors can allow similar level of descriptive ability comparing with the computationally expensive physical organic descriptors.

We next analyzed the feature importance of the trained RF model using RFECV method. The Top10 features are shown in Figure 4. It is interesting that seven of the Top10 features are spectrum- based descriptors, despite that there are only 26 spectrum descriptors in the 119-dimensional reaction encodings. This suggests that the spectrum descriptors are critical for the success of the model training. The most significant contribution to the model among these spectrum descriptors is the ^{1}H NMR absorption in the range of 2.5 to 3.5 (δ_H (2.5–3.5)), which corresponds to the absorption intensities on saturated hydrogen atoms can contribute to the yield prediction. In addition, the ^{1}H NMR absorption in the range of 7.5 to 8.5 (δ_H (7.5–8.5)) represents the characteristic absorption of α-hydrogen on the pyridine ring, and its absorption intensity expresses the substitution information of this site. In contrast to the widely used approaches[41] of labeling the NMR shift of certain site use as a physical organic descriptor, our spectrum descriptors highlight the presence or absence of absorption in a particular region of the spectrum image, which presents a new perspective on the chemical environment of the substructure in a molecule.

Considering that the spectrum images may not all be directly available, we also explored whether spectrum descriptors could be useful in a fully automatic process using software-

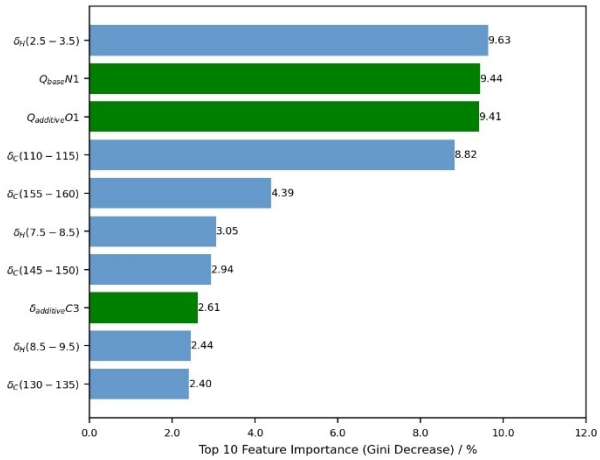


Figure 4. Ranking of feature importance. Blue labels the spectrum descriptors.

Table 1. Performance of automatically generated spectrum descriptors.

entry	Aryl Halide des	Additive des	Ligand des	Base des	R^2	Pearson R	MAE	RMSE
1	PhysOrg*	PhysOrg*	PhysOrg*	PhysOrg*	0.921*	/	/	7.81*
2	spec	PhysOrg*	PhysOrg*	PhysOrg*	0.929	0.965	4.99	7.28
3	nmrdb spec	PhysOrg*	PhysOrg*	PhysOrg*	0.907	0.952	5.26	8.15
4	nmrdb spec	nmrdb spec	nmrdb spec	nmrdb spec	0.912	0.956	5.50	8.14

* Reported in Ref. [41].

generated spectrum images. We used the open-source NMR prediction tool NMRDB[46–48] to rapidly predict the ^{1}H NMR and ^{13}C NMR spectrum of the molecules. Using the automatically generated spectrum, we explored their performances in the same dataset of Pd-catalyzed Buchwald-Hartwig cross-coupling of aryl halides (Table 1). Entry 1 is original prediction performance from Doyle's work using physical organic descriptors. Entry 2 is the averaged performance of ten trials in Figure 3d. Replacing the experimental spectrums of aryl halides with the NMRDB-predicted spectrums lower the regression performance limitedly (entry 3); the R^2 decreases from 0.929 to 0.907. If all physical organic descriptors were replaced by spectrum descriptors generated from the NMRDB-predicted spectrums, the regression performance is still acceptable with a R^2 value of 0.912 (entry 4). Further evaluations of the NMRDB-predicted spectrums were performed on all four reaction components and found comparable regression performances (Table S5). These results collectively suggest that the automatically generated spectrum can serve as an external knowledge source, which would be useful in large-scale virtual screening.

