THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Letter

Cite This: J. Phys. Chem. Lett. 2019, 10, 6026−6031

pubs.acs.org/JPCL

Machine Learning Protocol for Surface-Enhanced Raman Spectroscopy

Wei Hu,[†,‡,⊥] Sheng Ye,[‡,⊥] Yujin Zhang,[§,⊥] Tianduo Li,[†] Guozhen Zhang,[‡] Yi Luo,[‡] Shaul Mukamel,[∥] and Jun Jiang*[,‡]

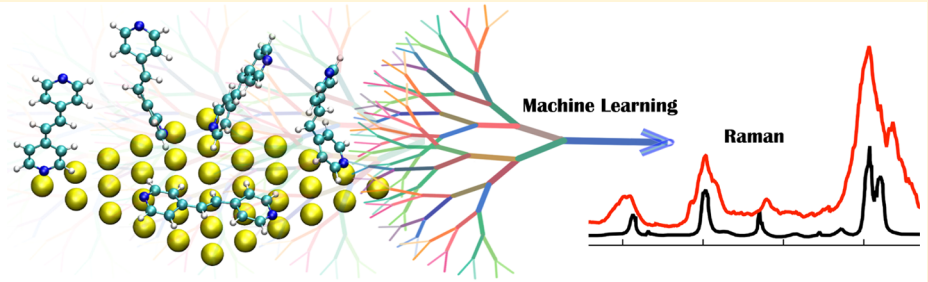
[†]Shandong Provincial Key Laboratory of Molecular Engineering, School of Chemistry and Pharmaceutical Engineering, Qilu University of Technology, Jinan, Shandong 250353, P.R. China

[‡]Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P.R. China

[§]School of Electronic and Information Engineering (Department of Physics), Qilu University of Technology, Jinan, Shandong 250353, P.R. China

[∥]Departments of Chemistry and Physics and Astronomy, University of California, Irvine, California 92697, United States

Supporting Information



ABSTRACT: Surface-enhanced Raman spectroscopy (SERS) is a powerful technique that can capture the electronic−vibrational "fingerprint" of molecules on surfaces. Ab initio prediction of Raman response is a long-standing challenge because of the diversified interfacial structures. Here we show that a cost-effective machine learning (ML) random forest method can predict SERS signals of a *trans*-1,2-bis (4-pyridyl) ethylene (BPE) molecule adsorbed on a gold substrate. Using geometric descriptors extracted from quantum chemistry simulations of thousands of ab initio molecular dynamics conformations, the ML protocol predicts vibrational frequencies and Raman intensities. The resulting spectra agree with density functional theory calculations and experiment. Predicted SERS responses of the molecule on different surfaces, or under external fields of electric fields and solvent environment, demonstrate the good transferability of the protocol.

Surface-enhanced Raman spectroscopy (SERS) is a powerful analytical tool for probing interfacial structures in situ at the molecular level.[1−5] Theoretical modeling is commonly used for interpreting SERS signals. Jensen and Schatz et al. have applied atomistic electrodynamics to study SERS response.[6,7] We have developed an interaction Hamiltonian model to predict SERS signals of molecules on surfaces.[8,9] However, because of elusive conformational variations of absorbed molecule and environmental fluctuations, one has to carry out quantum mechanical (QM) calculations for thousands of molecular dynamics (MD) conformations. A cost-effective approach to calculate SERS spectra with QM accuracy is desirable.

Machine learning (ML) is a family of statistics-based methods that can make predictions of properties of molecules and materials without invoking computationally demanding electronic structure calculations.[10] It has been applied to predict energy band gap,[11] potential energy surfaces,[12−17] molecular atomization energies,[18] dielectric response properties,[19−23] intrinsic bond energy,[24] and so on. A subclass of ML algorithms, known as random forest, can construct a multitude of decision trees for classification, regression, and other tasks.[25,26] Very recently, we have applied it to identify structural descriptors determining the electronic excitation of peptide bonds for predicting ultraviolet absorption spectra of proteins.[27]

In this work, we employ the random forest technique to predict the SERS signal of a *trans*-1,2-bis (4-pyridyl) ethylene (BPE) molecule (Figure 1a), which is widely used in SERS measurements because of its high-quality signals, high sensitivity, and good performance.[28,29] Based on iterative learning of QM data of electronic and vibrational structures for thousands of ab initio molecular dynamics (AIMD) conformations, the machine learning protocol (details are

Received: August 28, 2019
Accepted: September 20, 2019
Published: September 20, 2019

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026−6031

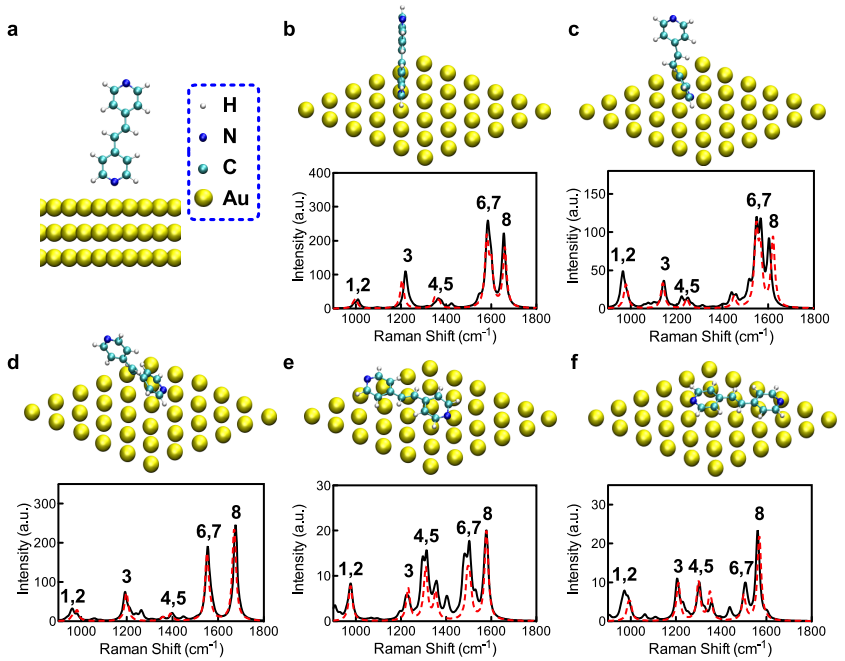


Figure 1. (a) Atomic model of BPE on Au(111) surface. The BPE/Au configuration taken at 0 (b), 1 ps (c), 2 ps (d), and 3 ps (e) of AIMD evolution and the stable structure (f) from DFT optimization, together with DFT-calculated (solid black line) and the ML-predicted SERS spectra (dashed red line).

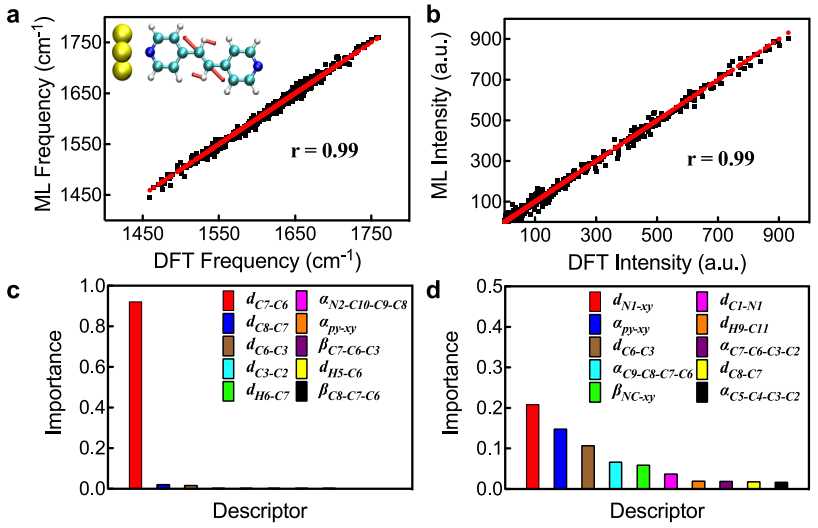


Figure 2. (a and b) Comparison of DFT-computed and ML-predicted vibrational frequencies and Raman intensities for mode 8 (shown in the inset of panel a). The Pearson correlation coefficient (r) of ML reflects the agreement. (c and d) Descriptor importance analysis of frequencies and Raman intensities.

provided in the Supporting Information) was applied. We then predicted the frequencies and Raman intensities of each dominant vibrational mode with its mode-specific descriptors. The final ML model was successfully applied to predict SERS of BPE in the presence of an explicit solvation environment, electric fields, and other surfaces, suggesting good transferability.

The conformational evolution of BPE adsorbed on the Au surface is simulated using AIMD implemented in the Vienna ab initio simulation package (VASP).[30,31] A unit cell (6 × 6) of Au(111) was selected to model the substrate. Starting with a vertical configuration where BPE perpendicularly adsorbed on the Au(111) surface in Figure 1b, the AIMD simulation at 300 K shows the evolution from the tilted BPE to the lying down configuration, as reflected by three configurations extracted at 1, 2, and 3 ps (Figure 1c–e). The equilibrium AIMD configuration at 3 ps agrees with the most stable configuration from DFT geometry optimization (Figure 1f).

To simulate the SERS of different interfacial structures, we used 4000 AIMD configurations with a 1 fs interval for QM calculations. The calculated root-mean-square deviation of BPE is 4.5 Å (Figure S1), indicating the big conformational changes and small correlation. The frequency analysis and Raman calculations are performed using the Gaussian 16 package[32] at the hybrid B3LYP functional level,[33] with 6-31G(d,p) and LANL2DA basis sets. In Figure 1b, eight vibrational modes dominate the computed SERS spectra because of the D_{2h} symmetry of the BPE molecule (spectra

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026−6031

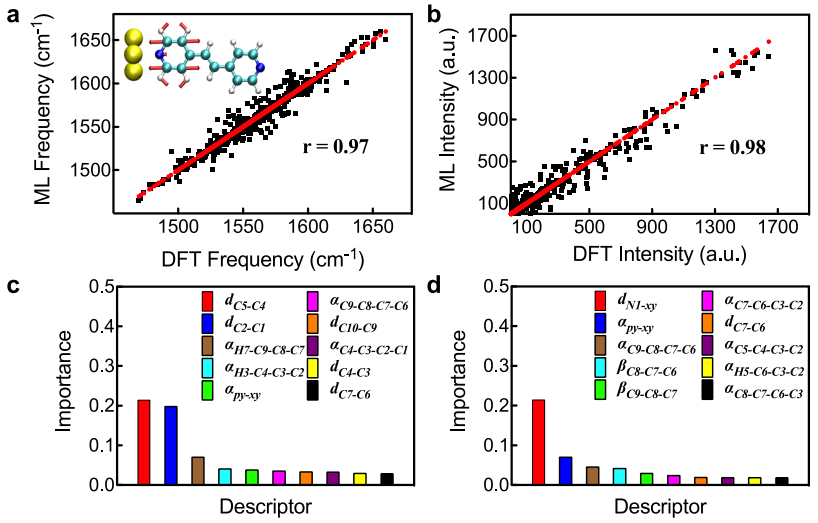


Figure 3. (a and b) Comparison of DFT-computed and ML-predicted vibrational frequencies and Raman intensities for mode 6 (shown in the inset of panel a). The Pearson correlation coefficient (r) of ML reflects the agreement. (c and d) Descriptor importance analysis of frequencies and Raman Intensities, respectively.

