SCIENTIFIC REPORTS
natureresearch

OPEN

Prediction of Absorption Spectrum Shifts in Dyes Adsorbed on Titania

Vishwesh Venkatraman*, Amsalu Efrem Yemene & John de Mello

Dye adsorption on metal-oxide films often results in small to substantial absorption shifts relative to the solution phase, with undesirable consequences for the performance of dye-sensitized solar cells and optical sensors. While density functional theory is frequently used to model such behaviour, it is too time-consuming for rapid assessment. In this paper, we explore the use of supervised machine learning to predict whether dye adsorption on titania is likely to induce a change in its absorption characteristics. The physicochemical features of each dye were encoded as a numeric vector whose elements are the counts of molecular fragments and topological indices. Various classification models were subsequently trained to predict the type of absorption shift i.e. blue, red or unchanged ($|\Delta\lambda| \leq 10$ nm). The models were able to predict the nature of the shift with a good likelihood (~80%) of success when applied to unseen data.

The light-harvesting properties of dye-sensitized metal oxides find a number of applications in photonic devices and chemical probes. They constitute a significant area of current research, and are key components of many devices including dye-sensitized solar cells[1], photo-electrochemical water splitters[2] and optical filters[3–5]. General requirements of the constituent dyes include broad absorption spectra (preferably extending into the near-infrared portion of the solar spectrum) accompanied by large extinction coefficients[1]. To meet these objectives, numerous organic[6] and metal-based[7] dyes have been designed[8], employing varying donors (D), π-bridges, and acceptors (A) including but not limited to the following configurations: D-π-A, D-A-π-A[9] and D-D-π-A[10]. Each dye is further chemisorbed onto a semiconducting metal oxide photoanode (usually TiO_2[11]), to provide a mesoporous metal-oxide dye interface at which efficient separation of photo-generated electron-hole pairs can occur. The nature of the donors, acceptor/anchoring groups and the strength of the dye-semiconductor coupling, all have a significant impact on the photostability and photochemical behaviour.

Broad absorption spectra are desirable for light harvesting. However, it is often seen (particularly for metal-free dyes) that the UV-vis absorption peaks of the dyes adsorbed on TiO_2 photoelectrodes are substantially shifted, compared to those in solution. While for some dyes, there is little or no change (for some a broadening of the peak is seen), peaks in other cases can be shifted by 100 nm or more[12,13] in either direction, greatly complicating the design and selection of candidate dyes. Reasons attributed to such phenomena include the deprotonation of the carboxylic/cyanoacrylic anchoring group, π-stacking interactions, complexation with metal ions and dye aggregation[9,14–19]. On adsorption, deprotonation can result in a carboxylate-TiO_2 unit that is a weaker electron acceptor than the native carboxylic acid[20,21]. Furthermore, during the sensitization process, significant dye aggregation can occur[22]. While J-aggregates lead to a red-shift, formation of H-aggregates causes a blue-shift, leading to a damping of the absorption efficiency. The aggregation behaviour is however very dye-specific. Complexation with metal ions such as aluminium, iron, tin, titanium and chromium are also seen to induce red-shifts particularly for anthocyanin dyes[23–25] owing to the suspected formation of a quinoidal structure[14]. In the case of catechol anchoring groups[26], absorption shifts have been attributed to the increased dipole moment of the Ti-ligand complex via an induced charge transfer dipole under excitation[27]. Solvatochromism also has a significant impact on the relative spectral shifts. For instance, it has been shown that in polar solvents, the electron-withdrawing power of the carboxylic acid decreases as a result of a partial deprotonation in the excited state[28]. The use of different sensitization solvents is also seen to affect the adsorption characteristics of dyes[29,30].

