



RESEARCH ARTICLE

Predictions of photophysical properties of phosphorescent platinum(II) complexes based on ensemble machine learning approach

Shuai Wang¹  | ChiYung Yam^{2,3} | Shuguang Chen^{1,2} | LiHong Hu⁴  |
Liping Li² | Faan-Fung Hung^{1,2,5} | Jiaqi Fan² | Chi-Ming Che^{1,2,5} |
GuanHua Chen^{1,2}

¹Department of Chemistry, The University of Hong Kong, Hong Kong, China

²Hong Kong Quantum AI Lab Limited, Hong Kong, China

³Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China

⁴School of Information Science and Technology, Northeast Normal University, Changchun, China

⁵State Key Laboratory of Synthetic Chemistry, HKU-CAS Joint Laboratory on New Materials, The University of Hong Kong, Hong Kong, China

Correspondence

GuanHua Chen, Department of Chemistry, The University of Hong Kong, Hong Kong, China.

Email: ghc.hku@gmail.com

ChiYung Yam, Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China.

Email: yamcy@uestc.edu.cn

Chi-Ming Che, Department of Chemistry, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.

Email: cmche@hku.hk

Funding information

Hong Kong Quantum AI Lab Ltd.; RGC General Research Fund, Grant/Award Number: 17309620; Guangdong Major Project of Basic and Applied Basic Research, Grant/Award Number: 2019B030302009; National Natural Science Foundation of China, Grant/Award Numbers: 22073007, 22273010; Shenzhen Basic Research Key Project Fund, Grant/Award Number: JCYJ20220818103200001

Abstract

Cyclometalated Pt(II) complexes are popular phosphorescent emitters with color-tunable emissions. To render their practical applications as organic light-emitting diodes emitters, it is required to develop Pt(II) complexes with high radiative decay rate constant and photoluminescence (PL) quantum yield. Here, a general protocol is developed for accurate predictions of emission wavelength, radiative decay rate constant, and PL quantum yield based on the combination of first-principles quantum mechanical method, machine learning, and experimental calibration. A new dataset concerning phosphorescent Pt(II) emitters is constructed, with more than 200 samples collected from the literature. Features containing pertinent electronic properties of the complexes are chosen and ensemble learning models combined with stacking-based approaches exhibit the best performance, where the values of squared correlation coefficients are 0.96, 0.81, and 0.67 for the predictions of emission wavelength, PL quantum yield and radiative decay rate constant, respectively. The accuracy of the protocol is further confirmed using 24 recently reported Pt(II) complexes, which demonstrates its reliability for a broad palette of Pt(II) emitters.

KEYWORDS

DFT, machine learning, OLED, phosphorescent emitters, photophysical properties

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 Hong Kong Quantum AI Lab. *Journal of Computational Chemistry* published by Wiley Periodicals LLC.

1 | INTRODUCTION

Organic light-emitting diodes (OLEDs) are sustainable light sources emergingly used in displays and many other fields.¹ While the first-generation fluorescence-based OLEDs are limited to 25% internal quantum efficiency (IQE) as the ratio of singlet and triplet excitons is in 1:3 according to spin statistics,² one can overcome this limit by utilizing triplet excitons via phosphorescent heavy metal-based emitters. Heavy metal atoms such as iridium and platinum can induce strong spin-orbit coupling (SOC) to facilitate the intersystem crossing process from the singlet to triplet excited states, and to promote radiative deactivation from the triplet excited state through phosphorescence.³ Thus, the second-generation phosphorescent OLEDs (PhOLEDs) based on iridium/platinum emitters can achieve IQE up to 100%. Over the last decade, the research on platinum-based PhOLEDs has steadily increased in popularity and the device performances have been improved with the introduction of tetradentate cyclometalating ligands.^{4–8} In fact, however, it would require substantial cost and efforts to develop high-performance PhOLED emitters experimentally, as the relationship between molecular structures and photophysical properties with interest is complicated and has not been elucidated adequately.

Theoretically, density functional theory⁹ (DFT) and time-dependent density-functional theory¹⁰ (TDDFT) are the widely-used tools to predict material properties. It enables the simulations of photophysical properties of phosphorescent emitters with reasonable balance between accuracy and efficiency. However, the results obtained from DFT or TDDFT calculations on phosphorescent emitters are still not accurate enough compared to the experimental ones.¹¹ To tackle this issue, machine learning (ML) algorithms, which map the complex relationship between properties and structures, could be applied.^{12–14} Simplified molecular-input line-entry system (SMILES)¹⁵ is a common tool to represent each molecule as a one-to-one string. However, it is not applicable for organometallic complexes due to the metal–ligand coordination bonds. Hence, common fingerprints/descriptors obtained via SMILES are no longer available as usual, and this poses difficulties in developing ML model for organometallic complexes. Instead, the structural parameters together with the properties based on the first-principles results are employed as features or descriptors in this work. We demonstrate that the obtained ML models improve the accuracy of predictions and thus provide an efficient tool to design novel OLED materials.¹⁶

Over the last decade, there were studies applying ML to predict OLED-related properties via SMILES. Alán et al.¹⁷ used linear regression method to predict emission energies based on absorption energies for thermally activated delayed fluorescence (TADF) molecules. Woon et al.¹⁸ established a random forest model for efficiency predictions of blue OLEDs. Lu et al.¹⁹ set up light gradient boosting machine (LightGBM) models for glass transition temperature (T_g) and decomposition temperature (T_d) prediction for pure organic OLED materials. In this work, we choose features based on first-principles simulations to set up ML models. In particular, ensemble learning algorithms such as random forest²⁰ (RF), adaptive boosting²¹ (Adaboost), light gradient

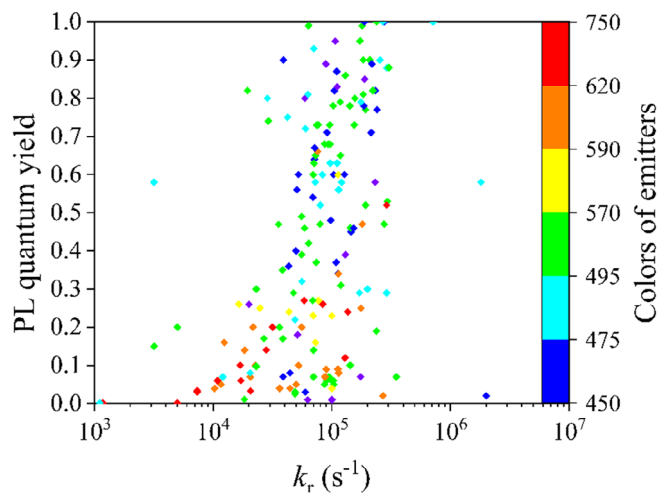


FIGURE 1 Photophysical properties distributions for PhOLED emitters.

boosting machine²² (LightGBM), and extreme gradient boosting²³ (XGB) are considered and compared with conventional ML algorithms. To further improve our ML models, stacking-based techniques²⁴ are employed to enhance the generality of ML models. Finally, the prediction performances of our models were further verified by comparing to recently reported experimental data, which demonstrates reasonably good agreement and thus confirms the robustness of the models. We expect that this protocol will be beneficial for the evaluation and discovery of new PhOLED emitters efficiently and accurately.