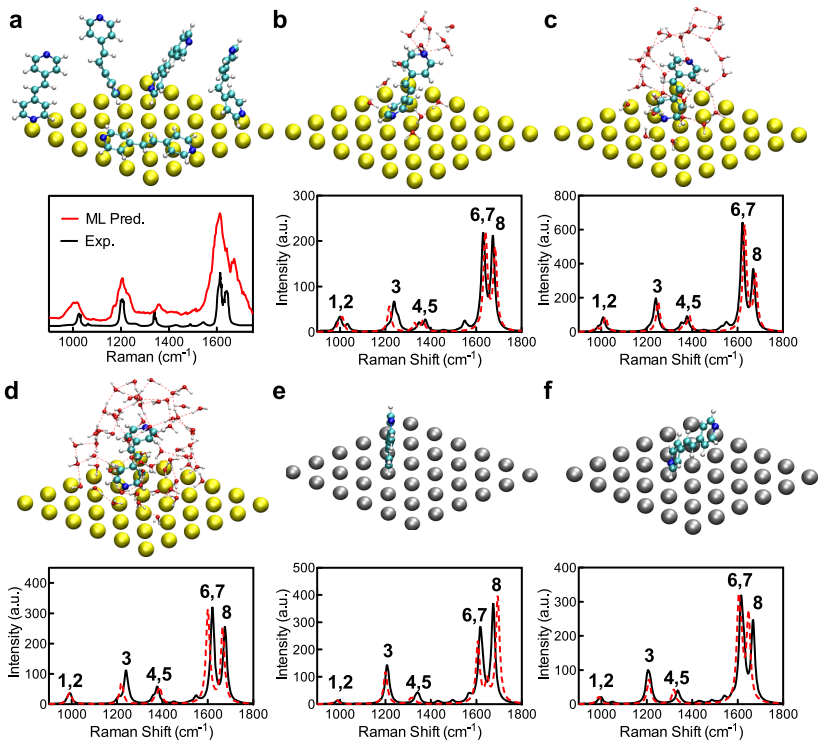


Figure 4. (a) Varying configurations of BPE on the Au(111) surface, based on which the ML-predicted (statistical result from 400 randomly selected testing configurations) SERS agrees with experiment. (b–d) Optimized structure and SERS spectra of BPE/Au system surrounded by 14, 32, and 64 water molecules. (e and f) Optimized structure and SERS spectra of BPE adsorbed on Ag and Pt surfaces.

were smoothed using a Lorentzian convolution of 20 cm^{-1}). These are assigned to the pyridyl ring breathing (modes 1 and 2), pyridyl ring twisting (mode 3), δ(C−H)py (modes 4 and 5), pyridyl ring stretching (modes 6 and 7) and ν(C=C) (mode 8) (Figure S2). The breaking symmetry of the BPE molecule results in three pairs of doubly degenerate vibrational modes (1 and 2, 4 and 5, and 6 and 7) (Figure S2).

During the structure evolution, the interfacial geometry varies greatly. After 1 ps, the dihedral angle between two pyridine rings becomes almost 90°, and almost all SERS peaks are red-shifted (Figure 1c) compared to the initial structure. Moreover, the splitting of modes 6 and 7 become more obvious. After 2 ps of evolution, the splitting of modes 6 and 7 vanishes (Figure 1d). The relative intensity of mode 8 with respect to modes 6 and 7 (I8/I6&7) is less than 1 in the beginning, which becomes larger than 1 after 2 ps. SERS spectra become more complicated when BPE lies on the surface (Figure 1e,f). Mode 3 shows a blue shift, while modes 6, 7, and 8 show red shifts.

Vibrational frequencies and the Raman intensities of these eight vibrational modes are the targets for ML training. The input variables (descriptors) contain the distance between BPE

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026−6031

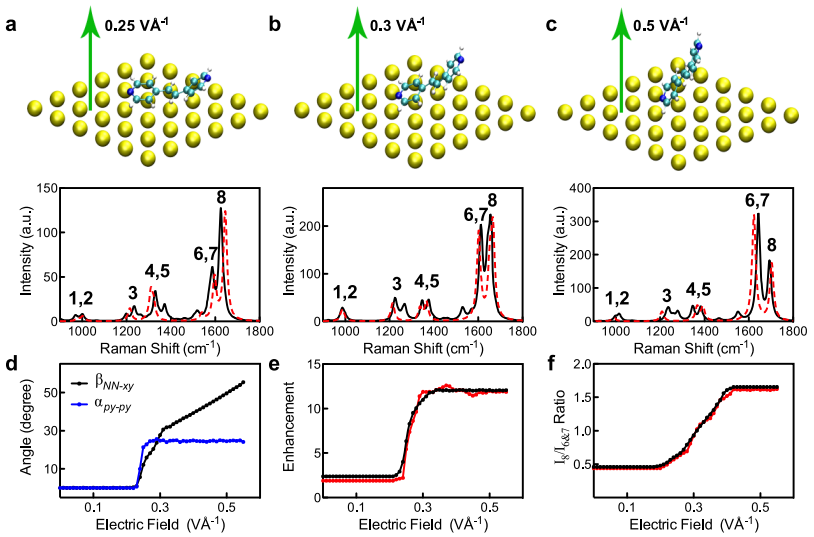


Figure 5. Optimized structure and SERS spectra of the BPE/Au system with external electric fields of (a) 0.25 V/Å, (b) 0.3 V/Å, and (c) 0.5 V/Å. (d) Variations of the dihedral angle between two pyridine (α_{py-py}) and the angle between BPE and the Au surface (β_{NN-xy}) subjected to different electric fields. (e and f) Comparison of the ML (red line) and DFT (black line) predicted SERS enhancement and the I8/I6&7 ratio under different electric fields.

and the gold surface d_{N1-xy}, the angle between BPE and gold β_{NC-xy}, dihedral between pyridine ring and gold α_{py-xy}, and the internal coordinates[34] of BPE (Figure S3). Ten key geometrical descriptors were selected by random forest for each vibrational mode (Table S1 and Figures S6 and S8) from initially 69 descriptors. Learning curves indicate that ML converges as samples exceed 3000 (Figure S5). The small autocorrelation functions along the trajectory also indicate the low-correlation of sample sets (Figure S7). The ML results after cross-validation for the frequency and Raman intensity of mode 8 are compared with DFT data in Figure 2a,b. The Pearson correlation coefficient (r), which measures the linear correlation between ML-predicted and DFT-calculatd values, was found to be 0.99 after cross-validation, demonstrating the prediction accuracy of ML. The ML treatment is obviously much more efficient by taking only 10^{-4} of the time of QM (Figure S9), significantly speeding up tedious conformation search for identification of the interfacial structures.

The random forest test reveals that the C_7-C_8 bond length is the dominant descriptor (Figure 2c) for predicting the frequency of mode 8. This can be explained by the fact that mode 8 is mainly the stretching vibration of central ethylene, as shown in the inset in Figure 2a. On the other hand, we find that there are several dominant descriptors for the Raman intensity. Both the relative position of the molecules with respect to the surface, such as d_{N1-xy}, α_{py-xy}, and β_{NC-xy}, and the molecular inner geometrical parameters, such as d_{C6-C3}, $\alpha_{C9-C8-C7-C6}$, and d_{C1-N1}, determine the Raman intensity (Figure 2d).

The ML prediction of mode 6 also achieves high accuracy (r = 0.97−0.98 in Figure 3a,b) after cross-validation, owing to its own structural descriptors (Figure 3c). Two bond lengths of d_{C5-C4} and d_{C2-C1} are equally important, together with several other descriptors with importance at around 5%. We attribute the diverse nature of important descriptors to the fact that mode 6 is more delocalized and contributed by d_{C5-C4} and d_{C2-C1} bond stretching and C−H in-plane bending. This is further verified by random forest analysis on modes 3 and 7 (Figures S10 and S11). For instance, the frequency of mode 7 is largely dependent on bond lengths of $d_{C12-C11}$ and d_{C10-C9}. Compared to the ML predicted frequencies, the Raman intensities depend on many additional descriptors (Figure 3d). However, the descriptors associated with the relative position of the adsorbate on the substrate become less important for modes 6 and 7 in comparison to mode 8.

We then extended the trained ML protocol to simulate SERS signals in different environments. We first show that our ML-predicted SERS based on 400 randomly selected AIMD structures can reproduce the previously reported experimental data[35] (Figure 4a), demonstrating the good accuracy of the ML model. The experiment focuses on a "hot spot" where only a few molecules are located, resulting in the narrower peaks than the ML prediction. The impact of the explicit solvent on the SERS signals was examined as well. An explicit model with 14, 32, and 64 water molecules was employed, with optimized geometries and DFT-computed SERS spectra shown in Figure 4b−d. Because the solvent molecules barely affect the HOMO−LUMO gaps and atomic charges of BPE (Table S2 and Figure S12), we did not include them as descriptors for ML training. We further applied the ML model to predict the SERS response of BPE on a couple of different metal surfaces (see Figure 4e,f for Ag(111) and Pt(111) and Figure S13 for Cu(111), Pd(111), Au(100) ,and Au(110)). All these comparisons show reasonable agreement between ML results and DFT results. The substrate and surface are crucial for SERS. The transferability of our ML model from one substrate to another is a substantial step toward building an ML protocol for various substrates and molecules.

Figure 5a−c illustrates the conformational and spectral changes induced by varying the electric field along the normal direction, based on QM DFT calculation of optimized structure and electronic states for BPE/Au under external bias. Three characteristic stages appear during the configuration evolution with increasing electric field (Figure 5d). In the first stage, the BPE molecule prefers to lie on the Au surface when the applied field is smaller than 0.23 V/Å. The second stage appears when the electric field increases from 0.23 to 0.30 V/Å, where one pyridine ring is lifted up (Figure

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026−6031

5a). This results in a dihedral angle between the two pyridine, i.e., α_{py-py} increases from 0° to a converged value of 25° (Figure 5d). In the third stage of the field above 0.30 V/Å, the α_{py-py} has nearly no change, while the β_{NN-xy} angle between BPE and the Au surface increases with increasing field.