So far, understanding the origin of these spectral changes and their possible effects, has largely been based on comparative studies of the dye in solution and in its adsorbed state. Selecting a dye purely on the basis of its solution-phase properties has until now been a unreliable task. Theoretical investigations have focused on analysing the aggregation behaviour[31–35] and impact of the anchoring groups[29], using density functional theory (DFT) and *ab initio* methods[36]. Calculations of excited-state properties using time-dependent DFT (TD-DFT)

Department of Chemistry, Norwegian University of Science and Technology, 7491, Trondheim, Norway. *email: vishwesh.venkatraman@ntnu.no

methods[35,37,38] are generally seen to agree well with the experimental measurements[39,40]. However, such tasks require considerable time and are therefore, not suitable for rapid screening tasks involving a large number of molecules. An additional challenge is to identify the dye-oxide binding mode which will change depending on the structure of the dye and the binding groups. For example, the —COOH group can form monodentate ester-like, bidentate chelate or bidentate bridging linkages[41,42]. In the absence of any prior knowledge, multiple combinations must be tested, thereby adding to the computational effort.

Machine learning (ML) approaches capable of identifying embedded correlations between structure (represented appropriately) and property have been successfully used in materials science and computational chemistry[43–47]. We therefore ask the question: based only on the knowledge of a dye molecule's chemical structure and its absorption spectrum in a given solvent, can we use data-driven ML techniques to predict the type of absorption shift? To this purpose, the UV-Vis absorption peaks in solution and on a metal oxide were extracted from literature for ~2000 dyes. The change in the maximum absorption wavelength from solution-phase to metal-oxide-supported was used to categorise the dyes as blue-shifted, red-shifted or unchanged. Supervised machine learning models were then trained to distinguish between the classes using descriptors such as molecular fragment counts and topological indices that are easily calculated from the dye structure. Our results show that using cheaply derived structure descriptors, the classification models can achieve 80% success in predicting the type of absorption shift.

Methods

Data curation. Absorption data in solution and on TiO_2 for ~2000 metal-free dyes were extracted from around 500 literature articles. Other metal oxides such as ZnO, NiO and SnO_2 were not considered as the available data was too limited. For some dyes, the reported absorption peaks (from different studies) in the same solvent were found to be significantly different and were therefore omitted. We further considered only those cases for which values were recorded in pure solvents and without any additives such as chenodeoxycholic acid. In the end, a total of 1961 observations corresponding to 1861 unique dyes were obtained. For these compounds, the difference in the absorption maxima (λ) i.e. $\Delta\lambda = \lambda_{\max}^{\text{soln}} - \lambda_{\max}^{\text{TiO}_2}$ ranged between −220 to +190 nm (see Fig. F1 in the Supplementary Material-II). The structures spanned various donor classes such as triphenylamines, phenothiazines, carbazoles, coumarins etc. with varying numbers and types of anchoring groups — catechol, hydroxylpyridium[48], cyanoacrylic, pyrimidine (see Fig. 1). In the assembled data, the dyes were divided into ten separate groups based on the solvent (see Table 1). The molecular structures (SMILES format), absorption properties and associated references are provided in Table S1 in the Supplementary Material-I.

Nature of the shift. The nature of the spectral shift was determined (see Fig. 2) by thresholding the difference between the solution phase and solid-state maxima ($\Delta\lambda = \lambda_{\max}^{\text{soln}} - \lambda_{\max}^{\text{TiO}_2}$) with respect to the following criteria:

$$S_{B:NR} = \begin{cases} \text{B}, & \text{if } \Delta\lambda > 10 \text{ nm} \\ \text{NR}, & \text{otherwise} \end{cases}$$