Conclusion

In summary, an OCR-based approach that automatically process spectrum images to molecular descriptor was developed. By shift-intensity and subsequent projecting them onto grids, the developed spectrum descriptors can exact the spectrum information from images. The machine learning application of the designed spectrum descriptors was further demonstrated in the yield prediction task of Pd-catalyzed Buchwald-Hartwig cross-coupling of aryl halides, which provided a machine learning model that has similar predictive ability comparing with the physical organic descriptors. In addition, analysis of the feature importance ranking revealed the key sites that are important for the determination of reaction yield. With the help of existing tools for rapid simulation of compound spectrum, we further demonstrated that the spectrum descriptors can be generated using virtual spectrums. This suggested that the spectrum descriptors can be easily integrated into a fully automatic machine learning workflow. We envision that this tool and the spectrum descriptors can provide useful support for future machine learning tasks of reaction performance prediction.

Acknowledgements

Financial support from the National Key R&D Program of China (2022YFA1504301, X. H.), National Natural Science Foundation of China (21873081 and 22122109, X. H.; 22103070, S.-Q. Z.), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-006, X. H.), Beijing National Laboratory for Molecular Sciences (BNLMS202102, X. H.), CAS Youth Interdisciplinary Team (JCTD-2021-11, X. H.), Fundamental Research Funds for the Central Universities (226-2022-00140 and 226–2022-00224, X. H.), the Center of Chemistry for Frontier Technologies and Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province (PSFM 2021–01, X. H.), the State Key Laboratory of Clean Energy Utilization (ZJU-CEU2020007, X. H.) are gratefully acknowledged. Machine learning modelling and calculations were performed on the high-performance computing system at Department of Chemistry, Zhejiang University.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords: molecular descriptor · spectrum · machine learning · yield prediction · quantitative structure-activity relationship

[1] A. F. Zahrt, S. V. Athavale, S. E. Denmark, *Chem. Rev.* **2020**, *120*, 1620–1689.
[2] K. W. Lexa, K. M. Belyk, J. Henle, B. Xiang, R. P. Sheridan, S. E. Denmark, R. T. Ruck, E. C. Sherer, *Org. Process Res. Dev.* **2022**, *26*, 670–682.
[3] S. Zhang, L. Xu, S. Li, J. C. A. Oliveira, X. Li, L. Ackermann, X. Hong, *Chemistry A European J* **2022**, DOI 10.1002/chem.202202834.
[4] Y. Shen, J. E. Borowski, M. A. Hardy, R. Sarpong, A. G. Doyle, T. Cernak, *Nat Rev Methods Primers* **2021**, *1*, 23.
[5] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
[6] R. W. Taft, *J. Am. Chem. Soc.* **1952**, *74*, 2729–2732.
[7] C. Duan, A. Nandy, H. Adamji, Y. Roman-Leshkov, H. J. Kulik, *J. Chem. Theory Comput.* **2022**, *18*, 4282–4292.
[8] A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Acc. Chem. Res.* **2021**, *54*, 1856–1865.
[9] H. J. Kulik, *Isr. J. Chem.* **2022**, *62*, DOI 10.1002/ijch.202100016.