Using the geometrical descriptors taken from QM-optimized structures, we successfully reproduced the SERS spectra with the previously trained ML protocol, as reflected by the ML SERS curves for BPE/Au subjected to 0.25, 0.30, and 0.50 V/Å fields in panels a, b, and c of Figure 5, respectively. We can further test the ML accuracy by examining the absolute and relative intensities of SERS peaks. The absolute intensity was represented by the variation curve of total SERS enhancement factor from 0.0 to 0.55 V/Å, which was reproduced by the ML predictions (Figure 5e). The relative intensity of mode 8 with respect to modes 6 and 7 (I8/I6&7) is examined in Figure 5f. Modes 6 and 7 are known to be very sensitive SERS fingerprints of the BPE/Au interface.[36,37] In Figure 5e,f, the ML-predicted dependence curves on electric fields are almost the same as the QM results. To be specific, there are also three stages for the SERS enhancement and I8/I6&7 ratio variation as electric field increases. These stages synchronize with the configuration evolution, except for the I8/I6&7 ratio increasing step ranging from 0.19 to 0.43 V/Å. In summary, both the absolute and relative intensities of the SERS spectra with electric filed and solvent environment can be predicted using the ML protocol trained with QM data sets without external fields, demonstrating good transferability.

In conclusion, an ML protocol based on a large data set of QM/AIMD calculations was developed for the SERS spectra of the BPE molecule. We analyzed the vibrational frequencies and the Raman intensities based solely on the interfacial geometrical information. ML predictions are accurate enough at a much lower cost than QM simulations. The ML structure–property relationships built on QM training sets without an external field were applied to predict experimental observations including effects of external fields. ML techniques based on quantum mechanical calculations should provide a cost-effective transferable tool for assigning experimental optical spectroscopy signals to molecular geometry.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.9b02517.

Computational details; root mean square deviation; eight most important vibrational modes; descriptors for predicting frequencies and Raman intensities; 10 descriptors in ML prediction; ML results for modes 6 and 8; learning curves for modes 6 and 8; heat map of the Pearson correlation coefficient among descriptors for modes 6 and 8; autocorrelation functions of the four most important descriptors for mode 8; comparison between mode projection and random forest importance analysis; partition of time consumption for computing Raman spectra for one structure; ML results for mode 3; ML results for mode 7; effect of solvent molecules and electric field on the atomic partial charges; effect of solvent molecules and electric field on the energy levels of BPE molecule; SERS spectra of BPE adsorbed on Cu(111), Pd(111), Au(100), and Au(110) surfaces (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jiangj1@ustc.edu.cn.

ORCID

Wei Hu: 0000-0002-7467-4783
Guozhen Zhang: 0000-0003-0125-9666
Shaul Mukamel: 0000-0002-6015-3135
Jun Jiang: 0000-0002-6116-5605

Author Contributions

⊥W.H., S.Y., and Y.Z. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (2018YFA0208603 and 2017YFA0303500), the National Natural Science Foundation of China (21703223, 11704209, 21633006, 21633007, and 21790350), the Anhui Initiative in Quantum Information Technologies (AHY090200), the National Science Foundation Grant CHE-1663822, and Taishan Scholar Program of Shandong Province.

■ REFERENCES

(1) Benz, F.; Schmidt, M.; Dreismann, A.; Chikkaraddy, R.; Zhang, Y.; Demetriadou, A.; Carnegie, C.; Ohadi, H.; de Nijs, B.; Esteban, R.; et al. Single-Molecule Optomechanics in "Picocavities. *Science* **2016**, *354*, 726–729.

(2) Zhang, Y.; Luo, Y.; Zhang, Y.; Yu, Y.-J.; Kuang, Y.-M.; Zhang, L.; Meng, Q.-S.; Luo, Y.; Yang, J.-L.; Dong, Z.-C.; et al. Visualizing Coherent Intermolecular Dipole–Dipole Coupling in Real Space. *Nature* **2016**, *531*, 623.

(3) Hackler, R.-A.; Kang, G.; Schatz, G.-C.; Stair, P.-C.; Van Duyne, R.-P. Analysis of Tio2 Atomic Layer Deposition Surface Chemistry and Evidence of Propene Oligomerization Using Surface-Enhanced Raman Spectroscopy. *J. Am. Chem. Soc.* **2019**, *141*, 414–422.

(4) Li, J.-F.; Huang, Y.-F.; Ding, Y.; Yang, Z.-L.; Li, S.-B.; Zhou, X.-S.; Fan, F.-R.; Zhang, W.; Zhou, Z.-Y.; Ren, B.; et al. Shell-Isolated Nanoparticle-Enhanced Raman Spectroscopy. *Nature* **2010**, *464*, 392.

(5) Hackler, R.-A.; McAnally, M.-O.; Schatz, G.-C.; Stair, P.-C.; Van Duyne, R.-P. Identification of Dimeric Methylalumina Surface Species During Atomic Layer Deposition Using Operando Surface-Enhanced Raman Spectroscopy. *J. Am. Chem. Soc.* **2017**, *139*, 2456–2463.

(6) Liu, P.-C.; Chulhai, D.-V.; Jensen, L. Single-Molecule Imaging Using Atomistic near-Field Tip-Enhanced Raman Spectroscopy. *ACS Nano* **2017**, *11*, 5094–5102.

(7) Jensen, L.; Aikens, C.-M.; Schatz, G.-C. Electronic Structure Methods for Studying Surface-Enhanced Raman Scattering. *Chem. Soc. Rev.* **2008**, *37*, 1061–1073.

(8) Duan, S.; Tian, G.-J.; Luo, Y. Visualization of Vibrational Modes in Real Space by Tip-Enhanced Non-Resonant Raman Spectroscopy. *Angew. Chem.* **2016**, *128*, 1053–1057.

(9) Duan, S.; Tian, G.-J.; Ji, Y.-F.; Shao, J.-S.; Dong, Z.-C.; Luo, Y. Theoretical Modeling of Plasmon-Enhanced Raman Images of a Single Molecule with Subnanometer Resolution. *J. Am. Chem. Soc.* **2015**, *137*, 9515–9518.

(10) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O.-A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

(11) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction Model of Band Gap for Inorganic Compounds by

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026–6031

Combination of Density Functional Theory Calculations and Machine Learning Techniques. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *93*, 115104.
(12) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73.
(13) Toyoura, K.; Hirano, D.; Seko, A.; Shiga, M.; Kuwabara, A.; Karasuyama, M.; Shitara, K.; Takeuchi, I. Machine-Learning-Based Selective Sampling Procedure for Identifying the Low-Energy Region in a Potential Energy Surface: A Case Study on Proton Conduction in Oxides. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *93*, 054112.
(14) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(15) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(16) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 094203.
(17) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978−10985.
(18) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O.-A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(19) Ma, H.; Saha, T. K.; Ekanayake, C. Statistical Learning Techniques and Their Applications for Condition Assessment of Power Transformer. *IEEE Trans. Dielectr. Electr. Insul.* **2012**, *19*, 481−489.
(20) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data* **2016**, *3*, 160012.
(21) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R. Critical Assessment of Regression-Based Machine Learning Methods for Polymer Dielectrics. *Comput. Mater. Sci.* **2016**, *125*, 123−135.
(22) Wu, K.; Sukumar, N.; Lanzillo, N. A.; Wang, C.; Ramprasad, R. R.; Ma, R.; Baldwin, A. F.; Sotzing, G.; Breneman, C. Prediction of Polymer Properties Using Infinite Chain Descriptors (Icd) and Machine Learning: Toward Optimized Dielectric Polymeric Materials. *J. Polym. Sci., Part B: Polym. Phys.* **2016**, *54*, 2082−2091.
(23) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.
(24) Yao, K.; Herr, J.-E.; Brown, S.-N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689−2694.
(25) Harmancı, A.-S.; Youngblood, M.-W.; Clark, V.-E.; Coşkun, S.; Henegariu, O.; Duran, D.; Erson-Omay, E.-Z.; Kaulen, L.-D.; Lee, T.-I.; Abraham, B.-J.; et al. Integrated Genomic Analyses of De Novo Pathways Underlying Atypical Meningiomas. *Nat. Commun.* **2017**, *8*, 14433.
(26) Capper, D.; Jones, D. T. W.; Sill, M.; Hovestadt, V.; Schrimpf, D.; Sturm, D.; Koelsche, C.; Sahm, F.; Chavez, L.; Reuss, D. E.; et al. DNA Methylation-Based Classification of Central Nervous System Tumours. *Nature* **2018**, *555*, 469.
(27) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A Neural Network Protocol for Electronic Excitations of N-Methylacetamide. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11612−11617.
(28) Hu, Y.-S.; Jeon, J.; Seok, T.-J.; Lee, S.; Hafner, J.-H.; Drezek, R.-A.; Choo, H. Enhanced Raman Scattering from Nanoparticle-Decorated Nanocone Substrates: A Practical Approach to Harness in-Plane Excitation. *ACS Nano* **2010**, *4*, 5721−5730.
(29) Freeman, R.-G.; Grabar, K.-C.; Allison, K.-J.; Bright, R.-M.; Davis, J.-A.; Guthrie, A.-P.; Hommer, M.-B.; Jackson, M.-A.; Smith, P.-C.; Walter, D.-G.; et al. Self-Assembled Metal Colloid Monolayers: An Approach to Sers Substrates. *Science* **1995**, *267*, 1629−1632.
(30) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169.
(31) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(32) Frisch, M.-J.; Trucks, G.-W.; Schlegel, H.-B.; Scuseria, G.-E.; Robb, M.-A.; Cheeseman, J.-R.; Scalmani, G.; Barone, V.; Petersson, G.-A.; Nakatsuji, H. *Gaussian 16*, revision A; 2016; 3.
(33) Stephens, P.-J.; Devlin, F.-J.; Chabalowski, C.-F.; Frisch, M.-J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(34) Panapitiya, G.; Avendaño-Franco, G.; Ren, P.; Wen, X.; Li, Y.; Lewis, J. P. Machine-Learning Prediction of Co Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140*, 17508−17514.
(35) Sprague-Klein, E.-A.; Negru, B.; Madison, L.-R.; Coste, S.-C.; Rugg, B.-K.; Felts, A.-M.; McAnally, M.-O.; Banik, M.; Apkarian, V.-A.; Wasielewski, M.-R.; et al. Photoinduced Plasmon-Driven Chemistry in Trans-1,2-Bis(4-Pyridyl)Ethylene Gold Nanosphere Oligomers. *J. Am. Chem. Soc.* **2018**, *140*, 10583−10592.
(36) Hu, W.; Tian, G.; Duan, S.; Lin, L.-L.; Ma, Y.; Luo, Y. Vibrational Identification for Conformations of Trans-1,2-Bis (4-Pyridyl) Ethylene in Gold Molecular Junctions. *Chem. Phys.* **2015**, *453*−*454*, 20−25.
(37) Hu, M.; Ou, F. S.; Wu, W.; Naumov, I.; Li, X.; Bratkovsky, A. M.; Williams, R. S.; Li, Z. Gold Nanofingers for Molecule Trapping and Detection. *J. Am. Chem. Soc.* **2010**, *132*, 12820−12822.

