Journal of Cheminformatics

RESEARCH

Open Access

Predictive modeling of visible-light azo-photoswitches' properties using structural features

Said Byadi[1], P. K. Hashim[2,3] and Pavel Sidorov[1,4*]

Abstract

In this manuscript we present the strategy for modeling photoswitch properties (maximum absorption wavelength and thermal half-life of photoisomers) of visible-light azo-photoswitches using structural data. We compile a comprehensive data set from literature sources and perform a rigorous benchmark to select the best feature type and modeling approach. The fragment counts have demonstrated the best performance in the benchmark for both properties. We validate the models in cross-validation and on an external set. The predictions of absorption wavelengths for this set are highly accurate; on the other hand, the model for thermal half-life is less reliable, likely due to the modest size of the data set related to half-life of photoisomers, although consensus modeling approach allows to improve the predictivity. We also provide an interpretation of the modeling results using ColorAtom approach and the insights into the chemical space covered by the data set.

Scientific contribution The paper provides a machine learning approach based only on structural features to predict two important photoswitch properties. Unlike previous studies, we do not use any quantum chemical features which accelerates the modeling procedure, while the accuracy of models remains high. Moreover, the fragment counts offer unique approach to model interpretation that is useful for rational design of photoswitches with desired properties.

Keywords Photoswitches, Azobenzene, Quantitative structure–property relationship, Machine learning, Molecular descriptors

Introduction

Photoswitches are a category of compounds whose chemical structures and properties can be changed by light irradiation [1]. The photochemically generated species (photoisomers) are less stable and hence can be reverted to more stable species thermally or via a reverse photoisomerization. In the field of photoswitches, the wavelength with the maximum light absorption (λ_{max}) and the thermal half-life of a metastable photoisomer ($t_{1/2}$) decide their use in various applications in material science and biology [2]. For instance, the well-known azobenzene photoswitch has λ_{max} of 320 nm and hence requires UV light for light-induced changes (*trans* to *cis* photoisomerization). However, the UV light is harmful to the living beings as it damages biological cells and its penetration

*Correspondence:
Pavel Sidorov
pavel.sidorov@icredd.hokudai.ac.jp
[1] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan
[2] Research Institute for Electronic Science, Hokkaido University, Kita 20, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan
[3] Graduate School of Life Science, Hokkaido University, Kita 10, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0810, Japan
[4] List Sustainable Digital Transformation Catalyst Collaboration Research Platform, Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan

ability to the deep tissue is very low compared to green and yellow lights in the visible region. Thus, for biological applications such as photopharmacology, where the drug action is controlled by the light, visible-light responsive photoswitches are required [3]. Moreover, photoswitches having long λ_{max} and long $t_{1/2}$ of meta-stable photoisomers are useful for various applications. However, some studies [4, 5] show that introducing particular substitution patterns may lead to inverse proportionality of λ_{max} and $t_{1/2}$, thus, the design of visible-light photoswitches would require careful consideration of structural patterns to avoid this effect. So far, experimentalists heavily rely on the time-consuming density functional theory calculations to estimate the λ_{max} and $t_{1/2}$ properties of photoswitches [6–8]. We envisioned to develop a machine learning model to accurately predict the λ_{max} and $t_{1/2}$ properties focusing on the visible-light photoswitches.

Previously constructed machine learning (ML) models were focused on the broad category of azo-dyes with the prediction of single parameter such as λ_{max} [9]. Another important parameter of photoswitches is the photoisomerization quantum yield, which can be predicted by a neural network machine learning model [10]. Very recently, ML-based model also developed for the prediction of thermal half-lives of azobenzene derivatives, though the authors used the data obtained in quantum mechanical calculations as a training set [11]. In the category of photoswitches, prediction of multiple parameters such as λ_{max} at two electronic excitations ($\pi \rightarrow \pi^*$ transition of *trans* and *cis* isomers) has been investigated using a multi-output Gaussian process model and molecular fingerprints (FP) [12]. Such structural descriptors derived directly from 2D representations of molecules allow for a significant acceleration of the modeling process. Indeed, the authors report to have achieved comparable predictive accuracy to time-dependent density functional theory (TD-DFT) calculation with reduced inference time on a curated data set consisting of the λ_{max} of *trans* and *cis* isomers of various photoswitches.

However, to the best of our knowledge, accurate prediction of two different parameters such as λ_{max} and $t_{1/2}$ properties in a standalone machine learning model based on structural features has not been achieved. In this paper, we present a computational study on visible-light photoswitches, employing quantitative structure–property relationship (QSPR) modeling to predict two crucial properties, λ_{max} and $t_{1/2}$ (Fig. 1). We compile a comprehensive data set from literature sources and perform a rigorous benchmark study in order to select the structural descriptors that would be the most appropriate to predict these properties. The accuracy of obtained ML models is also validated externally on a pre-selected test set. We also discuss the interpretability and the applicability domains of these models, as well as the challenges the endeavors of such modeling face.

Data and methods

Data collection and curation

The basis for the predictive models in this work is literature data. The data on the absorption properties is the

