Predicting Two-Photon Absorption Cross Sections with Experimental Accuracy Using Only Four Molecular Features Revealed by Interpretable Machine Learning

Yuming Su,‡ Yiheng Dai,‡ Yifan Zeng, Caiyun Wei, Yangtao Chen, Fuchun Ge, Peikun Zheng, Da Zhou,* Pavlo O. Dral,* Cheng Wang*

[a]Department of Chemistry, College of Chemistry and Chemical Engineering, iChem, Xiamen University, Xiamen 361005, P.R. China

[b]Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen 361005, P.R. China.

[c]Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Xiamen University, Xiamen 361005, P.R. China

[d]School of Mathematical Sciences and Fujian Provincial Key Laboratory of Mathematical Modeling and High-Performance Scientific Computation, Xiamen University, Xiamen 361005, P.R. China

Two-photon absorption, machine learning, rational design, high-throughput virtual screening.

ABSTRACT: Two-photon absorption has wide applications in bioimaging, photodynamic therapy, and three-dimensional printing. Designing molecules with a large two-photon absorption cross section (TPACS) is thus highly desirable for advancing these technologies. Here we used machine learning to analyze a TPACS dataset of ca. one thousand molecules collected from literature reports. We found that the length of the conjugated structure is the most important feature to determine the TPACS in a power law of ~1.8 order. The effect of donor and acceptor substitutions and structural coplanarity on the TPACS can be adequately addressed by another three features obtained by empirical rules. Combining these four molecular features with the experimental wavelength and solvent used in the measurement, we derived an interpretable model to predict molecular TPACS with an accuracy comparable to that of experimental measurements and common theoretical calculations. Our approach not only provides insight into the factors that are critical to TPACS but, as we demonstrate, also allows high-throughput screening of new TPA molecules.

Introduction

Two photon-absorption (TPA) is a nonlinear coherent process in which a molecule simultaneously absorbs two photons.[1, 2] TPA has been crucial in many technologies, including upconverted laser[3, 4], two-photon bioimaging[5-7], two-photon photodynamic therapy[8-10], and three-dimensional printing[11-13]. A range of molecules and materials with high TPA cross sections (TPACS, σ) were discovered,[14-17] as determined by the Z-scan[18] and two-photon excited fluorescence methods[19].

A general design principle is established for constructing TPA molecules: creating donor (D)-acceptor (A) push-pull structure together with a long π-conjugation in the molecule.[14, 17] Both features can lead to large transition dipole moments. In addition, quadrupolar D-π-A-π-D / A-π-D-π-A or multipolar DA_n / AD_n structures are also considered to be beneficial to obtaining large TPACS according to a Frenkel exciton model[17, 20]. However, these observations were made on a limited selection of systems and were not extensively tested considering all the experimental results obtained by the research community over the past years.

High-accuracy quantum chemical (QC) models can test the validity of these empirical design rules from the first principles.[21, 22] However, most QC methods still suffer from poor performance in predicting TPACS,[23-25] and the high-level QC calculations are usually expensive for examining many molecules with diverse structures and often, of considerably large size.

Machine learning (ML) can complement the QC methods to accelerate materials discovery[26-29]. Here we used an ML approach to study the structure-property relationship of TPA molecules based on reported experimental data containing TPACS of 856 molecules. In this study, we emphasize the interpretability of ML and aim to answer the following questions:

(1) Is there a quantitative relationship between the TPACS and the conjugation length of a molecule ?[14]

(2) Does a branched DA_n or AD_n structure have an edge over a simple D-A conjugation after eliminating the contribution of elongated conjugation length?[30]

(3) Are there other critical structural features beyond the donor-acceptor, conjugation, and multipolar to determine the TPACS?

Besides these scientific questions, our ML model also targets efficient high-throughput virtual screening (HTVS) to identify lead TPA compounds. This approach provides new opportunities for designing molecules with high TPACS.

Results

Dataset

An experimental dataset of 929 unique organic chromophores was collected from 275 literature reports (see the Supplementary Information, SI, Section 1). The dataset contains the TPACSs, the SMILES, names of the molecules, wavelengths of the TPA test, TPA measurement methods, solvents, and DOI number of the source publication. 443 molecules have only one TPACS value measured at a single wavelength, while the remaining 486 molecules have 2–11 TPACS values measured at different wavelengths (Figure 1a). The accuracy of the reported TPACS is difficult to check, but the level of accuracy is partially reflected by comparing the TPACS values of Rhodamine B at 798–802 nm from seven different sources[31], which are 52±41 GM (or 1.7±0.3 in logarithm) (Figure S1). We use lg(TPACS) in the following studies considering this level of accuracy. In addition, we also put all the molecular features for ML in the datasets, which are described in the next section.

The distribution of the molecular weights is shown in Figure 1c, while the distribution of the logarithm of TPACS per molecular weight is shown in Figure 1d; both are close to the normal distribution. The molecules in the dataset contain many elements, C, H, N, O, S, F, B, Cl, Br, P, Si, and I (in the order of their abundance), but the majority of molecules (564) contains only C, H, N, O (Figure 1b). The count of molecules measured in each solvent is shown in Figure 1e; many of the molecules (273) were measured in toluene, while altogether 21 different solvents were used. In order to avoid inaccuracy due to sparse data near the boundaries, only data points measured at wavelengths from 600 to 1100 nm were used, and the molecules containing P, Si, I elements were eliminated. A dataset containing 856 molecules were used in the following study.

Featurization of molecules

The wavelengths and solvents used in the TPA measurements were extracted as part of the features. The solvent information is encoded by three descriptors (ET(30), dielectric constant, and dipole moment). The ET(30)[32, 33] is defined by electronic transition energy of betaine 30 in different solvents to parameterize effect of solvent polarity. The information of the measurement methods is not used in the ML study, as many entries lack this information.