DOI: 10.1021/acs.jpclett.9b02517
J. Phys. Chem. Lett. 2019, 10, 6026−6031

Supporting Information

Machine Learning Protocol for Surface Enhanced Raman Spectroscopy

Wei Hu[1,2]†, Sheng Ye[2,†], Yujin Zhang[3,†], Tianduo Li[1], Guozhen Zhang[2], Yi Luo[2], Shaul Mukamel[4], Jun Jiang[2,*]

*Corresponding author. E-mail: jiangj1@ustc.edu.cn

Table of Contents

目录

1. Shandong Provincial Key Laboratory of Molecular Engineering, School of Chemistry and Pharmaceutical Engineering, Qilu University of Technology, Jinan, Shandong 250353, P. R. China. 2. Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China. 3. School of Electronic and Information Engineering (Department of Physics), Qilu University of Technology, Jinan, Shandong 250353, P. R. China. 4. Departments of Chemistry, and physics and astronomy, University of California, Irvine, CA 92697, USA. † These authors contributed equally to this work.

Computational Details

Molecular Dynamics Simulations

Molecular dynamics Simulations with 1 fs time step at temperatures of 300K were performed using *ab-initio* molecular dynamics (AIMD) implemented in Vienna *ab-initio* simulation package (VASP).[1] The Au surfaces were modeled with five atomic layers of (6 × 6) unit cell at the equilibrium lattice constant of 4.078 Å plus a 30 Å thick vacuum region. During the AIMD evolution, the three outermost layers of the junctions were fixed to mimic Au bulk, while all the other atoms were allowed to relax at all dimensions. For the k-point sampling, only the gamma point was adopted. The exchange-correlation effects were described by the Perdew−Burke−Ernzerhof generalized-gradient approximation (GGA-PBE)[2] with a plane-wave basis cutoff of 400 eV.[3] Meanwhile, the vdW-D3 method was employed to consider the dispersion correction.

Geometrical optimization shows that the flat configuration is the most stable one for the BPE/Au system. However, if the molecular dynamics simulation is started with flat configuration, the molecule will be most likely to stay in that orientation on the surface, resulting in narrow and biased sampling of conformations. The lack of representation of various conformations of BPE on the surface will undermine the prediction power of ML approach. We thus chose a vertical configuration as the initial condition of MD evolutions so as to achieve a diversified sampling of structures. After the trajectory was generated, 4000 configurations were extracted with a 1 fs interval. All the structures were used for including quantum chemistry calculations and machine learning. To evaluate the correlation of the configurations, we calculated the root mean square deviation (RMSD) of the adsorbed molecule. The RMSD was found to be 4.5 Å (Figure S1), indicating the big conformational changes and small correlation.

Raman Calculations

After AIMD simulation, a cluster model with the first slab layer of the supercell and the adorbate was taken out to perform Raman spectra calculations. For further calculations, all the atoms of the substrates were fixed. Analytical frequency and Raman calculations were performed at hybrid B3LYP functional with 6-31G(d,p) basis set[4] for the main elements and pseudo LANL2DA basis set[5-6] for Au atoms. All calculations were performed using Gaussian 09 program.[7] Then the Raman cross section was calculated by[8]

$$\left(\frac{d\alpha}{d\Omega}\right)_k = \frac{\pi^2}{\varepsilon_0^2}(\tilde{\nu}_{in} - \tilde{\nu}_k)^4 \frac{h}{8\pi^2 c\tilde{\nu}_k}\left(\frac{45\alpha_k'^2 + 7\gamma_k'^2}{45}\right)\frac{1}{1 - exp(-hc\tilde{\nu}_k/k_BT)}$$

Electronic Structure and Atomic Partial Charge Calculations

We calculated the electronic structures and atomic partial charges of BPE to investigate the environmental effects. We extracted the geometry of BPE alone from the optimized BPE/Au system surrounded by 14 explicit water molecules. Then we performed natural orbital analysis to obtain the information of the frontier orbitals and atomic partial charges. The results were then compared to the ones with 14 water molecules. The same procedure was also performed to study the effect of the external electric field. In this part, we used different functionals (B3LYP and wB97XD) and basis set (6-31G* and 6-31+G*) to avoid the method dependence.

As shown in Table S2 and Figure S12, we can see that both the HOMO-LUMO gaps and the atomic partial charges are barely changed, indicating that removing the solvent and electric field from the pool of descriptors is a reasonable approximation. Therefore, we can confirm that the SERS responses are not very sensitive to the descriptors directly related to solvent positions or electric field intensities, while the inclusion of BPE geometric descriptors are sufficient to indirectly reflect the presence of solvent and field.

The Machine Learning Protocol

Descriptors and ML targets. It noted that because of the D2h symmetry of BPE molecule, only 8 vibrational modes are relevant to the SERS spectra. As a result, we focus only on the 8 modes in the ML prediction to simplify the investigation (Figure S2). To establish a structure-spectra relationship, the relative position of BPE with respect to surface (including the distance between BPE and the gold surface $d_{N1\text{-}xy}$, the angle between BPE and gold $\beta_{NC\text{-}xy}$, and dihedral between pyridine ring and gold $\alpha_{py\text{-}xy}$) and internal coordinates were selected as descriptors for ML (Figure S3). To be specific, $d_{N1\text{-}xy}$ is calculated as

$$z_{N1} - z_{substrate}$$

$\beta_{NC\text{-}xy}$ as

$$\operatorname{atan}\left[\frac{(z_{C3} - z_{N1})}{\sqrt{(x_{C3} - x_{N1})^2 + (y_{C3} - y_{N1})^2}}\right]$$

and $\alpha_{py\text{-}xy}$ as

$$\operatorname{acos}\left(\frac{c}{\sqrt{a^2 + b^2 + c^2}}\right)$$

where $a = (y_{C3} - y_{N1})(z_{C1} - z_{N1}) - (z_{C3} - z_{N1})(y_{C1} - y_{N1})$

$$b = (z_{C3} - z_{N1})(x_{C1} - x_{N1}) - (z_{C1} - z_{N1})(x_{C3} - x_{N1})$$

$$c = (x_{C3} - x_{N1})(y_{C1} - y_{N1}) - (x_{C1} - x_{N1})(y_{C3} - y_{N1})$$

To eliminate of different range of input values which may undermine the robustness final ML model and speed up the training of random forests, we have normalized the input data by transformed with

$$m' = \frac{(m_i - m_{min})}{(m_{max} - m_{min})}$$

where m_i are the input data of random forest, and $m_{min}$and m_{max} are minimum and maximum values of the input data, respectively. m' are the normalized data.

Training and testing set. The random forest algorithm composed of 300 decision trees with 3 depths was used to predict frequencies and Raman intensities of system. A total number of 4000 data were randomly divided into two parts: 3600 were used for training and the rest (400) were used for testing with scikit-learn frame.[9] We have also used the first 80% of the time series data as train set and the other 20% as test set to verify the correlation of descriptors. We note that the Pearson correlation coefficient (r) between the DFT calculation and ML prediction are still as high as 0.95 (Figure S4). In addition, the final ML model can always reproduce the good results.

Self-correlation calculations. For the self-correlation for each of the features, we have calculated the Pearson correlation coefficient (r) between each descriptor. The results shows that most features have low linear correlations (Figure S6), which significantly improve the performance of final ML model prediction.[10]

Auto-correlation function calculations. The autocorrelation function of the descriptors along the trajectory has been calculated to evaluate the correlation of the descriptors between adjacent frames. Figure S7 shows the autocorrelation functions of the four most important descriptors for the frequency (a-d) and Raman intensity (e-h) respectively for the mode 8. For most descriptors, the autocorrelation functions decay to zero very quickly and oscillate near zero. The small autocorrelation functions indicate the low-correlation of the descriptors between adjacent frames.

Cross-validation. The accuracy and robustness of the final machine learning results was verified by the cross-validation technique[11]. A total number of 4000 sets of data were randomly and evenly

distributed into 10 bins in this procedure. Each bin (400) was used as a test set while the remaining nine bins (3600) as training set.

Learning curves. For the ML training of the BPE/Au system, we have calculated the learning curves of whole process. The mean absolute error approaches to convergence as the size of sample in ML training exceeds 3000, as Figure S5 suggest.

Importance analysis. Random forest is a machine learning algorithm frequently used for analyzing the importance of each feature.[9] As each tree evolves, predictions are made based on the Out-Of-Bag (OOB) data for that tree. At the same time, each descriptor in the OOB data is randomly permuted, one at a time, and each such modified data set is also predicted by the same tree. At the end of the model training process, the margins for each sample are calculated based on the OOB prediction as well as the OOB predictions with each descriptor permuted. Let M be the average margin based on the OOB prediction and M_j the average margin based on the OOB prediction with the j-*th* descriptor permuted. The difference between M and M_j ($M-M_j$) reflects the importance for the j-*th* descriptor. For regression problems, addressed here.

Extrapolation of the ML model. To demonstrate the extrapolation of the ML model, we studied the effects of explicit solvents, electric field, metal surfaces (including crystal faces) on the SERS of BPE. For the solvent effects, we added 14, 32 and 64 water molecules in the BPE/Au(111) system and optimized the structures (Figure 4b-d). Based on the optimized structures, we fixed Au atoms and water molecules and relaxed BPE molecule alone to perform the Raman calculations. For the electric field effect, we applied an electric field of 0~0.50 V/Å on the BPE/Au(111) system and perform the geometrical optimization and Raman calculations (Figure 5a-c). To study the SERS spectra of BPE on different metals surfaces, we built Ag(111), Pt(111), Cu(111) and Pd(111) with lattice constants of 4.086, 3.924, 3.615 and 3.891 Å (Figure 4e-f and Figure S13a-b). Furthermore, we also constructed Au(100) and Au(110) surface to study the effect of the crystal faces on the SERS simulations (Figure S13c-d). Considering the substrate/surface is very crucial for SERS, the transferability of our ML model from one substrate to others is a big step toward building a ML protocol for various substrates and molecules.

In the simulation of SERS using the trained ML model, we have chosen the inner coordinates of BPE and its relative position with respect to metal surface as the descriptors. Good agreement has been obtained between the DFT calculation and ML prediction (Figure 4b-f, Figure 5a-c and Figure S13a-d), indicating the transferability of the ML model from simple system to complicated systems. Over all, the ML model trained from gas phase structural data is able to predict Raman spectra of BPE under versatile environments.

The Root Mean Square Deviation (RMSD)

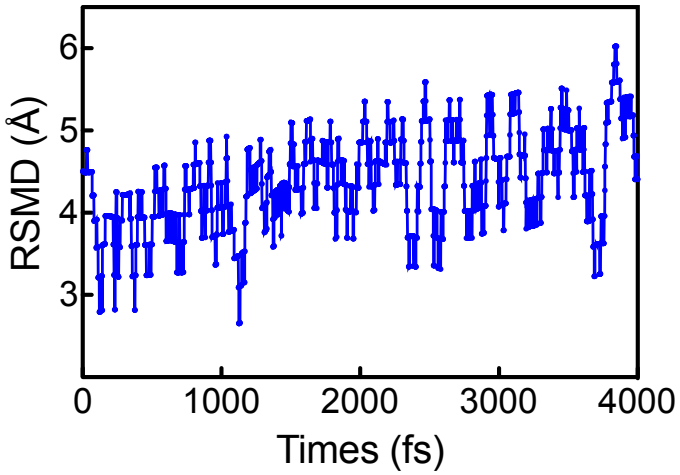


Figure S1. The root mean square deviation (RMSD) of the BPE molecule adsorbed on Au surfaces.