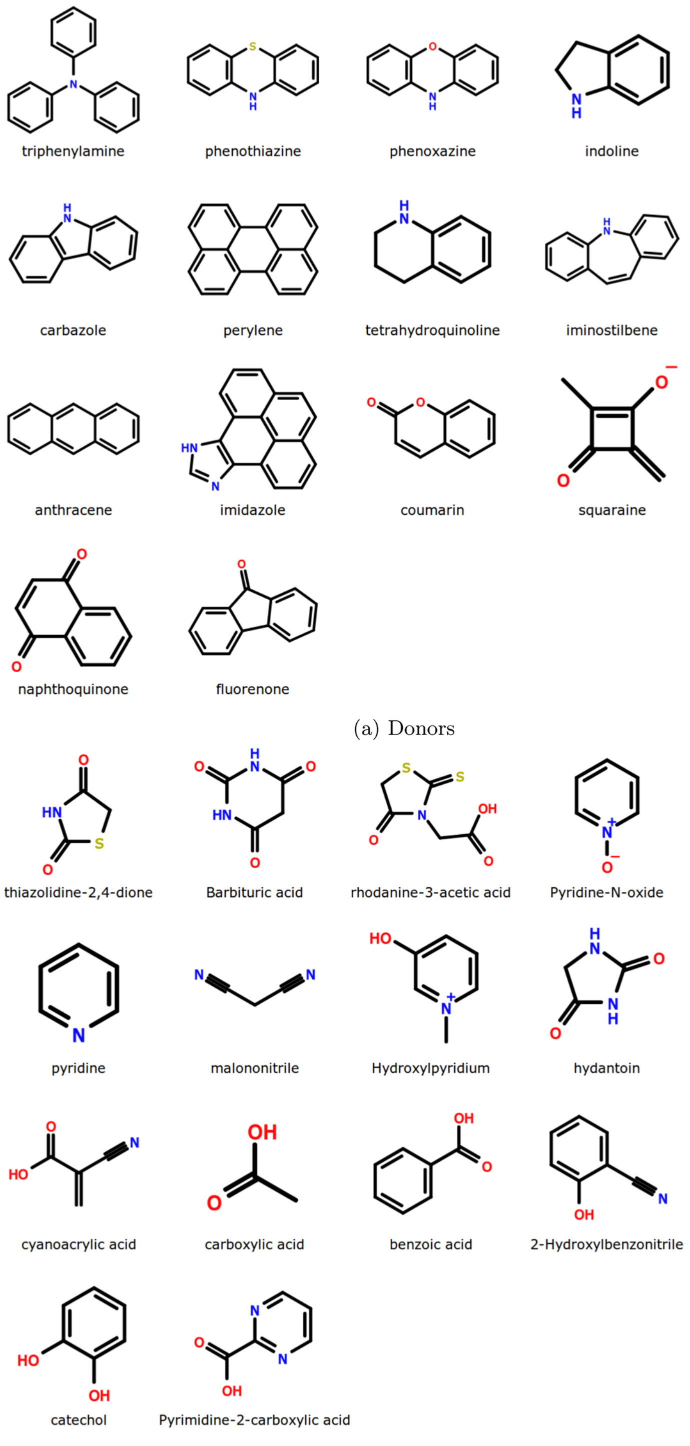
$$S_{B:N:R} = \begin{cases} \text{N}, & \text{if } -10 \text{ nm} \le \Delta\lambda \le 10 \text{ nm} \\ \text{R}, & \text{if } \Delta\lambda < -10 \text{ nm} \\ \text{B}, & \text{otherwise} \end{cases}$$

$$S_{BN:R} = \begin{cases} \text{BN}, & \text{if } \Delta\lambda < -10 \text{ nm} \\ \text{R}, & \text{otherwise} \end{cases}$$

where *B*, *R*, and *N* indicate a blue shift, red shift and little or no change respectively. Instead of using a strict cut-off of 0, deviations of 10 nm or less were designated *N*. Figure 3 provides a solvent-wise distribution of the experimentally derived categories. For a majority of the cases associated with weakly polar solvents such as DCM, THF and CHCl3, a blue-shifted absorption is seen, while red-shifted behaviour becomes more prominent as polarity increases. This may be attributed to a more efficient solvation of the dyes in the polar solvents[49].

The categories can be further grouped into *NR* (no change + red shift) or *BN* (blue-shift + no change). In the context of machine learning, a balanced distribution of the instances across the classes (50% to class A and 50% to class B) is preferred. Owing to the presence of higher number of instances for a given category, classification learners run the risk of predicting everything as one or the other class[50]. To this end, we analysed three different schemes (I) the first where three different groups *B*, *N* and *R* are established, (II) the second merges the red-shifted dyes with those indicating little or no change — *B*/*NR* and lastly, (III) merging the blue-shifted and no change — *BN*/*R*. When considering three independent groups — *B*,*N*,*R*, the class distributions (2:1:1) are slightly skewed towards the blue. For the second case (*B*:*NR*), a near 1:1 ratio is seen for most solvents is seen, with the exception of DMF for which a majority of the cases belong to the *NR* category. The converse holds true for the third case (*NB*:*R*), where the distribution of the categories is found to be significantly skewed (2.5:1) in favour of *NB* for a majority of the solvents. Consequently, in this paper, we focus mainly on the *B*:*NR* cases where the data is well balanced and more likely to yield a more effective classification.