564 of the features come from molecular fragment fingerprint (MFF) featurization.[34] In MFF, molecular fragments were generated by the extended-connectivity fingerprints (ECFP) method using a radius of 436 supported by the Deepchem python toolkit.[35] A vector recording the appearance times of each fragment in a molecule[36] was then created (Figure 2). Note that this is different from the unhashed Morgan fingerprints, as the MFF counted fragment structures without considering their further linkage to other parts of the molecule, while the Morgan fingerprints contain this information. The MFF is thus a simplification of the Morgan fingerprints to fit into the needs of analyzing a small dataset. An additional 107 features were generated by the RDKit Python toolkit[36, 37], which provide geometrical and electronic structural information of the molecule.

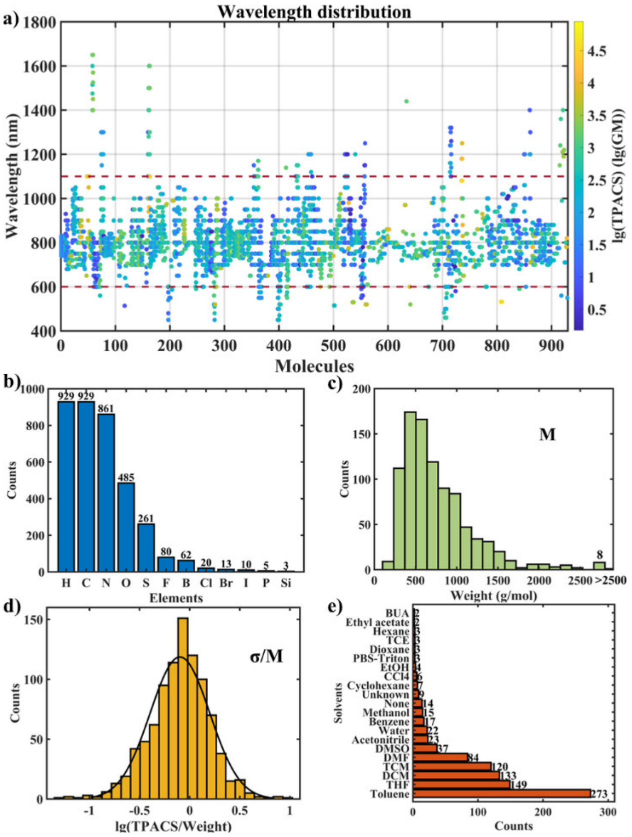


Figure 1. Dataset of TPACS of organic compounds. (a) Scatter plot of the distribution of wavelengths at which TPACS was measured; markers are color-coded according to lg(TPACS). Histograms of (b) elements contained in this dataset, (c) molecular weight, (d) the lg(TPACS) per molecular weight, and (e) solvents.

Given the well-known importance of conjugation for TPACS[38], we also created 21 conjugation features to describe the size, shape, and electronic properties of the conjugation structure (SI Section 2).

Overall, we obtained 696 initial features (Table 1), all of which have clear physical definitions and can be calculated very fast. More features are introduced in the following sections.

Feature selection

We assess the importance of these features using three ML models: Least Absolute Shrinkage and Selection Operator[39] (LASSO), Gradient Boosting Regression Tree[40] (GBRT), and Extreme Gradient Boosting[41] (XGBoost) regressor. In the ML process, the datasets were randomly split into the training set and the test set via cross-validation (CV), and the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score of the test sets were calculated to evaluate model performance.

For LASSO, importance of a given feature is manifested by the magnitude of the regression coefficient of the feature. For the GBRT and XGB Regressor, SHAP[42], a Python toolkit to calculate Shapley values, was implemented to generate more interpretable feature importance. We then combined the feature importance indexes of the three regressors

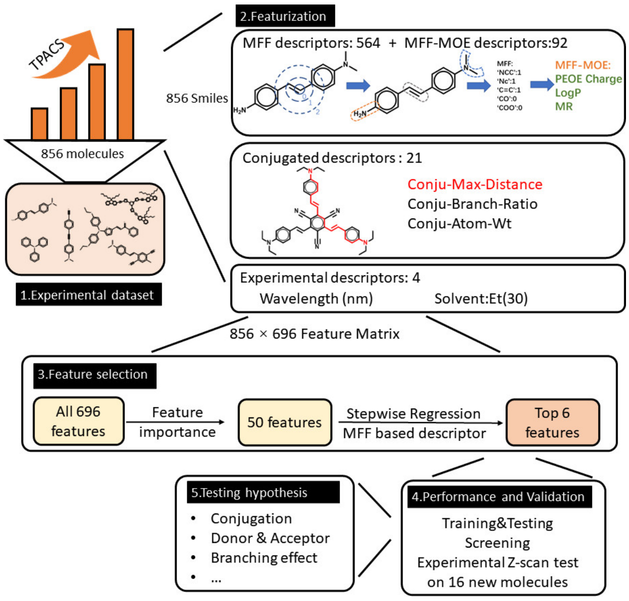


Figure 2. Featurization and feature selection. A scheme explaining molecular fragment fingerprint (MFF) featurization, conjugation features, experimental features, and the procedure of the feature selection. (MFF-MOE: MFF-based Molecular Operating Environment features. Conju-Max-Distance: The maximum conjugated length in one molecule. Conju-Branch-Ratio: a parameter to describe branching in the conjugated system. Conju-Atom-Wt: The atomically averaged weight in one molecule.)

(averaged over 240 CV runs) into a weighted one (SI Section 3, Figures S2), which was used to remove the least important features one at a time from the feature matrix.

Through this deletion process, we obtained 50 features that can retain the performance of the models (Table S1, Figure 3a), as shown by the scatter plots of true vs predicted values of testing sets (Figures 3b).