The Eight Most Important Vibrational Modes

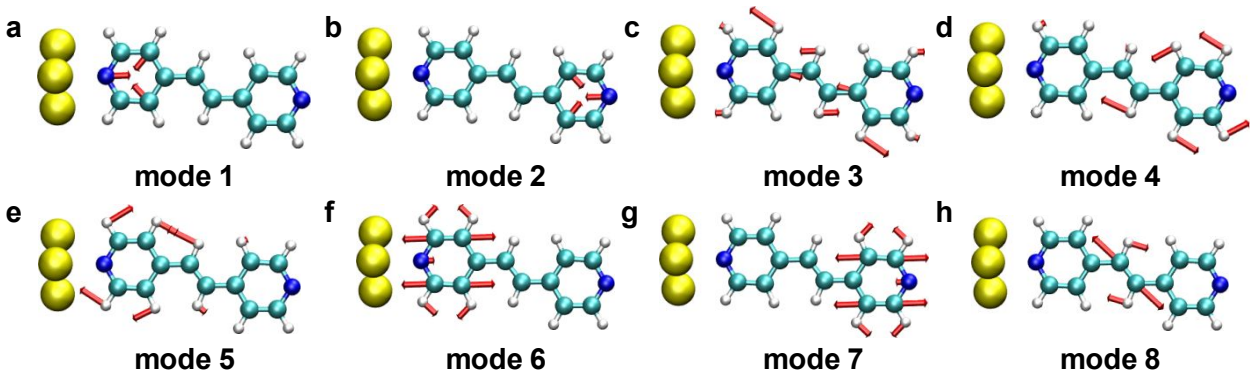
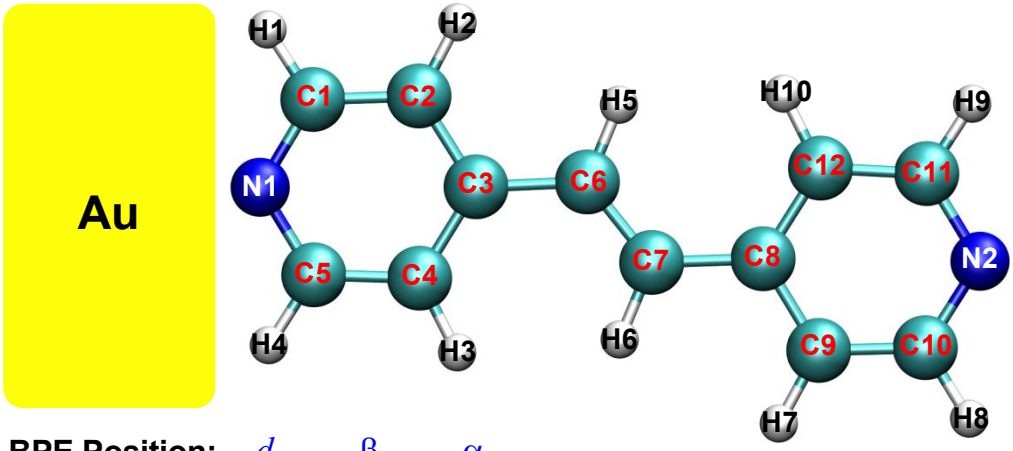


Figure S2. The 8 important vibrational modes that the present random forest method attempt to reproduce.

Descriptors for Predicting Frequencies and Raman Intensities



BPE Position: $d_{N1\text{-}xy}$, $\beta_{NC\text{-}xy}$, $\alpha_{py\text{-}xy}$

Bonds: $d_{C1\text{-}N1}$, $d_{C2\text{-}C1}$, $d_{C3\text{-}C2}$, $d_{C4\text{-}C3}$, $d_{C5\text{-}C4}$, $d_{C6\text{-}C3}$, $d_{C7\text{-}C6}$, $d_{C8\text{-}C7}$, $d_{C9\text{-}C8}$, $d_{C10\text{-}C9}$, $d_{N2\text{-}C10}$, $d_{C11\text{-}N2}$, $d_{C12\text{-}C11}$, $d_{H1\text{-}C1}$, $d_{H2\text{-}C2}$, $d_{H3\text{-}C4}$, $d_{H4\text{-}C5}$, $d_{H5\text{-}C6}$, $d_{H6\text{-}C7}$, $d_{H7\text{-}C9}$, $d_{H8\text{-}C10}$, $d_{H9\text{-}C11}$, $d_{H10\text{-}C12}$,

Angles: $\beta_{C2\text{-}C1\text{-}N1}$, $\beta_{C3\text{-}C2\text{-}C1}$, $\beta_{C4\text{-}C3\text{-}C2}$, $\beta_{C5\text{-}C4\text{-}C3}$, $\beta_{C6\text{-}C3\text{-}C2}$, $\beta_{C7\text{-}C6\text{-}C3}$, $\beta_{C8\text{-}C7\text{-}C6}$, $\beta_{C9\text{-}C8\text{-}C7}$, $\beta_{C10\text{-}C9\text{-}C8}$, $\beta_{N2\text{-}C10\text{-}C9}$, $\beta_{C11\text{-}N2\text{-}C10}$, $\beta_{C12\text{-}C11\text{-}N2}$, $\beta_{H1\text{-}C1\text{-}N1}$, $\beta_{H2\text{-}C2\text{-}C1}$, $\beta_{H3\text{-}C4\text{-}C3}$, $\beta_{H4\text{-}C5\text{-}C4}$, $\beta_{H5\text{-}C6\text{-}C3}$, $\beta_{H6\text{-}C7\text{-}C6}$, $\beta_{H7\text{-}C9\text{-}C8}$, $\beta_{H8\text{-}C10\text{-}C9}$, $\beta_{H9\text{-}C11\text{-}N2}$, $\beta_{H10\text{-}C12\text{-}C11}$

Bonds: $\alpha_{C3\text{-}C2\text{-}C1\text{-}N1}$, $\alpha_{C4\text{-}C3\text{-}C2\text{-}C1}$, $\alpha_{C5\text{-}C4\text{-}C3\text{-}C2}$, $\alpha_{C6\text{-}C3\text{-}C2\text{-}C1}$, $\alpha_{C7\text{-}C6\text{-}C3\text{-}C2}$, $\alpha_{C8\text{-}C7\text{-}C6\text{-}C3}$, $\alpha_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\alpha_{C10\text{-}C9\text{-}C8\text{-}C7}$, $\alpha_{N2\text{-}C10\text{-}C9\text{-}C8}$, $\alpha_{C11\text{-}N2\text{-}C10\text{-}C9}$, $\alpha_{C12\text{-}C11\text{-}N2\text{-}C10}$, $\alpha_{H1\text{-}C1\text{-}N1\text{-}C2}$, $\alpha_{H2\text{-}C2\text{-}C1\text{-}N1}$, $\alpha_{H3\text{-}C4\text{-}C3\text{-}C1}$, $\alpha_{H4\text{-}C5\text{-}C4\text{-}C3}$, $\alpha_{H5\text{-}C6\text{-}C3\text{-}C2}$, $\alpha_{H6\text{-}C7\text{-}C6\text{-}C3}$, $\alpha_{H7\text{-}C9\text{-}C8\text{-}C7}$, $\alpha_{H8\text{-}C10\text{-}C9\text{-}C8}$, $\alpha_{H9\text{-}C11\text{-}N2\text{-}C10}$, $\alpha_{H10\text{-}C12\text{-}C11\text{-}N2}$,

Figure S3. Descriptors used for predicting frequencies and Raman intensities.

Ten Descriptors in ML Predition

Table S1. Ten Descriptors used in ML to predict the frequency and intensity of the 8 most important vibrational modes