Molecular descriptors. For a statistical structure-property relationship to be established, the molecules need to be represented in a way that captures their physicochemical characteristics. In cheminformatics, these representations are typically referred to as molecular descriptors[51] i.e. a vector of numbers that captures the



(a) Donors

(b) Anchoring groups

Figure 1. Prominent donors and anchoring groups present in the dyes included in the data set.

chemical information in a computer-interpretable form. Machine learning approaches use these vectors to infer a predictive model from the training data. Here, we have employed the following schemes to encode each dye:

Atom-bond sequences. The first set of descriptors are extracted by enumerating all the shortest paths (successive connected atom-bond sequences) between each pair of atoms[52]. The descriptor calculation was carried out using the ISIDA Fragmentor2017 software[53]. The minimum and maximum length of the atom-bond sequences were set to 3 and 6 respectively.

Solvent	N_{Obs}	Range (nm)	P'	Distribution		
				B:NR	B:N:R	NB:R
Chloroform (CHCl3)	258	−94–134	4.1	141:117	141:67:50	208:50
Dichloromethane (DCM)	753	−219–190	3.1	429:324	429:188:136	617:136
Tetrahydrofuran (THF)	493	−116–140	4	205:288	205:117:171	322:171
Ethanol (EtOH)	174	−207–102	5.2	47:127	47:51:76	98:76
Methanol (MeOH)	39	−21–122	5.1	15:24	15:12:12	27:12
Acetonitrile (MeCN)	55	−109–108	5.8	22:33	22:7:26	29:26
Toluene	36	−22–64	2.4	16:20	16:11:9	27:9
1,4-dioxane	19	−71–33	4.8	3:16	3:3:13	6:13
Dimethylformamide (DMF)	114	−92–115	6.4	20:94	20:35:59	55:59
Dimethylsulfoxide (DMSO)	20	−38–62	7.2	11:9	11:3:6	14:6
Overall	1915	−219–190	—	909:1052	909:494:558	1403:558

Table 1. Number of data points (N_{Obs}) grouped according to the solvent in which the spectra were measured. The third column shows range of the shift (in nm) calculated as $\lambda_{\max}^{\mathrm{soln}} - \lambda_{\max}^{\mathrm{TiO_2}}$. The relative polarities (P') for the solvents are taken from Snyder[104]. The last column shows the distribution of the categories formed by combining dyes that show no/little change (N), red-shift (R) or blue-shift (B) after adsorption on TiO_2.

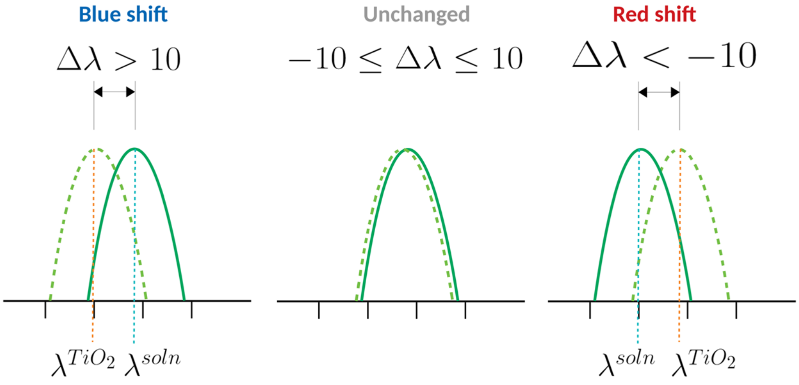

Blue shift
$\Delta\lambda > 10$
Unchanged
$-10 \leq \Delta\lambda \leq 10$
Red shift
$\Delta\lambda < -10$
λ^{TiO_2} λ^{soln}
λ^{soln} λ^{TiO_2}