2 | METHODOLOGIES

2.1 | Dataset construction and division

In this work, we mainly focus on cyclometalated Pt(II) complexes with tridentate or tetradentate ligands. Photophysical data of 206 phosphorescent Pt complexes reported in the literature are collected, including emission wavelength, photoluminescence (PL) quantum yield, and radiative decay rate constant k_r . They are mostly measured in degassed solutions⁵ under ambient conditions and their photophysical properties distributions are shown in Figure 1 and Section S1. For molecules with structured emission band and multiple emission maxima,²⁵ we select the lowest emission wavelength. The details of data preprocessing can be found in Section S2 and Table S3. The emission wavelengths span a range from 430 to 661 nm, with a mean value of 526 nm. Figure 1 shows that the values of the reported compounds distribute evenly. Similarly, the PL quantum yields of the compounds distribute evenly with a mean value of 0.421. On the other hand, k_r values span over three orders of magnitude. The majority of k_r values are around $1.00 \times 10^5 \text{ s}^{-1}$ with a mean value of $1.20 \times 10^5 \text{ s}^{-1}$. This poses a great difficulty in constructing ML models for k_r prediction. To remedy this situation, care has to be taken in the data division. To maintain a balance between training and

TABLE 1 Features for machine learning models.

Features	Description
v	Emission energy from the T_1 state to the S_0 state
coor_bond_length (N)	Coordinate bond lengths for complexes, the shortest one to longest corresponds with N from 1 to 4
coor_bond_type (N)	Types of coordination (Pt-C, Pt-N, Pt-O, and Pt-Cl) The order of this series of features is correlated to coor_bond_length (N)
ρ_{Pt}	Average electron density at Pt atom
ρ_{coor} (N)	Average electron density at the four coordination atoms
H_ T_1 _S $_0$	Spin-orbit coupling constant between T_1 state and S_0 state
H_ T_1 _S $_1$	Spin-orbit coupling constant between T_1 state and S_1 state
R_EH_excited state ^{a/b}	Charge-transfer descriptor interpreted in terms of the electron-hole distance in a given excitation. "a" means calculation based on literatures, ²⁸⁻³¹ "b" means calculation based on reference. ³² Small value indicates short-range excitations.
LAMBDA_excited state	Charge-transfer descriptor measures the spatial overlap in a given excitation. ³³ Small value signifies a long-range excitation
CT_excited state	Charge-transfer character, ²⁸⁻³¹ 1 for completely charge-separated states; 0 for locally excited excitonic states
HOMO	Highest occupied molecular orbital energy
LUMO	Lowest unoccupied molecular orbital energy
μ	Molecular dipole moment
f	Oscillator strength of radiative transition from T_1 state to S_0 state
Calc_ λ/k_r	Calculated emission wavelength/radiative decay rate constant
Refractive index	Refractive index to reflect the experiment testing condition

testing sets within the limited data, an improved Kennard Stone algorithm,²⁶ which partitions the sample set based on maximum-minimum X-Y distance (SPXY),²⁷ is adopted,

$$d_{xy}(p, q) = \frac{d_x(p, q)}{\max_{p, q \in [1, N]} d_x(p, q)} + \frac{d_y(p, q)}{\max_{p, q \in [1, N]} d_y(p, q)} \quad (p, q \in [1, N]) \quad (1)$$

where x and y represent the features and the target properties, respectively. p and q denote p th and q th samples in the whole dataset with N samples.

2.2 | Features for ML

The common descriptors/fingerprints computed through SMILES are not applicable, due to their poor descriptions of coordinate bonds in organometallic complexes. Instead, in addition to structural features, first-principles simulations are employed to calculate the photophysical properties of complexes to generate molecular features in the dataset. Overall, we choose the features that are associated with the metal coordination, together with those related to the photophysical properties of these complexes, for training the ML models. As shown in Equation (2), the rate constant is determined by the emission energy and oscillator strength of the phosphorescence transition. This is in turn highly related to the frontier orbitals of the organometallic complex. In addition, for spin-flip transition, SOCs are essential for the transition to occur. In a Pt-emitter, the metal ion and its surrounding atoms play key roles in the phosphorescence process, thus average electron densities of these atoms are taken as features or descriptors. Besides, the coordinate bond type and coordinate bond length are considered as well. Table 1 lists all the descriptors utilized in the ML models. Details of each feature/descriptor can be found in Section S3.

2.3 | Machine learning algorithms

Ensemble learning strategies are employed to construct comprehensive models, including RF,²⁰ Adaboost,²¹ LightGBM,²² and XGB.²³ In contrast to single weak ML, ensemble learning can construct faster and more accurate ML models with limited data. To assess their performance, these ensemble learning models are compared with three types of conventional ML algorithms, including support vector machines³⁴ (SVM), k-nearest neighbors³⁵ (KNN), and kernel ridge regression³⁶ (KRR). In addition, stacking techniques are adopted to enhance the generality of ML models.²⁴ After SPXY data division, 80% of the dataset was selected for model training, and the remaining 20% was used as an independent test set. Ten-fold cross-validation was adopted to improve the stability of the obtained ML models. The hyperparameters are tuned by python library hyperopt.³⁷ Details of the hyperparameter optimization can be found in Section S4. Performances on the models were evaluated based on the squared correlation coefficient (R^2), the root mean square error (RMSE), and mean absolute error (MAE). Feature importance analysis for the optimal models with RF and LightGBM is conducted by utilizing their built-in attributes and averaging the results from 10 models obtained through cross-validation, as explained in Section S4.

2.4 | First-principles simulations

Gaussian16 program package³⁸ was utilized for all geometry optimizations, and ADF2021 package³⁹ was employed to calculate the phosphorescence of the Pt(II) complexes. The optimized geometries of ground state (S_0) and excited states (T_1) were calculated by DFT and

TABLE 2 Performance of each ML algorithm on the prediction of emission wavelength.

ML models	Independent testing set ^a		
	MAE (nm)	RMSE (nm)	R ²
KNN_Uniform	27.38 ± 1.17	40.90 ± 1.52	0.57 ± 0.04
KNN_Distance	17.15 ± 1.03	28.33 ± 2.40	0.80 ± 0.04
SVM	20.52 ± 2.12	32.00 ± 2.91	0.76 ± 0.05
KRR	22.62 ± 0.67	28.55 ± 0.80	0.82 ± 0.02
RF	7.98 ± 0.72	11.81 ± 1.15	0.97 ± 0.01
LightGBM	5.57 ± 0.63	8.73 ± 1.24	0.98 ± 0.01
Adaboost	13.28 ± 0.47	16.01 ± 0.65	0.95 ± 0.01

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; RF, random forest; SVM, support vector machine.