Mode		Descriptors
1	Frequency	$\boldsymbol{\beta}_{NC\text{-}xy}$, $d_{C1\text{-}N1}$, $d_{H1\text{-}C1}$, $d_{C4\text{-}C3}$, $d_{C2\text{-}C1}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $d_{H4\text{-}C5}$, $d_{N1\text{-}xy}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\beta}_{C5\text{-}C4\text{-}C3}$
	Intensity	$d_{N1\text{-}xy}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $\boldsymbol{\beta}_{C9\text{-}C8\text{-}C7}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $d_{C5\text{-}C4}$, $d_{C3\text{-}C2}$, $d_{C4\text{-}C3}$, $\boldsymbol{\alpha}_{C5\text{-}C4\text{-}C3\text{-}C2}$
2	Frequency	$d_{C11\text{-}N2}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $d_{C9\text{-}C8}$, $d_{N2\text{-}C10}$, $\boldsymbol{\alpha}_{N2\text{-}C10\text{-}C9\text{-}C8}$, $d_{C12\text{-}C11}$, $d_{C7\text{-}C6}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $d_{C10\text{-}C9}$, $d_{H9\text{-}C11}$
	Intensity	$\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{H9\text{-}C11}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $d_{N1\text{-}xy}$, $\boldsymbol{\beta}_{C9\text{-}C8\text{-}C7}$, $\boldsymbol{\beta}_{C6\text{-}C3\text{-}C2}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $\boldsymbol{\alpha}_{C5\text{-}C4\text{-}C3\text{-}C2}$, $\boldsymbol{\alpha}_{N2\text{-}C10\text{-}C9\text{-}C8}$
3	Frequency	$\boldsymbol{\alpha}_{py\text{-}xy}$, $d_{C8\text{-}C7}$, $d_{C2\text{-}C1}$, $d_{C6\text{-}C3}$, $\boldsymbol{\alpha}_{C12\text{-}C11\text{-}N2\text{-}C10}$, $\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $\boldsymbol{\alpha}_{H4\text{-}C5\text{-}C4\text{-}C3}$, $\boldsymbol{\beta}_{C6\text{-}C3\text{-}C2}$, $\boldsymbol{\beta}_{C10\text{-}C9\text{-}C8}$, $\boldsymbol{\beta}_{C3\text{-}C2\text{-}C1}$
	Intensity	$\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{N1\text{-}xy}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $\boldsymbol{\alpha}_{C4\text{-}C3\text{-}C2\text{-}C1}$, $\boldsymbol{\beta}_{N2\text{-}C10\text{-}C9}$, $\boldsymbol{\beta}_{C2\text{-}C1\text{-}N1}$, $d_{C8\text{-}C7}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $d_{H9\text{-}C11}$, $\boldsymbol{\alpha}_{H1\text{-}C1\text{-}N2\text{-}C2}$
4	Frequency	$d_{C7\text{-}C6}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $d_{C6\text{-}C3}$, $d_{C12\text{-}C11}$, $\boldsymbol{\beta}_{C5\text{-}C4\text{-}C3}$, $\boldsymbol{\beta}_{C6\text{-}C3\text{-}C2}$, $\boldsymbol{\beta}_{C7\text{-}C6\text{-}C3}$, $d_{C4\text{-}C3}$, $d_{H10\text{-}C12}$, $\boldsymbol{\beta}_{C3\text{-}C2\text{-}C1}$
	Intensity	$\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{C7\text{-}C6}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\alpha}_{C5\text{-}C4\text{-}C3\text{-}C2}$, $d_{H9\text{-}C11}$, $d_{H10\text{-}C12}$, $d_{C2\text{-}C1}$, $d_{H8\text{-}C10}$, $d_{H7\text{-}C9}$, $d_{N1\text{-}xy}$
5	Frequency	$d_{C7\text{-}C6}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $d_{C10\text{-}C9}$, $d_{H1\text{-}C1}$, $d_{H3\text{-}C4}$, $\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $\boldsymbol{\alpha}_{C11\text{-}N2\text{-}C10\text{-}C9}$, $d_{C3\text{-}C2}$, $d_{C2\text{-}C1}$, $d_{C9\text{-}C8}$
	Intensity	$\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{C7\text{-}C6}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $d_{H1\text{-}C1}$, $d_{H3\text{-}C4}$, $d_{N1\text{-}xy}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $\boldsymbol{\alpha}_{C3\text{-}C2\text{-}C1\text{-}N1}$, $\boldsymbol{\alpha}_{C4\text{-}C3\text{-}C2\text{-}C1}$, $d_{H5\text{-}C4}$
6	Frequency	$d_{C5\text{-}C4}$, $d_{C2\text{-}C1}$, $\boldsymbol{\alpha}_{H7\text{-}C9\text{-}C8\text{-}C7}$, $\boldsymbol{\alpha}_{H3\text{-}C4\text{-}C2\text{-}C3}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $d_{C10\text{-}C9}$, $\boldsymbol{\alpha}_{C4\text{-}C3\text{-}C2\text{-}C1}$, $d_{C4\text{-}C3}$, $d_{C7\text{-}C6}$
	Intensity	$d_{N1\text{-}xy}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$, $\boldsymbol{\beta}_{C9\text{-}C8\text{-}C7}$, $\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{C7\text{-}C6}$, $\boldsymbol{\alpha}_{C5\text{-}C4\text{-}C3\text{-}C2}$, $\boldsymbol{\alpha}_{H5\text{-}C6\text{-}C3\text{-}C2}$, $\boldsymbol{\alpha}_{C8\text{-}C7\text{-}C6\text{-}C3}$
7	Frequency	$d_{C12\text{-}C11}$, $d_{C10\text{-}C9}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $d_{C7\text{-}C6}$, $d_{C9\text{-}C8}$, $d_{C8\text{-}C7}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $d_{C3\text{-}C2}$, $d_{C4\text{-}C3}$, $d_{N1\text{-}xy}$
	Intensity	$d_{N1\text{-}xy}$, $d_{C6\text{-}C3}$, $\boldsymbol{\beta}_{C5\text{-}C4\text{-}C3}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\beta}_{H6\text{-}C7\text{-}C6}$, $d_{C8\text{-}C7}$, $\boldsymbol{\beta}_{C9\text{-}C8\text{-}C7}$, $\boldsymbol{\beta}_{C4\text{-}C3\text{-}C2}$, $d_{C11\text{-}N2}$, $d_{C12\text{-}C11}$
8	Frequency	$d_{C7\text{-}C6}$, $d_{C8\text{-}C7}$, $d_{C6\text{-}C3}$, $d_{C3\text{-}C2}$, $d_{H6\text{-}C7}$, $\boldsymbol{\alpha}_{N2\text{-}C10\text{-}C9\text{-}C8}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $\boldsymbol{\beta}_{C7\text{-}C6\text{-}C3}$, $d_{H5\text{-}C6}$, $\boldsymbol{\beta}_{C8\text{-}C7\text{-}C6}$
	Intensity	$d_{N1\text{-}xy}$, $\boldsymbol{\alpha}_{py\text{-}xy}$, $d_{C6\text{-}C3}$, $\boldsymbol{\alpha}_{C9\text{-}C8\text{-}C7\text{-}C6}$, $\boldsymbol{\beta}_{NC\text{-}xy}$, $d_{C1\text{-}N1}$, $d_{H9\text{-}C3}$, $\boldsymbol{\alpha}_{C7\text{-}C6\text{-}C3\text{-}C2}$, $d_{C8\text{-}C7}$, $\boldsymbol{\alpha}_{C5\text{-}C4\text{-}C3\text{-}C2}$

ML Results for Mode 6 and 8 (First 80% as Train and the Left 20% as Test)

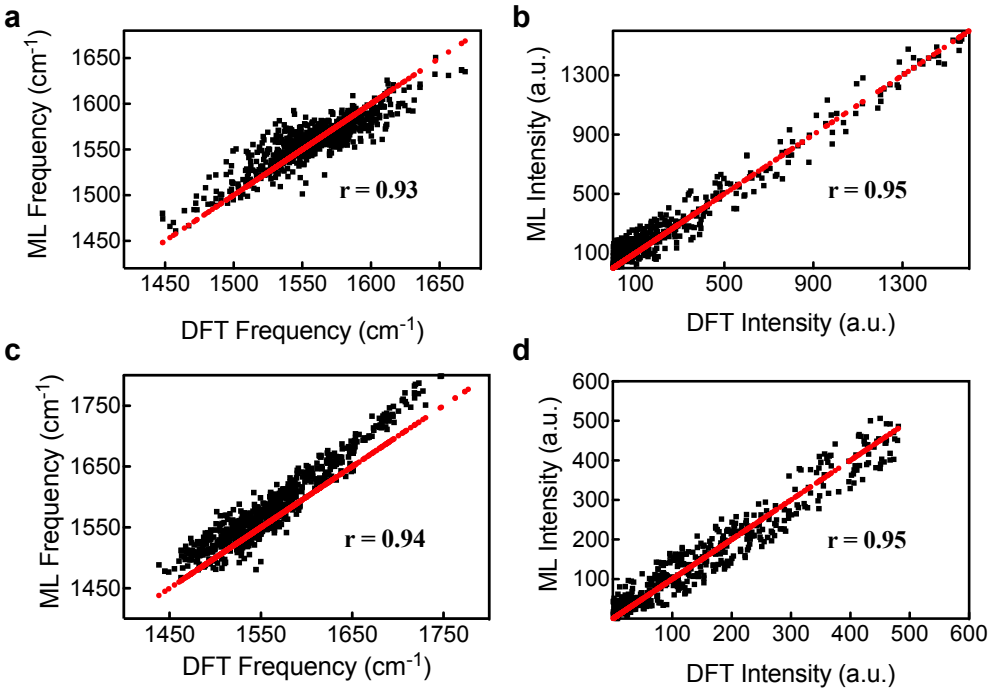


Figure S4. a,b, Comparison of DFT-computed and ML-predicted (first 80% of the time series data as train sets and the other 20% as test sets) normal vibrational frequencies and Raman intensities for mode 6. Pearson correlation coefficient (*r*) of ML reflects the agreement. **c,d,** Same with **a,b** but for Mode 8.

The Learning Curves for Mode 6 and 8

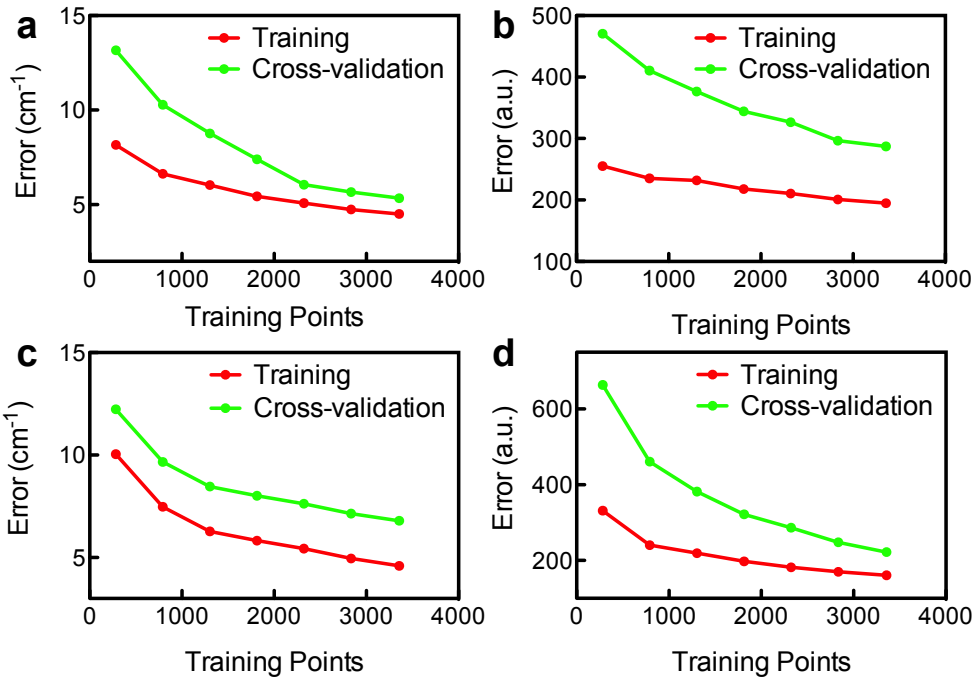


Figure S5 a,b, The learning curves for predicting the frequencies and Raman intensities of Mode 6. **c,d,** Same with **a,b** but for Mode 8.