As there were still highly correlated features within the 50 features shown by correlation coefficients matrix (Figure S3), we further reduced the number of features by stepwise regression. The most important feature among the 50 ones was the conjugation length that is measured by the number of bonds linking the farthest atom pair in a conjugation system ("Conju-Max-Distance"). We then calculated the performance gain after adding each of the rest 49 features using the XGBoost model. The feature providing the highest performance gain was added to the selected feature set. This procedure was repeated to select the third feature, and so on. We found that a minimum of an additional 9 features plus the "Conju-Max-Distance" can retain the performance of the XGBoost model: "MaxPartialCharge", "MaxAbsPartialCharge", "SMR_VSA10", "VSA_EState3", "MaxEStateIndex", "VSA_Estate1", "VSA_EState2", "Wavelength (Exp nm)", "ET(30) (Solvent)".

Table 1. The features used in this study.

Name	Number	Description
Initial features for model screening and feature selection (696)		
MFF	564	Describing molecular structure and functional groups
RDKit	107	Describing molecular shape and electronic structure
Conjugation	21	Describing the properties of conjugation structure
Solvent	3	Describing the polarity of solvents
Wavelength	1	Experimental TPA wavelength in nm
Adding MFF-based features to enhance interpretability		
MFF-MOE features	80	Atom-attributed properties summed up to MFF
Other features for SHAP analysis		
DAratio	1	Distance between donor and acceptor divided by conjugation length

The wavelength and solvent index are molecule-independent features that are related to the experimental measurements. The other seven features are all "Molecular Operating Environment" (MOE) features describing the local environment of atoms in a molecule. Many of these MOE features are additive. As we would like to attribute the molecular properties to fragments containing functional groups that are familiar to chemists, we established MFF-based MOE features (MFF-MOE) by a simple summation to replace the atom-based ones, including the "PEOE Charge"[43], "LogP"[44] and "MR". "PEOE Charge" is obtained by summing up the Gasteiger charges of atoms in an MFF fragment. LogP is the logarithm of oil (octanol)–water partition coefficient of a molecule. The summation of atomic attribution of LogP to MFF can identify polar groups in the molecule. Similarly, the MR is the polarizability of the molecule determined by molar refractivity, and the summation of its atomic attribution to the MFF level can describe polarizability of a molecular fragment.

After adding a series of these new MFF-MOE features to replace the seven atomic MOE features, we obtained a new feature matrix containing 94 features (Table S2). To our surprise, after the stepwise regression, we obtained a feature set with only 6 features to give quite good performance of the XGBoost model, and only four of them are molecule-based features while the other two are the measurement wavelength and solvent feature. Besides the "Conju-Max-Distance", "Wavelength (Exp nm)", and "ET(30) (Solvent)", the newly selected MFF-MOE features are "PEOE-Charge-Max", "LogP-Min", and "MR-Max".

Table 2. Performance of 10 regressors on different feature matrices.

Performance / Regressor	[856 × 696]			[856 × 6]		
	MSEa	MAE	R2	MSE	MAE	R2
AdaBoost	0.32	0.44	0.40	0.36	0.48	0.31
DNN	0.30	0.40	0.44	0.36	0.45	0.33
Decision Tree	0.41	0.48	0.22	0.37	0.46	0.29
ElasticNet	0.30	0.41	0.43	0.38	0.48	0.28
GBRT	0.22	0.34	0.59	0.23	0.37	0.56
LASSO	0.30	0.41	0.43	0.39	0.48	0.27
MLPRegressor	0.33	0.42	0.37	0.40	0.50	0.23
k-nearest neighbor	0.39	0.47	0.25	0.42	0.49	0.19
Random Forest	0.23	0.34	0.56	0.23	0.35	0.57
XGBoost	0.22	0.33	0.59	0.21	0.35	0.58