^aThe standard deviations are calculated by the difference in the prediction of each fold.

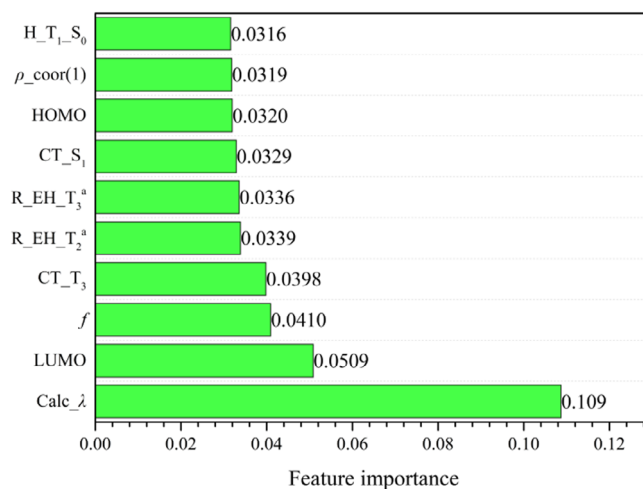
TDDFT, respectively, with B3LYP functional.⁴⁰ Relativistic effects were considered for Pt atom using the Stuttgart basis set⁴¹ and pseudopotential. 6-31G* atomic basis set^{42,43} was used for all other atoms. Solvation effects were taken into account using polarizable continuum model (PCM).⁴⁴ With the optimized structures, emission properties were calculated using the ADF2021 package.³⁹ For phosphorescence transitions, SOC was treated as a perturbation based on the scalar relativistic orbitals.⁴⁵ Triple-zeta polarized Slater-type basis set for all atoms and PBE0 functional were used in the TDDFT calculations.⁴⁶ In addition, matrix effects were considered using the COSMO continuum solvation model in relevant experimental testing medium.⁴⁷ With the emission energy and oscillator strength obtained, k_r can be calculated as follows.

$$k_r = \frac{2\pi\nu^2 e^2}{\epsilon_0 m c^3} f \quad (2)$$

where ν is the emission energy from the lowest triplet state (T_1) to the ground state (S_0); e denotes the elementary electric charge; ϵ_0 is the vacuum permittivity; m represents the mass of electrons; c is the speed of light, and f is the oscillator strength of $T_1 \rightarrow S_0$ transition.

3 | RESULTS AND DISCUSSION

ML models are constructed using different algorithms, including RF,²⁰ AdaBoost,²¹ LightGBM,²² XGB,²³ SVM,³⁴ KNN,³⁵ and KRR.³⁶ The performance of each model regarding emission wavelength predictions after 10-fold cross-validation can be found in Table 2. Apparently, LightGBM gives the best performance in terms of both correlation coefficient and errors. Basically, LightGBM improves the efficiency and scalability of gradient boosting (GB) algorithm without sacrificing its inherited effective performance, which is suitable for rapid assessment of the dataset. See details of the performance of each ML algorithm in Tables S4–S6.

**FIGURE 2** Ten most important features for emission wavelength extracted from LightGBM-based machine learning model.

We further analyze the importance of each feature as listed in Figure 2. It can be seen that the calculated emission wavelength shows the most remarkable contribution in determining the target property. In addition, LUMO/HOMO energy, oscillator strength and SOC constant contribute significantly to the model as well. Others are primarily transition-related features. For instance, the charge-transfer features^{28–31} (CT series in Figure 2) at the transitions from the third triplet state and first singlet state to the ground state both contribute to the optimal model.

Different stacking strategies with a two-layer architecture are used to further improve the generality and stability of the emission wavelength model. One involves concatenating the initial features with a superior prediction result to use as the features for stacking, while the other solely selects several best prediction results as the features. See stacking details in Table S7. The optimal stacking architecture is shown in Figure 3 where the wavelength predicted by LightGBM model is concatenated with all the features utilized in the base-learner layer. Utilizing the new features, four different types of meta-learners are tested and the results are listed in Table 3 and Table S10. SVM as meta-learner shows the highest correlation coefficient and lowest errors, and it is selected as the optimal meta-learner. Finally, Figure 4 plots the performance of the stacking model in predicting emission wavelength on the independent testing set. It is clear that the ML predicted results agree well with experiments and show the substantial improvement over simulation results. Excellent consistency between ML predicted results and experimental results with R² of 0.96, MAE of 7.21 nm and RMSE of 13.00 nm is obtained.

Radiative decay rate constant k_r of OLED emitter is another fundamental quantity that determines the performance of OLEDs. Emitters with large k_r values are preferred to avoid bimolecular quenching processes in OLED that could lead to severe efficiency roll-off at high luminance and material degradation. To predict the radiative decay rate constant k_r , eight models are constructed using different ML algorithms whose performance are listed in Table 4. It is found that KNN,

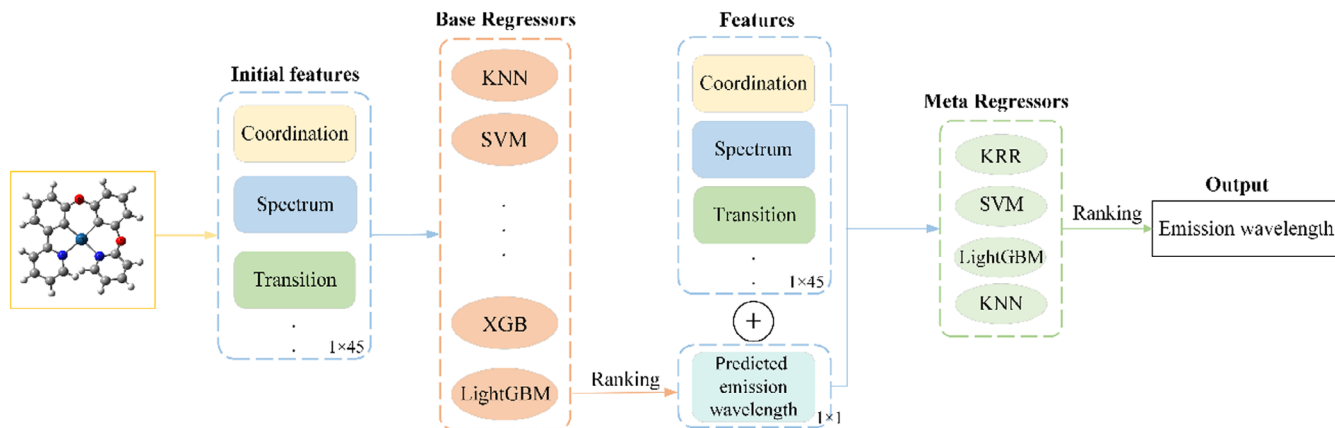


FIGURE 3 Proposed stacking architecture for emission wavelength prediction.