Heat Map of the Pearson Correlation Coefficient among Descriptors for Mode 6 and 8

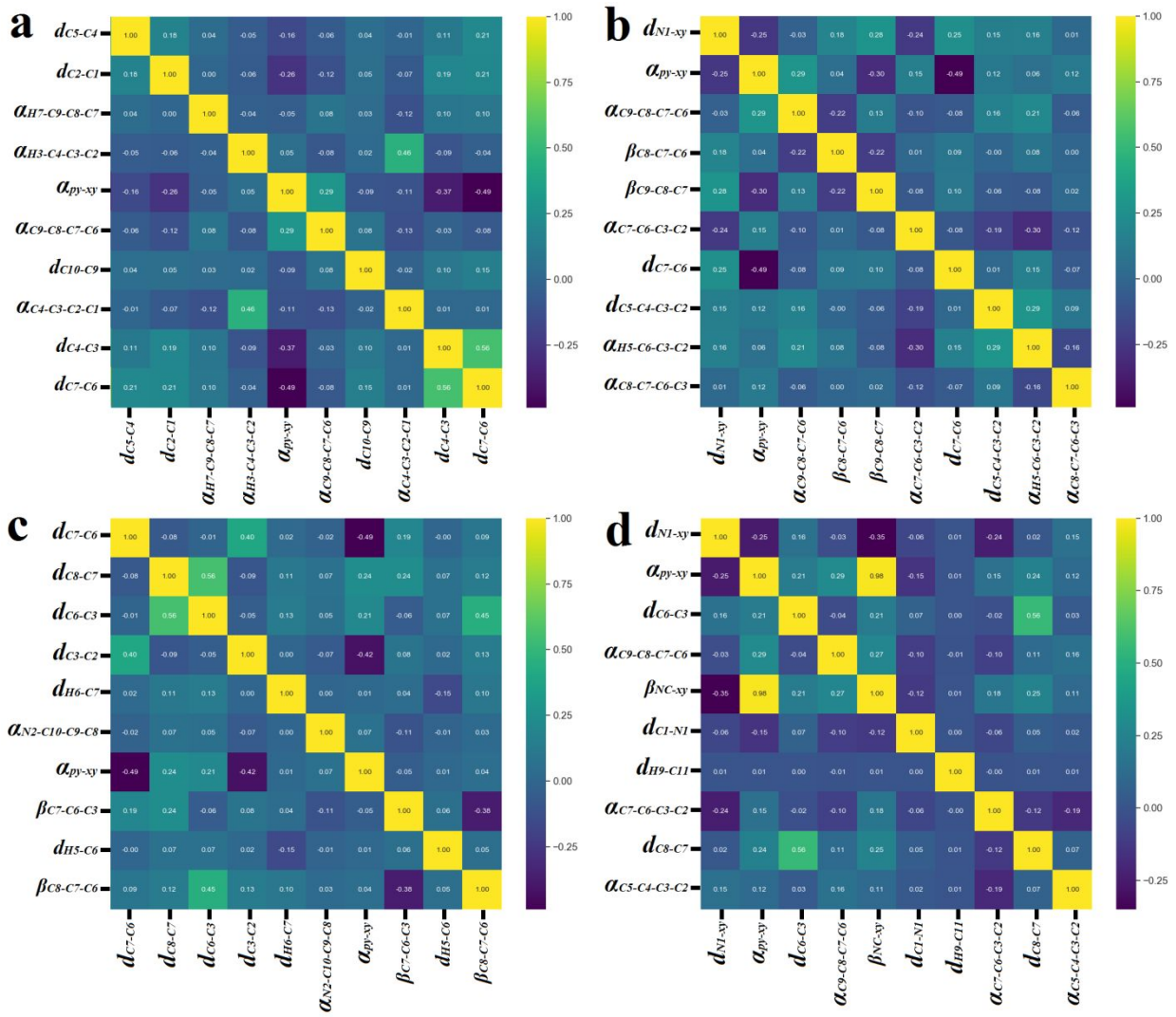


Figure S6. a,b, The Pearson correlation coefficient (r) among the descriptors for predicting the vibrational frequencies and Raman intensities of Mode 6. **c,d,** Same with **a,b** but for Mode 8.

Autocorrelation Functions of the Four Most Important Descriptors for Mode 8

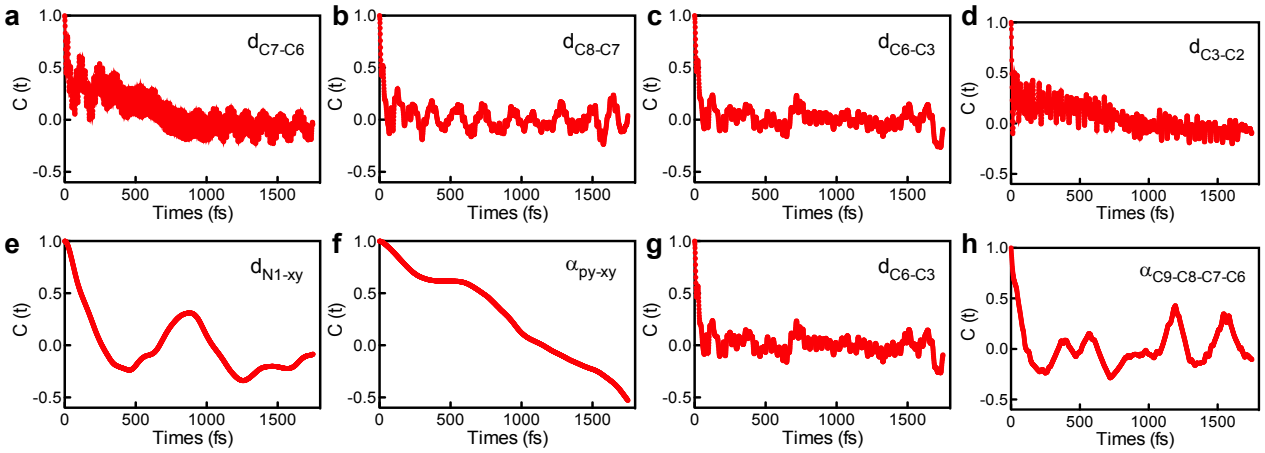


Figure S7. The autocorrelation functions of the four most important descriptors for the frequency (a-d) and Raman intensity (e-h) respectively for the mode 8.

Comparison between Mode Projection and Random Forest Importance Analysis

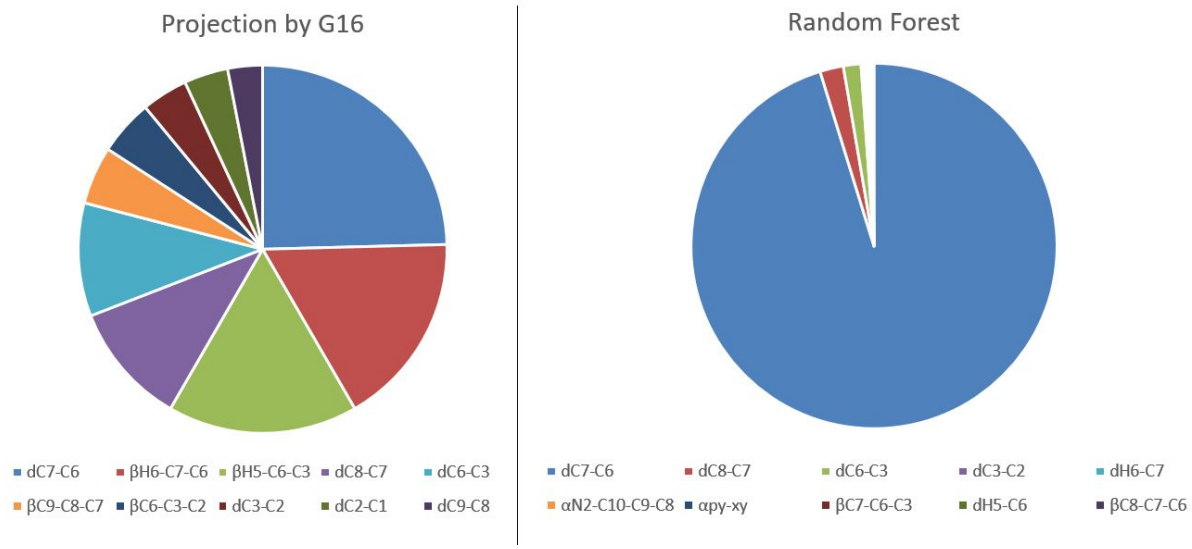


Figure S8. Comparison of the ten most important internal coordinates obtained from projection of the vibrational mode and random forest important analysis for mode 8.

We obtained the most important internal coordinate by projecting the vibrational modes onto the internal coordinates using the keyword "freq=intmodes" in Gaussian 16 package. The distribution of important internal coordinates given by the projection method has some overlap with that given by the importance analysis result. Taking the 8th vibrational modes for instance, three of first five variables in both sets are consistent, especially both methods predict that (dC7-C6) is the most important one. Meanwhile, as shown in Figure S8 the distribution of 10 most important internal coordinates is more even by projection method of Gaussian 16, while the distribution is much more uneven by the importance analysis. Furthermore, not only the vibration mode but also the Raman intensities need important analysis, so the time saving with feature engineering may not be as expected.

Partition of Time Consumption for Computing Raman Spectra for One Structure

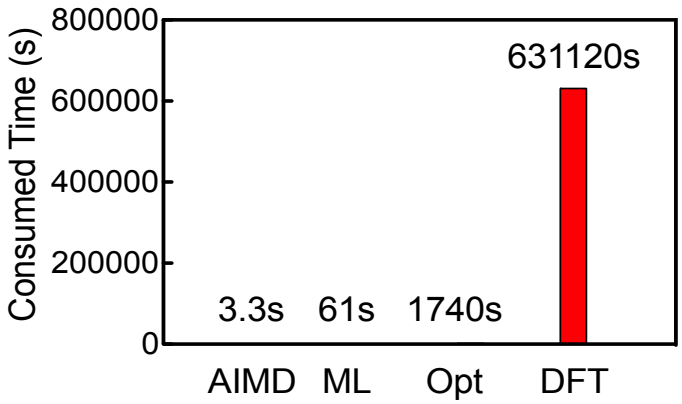


Figure S9. Partition of time consumption for computing the Raman spectra for one structure. Structure determination for one frame: 3.3s for ab initio molecular dynamics (AIMD), 1740s for geometrical optimization (Opt). Raman spectra simulation for one frame: 61s for machine learning (ML), 631120s for DFT calculation.

From Figure S9, we can see that the CPU time needed for calculating Raman spectra of one structure using quantum chemistry method is ~632863 seconds (structure sampling by AIMD: 5.2×10^{-4} %, structure optimization by DFT: 0.28 %, Raman spectra calculation by DFT: 99.72 %), and the total time for ML approach is ~1804 seconds (structure sampling by AIMD: 0.18 %, structure optimization by DFT: 96.45 %, Raman spectra by ML: 3.38 %). Noticeably, the time for DFT calculations (including SCF part and Raman spectra simulation) of a single BPE/Au configuration is 175 CPU hours, which is 7.3 hours for a 24-core Intel Xeon E7-8860 v4 node. To complete the DFT calculations of 4000 different configurations, it would take 1216 days in a 24-core node. While for ML simulations of these structures, it would just take a few hours. Furthermore, each time one changes the surface or other external conditions, one has to spend appreciable time in DFT calculations to test different geometric structures. From this point of view, the ML method can significantly speed up tedious conformation search for identification of the interfacial structures of adsorbates on a specific surface.

ML Results for Mode 3

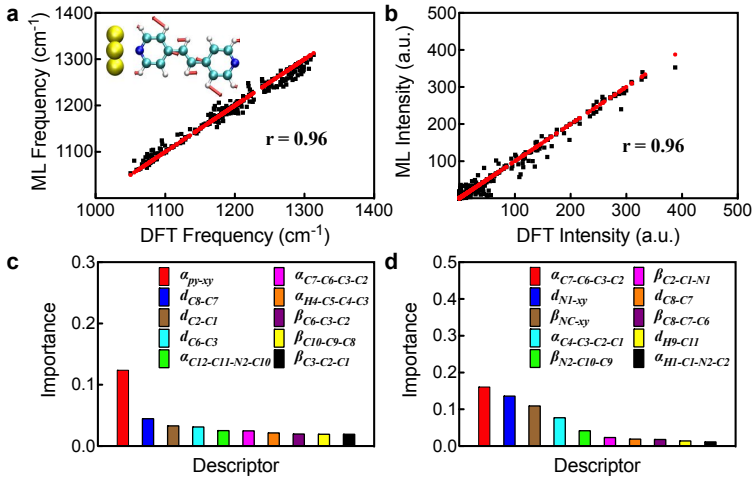


Figure S10 Comparison of DFT-computed and ML-predicted normal vibrational frequencies (a) and Raman intensities (b) for mode 3 (shown in the inset of **a**). Pearson correlation coefficient (*r*) of ML reflects the agreement. Descriptor importance analysis of frequencies (c) and Raman Intensities (d), respectively.