[10] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Cent. Sci.* **2021**, *7*, 1622–1637.
[11] Y. Liu, Q. Yang, Y. Li, L. Zhang, S. Luo, *Chin. J. Org. Chem.* **2020**, *40*, 3812.
[12] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281–1289.
[13] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.
[14] S.-Q. Zhang, X. Hong, *Acc. Chem. Res.* **2021**, *54*, 2158–2171.
[15] P. H.-Y. Cheong, C. Y. Legault, J. M. Um, N. Çelebi-Ölçüm, K. N. Houk, *Chem. Rev.* **2011**, *111*, 5042–5137.
[16] X. Qi, Y. Li, R. Bai, Y. Lan, *Acc. Chem. Res.* **2017**, *50*, 2799–2808.
[17] T. Sperger, I. A. Sanhueza, I. Kalvet, F. Schoenebeck, *Chem. Rev.* **2015**, *115*, 9532–9586.
[18] L. Xue, J. Bajorath, *CCHTS* **2000**, *3*, 363–372.
[19] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.
[20] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
[21] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
[22] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2020**, *247*, 106949.
[23] M. Gastegger, K. T. Schütt, K.-R. Müller, *Chem. Sci.* **2021**, *12*, 11473–11483.
[24] M. Haghighatlari, J. Li, F. Heidar-Zadeh, Y. Liu, X. Guan, T. Head-Gordon, *Chem* **2020**, *6*, 1527–1542.
[25] C. Cobas, *Magn. Reson. Chem.* **2020**, *58*, 512–519.
[26] A. Kurotani, T. Kakiuchi, J. Kikuchi, *ACS Omega* **2021**, *6*, 14278–14287.
[27] C. P. Gordon, C. Raynaud, R. A. Andersen, C. Copéret, O. Eisenstein, *Acc. Chem. Res.* **2019**, *52*, 2278–2289.
[28] R. P. Verma, C. Hansch, *Chem. Rev.* **2011**, *111*, 2865–2899.
[29] A. K. C. A. Reis, R. Rittner, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2007**, *66*, 681–685.
[30] R. Bursi, T. Dao, T. Van Wijk, M. De Gooyer, E. Kellenbach, P. Verwer, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861–867.
[31] E. L. Willighagen, H. M. G. W. Denissen, R. Wehrens, L. M. C. Buydens, *J. Chem. Inf. Model.* **2006**, *46*, 487–494.
[32] H. Ren, Q. Zhang, Z. Wang, G. Zhang, H. Liu, W. Guo, S. Mukamel, J. Jiang, *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2202713119.
[33] W. Hu, S. Ye, Y. Zhang, T. Li, G. Zhang, Y. Luo, S. Mukamel, J. Jiang, *J. Phys. Chem. Lett.* **2019**, *10*, 6026–6031.
[34] H. Ren, H. Li, Q. Zhang, L. Liang, W. Guo, F. Huang, Y. Luo, J. Jiang, *Fundam. res.* **2021**, *1*, 488–494.
[35] L. Zhao, J. Zhang, Y. Zhang, S. Ye, G. Zhang, X. Chen, B. Jiang, J. Jiang, *JACS Au* **2021**, *1*, 2377–2384.
[36] S. Ye, G. Zhang, J. Jiang, *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2025879118.
[37] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*, 6924–6935.
[38] G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, R. Car, *Phys. Chem. Chem. Phys.* **2020**, *22*, 10592–10602.
[39] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, P. Rinke, *Adv. Sci.* **2019**, *6*, 1801367.
[40] B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang, Y. Luo, *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2212711119.
[41] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186–190.
[42] SciFinder®, https://scifinder.cas.org/.
[43] L. P. J., M. W. G., *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, National Institute Of Standards And Technology, Gaithersburg MD, **2022**.
[44] R. S. McDonald, P. A. Wilks, *Appl. Spectrosc.* **1988**, *42*, 151–162.
[45] A. N. Davies, P. Lampen, *Appl. Spectrosc.* **1993**, *47*, 1093–1099.
[46] A. M. Castillo, L. Patiny, J. Wist, *J. Magn. Reson.* **2011**, *209*, 123–130.
[47] D. Banfi, L. Patiny, *Chimia* **2008**, *62*, 280.
[48] J. Aires-de-Sousa, M. C. Hemmer, J. Gasteiger, *Anal. Chem.* **2002**, *74*, 80–90.

Manuscript received: January 5, 2023
Revised manuscript received: February 10, 2023
Accepted manuscript online: February 10, 2023
Version of record online: February 24, 2023