TABLE 3 Meta-regressors comparison on emission wavelength in stacking.

ML models	Independent testing set ^a		
	MAE	RMSE	R ²
KRR	13.10 ± 0.74	15.85 ± 0.78	0.94 ± 0.01
SVM	7.22 ± 0.77	13.00 ± 1.33	0.96 ± 0.01
LightGBM	12.37 ± 0.70	16.09 ± 1.13	0.94 ± 0.01
KNN_Distance	11.96 ± 1.01	17.67 ± 2.76	0.92 ± 0.03

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; SVM, support vector machine.

^aThe standard deviations are calculated by the difference in the prediction of each fold.

TABLE 4 Performance of each ML algorithm on the prediction of radiative decay rate constants.

ML models	Independent testing set ^a		
	MAE	RMSE	R ²
KNN_Uniform	0.28 ± 0.01	0.32 ± 0.01	0.23 ± 0.05
KNN_Distance	0.19 ± 0.01	0.22 ± 0.01	0.74 ± 0.06
SVM	0.22 ± 0.01	0.26 ± 0.01	0.71 ± 0.08
KRR	0.22 ± 0.01	0.27 ± 0.01	0.57 ± 0.04
RF	0.16 ± 0.01	0.22 ± 0.01	0.74 ± 0.05
LightGBM	0.17 ± 0.01	0.24 ± 0.01	0.65 ± 0.06
Adaboost	0.23 ± 0.01	0.27 ± 0.02	0.63 ± 0.09
XGB	0.16 ± 0.01	0.21 ± 0.03	0.73 ± 0.07

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; RF, random forest; SVM, support vector machine.

^aThe standard deviations are calculated by the difference in the prediction of each fold.

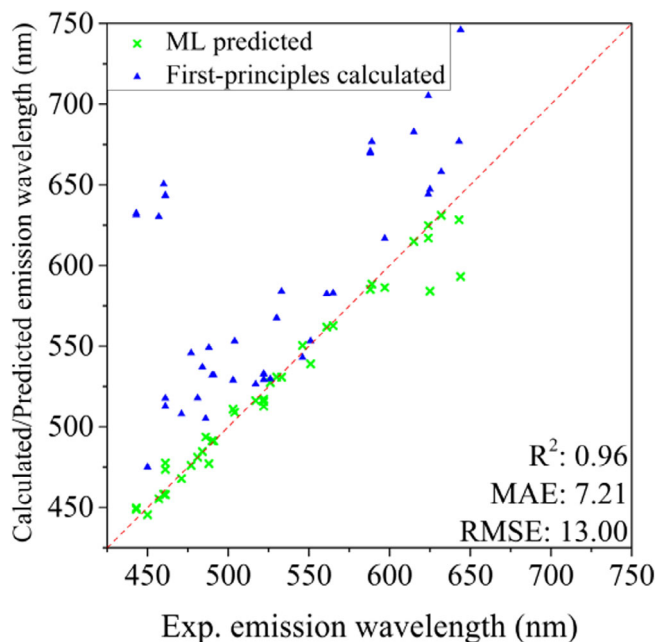


FIGURE 4 Support vector machine performances of emission wavelength calibration model on the independent testing set.

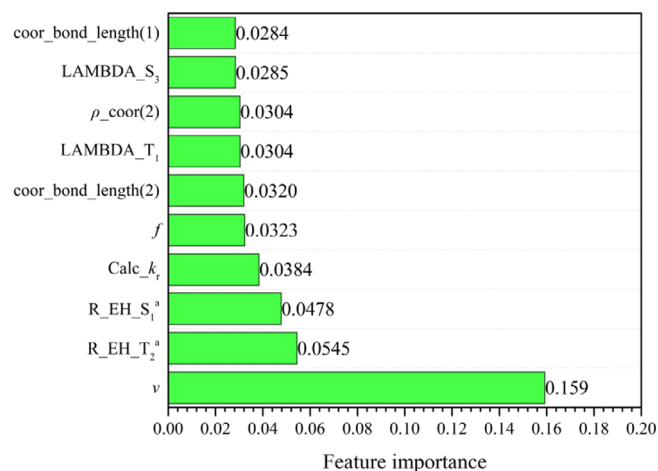


FIGURE 5 Ten most important features for radiative decay rate constants extracted from random forest-based machine learning model.

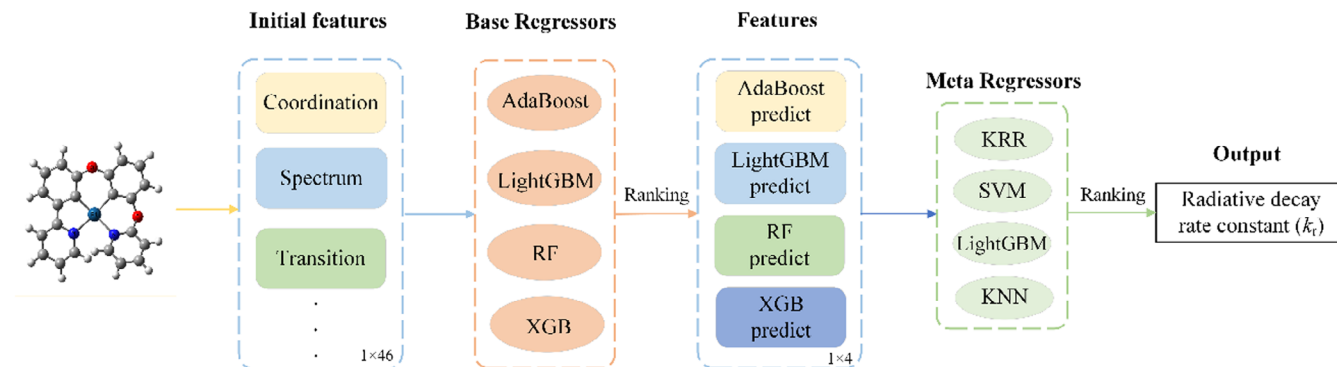


FIGURE 6 Proposed stacking architecture for k_r prediction.

TABLE 5 Meta-regressors comparison on radiative decay rate constants in stacking.

ML models	Independent testing set ^a		
	MAE	RMSE	R^2
KRR	0.95 ± 0.01	1.26 ± 0.01	0.77 ± 0.00
SVM	0.45 ± 0.05	0.51 ± 0.05	0.01 ± 0.01
LightGBM	0.22 ± 0.01	0.27 ± 0.01	0.53 ± 0.06
KNN_Distance	0.21 ± 0.01	0.25 ± 0.01	0.67 ± 0.04

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; SVM, support vector machine.