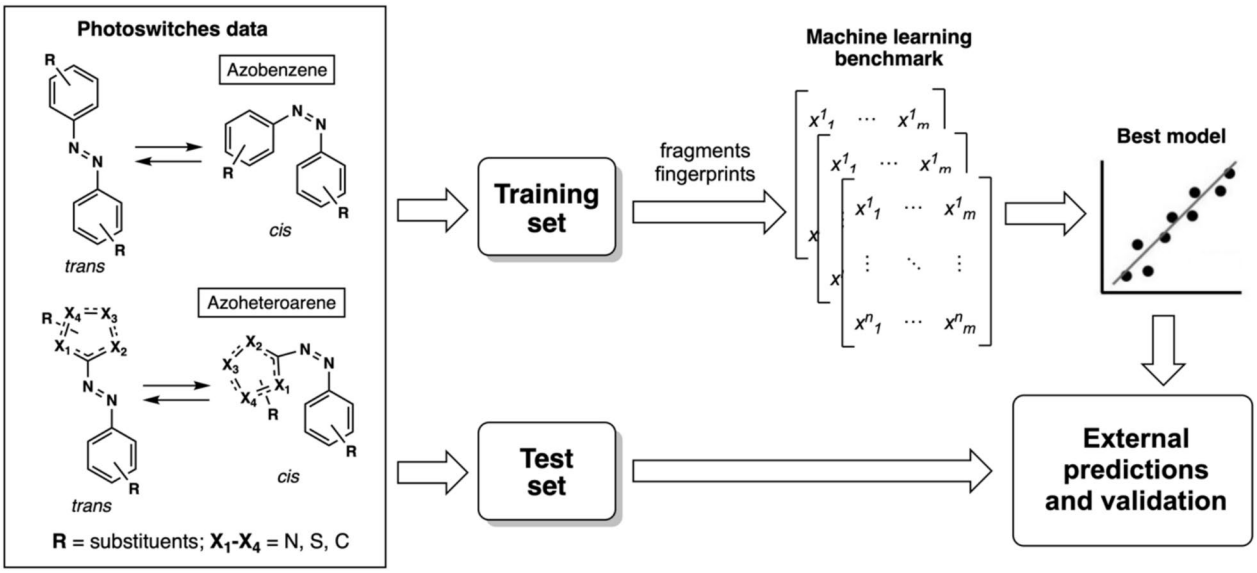

Fig. 1 General workflow of the study. Literature data set of azobenzenes and azoheteroarenes is used for the benchmark of models built on a variety of molecular fingerprints and fragment count descriptors. The best performing model is selected via cross-validation and validated on an external test set

most abundant; on the other hand, there are no large publicly available datasets that contain curated experimental data on the isomerization of azo-photoswitches. Following entries make part of the integrated data set: (1) 191 azobenzene dyes reported in [13, 14]; (2) 212 various heteroaryl dyes with spectra measured in ethanol, reported in [9]; (3) 91 heteroaryl dyes with spectra measured in acetonitrile, reported in [15–21]; (4) 80 molecules with both λ_{max} and half-life time measured, reported in [5, 22, 23]; (5) photoswitch data set reported in [12], containing azo-photoswitches with a variety of experimentally measured properties, including absorption maxima and isomerization rate. All data in the sources is annotated by room temperature or explicitly 25 ℃. Moreover, while most data sources explicitly indicate the stable isomer, some entries lack this information. For modeling purposes, we assume that the E-isomer is the more stable form for all azo-photoswitches in the dataset. The structures where this is not explicitly stated are labeled as "*E**". More detailed description of sources and data is available in SI.

It should be noted that not all properties are reported for each compound in these sets. We have performed an additional curation to merge all available data into a single data set. For merging, chemical structures were compared between all sets. In the case of duplicates, the most recent or the most complete (containing more reported properties) data was kept. In total, there are 798 unique compounds with measured λ_{max}, and 134 with measured $t_{1/2}$. We have divided our data based on their properties (wavelength and half-life) into two sets: 90% for cross-validation (CV), hyperparameter optimization and training (718 and 120 data points for λ_{max} and $t_{1/2}$, respectively), and 10% for external validation (80 and 14 data points for λ_{max} and $t_{1/2}$, respectively). The separation of data sets was performed randomly, with stratification by property values, so that the full range of values would be present in both training and test sets (see Fig. 2 for the property distribution in each set). The full data set is available in Additional file 2.

Model hyperparameter optimization and modeling

To build a robust QSPR model, compounds need to be represented as relevant numerical parameters—molecular descriptors. It is impossible to know in advance which descriptors will provide the best predictivity. Most often, the choice of the best descriptors is done via a rigorous benchmark study. While prediction of photoswitch properties, especially the isomerization rate and half-life, would intuitively benefit from the knowledge of 3D structures, the data that is used in this work does not reflect any conformational information. Therefore, we only consider several types of 2D structural descriptors, including CircuS and ChyLine fragments [24], Morgan [25], Atom pairs [26], Avalon [27], and native RDKit FP [28]. Molecular FP are binary features (they only take values 0 or 1) commonly used in QSPR modeling, especially for biological activities [29]. Morgan, or extended circular, fingerprints encode a structure by presence/absence of substructures of circular topology (atom and neigboring environment of a certain radius). It is also possible to generalize such representation further by using chemical features (aromaticity, H-bond donor/acceptors, etc.) instead of atom symbols, herein called Morgan features. Native RDKit FP similarly encode the substructures of linear and branched topologies. Like Morgan FP, aditional degree of generalization may be introduced by "layering" the substructures with different information (topology, aromaticity, bond orders, ring sizes, etc.), further referenced as "RDKit FP layered". Atom pairs generally contain the least