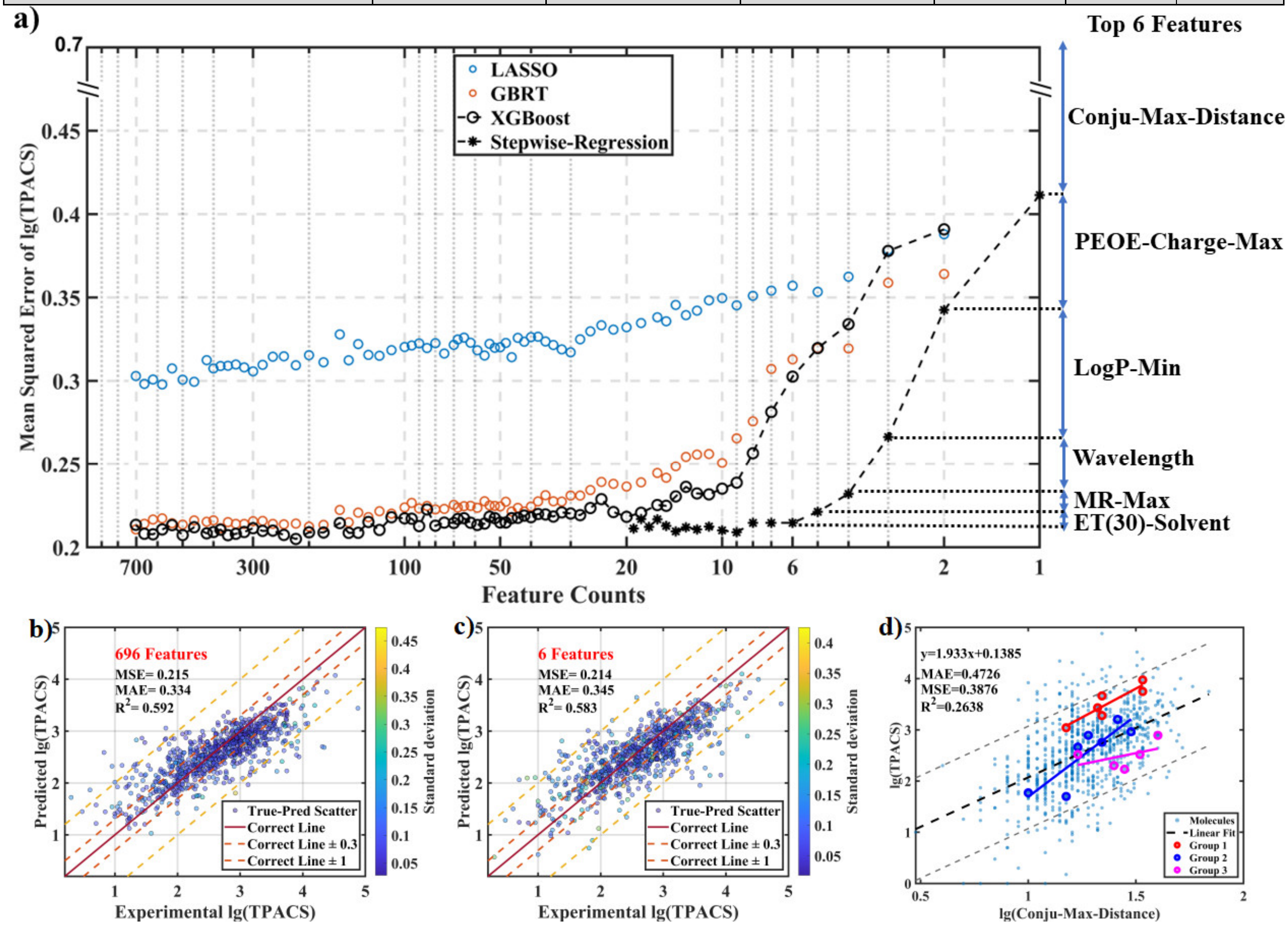


Figure 3. Model performance during feature selection procedure. (a) Mean squared error (MSE) against feature selection procedure: feature importance-based feature selection in LASSO, GBRT and XGBoost were denoted as blue, red, and black circles,

respectively; black stars represent the stepwise regression; top 6 features selected by stepwise regression were shown on the right of the y-axis; the number of features, the metrics and the TPACS are all shown in log scale. Scatter plot of models using (b) [856 (number of molecules) × 696 (number of features)] and (c) [856 × 6] feature matrices and the XGBoost regressor: the standard deviations of the predicted values in the 240 CV runs of the model is represented by the color axis in log scale. (d) The parity plot of experimental lg(TPACS) *vs.* lg(Conju-Max-Distance): three groups of structurally related compounds of different conjugation lengths in the dataset are also highlighted by red, blue, and purple circles; note that it is difficult to find examples of homologous structures, and these selected series also differ in functional groups.

Performance of machine learning models

240 splits of training and testing sets were randomly generated to evaluate the model performances with a train-test ratio of 85:15 (728 samples for training and 128 samples for testing). Table 2 listed the average MSE, MAE, R2 scores of the testing sets using the full feature matrix [856 × 696] and the selected feature matrix [856 × 6] with a bunch of different ML models (Table S3). The MAE value representing the error of the prediction was as low as 0.33 in lg(TPACS) units, which corresponds to an accuracy within a factor of 2. The true-predict scatter plots (Figures 3b&3c, S4&5) further confirm this performance. This level of accuracy is already comparable to the accuracy of experimental measurement.

Meanwhile, theoretical calculations of TPACS suffer from large uncertainty[23-25, 45, 46]. Even comparing the popular density functional theory (DFT) results to the benchmark calculation by coupled cluster (CC) high-level QC method gave MAE > 0.334 in logarithm (Figure S6)[23]. Our simple ML model with only four molecular features thus has comparable accuracy to that of commonly used DFT methods

Interpretation of the machine-learning model

We used the SHAP value[47] as a guide to interpret the ML model (SI Section 5). The SHAP value measures in the ML model how a specific feature contributes to the predicted TPACS of each sample. The SHAP values of different features of one sample sum up to its TPACS subtracting the mean TPACS. For a given feature, a plot of SHAP values against the feature values of different samples (SHAP plot) maps out the contribution of the feature in determining TPACS (Figure 4). To analyze other feature of interest that is not included in the selected 6 features, we added the feature to the feature matrix and refit the model [856 × (6+1)] to calculate its SHAP value.

These SHAP plots allow us to test established concepts of the TPA structure-property relationship. We found that many of the concepts are consistent with the experimental statistics, but a few of them are not strongly supported.

1. Conjugation length *vs.* conjugation area

It has long been noticed that a larger molecule with a larger conjugated structure has a higher TPACS. Noticeably, the ML model selected the conjugation length rather than the conjugation area (Conju-Stru-VSA) as the most critical feature. Conju-Stru-VSA is poorly related to the TPACS (Figure S7).

The area of a conjugated system is also closely related to its molecular weight. Practically, in many applications, the specific TPACS per molecular weight is of interest. It is thus important to know whether the TPACS linearly depends on the molecular weight. The plot of lg(TPACS) against the logarithm of either the whole molecular weight (Full-Wt) or the weight of the conjugated systems in the molecule (conju-Wt) (Figure S8) showed a weak correlation.

2. Is there a quantitative relationship between the TPACS and the conjugated length of a molecule?

From the SHAP plot of the 'Conju-Max-Distance' (Figure 4a), we observed a linear correlation between the logarithm of this feature and the SHAP value with a slope of 1.79 ± 0.05. As the SHAP value is logarithm to the TPACS, this slope corresponds to a power law of TPACS depending on the conjugation length:

$$\text{TPACS} \propto (\text{Conju-Max-Distance})^{1.79 \pm 0.05}$$

This slope of 1.79 is roughly consistent with the linear fitting of lg(TPACS) against lg(Conju-Max-Distance) that happens to be 1.9±0.2 with a much larger error (Figure 3d). Three groups of structurally related compounds of different conjugation lengths in the dataset (Figure S9) revealed similar trend (Figure 3d). This slope is also confirmed by another data analysis method: accumulated local effects (SI Section 6 and Figure S10a). The SHAP method thus helps us to isolate the contribution of conjugation length and extracts the first quantitative relationship between the conjugation length and the TPACS.