ML Results for Mode 7

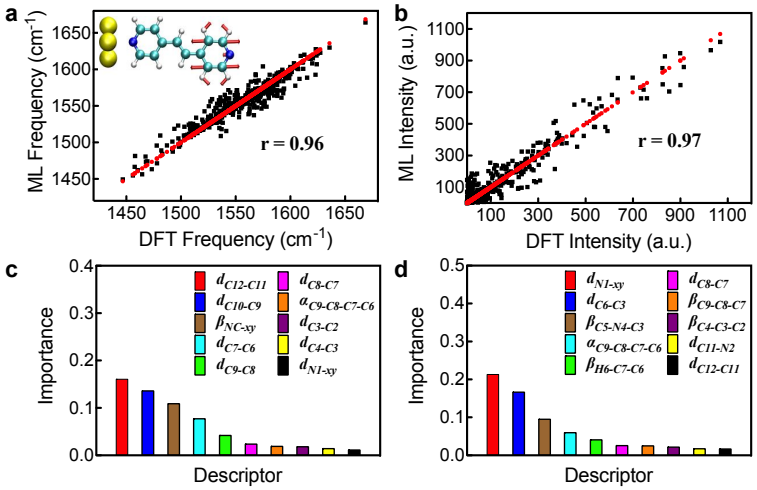


Figure S11 Comparison of DFT-computed and ML-predicted normal vibrational frequencies (a) and Raman intensities (b) for mode 7 (shown in the inset of **a**). Pearson correlation coefficient (*r*) of ML reflects the agreement. Descriptor importance analysis of frequencies **(c)** and Raman Intensities **(d)**, respectively.

Effect of Solvent Molecules and Electric Field on the Atomic Partial Charges

Table S2. Comparison of the atomic partial charges of BPE molecules with and without the solvent water and electric field at B3LYP/6-31G* and wB97XD/6-31+G* levels.

	Solvent Water Effect				Electric Field Effect			
	B3LYP/6-31G*		wB97XD/6-31+G*		B3LYP/6-31G*		wB97XD/6-31+G*	
	With Water	Without Water	With Water	Without Water	With Electric Field	Without Electric Field	With Electric Field	Without Electric Field
C1	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
C2	-0.25	-0.24	-0.27	-0.25	-0.25	-0.25	-0.27	-0.27
C3	-0.05	-0.03	-0.04	-0.03	-0.05	-0.05	-0.05	-0.05
C4	-0.26	-0.26	-0.27	-0.28	-0.25	-0.25	-0.27	-0.26
C5	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
C6	-0.20	-0.22	-0.21	-0.23	-0.19	-0.19	-0.20	-0.20
C7	-0.20	-0.21	-0.21	-0.22	-0.19	-0.18	-0.19	-0.19
C8	-0.05	-0.02	-0.04	-0.02	-0.05	-0.05	-0.05	-0.04
C9	-0.26	-0.26	-0.27	-0.27	-0.25	-0.25	-0.27	-0.27
C10	-0.26	-0.26	-0.27	-0.27	-0.25	-0.25	-0.27	-0.26
C11	0.02	0.03	0.01	0.02	0.02	0.02	0.01	0.01
C12	0.02	0.06	0.01	0.06	0.02	0.02	0.01	0.01
N1	-0.45	-0.50	-0.46	-0.51	-0.45	-0.46	-0.46	-0.47
N2	-0.45	-0.55	-0.46	-0.57	-0.45	-0.46	-0.46	-0.47
H1	0.23	0.23	0.24	0.24	0.23	0.23	0.24	0.24
H2	0.24	0.25	0.25	0.27	0.24	0.24	0.25	0.25
H3	0.24	0.24	0.25	0.26	0.24	0.24	0.25	0.25
H4	0.23	0.23	0.25	0.25	0.23	0.23	0.24	0.24
H5	0.23	0.27	0.24	0.28	0.23	0.23	0.24	0.24
H6	0.24	0.25	0.25	0.26	0.23	0.23	0.24	0.24
H7	0.24	0.26	0.25	0.27	0.24	0.24	0.25	0.25
H8	0.24	0.27	0.25	0.28	0.24	0.24	0.25	0.25
H9	0.23	0.23	0.24	0.24	0.23	0.22	0.24	0.24
H10	0.23	0.25	0.24	0.27	0.23	0.23	0.24	0.24

Effect of Solvent Molecules and Electric Field on the Energy Levels of BPE Molecule

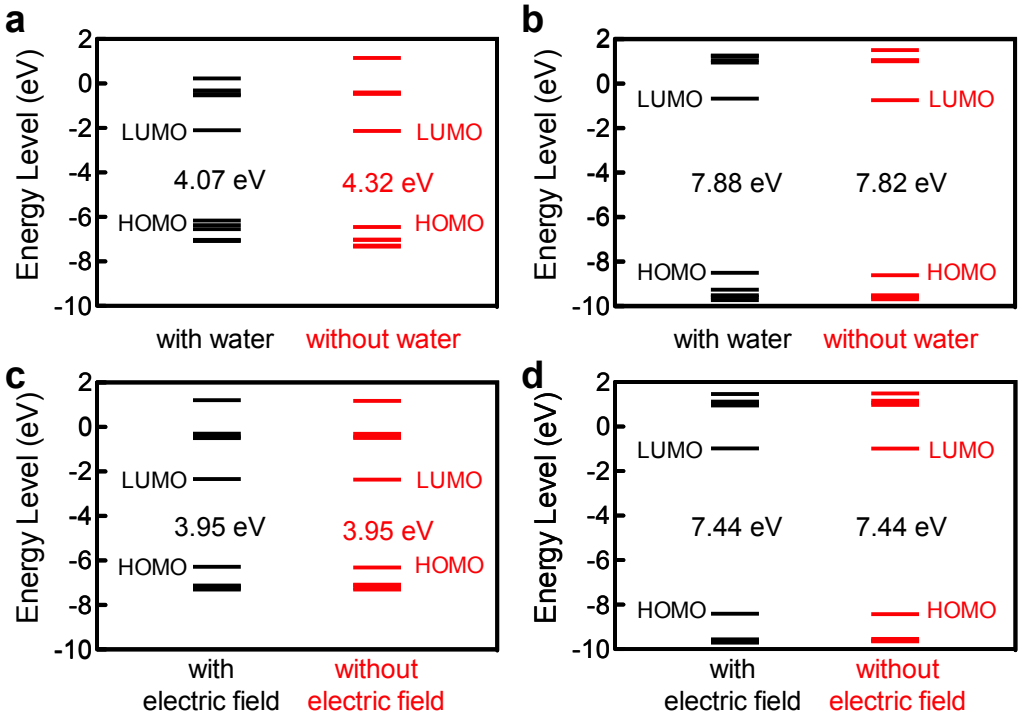


Figure S12. Comparison of the energy levels of BPE molecule with and without the solvent water (a,b) and electric field (c,d) at B3LYP/6-31G* (a,c) and wB97XD/6-31+G* levels (b,d).

SERS Spectra of BPE Adsorbed on Cu(111), Pd(111), Au(100), Au(110) Surface

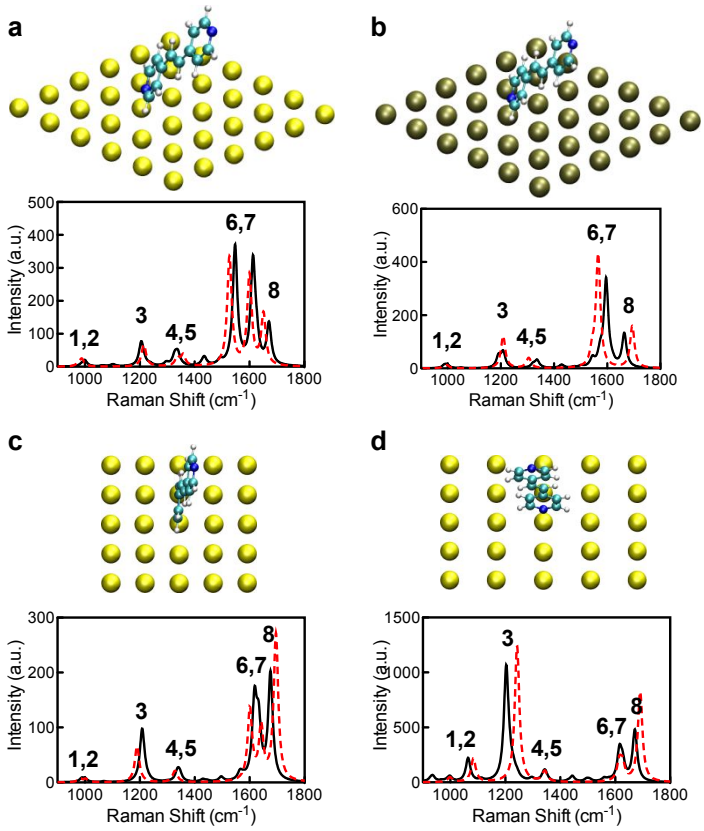


Figure S13. The optimized structure of BPE adsorbed on (a) Cu(111), (b) Pd(111), (c) Au(100), (d) Au(110), and their SERS predicted by DFT calculation (solid black line) and ML simulation (dashed red line), respectively.

Reference

(1) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169.
(2) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(3) Blöchl, P.-E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953.
(4) McLean, A. D.; Chandler, G. S. Contracted Gaussian Basis Sets for Molecular Calculations. I. Second Row Atoms, Z=11–18. *J. Chem. Phys.* **1980**, *72*, 5639-5648.
(5) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571-2577.
(6) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829-5835.
(7) Frisch, M.-J.; Trucks, G.-W.; Schlegel, H.-B.; Scuseria, G.-E.; Robb, M.-A.; Cheeseman, J.-R.; Scalmani, G.; Barone, V.; Petersson, G.-A.; Nakatsuji, H. Gaussian 16. *Revision A* **2016**, *3*.
(8) Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. Quantum Chemical Calculation of Vibrational Spectra of Large Molecules—Raman and Ir Spectra for Buckminsterfullerene. *Journal of Computational Chemistry* **2002**, *23*, 895-910.
(9) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947-1958.
(10) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O.-A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404-3419.
(11) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K. R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J Chem Theory Comput* **2013**, *9*, 3404-19.