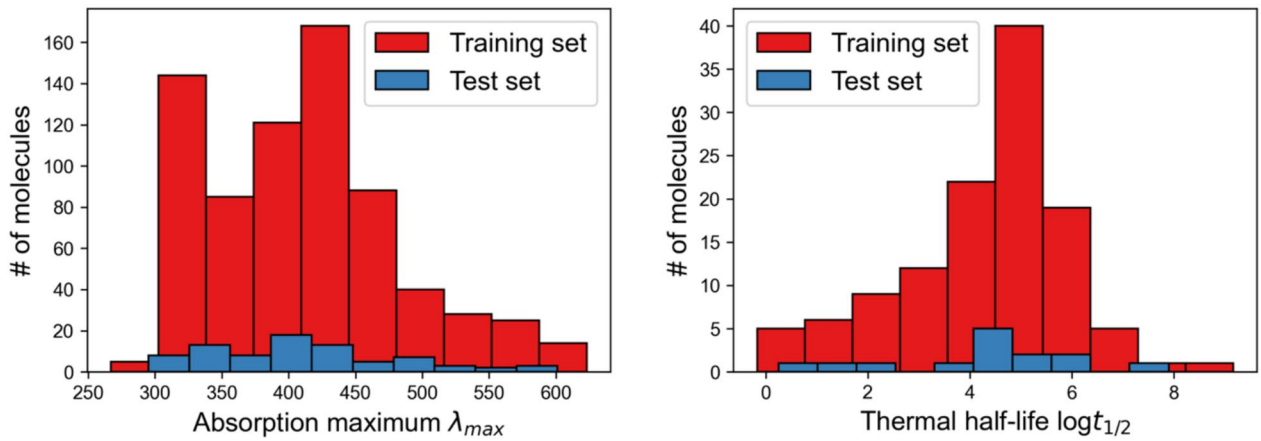

Fig. 2 The histograms of the distribution of λ_{max} and log$t_{1/2}$ in the training and test sets

information: only pairs of atoms with distances (number of bonds) between them are recorded. Avalon FP account for a variety of structural information, although this information is unavailable externally via the RDKit library. Fragment descriptors, like Circular Substructures (CircuS) and Chython Linear (ChyLine) fragments that we use here, improve on the information content of the features by recording the actual count of substructures that each descriptor represents. We have previously shown that CircuS descriptors outperform FP in other tasks [24]. Details on the parameters for each type of descriptors are available in Additional file 1. All models also include a preprocessing step that removes features with zero variance and scales the features to the range of 0 to 1.

The machine learning algorithm of choice here was Support Vector Machines (SVM) [30], as it is capable of handling smaller data sets with high precision due to high customizability of the fitting function (for results of other methods in the benchmarks, see Additional file 1). However, such adaptability also requires rigorous optimization of method hyperparameters, including the descriptor space. To tackle this, we employ the in-house optimization library—DOPtools [31]—that uses well-established optimization strategies to identify the optimal hyperparameters for the ML model. The construction and evaluation of these models adheres to best practices in cheminformatics [32]. A repeated 10-fold CV strategy was carefully applied during hyperparameter optimization, serving as a robust means of assessing the model's predictive performance. The performance is statistically evaluated by Root Mean Square Error (*RMSE*) and determination coefficient R^2 on the test set predictions:

$$RMSE = \frac{1}{N}\sum_{N}\left(y_{obs,i} - y_{pred,i}\right)^2$$

$$R^2 = 1 - \frac{\sum_N \left(y_{obs,i} - y_{pred,i}\right)^2}{\sum_N \left(y_{obs,i} - \hat{y}_{obs,i}\right)^2}$$

where N is the number of points in the set, $y_{obs,i}$ is the experimentally observed value of the ith data point, $y_{pred,i}$ is the predicted value of the ith data point, and $\hat{y}_{obs,i}$ is the average observed value across the set.

The analysis of chemical space is performed using UMAP method [33] in Python 3.10. The interpretation of the model is done using ColorAtom method [34] as implemented in the DOPtools library. ColorAtom calculates the weights of fragments as partial derivatives of the model predictions. The atoms accumulate the weights of the fragments they participate in, which eventually are summed up to calculate the atomic contribution. The visualization of these atomic contributions allows to see which parts of the molecule have the highest contribution to the prediction. The code for the analysis of benchmark and modeling results is available in GitHub repository.

Results and discussion

Prediction of λ_{max}

The model for the prediction of λ_{max} was based on the ensemble of data with experimentally measured spectra, 798 compounds in total. The benchmark study on molecular descriptors was performed along with the hyperparameter optimization, so that every type of descriptors achieves the best possible predictivity in CV. The results can be seen in the Fig. 3. Here, the CircuS and ChyLine fragments show the best performance. Both types are based on the fragment enumeration, with the difference in the substructure topologies. CircuS fragments count the number of occurrences of circular fragments, and ChyLine—of linear fragments, improving the information content over molecular fingerprints. The accuracy of predictions here is on par with the state-of-the-art approaches reported elsewhere (e.g., Griffiths et al. [12], which is the largest model to-date, report RMSE of 20.9 nm vs 21.6 nm in cross-validation for our model. Note that the scores are not directly comparable, as they report random train/test splits instead of CV).

Note that this model does not consider the physicochemical parameters of the solvents, although the solvents may have a considerable effect on the absorption of photoswitches [35]. Since not all data points in the data set presented here are annotated by the solvent in which the measurements are performed, building a model with solvent parameters included would reduce the size of the data set significantly. Still, we have performed such modeling, and the results are very similar to the model above ($R^2 = 0.907$, RMSE = 21.7 nm in CV); the details are available in Additional file 1.

Prediction of $t_{1/2}$

We use the half-life (in s) measurements in logarithmic scale in predictive models due to the extremely wide range of values in a non-logarithmic scale. The photoswitch data set provides the values of the isomerization rate, which were transformed into $t_{1/2}$ according to the first-order kinetic equation. The total number of data points for the model training was 134. Modeling followed the same protocol as described above. The benchmark was performed on models built only using structural parameters of the compounds, and the results are shown in Fig. 4 below. As the figure shows, the best model was built on linear ChyLine fragments, although the