Beyond the statistical analyses, we also try to rationalize such a dependence based on physical models. A simple model of conjugated parallel p-orbitals to form a linear π-system with alternative double and single bonds showed that the lg(TPACS) is linear to the lg(Conju-Max-Distance) with a slope close to 1.7 in a reasonable conjugation length range (SI Section 7 and Figures S10b, S11). More accurate time-dependent density functional theory (TDDFT) calculations[48] on a series of molecules of different conjugation lengths (SI Section 8, Table S5 and Figure S10c) gave a lg(TPACS)-lg(Conju-Max-Distance) slope of 2.4.

3. The degree of conjugation

The "MR-Max" uses molar refractivity to describe polarizability of a fragment. As the "MR-Max" is correlated to the conjugation length, we used principal component analysis (SI Section 9, Table S6, Figure S12) to remove interference from the latter and found that the "MR-Max" possibly manifests the degree of conjugation.

The conjugated C=C bonds, triphenylamine groups increase the SHAP value of "MR-Max", while a single bond connection between two aromatic rings or other substructure causingnonplanarity of the conjugation system has a negative effect (SI Section 10 and Figures S13c, d & S14). Some

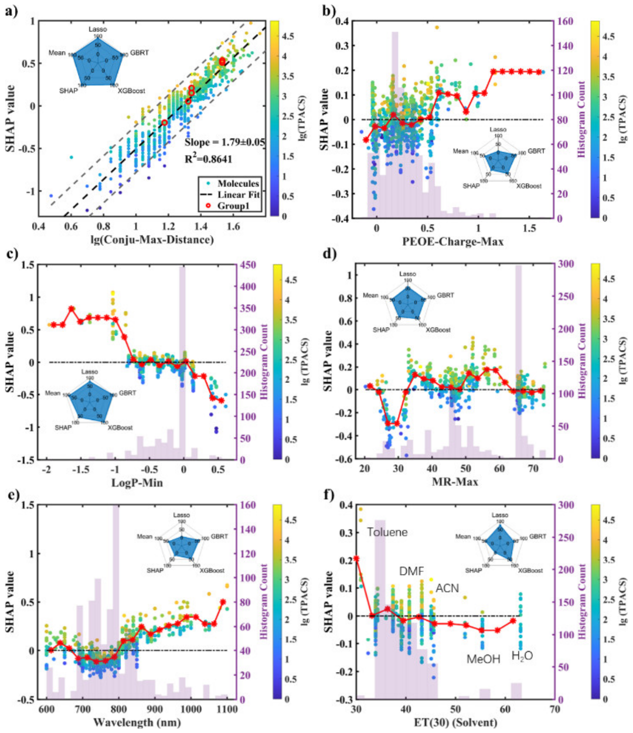


Figure 4. Chemical information extracted from machine learning models. (a) SHAP analysis of lg(Conju-Max-Distance): the linear fitting gives: SHAP value = lg(Conju-Max-Distance)×1.79 − 2.29, and a group of structurally related compounds selected from Figure S9 is highlighted in red circle. (b)–(d) SHAP analysis of MFF-MOE features. (f) SHAP analysis of Wavelength (Exp nm). (f) SHAP analysis of ET(30). The pentagon spider insets are showing five feature importance indexes: the normalized LASSO coefficients ([856 × 696] model), the GBRT feature importance ([856 × 696] model), the XGBoost feature importance ([856 × 696] model), the sum of SHAP values based on the XGBoost model ([856 × 6] model), and the mean value of the above four indexes (Table S4). The histogram of every figure shows the distribution of features.

heteroatoms in the conjugated system like an azo linkage between benzene rings also seem to have a negative effect.

The "MR-Max" using polarizability as a probe is thus complementary to the conjugation length to describe the degree of conjugation of the system.

4. How to quantify the impact of donor and acceptor substitution groups to TPACS?

Donor and acceptor substituents in the conjugated structure are critical to the TPA in a Donor–π-Acceptor design. The selected "LogP-Min" feature can mark the existence of highly polar groups on the molecules. These groups are usually also strongly electron-donating or electron-withdrawing groups (Figure S13b). The SHAP plot showed that the more negative this parameter (the more polar the group), the higher the TPACS, which is consistent with the push-pull design principle (Figure 4c).

However, polarity alone cannot adequately describe the electronic property of a functional group. "PEOE-Charge-Max" supplements the description by identifying positively charged conjugated carbon backbone that is connected to strong electron-withdrawing group (Figure S13a), as shown by the SHAP plot that adds correction to the positively valued region (positively charged backbone) (Figure 4b).

5. Is multipolar DA_n or AD_n structure from branching of the conjugated system beneficial for TPA?

We considered multipolar DA_n or AD_n branched structure and quadrupolar D-π-A-π-D / A-π-D-π-A linear structure by the features of "Conju-Branch-Ratio" and "DAratio", respectively. The Conju-Branch-Ratio is only weakly correlated to the conjugation length and adequately addresses the branching of a conjugation system in a multipolar structure (SI Section 11 and Figure S15a, S16). However, the absolute SHAP values of the Conju-Branch-Ratio are mostly smaller than 0.05, indicating that it only has a minor influence on the TPACS. No higher-order contribution of Conju-Branch-Ratio together with the conjugation length was observed either (Figure S15e).

Similarly, the SHAP plot of the DAratio (distance between the donor and acceptor divided by the conjugation length) showed absolute values mostly smaller than 0.1 (Figure S15f), suggesting a small effect. Moreover, the positive SHAP value at DAratio > 0.5 is against a beneficial effect of the quadrupolar structure, as the DAratio closer to 1 corresponds to a dipolar D-A structure rather than a quadrupolar D-A-D or A-D-A structure.

These statistical analyses of the multipolar or quadrupolar structures thus contradict the conventional wisdom about the importance of them in obtaining high TPACS. The observation of high TPACS in multipolar or quadrupolar molecules can be mainly attributed to their elongated conjugation length.