^aThe standard deviations are calculated by the difference in the prediction of each fold.

RF, and XGB have similar correlation coefficients and errors. However, RF holds the best prediction stability when compared to the other two models. RF is usually robust and stable for the outliers and noises, a property that is suitable for the case of k_r since the experimental values span over three orders of magnitude and are not balanced. Essentially, RF creates trees on the subset of the data and combines the output of all the trees, which reduces overfitting problem in decision trees and the variance. Similarly, we analyze the importance of different features as plotted in Figure 5. Again, the emission energy has the highest contribution as it is one of the key terms determining the rate constant. On the other hand, the features about electron-hole distance^{28–31} from T_2 and S_1 to ground state ($R_{EH_T_2}^a$, $R_{EH_S_1}^a$), which describe the average electron-hole separation, have a considerable importance compared to the other features. In contrast, the calculated rate constant ranks lower, which indicates the relative inaccurate description of singlet-triplet transition in TDDFT. Besides, oscillator strength, bond lengths of coordination bonds, and average electron densities of the coordination atoms are also important features in the model.

To further improve the prediction performance, stacking technique is adopted with two layers. In practice, the benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions more stable than the single model in the base-learner layer. Specially, stacking technique will perform more robust and reliable when

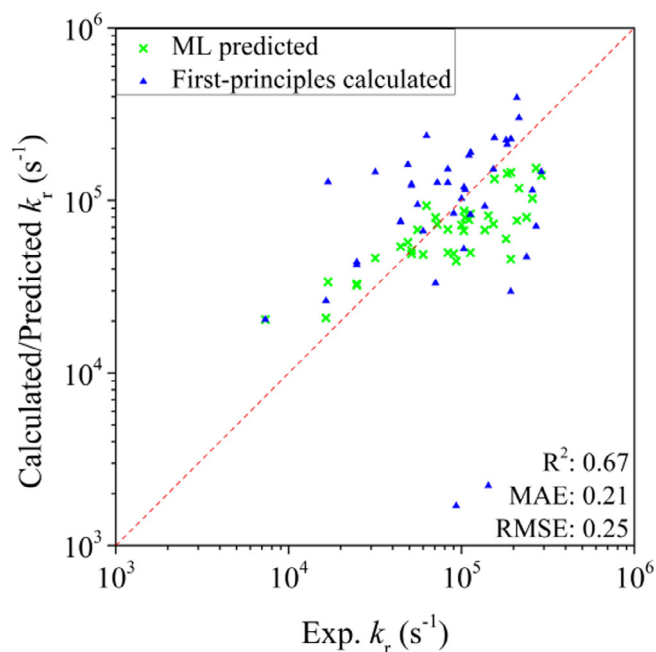


FIGURE 7 k-Nearest neighbor performances of radiative decay constant calibration model on the independent testing set.

changing different testing data. Stacking details for k_r can be seen in Table S8. As shown in Figure 6, the optimal stacking architecture includes the aggregation of AdaBoost, LightGBM, RF, and XGBoost as the base learner layer. These ensemble algorithms exhibit superior results on the independent testing set so that their predicted k_r values are selected as the features for the subsequent training. This kind of selection would take advantage of different algorithms to get more stable predictions. Then four different types of meta-learners are tested and compared. The results are shown in Table 5 and Table S11. Meta-learner KNN_Distance, among the four algorithms, exhibits the best performance on error items after stacking. The results with meta-learner KNN_Distance can be seen in Figure 7. Clearly, our ML model significantly improves the prediction of radiative decay rate constants with MAE of 0.21, RMSE of 0.25, and R^2 of 0.67.

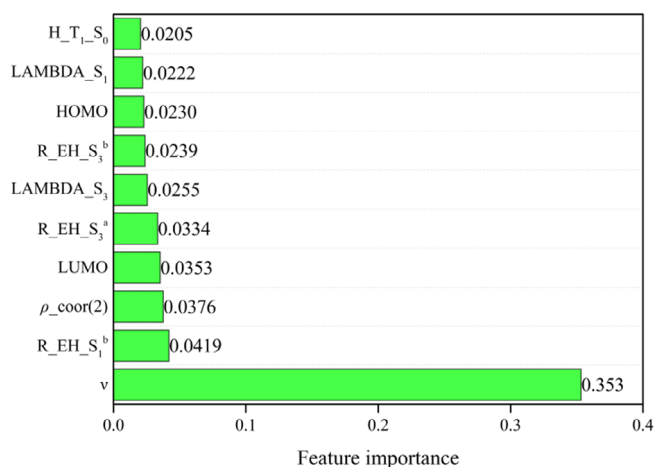
PL quantum yield is defined as the ratio of the number of photons emitted to the number of photons absorbed, which is a prominent

TABLE 6 Performance of each ML algorithm on the prediction of PL quantum yield.

ML models	Independent testing set ^a		
	MAE	RMSE	R ²
KNN_Uniform	0.21 ± 0.01	0.26 ± 0.01	0.37 ± 0.03
KNN_Distance	0.16 ± 0.00	0.21 ± 0.01	0.59 ± 0.03
SVM	0.21 ± 0.00	0.24 ± 0.01	0.63 ± 0.04
KRR	0.19 ± 0.00	0.23 ± 0.00	0.57 ± 0.03
LightGBM	0.09 ± 0.01	0.13 ± 0.01	0.83 ± 0.04
Adaboost	0.13 ± 0.00	0.17 ± 0.01	0.74 ± 0.02
RF	0.11 ± 0.00	0.15 ± 0.01	0.81 ± 0.02
XGB	0.13 ± 0.01	0.18 ± 0.01	0.74 ± 0.04

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; PL, photoluminescence; RF, random forest; SVM, support vector machine.