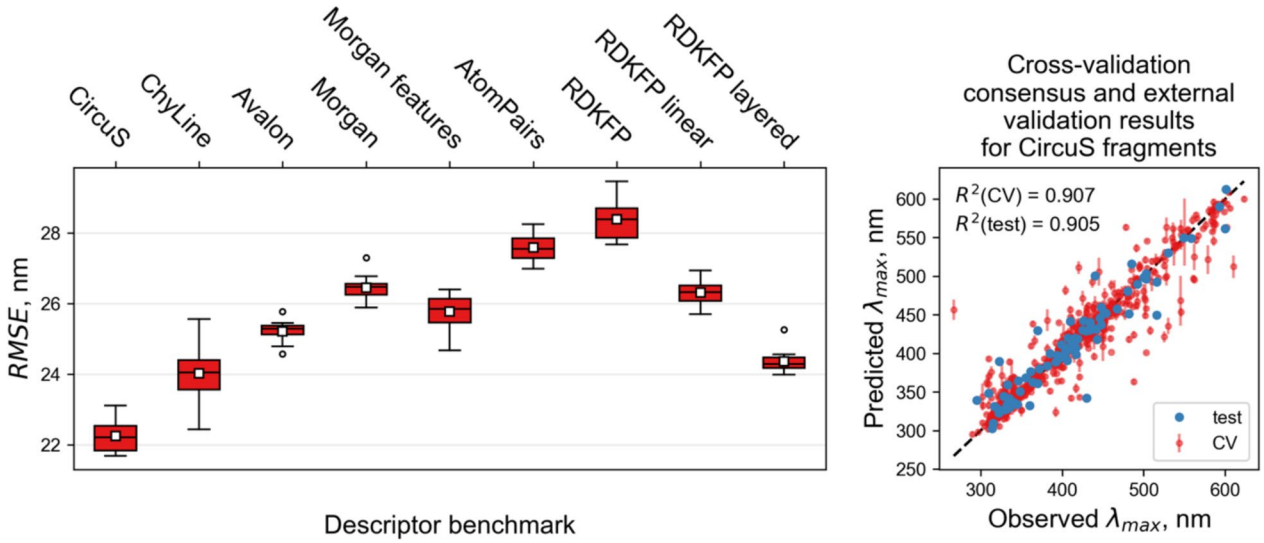

Fig. 3 Descriptor benchmark results (left) for the model predicting λ_{max}. Each boxplot represents the distribution of scores (RMSE, in nm) for 10 repeats of CV on the training set with random shuffling (white square for the mean score, the box for the interquartile range IQR, whiskers for 1.5 IQR, other points are outliers). The best model was obtained on CircuS fragments, achieving $R^2 = 0.907$ in CV and R^2(test) = 0.905 on an external test set, the observed *vs* predicted plot (right) shows the CV and external results for this model. Error bar on the points indicated the standard deviation of the prediction across CV repeats

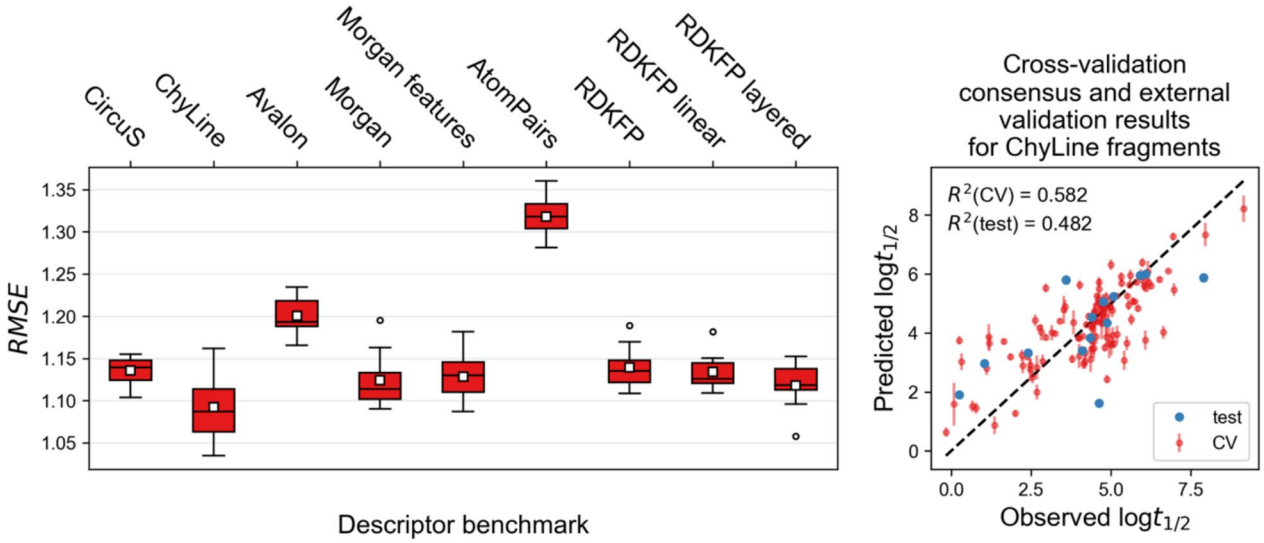

Fig. 4 Descriptor benchmark results (left) for the model predicting log$t_{1/2}$. Each boxplot represents the distribution of scores (in log units of $t_{1/2}$) for 10 repeats of CV on the training set with random shuffling (the boxplot elements follow the same designation as in Fig. 2). The best model was obtained on ChyLine fragments, the observed *vs* predicted plot (right) shows the CV results for this model. Error bar on the points indicated the standard deviation of the prediction across CV repeats

performance is overall quite poor ($R^2 = 0.584$ in CV and 0.482 in external validation).

One way to overcome weak predictions by an individual model is using consensus modeling [32], i.e., averaging the predictions made by several models. This allows to complement the initial model's bad predictions by the insights of another model, thus, hopefully, reducing the errors of predictions for outliers. In this case, we have started by calculating consensus predictions of the best model (ChyLine) with the second best (RDkit layered FP) and have observed a great improvement of the results, especially for two notable outliers by the middle of the

property range (Fig. 5, center). Adding a model built on RDkit linear FP further improved the predictions for these molecules (Fig. 5, on the right), however, a notable outlier by the lower end of the property had worse predictions here. It should be noted that this molecule is the only thiazole derivative in the test set, and all other thiazole-based azo-dyes in the training set have a much higher $\log t_{1/2}$ value, indicating that this compound is most likely outside of the applicability domain of these models.