We put more analyses on other features including aliphatic chain, testing method, and solvent polarity in the SI (SI Section 12 and Figures S17-18).

Screening

25006 commercially available molecules from Innochem website (inno-chem.com.cn) with more than 20 non-H atoms and at least one aromatic ring were collected for screening. The target TPA wavelength was set to 800 nm. The trained XGBoost model was used for predicting TPACS values (SI Section 13). Synthetic Accessibility Score (SAS)[49] and Synthetic Bayesian Accessibility (SYBA)[50] scores were also calculated to estimate their synthetic accessibility (Figure S19a-c).

Experimental validation

To test the predictive power of the ML model, we chose some of the commercially available molecules and measured their TPACS by the Z-scan technique to compare with the values predicted by the ML model. Two-photon absorption spectra of 16 molecules are shown in Figure S20. The ML model gives reasonable predictions of their TPACS as compared to the measured values at their peak wavelength (Figure S19d), which provides an independent validation of the accuracy of our approach.

Conclusion

We obtained a simple and interpretable model to predict two-photon absorption cross section (TPACS) of different chromophores based on experimental data from literature. Despite of containing only four molecule-based features, the model achieves a predictive accuracy comparable to both the experimental measurements and the popular density functional theory calculations. The model identifies the conjugation length as the most critical feature and gives the first quantitative relationship between the TPACS and the conjugation length. Based on this model, we also tested several popular observations in the field of two-photon absorption research. To our surprise, we found that a widely practiced approach to design DA_n or AD_n multipolar structure does not enhance the TPACS beyond the effect of conjugation lengthening. We envision that this simple ML model can allow fast screening of databases to accelerate the development of high-performance organic non-linear optical materials.

ASSOCIATED CONTENT

This material is available free of charge via the Internet at http://pubs.acs.org.

Detailed information for features used by machine learning and feature selection algorithm; dependencies, performance and parameters of 10 regressors; further validation of the quantitative dependence between Conju-Max-Distance and TPACS; screening and experimental validation of 16 new molecules. (PDF)

Raw collected data and feature matrix of machine learning; original codes of feature matrix generation, feature importance-based feature selection, stepwise regression, SHAP analysis. (ZIP)

AUTHOR INFORMATION

Corresponding Author

Cheng Wang – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China; orcid.org/0000-0002-7906-8061; Email: wangchengxmu@xmu.edu.cn

Pavlo O. Dral – State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China;

orcid.org/00000002-2975-9876; Email: dral@xmu.edu.cn; drdral.com

Da Zhou – School of Mathematical Sciences and Fujian Provincial Key Laboratory of Mathematical Modeling and High-Performance Scientific Computation, Xiamen University, Xiamen 361005, P. R. China; Email: zhouda@xmu.edu.cn

Authors

Yuming Su – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China

Yiheng Dai – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China

Yifan Zeng – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China

Caiyun Wei – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China

Yangtao Chen – State Key Laboratory of Physical Chemistry of Solid Surfaces, iChem, Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China

Fuchun Ge – State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China;

Peikun Zheng – State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China;

Author Contributions

‡Y.S. and Y.D. contributed equally to this work. contributed equally.

Funding Sources

C. W. acknowledges funding from the National Key R&D Program of China (2021YFA1502500), the National Natural Science Foundation of China (22125502, 21721001)

Y. D. and Y.Z. acknowledges funding from NFFTBS (No. J1310024)

P.O.D. acknowledges funding by the National Natural Science Foundation of China (No. 22003051), the Fundamental Research Funds for the Central Universities (No. 20720210092), and via the Lab project of the State Key Laboratory of Physical Chemistry of Solid Surfaces.

D.Z. acknowledges funding from the National Natural Science Foundation of China (11971405).

Notes

The authors declare no competing financial interest.

REFERENCES

1. Göppert-Mayer, M., Über Elementarakte mit zwei Quantensprüngen. *Annalen der Physik* 1931, *401* (3), 273-294.

2. Kaiser, W.; Garrett, C. G. B., Two-Photon Excitation in CaF_2: Eu^{2+}. *Physical Review Letters* 1961, *7* (6), 229-231.

3. Xu, Y.; Chen, Q.; Zhang, C.; Wang, R.; Wu, H.; Zhang, X.; Xing, G.; Yu, W. W.; Wang, X.; Zhang, Y.; Xiao, M., Two-Photon-Pumped Perovskite Semiconductor Nanocrystal Lasers. *Journal of the American Chemical Society* 2016, *138* (11), 3761-3768.

4. Sivaramakrishnan, S.; Muthukumar, V. S.; Sivasankara Sai, S.; Venkataramaniah, K.; Reppert, J.; Rao, A. M.; Anija, M.; Philip, R.; Kuthirummal, N., Nonlinear optical scattering and absorption in bismuth nanorod suspensions. *Applied Physics Letters* 2007, *91* (9), 093104.

5. Jiang, M.; Gu, X.; Lam, J. W. Y.; Zhang, Y.; Kwok, R. T. K.; Wong, K. S.; Tang, B. Z., Two-photon AIE bio-probe with large Stokes shift for specific imaging of lipid droplets. *Chemical Science* 2017, *8* (8), 5440-5446.

6. Yu, C.; Schimelman, J.; Wang, P.; Miller, K. L.; Ma, X.; You, S.; Guan, J.; Sun, B.; Zhu, W.; Chen, S., Photopolymerizable Biomaterials and Light-Based 3D Printing Strategies for Biomedical Applications. *Chem Rev* 2020, *120* (19), 10695-10743.

7. Kim, K. H.; Singha, S.; Jun, Y. W.; Reo, Y. J.; Kim, H. R.; Ryu, H. G.; Bhunia, S.; Ahn, K. H., Far-Red/Near-Infrared Emitting, Two-Photon Absorbing, and Bio-Stable Amino-Si-Pyronin Dyes. *Chemical Science* 2019.