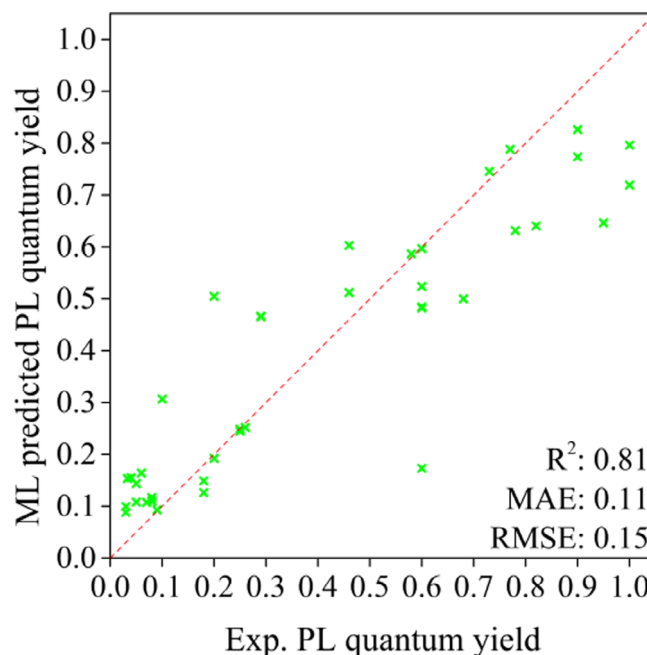
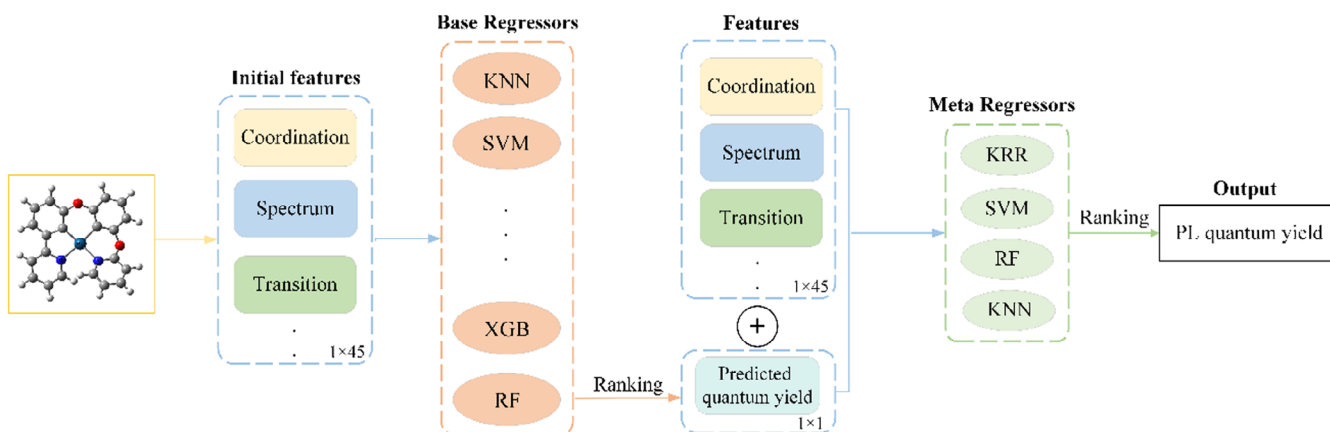
^aThe standard deviations are calculated by the difference in the prediction of each fold.

**FIGURE 8** Ten most important features for photoluminescence quantum yield extracted from random forest-based machine learning model.**TABLE 7** Meta-regressors comparison on quantum yield in stacking.

ML models	Independent testing set ^a		
	MAE	RMSE	R ²
KRR	0.13 ± 0.01	0.17 ± 0.00	0.78 ± 0.01
SVM	0.15 ± 0.01	0.17 ± 0.01	0.81 ± 0.04
RF	0.11 ± 0.00	0.15 ± 0.01	0.81 ± 0.02
KNN_Distance	0.15 ± 0.01	0.19 ± 0.01	0.65 ± 0.04

Abbreviations: KNN, k-nearest neighbor; KRR, kernel ridge regression; ML, machine learning; RF, random forest; SVM, support vector machine.

^aThe standard deviations are calculated by the difference in the prediction of each fold.

**FIGURE 10** Random forest performance of photoluminescence quantum yield prediction model on the independent testing set.**FIGURE 9** Stacking architecture for photoluminescence quantum yield prediction.

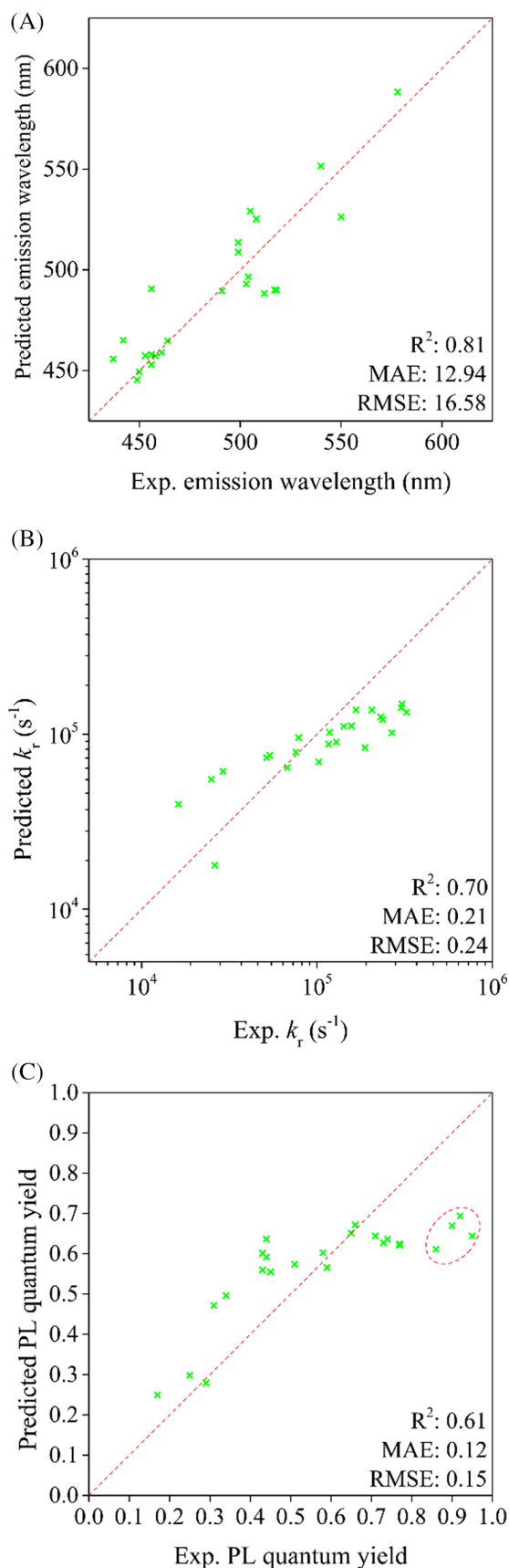


FIGURE 11 Performances of optimal machine learning models for (A) emission wavelength, (B) radiative decay rate constant, and (C) photoluminescence quantum yield on external samples.

property of OLED emitters. From the theoretical perspective, accurate prediction of PL quantum yield of phosphorescent Pt emitters using first-principles simulations is challenging.⁴⁸ In this work, quantum yield is therefore not chosen as one of the features. Instead, features listed in Table 1 are used, except calculated emission wavelength and k_r . Consequently, eight models are constructed using different ML algorithms. Table 6 illustrates the performance of each model. RF is selected and its feature importance analysis for PL quantum yield is shown in Figure 8. As expected, emission energy is the most important feature, followed by average electron density of coordination atoms, LUMO/HOMO energies, SOC constant, and other features that are closely related to the radiative transition.