Interpretation of predictions and analysis of chemical space

The validation of the model's accuracy on an external data set is supplemented by the interpretation of the model and the analysis of the chemical spaces of training and test sets for each property. Note that this is performed for the individual best model, as the descriptor space or a single hyperparameter setup for the consensus model cannot be defined. We have studied the chemical spaces of training and test set for both properties with UMAP method by projecting the descriptor spaces selected in the benchmark by the best model into a 2-dimensional space (see Fig. 6, on the left for both sets). The Fig. 6A shows that the large and diverse data set related to λ_{max} covers the chemical space quite well, and the molecules of the test set are close to the clusters of the training set. On the other hand, the data set related to $\log t_{1/2}$ is relatively small and, more importantly, less diverse. Most molecules of the test set are found outside of the initial clusters, thus, which increases the likelihood of them being outside of the applicability domain of the model.

The interpretation of the predictions here is done using ColorAtom methodology [34]. In ColorAtom, the weights of fragment descriptors are calculated as partial derivatives of prediction and define the atomic contributions of all atoms in a molecule. This allows to see how the presence of different functional groups affects the predicted values and guide the rational synthesis of compounds with desired properties. Generally speaking, the ColorAtom is applicable to one molecule at a time; however, here (Fig. 6) we have scaled the atomic contributions to the maximum in the subset, so that all molecules can be compared (full tables with the whole test set are available in SI, Additional files 1 and 3). It should be noted, however, that some limitations should be considered for ColorAtom. First, it is limited to fragments descriptors (CircuS, ChyLine, etc.) and would not work with fingerprints; luckily, the best models in both cases were based on fragments. Second, the algorithm that calculates the weights of descriptors assumes linear, or at least monotonous, relationship, as a partial derivative is used, so the interpretations should be used with caution. Finally, although it is true for any ML model, interpretations of the predictions are limited to the model and, thus, to the training data, meaning that if some effects are not accounted for in the data, they would not be relevant in the atomic contributions.

As far as the interpretations go here, we can observe some trends that are reported in literature. Functional groups (e.g., NO_2, CN, OMe) in their role of chromophore

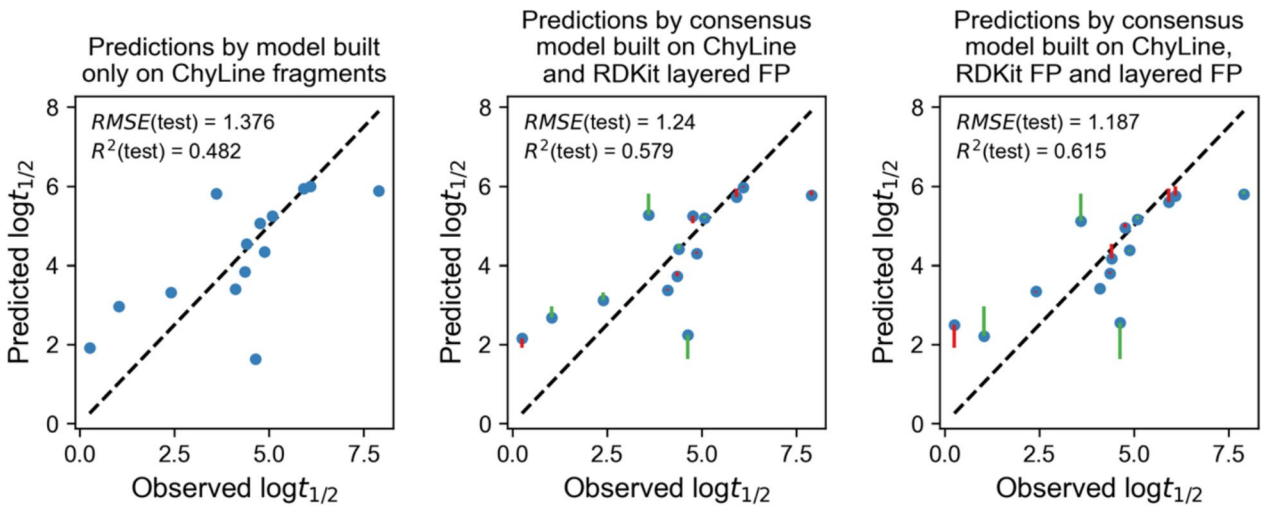

Fig. 5 Comparison of performance of the individual best model and its consensus with other models. Observed vs predicted are shown in each plot, with the vertical lines on the consensus plots showing improvement (in green) or deterioration (in red) of prediction compared to the individual model. Consensus models improve the predictions for most outliers of the test set, except the molecule with the lowest $\log t_{1/2}$ value which is most likely outside of the applicability domain