8. Xu, L.; Zhang, J.; Yin, L.; Long, X.; Zhang, W.; Zhang, Q., Recent progress in efficient organic two-photon dyes for fluorescence imaging and photodynamic therapy. *J Mater Chem C* 2020, *8* (19), 6342-6349.

9. Sun, Z.; Zhang, L.-P.; Wu, F.; Zhao, Y., Photosensitizers for Two-Photon Excited Photodynamic Therapy. *Adv. Funct. Mater.* 2017, *27* (48), 1704079.

10. Shen, Y.; Shuhendler, A. J.; Ye, D.; Xu, J.-J.; Chen, H.-Y., Two-photon excitation nanoparticles for photodynamic therapy. *Chem. Soc. Rev.* 2016, *45* (24), 6725-6741.

11. Lay, C. L.; Koh, C. S. L.; Lee, Y. H.; Phan-Quang, G. C.; Sim, H. Y. F.; Leong, S. X.; Han, X.; Phang, I. Y.; Ling, X. Y., Two-Photon-Assisted Polymerization and Reduction: Emerging Formulations and Applications. *ACS Appl Mater Interfaces* 2020, *12* (9), 10061-10079.

12. Xing, J.-F.; Zheng, M.-L.; Duan, X.-M., Two-photon polymerization microfabrication of hydrogels: an advanced 3D printing technology for tissue engineering and drug delivery. *Chem. Soc. Rev.* 2015, *44* (15), 5031-5039.

13. Taguchi, A.; Nakayama, A.; Oketani, R.; Kawata, S.; Fujita, K., Multiphoton-Excited Deep-Ultraviolet Photolithography for 3D Nanofabrication. *ACS Applied Nano Materials* 2020, *3* (11), 11434-11441.

14. Pawlicki, M.; Collins, H. A.; Denning, R. G.; Anderson, H. L., Two-Photon Absorption and the Design of Two-Photon Dyes. *Angew. Chem. Int. Ed.* 2009, *48* (18), 3244-3266.

15. Kang, D.; Zhu, S.; Liu, D.; Cao, S.; Sun, M., One- and Two-Photon Absorption: Physical Principle and Applications. *Chem Rec* 2020, *20* (9), 894-911.

16. He, G. S.; Tan, L.-S.; Zheng, Q.; Prasad, P. N., Multiphoton Absorbing Materials: Molecular Designs, Characterizations, and Applications. *Chem. Rev.* 2008, *108* (4), 1245-1330.

17. Terenziani, F.; Katan, C.; Badaeva, E.; Tretiak, S.; Blanchard-Desce, M., Enhanced Two-Photon Absorption of Organic Chromophores: Theoretical and Experimental Assessments. *Advanced Materials* 2008, *20* (24), 4641-4678.

18. Pálfalvi, L.; Tóth, B. C.; Almási, G.; Fülöp, J. A.; Hebling, J., A general Z-scan theory. *Applied Physics B* 2009, *97* (3), 679.

19. Xu, C.; Webb, W. W., Measurement of two-photon excitation cross sections of molecular fluorophores with data from 690 to 1050 nm. *J. Opt. Soc. Am. B* 1996, *13* (3), 481-491.

20. Xu, L.; Lin, W.; Huang, B.; Zhang, J.; Long, X.; Zhang, W.; Zhang, Q., The design strategies and applications for organic multi-branched two-photon absorption chromophores with novel cores and branches: a recent review. *J Mater Chem C* 2021, *9* (5), 1520-1536.

21. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning: Generative models for matter engineering. *Science* 2018, *361* (6400), 360-365.

22. Zaleśny, R.; Murugan, N. A.; Tian, G.; Medved', M.; Ågren, H., First-Principles Simulations of One- and Two-Photon Absorption Band Shapes of the Bis(BF2) Core Complex. *The Journal of Physical Chemistry B* 2016, *120* (9), 2323-2332.

23. Chołuj, M.; Alam, M. M.; Beerepoot, M. T. P.; Sitkiewicz, S. P.; Matito, E.; Ruud, K.; Zaleśny, R., Choosing Bad versus Worse: Predictions of Two-Photon-Absorption Strengths Based on Popular Density Functional Approximations. *Journal of Chemical Theory and Computation* 2022, *18* (2), 1046-1060.

24. Beerepoot, M. T. P.; Alam, M. M.; Bednarska, J.; Bartkowiak, W.; Ruud, K.; Zalesny, R., Benchmarking the Performance of Exchange-Correlation Functionals for Predicting Two-Photon Absorption Strengths. *J Chem Theory Comput* 2018, *14* (7), 3677-3685.

25. Beerepoot, M. T. P.; Friese, D. H.; List, N. H.; Kongsted, J.; Ruud, K., Benchmarking two-photon absorption cross sections: performance of CC2 and CAM-B3LYP. *Physical Chemistry Chemical Physics* 2015, *17* (29), 19306-19314.

26. Ryu, B.; Wang, L.; Pu, H.; Chan, M. K. Y.; Chen, J., Understanding, discovery, and synthesis of 2D materials enabled by machine learning. *Chem. Soc. Rev.* 2022.

27. Westermayr, J.; Marquetand, P., Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* 2021, *121* (16), 9873-9926.

28. Xu, S.; Liu, X.; Cai, P.; Li, J.; Wang, X.; Liu, B., Machine-Learning-Assisted Accurate Prediction of Molecular Optical Properties upon Aggregation. *Advanced Science* 2022, *9* (2), 2101074.

29. Zhang, Q.; Zheng, Y. J.; Sun, W.; Ou, Z.; Odunmbaku, O.; Li, M.; Chen, S.; Zhou, Y.; Li, J.; Qin, B.; Sun, K., High-Efficiency Non-Fullerene Acceptors Developed by Machine Learning and Quantum Chemistry. *Advanced Science* 2022, *n/a* (n/a), 2104742.