Next, similar to the emission wavelength, the stacking approach that involves concatenating the best-predicted result is used to further enhance the prediction capability of the PL quantum yield model. Different stacking strategies are compared in Table S9. Here, the quantum yield predicted by RF model is concatenated with all other features utilized in the base-learner layer. The best stacking architecture is shown in Figure 9. With the new features, different types of meta-learners are tested and the results are listed in Table 7 and Table S12. RF as meta-learner shows the highest correlation coefficient and lowest errors, and it is selected as the optimal meta-learner. Finally, Figure 10 plots the performance of the stacking model in predicting PL quantum yield on the independent testing set. Satisfactory consistency between ML predicted results and experimental measurements is achieved with R^2 of 0.81, MAE of 0.11 and RMSE of 0.15.

Realizing accurate photophysical properties predictions of new complexes prior to experimental synthesis will be of great importance for the development of OLED materials. To assess the performance of our optimal models, 24 recently reported complexes are collected as external samples to evaluate the generalization capacity of the three vital photophysical properties. More detailed information on these Pt-based complexes can be found in Section S5. Figure 11A–C shows respectively the prediction of the three photophysical properties. It can be seen that the ML models satisfactorily predict the phosphorescence properties of these latest studied Pt-based complexes over diverse scaffolds and therefore confirms the generalization abilities of our models.

For the emission wavelength predictions, R^2 of 0.81, RMSE of 16.58 nm and MAE of 12.94 nm are achieved. When compared with the mean value of 526 nm, the result is accurate enough for evaluation of the performance and screening of these emitters. On the other hand, regarding radiative decay rate constant k_r , MAE of 0.21 and RMSE of 0.24 (both in log scale) are very similar to those in the testing set and acceptable for k_r predictions. For the PL quantum yield, an outstanding performance with MAE of 0.12 and RMSE of 0.15 is realized, except that there are four outliers as highlighted in Figure 11C.

Only a small proportion of samples in the training dataset are with high PL quantum yields, exceeding 0.9. This scarcity of training data poses a challenge for the models to effectively learn and predict target values at such a high level. Consequently, it is relatively difficult to provide very accurate predictions for the samples with exceptionally high experimental PL quantum yields. Prediction details of the

external samples can be seen in Table S13. Overall, the developed ML models in this work not only give exceptional performance on the independent testing set, but also demonstrate satisfactory results on the external testing set. Comparison results between the optimal single models and stacking-based models are listed in Table S14, demonstrating a significant improvement in the robustness and stability of the stacking-based ML techniques.

4 | CONCLUSION

In summary, a general protocol of first-principles calculations and ML models is constructed to predict the photophysical properties of phosphorescent Pt(II) emitters. Ensemble ML models (XGBoost, LightGBM, RF, and AdaBoost) are utilized based on a dataset of 206 Pt-based emitters. These are compared with conventional ML models (SVM, KNN, and KRR) to demonstrate their performance. The feature analysis reveals that emission energy, coordination, and SOC constant show significant contributions in the prediction. To further improve the performance, the stacking methods are implemented. Finally, recently reported Pt-complexes are employed as external samples to evaluate the generalization capability of the ML models, which indicates the robustness of the protocol. This work presents the first ML protocol for predicting and evaluating three important photophysical properties of Pt-emitters by employing ensemble ML algorithms. We expect the protocol would be beneficial to scientists in designing novel Pt-emitters with superior performances and thus help discover novel OLED materials.

AUTHOR CONTRIBUTIONS

Shuai Wang: Writing—original draft, investigation, software, data curation, formal analysis. **ChiYung Yam:** Supervision, conceptualization, writing—review and editing. **Shuguang Chen:** Project administration, resources. **LiHong Hu:** Conceptualization, methodology. **Liping Li:** Investigation, software, data curation. **Faan-Fung Hung:** Data curation, writing—review and editing. **Jiaqi Fan:** Investigation, software, data curation. **Chi-Ming Che:** Supervision, conceptualization, funding acquisition. **GuanHua Chen:** Supervision, conceptualization, funding acquisition, writing—review and editing.

ACKNOWLEDGMENTS

This work is financially supported by the Hong Kong Quantum AI Lab Ltd. and RGC General Research Fund under grant no. 17309620. Chi-Ming Che and Faan-Fung Hung acknowledge Guangdong Major Project of Basic and Applied Basic Research (grant no. 2019B030302009). Chi Yung Yam acknowledges the support from National Natural Science Foundation of China (grant no. 22073007). LiHong Hu thanks financial support from National Natural Science Foundation of China (no. 22273010). Chi Yung Yam acknowledges the support from Shenzhen Basic Research Key Project Fund (grant no. JCYJ20220818103200001).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All data and code employed in this work are available from an open GitHub repository: https://github.com/JerryShuaiWANG/PhOLED_ML