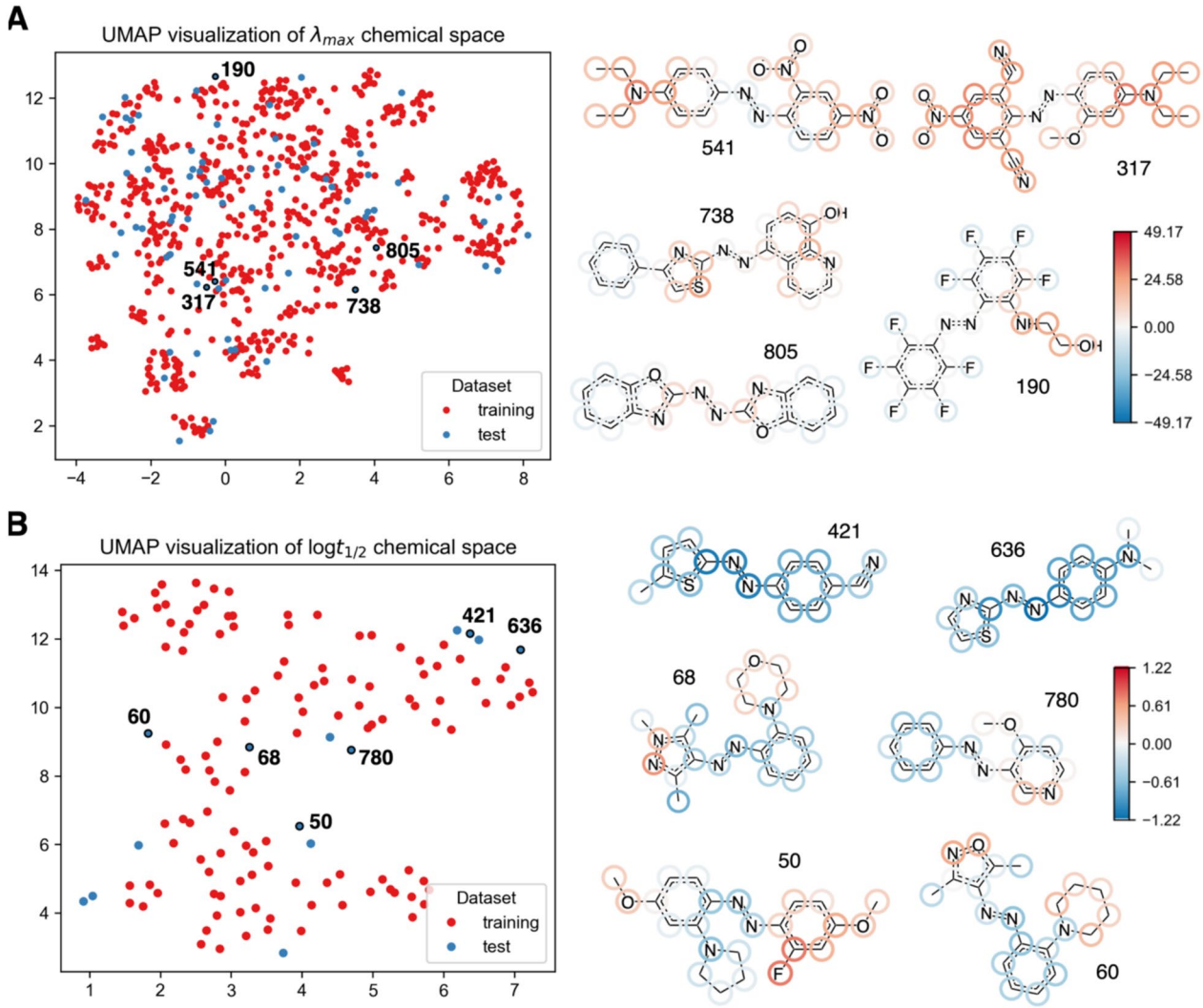

Fig. 6 Interpretation of models for prediction of (**a**) λ_{max} and (**b**) log$t_{1/2}$ by ColorAtom (on the right) and visualization of chemical space of each model (on the left). In ColorAtom, the contributions of atoms are coded blue for negative contributions and red for positive, with the intensity of color indicating the scale of the effect. White-coded atoms have virtually no contribution to the prediction. The contributions are scaled to the maximum in the test set for each property, to allow the comparison of the effect; the scale is indicated as a colorbar. Visualization of the chemical space is made by UMAP on the descriptor shown to be the best in the benchmark; molecules are colored red for training set and blue for test set. IDs are given for some molecules used for model interpretation; they follow the indexing in the Additional file 2

or auxochromes on the azo-dye affect λ_{max} and log$t_{1/2}$ of a photoisomer [17, 23]. For example, interpreting the model for λ_{max}, we observed high positive contribution of functional groups (e.g., NO_2 and NEt_2 in entries 541 and 317, Fig. 6A) for the prediction, which is consistent with previous reports [36, 37]. Similarly, introducing heteroarenes (entries 738, 805, Fig. 6A) into a dye has a positive, albeit smaller, effect [38, 39]. It should be noted that these effects are often associated with the push/pull character of interaction between electron-donating and withdrawing groups or other electronic effects; our models, however, only consider the structural descriptors, which may only implicitly contain such information. Similarly, log$t_{1/2}$ can be increased by the presence of heteroarenes [40] (entries 60, 68, Fig. 6B) or by substitution patterns (e.g., F in ortho position has been reported to increase half-life [38, 41], as in entry 50). Sometimes, however, ColorAtom may omit the expected effect from functional groups (e.g., OMe in entry 780, Fig. 6B), likely due to relatively small contribution compared to other fragments in a molecule.

An important limitation of the presented models is the omission of the solvent effects. While these are important for both λ_{max} and $t_{1/2}$, the currently collected dataset does

not contain sufficient data to account for them. Specifically, some of the data sources for λ_{max} lack solvent annotations. On the other hand, while all data points in the set related to half-life have those, their distribution is highly uneven, with the majority (95 entries) measured in acetonitrile. We report the models following different strategies—either isolating the dataset of the majority solvent or using solvent properties as descriptors—in SI (see Additional file 1). Neveretheless, netiher approach led to a significant improvement of predictions. We believe that a more systematic dataset is required to assess solvent effects properly and encourage the community to contribute to its creation and collection.