30. Kogej, T.; Beljonne, D.; Meyers, F.; Perry, J. W.; Marder, S. R.; Brédas, J. L., Mechanisms for enhancement of two-photon absorption in donor–acceptor conjugated chromophores. *Chemical Physics Letters* 1998, *298* (1), 1-6.

31. Makarov, N. S.; Drobizhev, M.; Rebane, A., Two-photon absorption standards in the 550–1600 nm excitation wavelength range. *Optics Express* 2008, *16* (6), 4029-4047.

32. Cerón-Carrasco, J. P.; Jacquemin, D.; Laurence, C.; Planchat, A.; Reichardt, C.; Sraïdi, K., Solvent polarity scales: determination of new ET(30) values for 84 organic solvents. *J. Phys. Org. Chem.* 2014, *27* (6), 512-518.

33. Solvent Effects on the Absorption Spectra of Organic Compounds. In *Solvents and Solvent Effects in Organic Chemistry*, 2010; pp 359-424.

34. Guo, Y.; He, X.; Su, Y.; Dai, Y.; Xie, M.; Yang, S.; Chen, J.; Wang, K.; Zhou, D.; Wang, C., Machine-Learning-Guided Discovery and Optimization of Additives in Preparing Cu Catalysts for CO2 Reduction. *J. Am. Chem. Soc.* 2021, *143* (15), 5755-5762.

35. Wu, Z.; Ramsundar, B.; Feinberg, Evan N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 2018, *9* (2), 513-530.

36. Labute, P., A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* 2000, *18* (4), 464-477.

37. Landrum, G. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org/.

38. Ohta, K.; Yamada, S.; Kamada, K.; Slepkov, A. D.; Hegmann, F. A.; Tykwinski, R. R.; Shirtcliff, L. D.; Haley, M. M.; Sałek, P.; Gel'mukhanov, F.; Ågren, H., Two-Photon Absorption Properties of Two-Dimensional π-Conjugated Chromophores: Combined Experimental and Theoretical Study. *The Journal of Physical Chemistry A* 2011, *115* (2), 105-117.

39. Kim, S.-J.; Koh, K.; Lustig, M.; Boyd, S.; Gorinevsky, D., An Interior-Point Method for Large-Scale -Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing* 2007, *1* (4), 606-617.

40. Ye, J.; Chow, J.-H.; Chen, J.; Zheng, Z., Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management*, Association for Computing Machinery: Hong Kong, China, 2009; pp 2061–2064.

41. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785–794.

42. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I., From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020, *2* (1), 56-67.

43. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 1980, *36* (22), 3219-3228.

44. Wildman, S. A.; Crippen, G. M., Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* 1999, *39*, 868-873.

45. Esipova, T. V.; Rivera-Jacquez, H. J.; Weber, B.; Masunov, A. E.; Vinogradov, S. A., Two-Photon Absorbing Phosphorescent Metalloporphyrins: Effects of pi-Extension and Peripheral Substitution. *J. Am. Chem. Soc.* 2016, *138* (48), 15648-15662.

46. Salek, P.; Vahtras, O.; Guo, J. D.; Luo, Y.; Helgaker, T.; Agren, H., Calculations of two-photon absorption cross sections by means of density-functional theory. *Chemical Physics Letters* 2003, *374* (5-6), 446-452.

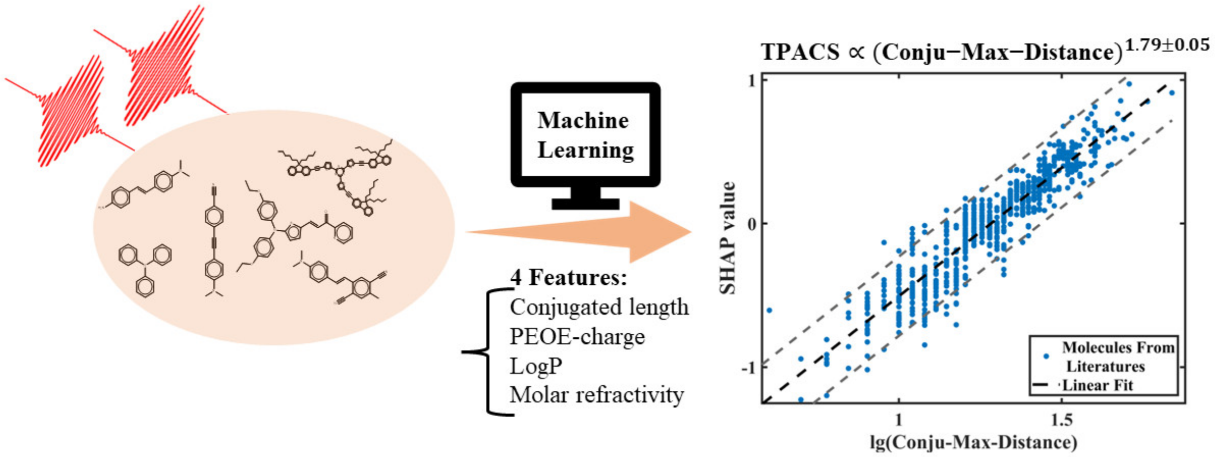
47. Lundberg, S. M.; Lee, S.-I., A Unified Approach to Interpreting Model Predictions. *ArXiv* 2017, *abs/1705.07874*.

48. Lu, T.; Chen, F., Multiwfn: a multifunctional wavefunction analyzer. *J. Comput. Chem.* 2012, *33* (5), 580-92.

49. Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* 2009, *1* (1), 8.

50. Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D., SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics* 2020, *12* (1).

SYNOPSIS TOC


Machine
Learning
4 Features:
Conjugated length
PEOE-charge
LogP
Molar refractivity
TPACS ∝ (Conju−Max−Distance)^1.79±0.05
SHAP value
lg(Conju-Max-Distance)
Molecules From Literatures
Linear Fit