ORCID

Shuai Wang  <https://orcid.org/0000-0002-3195-2691>

LiHong Hu  <https://orcid.org/0000-0003-3792-2917>

REFERENCES

- [1] A. Saeki, K. Kranthiraja, *Jpn. J. Appl. Phys.* **2019**, *59*, SD0801.
- [2] L. Paterson, F. May, D. Andrienko, *J. Appl. Phys.* **2020**, *128*, 160901.
- [3] G. Hong, X. Gan, C. Leonhardt, Z. Zhang, J. Seibert, J. M. Busch, S. Bräse, *Adv. Mater.* **2021**, *33*, 2005630.
- [4] G. Li, Y. She, *Tetradentate Cyclometalated Platinum(II) Complexes for Efficient and Stable Organic Light-Emitting Diodes*, IntechOpen, London **2018**.
- [5] K. Li, G. S. M. Tong, Q. Wan, G. Cheng, W.-Y. Tong, W.-H. Ang, W.-L. Kwong, C.-M. Che, *Chem. Sci.* **2016**, *7*, 1653.
- [6] H. Li, T.-L. Lam, X. Tan, L. Dai, C.-M. Che, *SID Symp Digest Techn Pap* **2021**, *52*, 328.
- [7] J. Sun, H. Ahn, S. Kang, S.-B. Ko, D. Song, H. A. Um, S. Kim, Y. Lee, P. Jeon, S.-H. Hwang, Y. You, C. Chu, S. Kim, *Nat. Photon.* **2022**, *16*, 212.
- [8] H. Li, T.-L. Lam, L. Yan, L. Dai, B. Choi, Y.-S. Cho, Y. Kwak, C.-M. Che, *Tetradentate Platinum(II) Emitters: Design Strategies, Photophysics, and OLED Applications*, IntechOpen, London **2020**.
- [9] J.-L. Calais, *Int. J. Quantum Chem.* **1993**, *47*, 101.
- [10] M. Marques, A. Rubio, E. K. Gross, K. Burke, F. Nogueira, C. A. Ullrich, *Time-Dependent Density Functional Theory*, Springer Science & Business Media, Dordrecht **2006**.
- [11] S. DiLuzio, V. Mdluli, T. U. Connell, J. Lewis, V. VanBenschoten, S. Bernhard, *J. Am. Chem. Soc.* **2021**, *143*, 1179.
- [12] J. Sun, J. Wu, T. Song, L. Hu, K. Shan, G. Chen, *J. Phys. Chem. A* **2014**, *118*, 9120.
- [13] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, *Chem. Rev.* **2021**, *121*, 9816.
- [14] L. Hu, X. Wang, L. Wong, G. Chen, *J. Chem. Phys.* **2003**, *119*, 11501.
- [15] M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, A. Vaitkus, *J. Cheminform* **2018**, *10*, 23.
- [16] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, *J. Phys. Mater.* **2019**, *2*, 032001.
- [17] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, *Nat Mater* **2016**, *15*, 1120.
- [18] M. A. B. Janai, K. L. Woon, C. S. Chan, *Org. Electron.* **2018**, *63*, 257.
- [19] Y. Zhao, C. Fu, L. Fu, Y. Liu, Z. Lu, X. Pu, *Mater Today Chem* **2021**, *22*, 100625.
- [20] L. Breiman, *Mach Learn* **2001**, *45*, 5.
- [21] I. Palit, C. K. Reddy, *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 1904.
- [22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, in *Advances in Neural Information Processing Systems* (Ed: T.-Y. Liu), Curran Associates, New York **2017**.
- [23] T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association For Computing Machinery, New York, NY **2016**, p. 785.
- [24] D. H. Wolpert, *Neural Netw* **1992**, *5*, 241.
- [25] K. Li, G. Cheng, C. Ma, X. Guan, W.-M. Kwok, Y. Chen, W. Lu, C.-M. Che, *Chem. Sci.* **2013**, *4*, 2630.
- [26] R. W. Kennard, L. A. Stone, *Technometrics* **1969**, *11*, 137.

- [27] R. K. H. Galvão, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, T. C. B. Saldanha, *Talanta* **2005**, *67*, 736.
- [28] F. Plasser, S. A. Bäßler, M. Wormit, A. Dreuw, *J. Chem. Phys.* **2014**, *141*, 024107.
- [29] F. Plasser, H. Lischka, *J. Chem. Theory Comput.* **2012**, *8*, 2777.
- [30] S. A. Mewes, J.-M. Mewes, A. Dreuw, F. Plasser, *Phys. Chem. Chem. Phys.* **2016**, *18*, 2548.
- [31] F. Plasser, M. Wormit, A. Dreuw, *J. Chem. Phys.* **2014**, *141*, 024106.
- [32] C. A. Guido, P. Cortona, B. Mennucci, C. Adamo, *J. Chem. Theory Comput.* **2013**, *9*, 3118.
- [33] M. J. G. Peach, P. Benfield, T. Helgaker, D. J. Tozer, *J. Chem. Phys.* **2008**, *128*, 044118.
- [34] C. Cortes, V. Vapnik, *Mach Learn* **1995**, *20*, 273.
- [35] T. Cover, P. Hart, *IEEE Trans Inf Theory* **1967**, *13*, 21.
- [36] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA **2012**.
- [37] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D. D. Cox, *Comput. Sci. Discov.* **2015**, *8*, 014008.
- [38] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16, Revision A.03*, Gaussian, Wallingford, CT **2016**.
- [39] E. J. Baerends, T. Ziegler, J. Autschbach, D. Bashford, A. Bérces, F. M. Bickelhaupt, C. Bo, P. M. Boerrigter, L. Cavallo, D. P. Chong, L. Deng, R. M. Dickson, D. E. Ellis, M. van Faassen, L. Fan, T. H. Fischer, C. F. Guerra, M. Franchini, A. Ghysels, A. Giammona, S. J. A. van Gisbergen, A. W. Götz, J. A. Groeneveld, O. V. Gritsenko, M. Grüning, S. Gusarov, F. E. Harris, P. van den Hoek, C. R. Jacob, H. Jacobsen, L. Jensen, J. W. Kaminski, G. van Kessel, F. Kootstra, A. Kovalenko, M. V. Krykunov, E. van Lenthe, D. A. McCormack, A. Michalak, M. Mitoraj, S. M. Morton, J. Neugebauer, V. P. Nicu, L. Noodleman, V. P. Osinga, S. Patchkovskii, M. Pavanello, P. H. T. Philipsen, D. Post, C. C. Pye, W. Ravenek, J. I. Rodríguez, P. Ros, P. R. T. Schipper, H. van Schoot, G. Schreckenbach, J. S. Seldenthuis, M. Seth, J. G. Snijders, M. Solà, M. Swart, D. Swerhone, G. te Velde, P. Vernooijs, L. Versluis, L. Visscher, O. Visser, F. Wang, T. A. Wesolowski, E. M. van Wezenbeek, G. Wiesenecker, S. K. Wolff, T. K. Woo, A. L. Yakovlev, *SCM, Vrije Universiteit, Amsterdam, The Netherlands* **2021**.
- [40] B. Miehlich, A. Savin, H. Stoll, H. Preuss, *Chem. Phys. Lett.* **1989**, *157*, 200.
- [41] J. M. L. Martin, A. Sundermann, *J. Chem. Phys.* **2001**, *114*, 3408.
- [42] R. Ditchfield, W. J. Hehre, J. A. Pople, *J. Chem. Phys.* **1971**, *54*, 724.
- [43] P. C. Hariharan, J. A. Pople, *Theor. Chim. Acta* **1973**, *28*, 213.
- [44] B. Mennucci, *WIREs Comput. Mol. Sci.* **2012**, *2*, 386.
- [45] F. Wang, T. Ziegler, *J. Chem. Phys.* **2005**, *123*, 154102.
- [46] E. Van Lenthe, E. J. Baerends, *J. Comput. Chem.* **2003**, *24*, 1142.
- [47] C. C. Pye, T. Ziegler, *Theor. Chem. Acc.* **1999**, *101*, 396.
- [48] I. Kim, W.-J. Son, Y.-S. Choi, A. Osipov, D. Lee, H. Lee, Y. Jung, J. Son, H. Choi, W. Sotoyama, A. Yakubovich, J. Shin, H. S. Lee, *J. Phys. Chem. C* **2019**, *123*, 11140.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: S. Wang, C. Yam, S. Chen, L. Hu, L. Li, F.-F. Hung, J. Fan, C.-M. Che, G. Chen, *J. Comput. Chem.* **2024**, *45*(6), 321. <https://doi.org/10.1002/jcc.27238>