Conclusions

In this manuscript we outline the modeling strategies for prediction of both λ_{max} and $t_{1/2}$ of photoisomers of visible-light azo-photoswitches. The structural features (molecular fingerprints and fragment counts) that we investigate provide a solid base for the accurate and fast modeling and virtual screening. Moreover, model interpretation by the ColorAtom approach allows for the rational design of new compounds, by pinpointing the structural motifs leading to the desired property values. The predictions of the model were validated on an external set of 80 and 14 for λ_{max} and $t_{1/2}$, respectivelly. The data for λ_{max} of azobenzene and azoheteroarene families of photoswitches is abundant, and the accuracy of predictive models, even built on purely 2D representations of molecules without any electronic information, is quite high, as we demonstrate here. The external test set of photoswitches has demonstrated relatively small errors of prediction of λ_{max}, despite the property not being directly correlated to the structural features we employ. On the other hand, the lack of data on the $t_{1/2}$, as well as the general complexity of the mechanism of isomerization in azo-photoswitches lead to not quite satisfactory performance of such models. While generally we observe quite good predictivity in the external set, several outliers with quite large prediction errors (up to 2 log units for $t_{1/2}$) are found. We would like to note that the chemical space covered by the data set for $t_{1/2}$ is quite narrow, which may lead to a limited reliability of predictions in the higher range of the property. Moreover, the application of consensus modeling allows to significantly reduce the error of prediction for most outliers. We provide the full integrated data set to encourage further investigations using other machine learning methods or features, including quantum chemical calculations, and the expansion of the chemical space of photoswitches. A more systematic data set for the half-life, incorporating a variety of both structural features and measurement conditions such as solvents, would benefit the future efforts in modeling this property with machine learning methods.

Abbreviations

ML	Machine learning
TD-DFT	Time-dependent density functional theory
QSPR	Quantitative structure–property relationship
FP	Fingerprints
SVM	Support vector machines
RMSE	Root mean squared error
CV	Cross-validation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00993-7.

Additional file 1.

Additional file 2.

Additional file 3.

Author contributions

P.K.H. performed the data collection and curation. S.B. performed the modeling benchmarks, predictive modeling, and model interpretation. P.S. designed the study and wrote the code used in benchmarking studies and model interpretation. P.K.H. and P.S. prepared the manuscript text, P.S. and S.B prepared the figures. All authors reviewed the manuscript.

Funding

S.B. and P.S. thank the financial support of the Institute for Chemical Reaction Design and Discovery (ICReDD), which was established by the World Premier International Research Initiative (WPI), MEXT, Japan, List Sustainable Digital Transformation Catalyst Collaboration Research Platform offered by Hokkaido University, and JSPS KAKENHI grant 23H03807. This work was partially supported by "Crossover Alliance to Create the Future with People, Intelligence and Materials" from MEXT, Japan to P. K. H.

Availability of data and materials

The detailed description of the computational procedures and the results of benchmark on other methods are available in the Additional file 1. The data used in this study is available as a Supplementary Material (see Additional file 2). The calculated molecular descriptors, the code to reproduce the modeling results and the detailed results of the model hyperparameter optimization are available via the GitHub repository: [https://github.com/icredd-cheminfo/Photoswitches].

Declarations

Competing interests

The authors declare no competing interests.

Received: 21 October 2024 Accepted: 20 March 2025
Published online: 01 April 2025

References

1. Göstl R, Senf A, Hecht S (2014) Remote-controlling chemical reactions by light: towards chemistry with high spatio-temporal resolution. Chem Soc Rev 43:1982. https://doi.org/10.1039/c3cs60383k
2. Beharry AA, Woolley GA (2011) Azobenzene photoswitches for biomolecules. Chem Soc Rev 40:4422. https://doi.org/10.1039/c1cs15023e
3. Zhang Z, Wang W, O'Hagan M et al (2022) Stepping out of the blue: from visible to near-IR triggered photoswitches. Angew Chemie. https://doi.org/10.1002/ange.202205758