Figure 2. Nature of the spectral shift based on the positions of the absorption peaks corresponding to the solution phase ($\lambda_{\max}^{\mathrm{soln}}$) and solid-state maxima ($\lambda_{\max}^{\mathrm{TiO_2}}$). The difference between the two values: $\Delta\lambda = \lambda_{\max}^{\mathrm{soln}} - \lambda_{\max}^{\mathrm{TiO_2}}$ is used to determine whether there is a blue shift, red shift or no change upon adsorption.

Topological indices. The second set of descriptors includes constitutional indices (number of hetero atoms and aromatic rings, hydrogen bond acceptors and donors) as well as topological indices (derived from chemical graph representations) that take into account the connectivity along with atom and bond labels. Popular descriptors include the electrotopological state[54] (EState) indices that encode the topology and electronic environment of molecular fragments. Other variables include MOE-type descriptors[55] that are based on an approximate accessible van der Waals surface area calculation for each atom, along with some other atomic property. Here, we have included properties such as *logP* (octanol/water), molar refractivity, and partial charge within a binned range (corresponding to a subdivision of the molecular surface area). The descriptors were computed using the open source cheminformatics toolkit RDKit[56]. For a preset bin size (k), the calculated descriptors include SlogP–VSA$_k$ (capture hydrophobic and hydrophilic effects), SMR–VSA$_k$ (polarizability) and PEOE–VSA$_k$ (capture electrostatic interactions). The descriptors were computed using the open source cheminformatics toolkit RDKit[56].

The descriptors were selected to capture relevant features of the dye's chemical structure and without resorting to DFT or other computationally intensive calculations. Each structure was therefore described by a vector of length 2060 (solvent polarity was added as an additional descriptor) with computations taking less than 3 minutes to calculate all descriptors for the entire data set.

Machine learning. In order to identify machine learning models capable of discriminating between the different types of shifts ($B/N/R$), six popular classification schemes were explored: linear discriminant analysis[57] (LDA), k-nearest neighbours (k-NN), kernel-based support vector machines[58] (SVM), and tree-based models such as classification and regression trees[59] (CART), random forests[60] (RF) and gradient boosting machines[61] (GBM). Linear discriminant analysis works by identifying a linear combination of the variables (projection onto a smaller subspace) that best separates the classes. The k-NN algorithm classifies an object based on a majority vote of its k nearest neighbours that are identified by calculating the Euclidean distance from the point of interest (the class of which is to be determined) to all the points in training set. Support vector machines perform classification by finding the hyperplane that maximizes the margin between the classes[58]. For a two-dimensional space, the hyperplane is a line that divides a plane into two parts such that each class lies on either side. The vectors that