4. Bléger D, Hecht S (2015) Visible-light-activated molecular switches. Angew Chemie Int Ed 54:11338–11349. https://doi.org/10.1002/anie.201500628
5. Lameijer LN, Budzak S, Simeth NA et al (2020) General principles for the design of visible-light-responsive photoswitches: tetra- ortho -chloro-azobenzenes. Angew Chemie Int Ed 59:21663–21670. https://doi.org/10.1002/anie.202008700
6. Mukadum F, Nguyen Q, Adrion DM et al (2021) Efficient discovery of visible light-activated azoarene photoswitches with long half-lives using active search. J Chem Inf Model 61:5524–5534. https://doi.org/10.1021/acs.jcim.1c00954
7. Raucci U, Sanchez DM, Martínez TJ, Parrinello M (2022) Enhanced sampling aided design of molecular photoswitches. J Am Chem Soc 144:19265–19271. https://doi.org/10.1021/jacs.2c04419
8. Adrion DM, Lopez SA (2023) Design rules for optimization of photophysical and kinetic properties of azoarene photoswitches. Org Biomol Chem 21:7351–7357. https://doi.org/10.1039/D3OB01298K
9. Mai J, Lu T, Xu P et al (2022) Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. Dye Pigment 206:110647. https://doi.org/10.1016/j.dyepig.2022.110647
10. Axelrod S, Shakhnovich E, Gómez-Bombarelli R (2022) Excited state non-adiabatic dynamics of large photoswitchable molecules using a chemically transferable machine learning potential. Nat Commun 13:3440. https://doi.org/10.1038/s41467-022-30999-w
11. Axelrod S, Shakhnovich E, Gómez-Bombarelli R (2023) Thermal half-lives of azobenzene derivatives: virtual screening based on intersystem crossing using a machine learning potential. ACS Cent Sci 9:166–176. https://doi.org/10.1021/acscentsci.2c00897
12. Griffiths RR, Greenfield JL, Thawani AR et al (2022) Data-driven discovery of molecular photoswitches with multioutput Gaussian processes. Chem Sci 13:13541–13551. https://doi.org/10.1039/d2sc04306h
13. Buttingsrud B, Alsberg BK, Åstrand P-O (2007) Quantitative prediction of the absorption maxima of azobenzene dyes from bond lengths and critical points in the electron density. Phys Chem Chem Phys 9:2226–2233. https://doi.org/10.1039/B617470A
14. Mustroph H, Epperlein J (1981) Untersuchungen zum UV/Vis-spektralverhalten von azofarbstoffen. V [16]. Quantitative beschreibung der absorptionsmaxima mehrfach substituierter azobenzene mit einem inkrementsystem. J für Prakt Chemie 323:755–775. https://doi.org/10.1002/prac.19813230508
15. Lin R, Hashim PK, Sahu S et al (2023) Phenylazothiazoles as visible-light photoswitches. J Am Chem Soc 145:9072–9080. https://doi.org/10.1021/jacs.3c00609
16 Heindl AH, Wegner HA (2020) Rational design of azothiophenes—substitution effects on the switching properties. Chem A Eur J 26:13730–13737. https://doi.org/10.1002/chem.202001148
17. Devi S, Saraswat M, Grewal S, Venkataramani S (2018) Evaluation of substituent effect in Z -isomer stability of arylazo-1 H -3,5-dimethylpyrazoles: interplay of steric, electronic effects and hydrogen bonding. J Org Chem 83:4307–4322. https://doi.org/10.1021/acs.joc.7b02604
18. Fang D, Zhang Z-Y, Shangguan Z et al (2021) (Hetero)arylazo-1,2,3-triazoles: "clicked" photoswitches for versatile functionalization and electronic decoupling. J Am Chem Soc 143:14502–14510. https://doi.org/10.1021/jacs.1c08704
19 Kumar P, Srivastava A, Sah C et al (2019) Arylazo-3,5-dimethylisoxazoles: azoheteroarene photoswitches exhibiting high Z -isomer stability, solid-state photochromism, and reversible light-induced phase transition. Chem A Eur J 25:11924–11932. https://doi.org/10.1002/chem.201902150
20. Hallas G, Choi J-H (1999) Synthesis and spectral properties of azo dyes derived from 2-aminothiophenes and 2-aminothiazoles. Dye Pigment 42:249–265. https://doi.org/10.1016/S0143-7208(99)00031-5
21. Weston CE, Richardson RD, Haycock PR et al (2014) Arylazopyrazoles: azoheteroarene photoswitches offering quantitative isomerization and long thermal half-lives. J Am Chem Soc 136:11878–11881. https://doi.org/10.1021/ja505444d
22. Ahmed Z, Siiskonen A, Virkki M, Priimagi A (2017) Controlling azobenzene photoswitching through combined ortho -fluorination and -amination. Chem Commun 53:12520–12523. https://doi.org/10.1039/C7CC07308A
23. Gaur AK, Kumar H, Gupta D et al (2022) Structure-property relationship for visible light bidirectional photoswitchable azoheteroarenes and thermal stability of Z -isomers. J Org Chem 87:6541–6551. https://doi.org/10.1021/acs.joc.2c00088
24. Tsuji N, Sidorov P, Zhu C et al (2023) Predicting highly enantioselective catalysts using tunable fragment descriptors. Angew Chemie—Int Ed 62:e202218659. https://doi.org/10.1002/anie.202218659
25. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
26. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
27. Gedeck P, Rohde B, Bartels C (2006) QSAR — how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. J Chem Inf Model 46:1924–1936. https://doi.org/10.1021/ci050413p
28. Landrum G (2006) RDKit: Open-source cheminformatics
29. Danishuddin KAU (2016) Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today 21:1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013
30. Drucker H, Burges CJC, Kaufman L, et al (1997) Support vector regression machines. In: Advances in neural information processing systems. pp 155–161
31. Byadi S, Gantzer P, Gimadiev T, Sidorov P (2024) DOPtools: a Python platform for descriptor calculation and model optimization. Overview and usage guide. ChemRxiv
32. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488. https://doi.org/10.1002/minf.201000061
33 McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. J Open Source Softw 3:861. https://doi.org/10.2110/joss.00861
34. Marcou G, Horvath D, Solov'Ev V et al (2012) Interpretability of SAR/QSAR models of any complexity by atomic contributions. Mol Inform 31:639–642
35. Rauf MA, Hisaindee S (2013) Studies on solvatochromic behavior of dyes using spectral techniques. J Mol Struct 1042:45–56. https://doi.org/10.1016/j.molstruc.2013.03.050
36. Choi J, Jeon J, Kim M et al (2008) Novel azo dyes derived from phthalimide. Part 1: synthesis and spectroscopic properties. Color Technol 124:92–99. https://doi.org/10.1111/j.1478-4408.2008.00127.x
37. Pithan PM, Kuhlmann C, Engelhard C, Ihmels H (2019) Synthesis of 5-alkyl- and 5-phenylamino-substituted azothiazole dyes with solvatochromic and DNA-binding properties. Chem A Eur J 25:16088–16098. https://doi.org/10.1002/chem.201903657
38. Jelínková V, Dellai A, Vachtlová M et al (2024) Molecular azo–imidazole photoswitches: property tuning by substitution. J Photochem Photobiol A Chem 449:115390. https://doi.org/10.1016/j.jphotochem.2023.115390
39. Gallardo-Rosas D, Guevara-Vela JM, Rocha-Rinza T et al (2024) Structure and isomerization behavior relationships of new push–pull azo-pyrrole photoswitches. Org Biomol Chem 22:4123–4134. https://doi.org/10.1039/D4OB00417E
40. Calbo J, Weston CE, White AJP et al (2017) Tuning azoheteroarene photoswitch performance through heteroaryl design. J Am Chem Soc 139:1261–1274. https://doi.org/10.1021/jacs.6b11626
41. Patel S, Das B, Mishra P, Chandra G (2024) Light triggered reversible aggregation/dispersion of hydroxy azo-benzenes during photo switching: solvent, ions assisted dispersion, and induced quenching emission. ChemPhotoChem. https://doi.org/10.1002/cptc.202300350

Publisher's Note