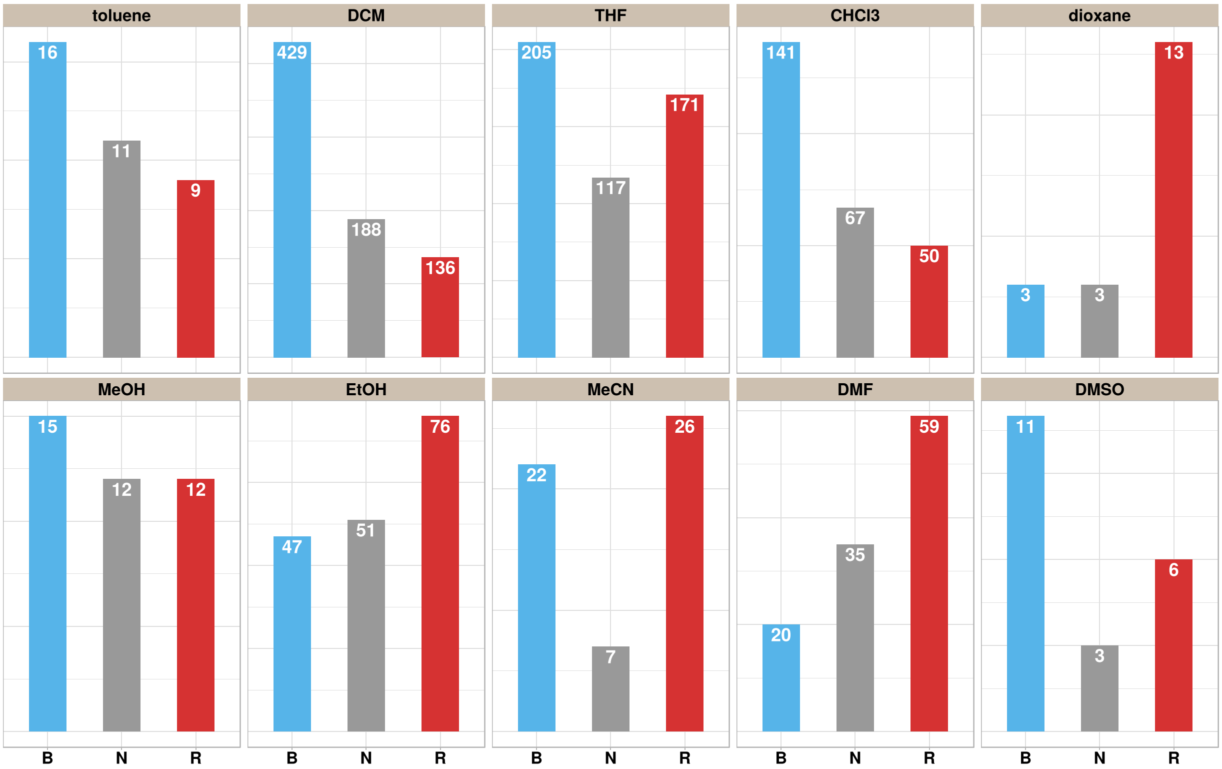


Figure 3. Distribution of the absorption shifts with respect to the solvents. The letter "B" indicates a blue shift, while "N" and "R" correspond to no change ($|\Delta\lambda| \leq 10$) or a red shift respectively. The solvents are sorted in increasing order of polarity from left to right. Polarity values are listed in Table 1. Additional plots are provided in the Supplementary Material-II.

define this hyperplane are the support vectors. The tree-based models output a series of if-then-else statements where the features are systematically checked to determine a final result. While the CART approach produces a single tree, both random forests and GBM are ensemble approaches where the outcome is the combination of the the decisions from multiple models. The difference lies in the way the trees are built: RF builds deep independent trees, while GBM creates successive models with each tree improving on the previous i.e. they seek to improve the result based on the current estimate.

Statistical modelling.

Analysis of the data started with the removal of descriptor columns with little or no variation and those containing missing values (due to an inability to calculate one or more descriptors). The data was then split randomly into independent calibration (75%) and test (25%) sets. The presence of highly correlated variables (multicollinearity) can affect predictive performance. Following previous studies[62,63], a pair-wise squared correlation cut-off of 0.90 was applied to the training set, whereby only one (arbitrarily determined) among the correlated pair of variables was retained. This reduced the number of variables from 2000 to around 200. In order to select the best model parameters (e.g. number of trees for RF, depth of the tree, number of neighbours to be considered (k-NN)), a five-fold cross-validation was employed, followed by randomization tests to reduce the risk of overfitting. A grid search was carried out to identify the optimal parameter combinations for the ML models used. The modelling was carried out using R[64]. Owing to the class imbalance in the data, models trained using performance metrics such as the accuracy are biased towards the more frequent class (sensitive to class skews), and may suffer from a lack of generalizability. We have therefore assessed classification performance using the balanced accuracy[65,66] which is defined as the average accuracy obtained on all classes:

$$BACC = \frac{1}{m}\sum_{i}^{m}\frac{k_i}{n_i} \tag{1}$$

where k_i is the number of correct predictions in class i, m is the number of classes and n_i is the number of examples in class i. Other metrics such as the average accuracy (the average per-class effectiveness of a classifier), sensitivity (the true positive rate - TPR) and specificity (the true negative rate - TNR) are also reported for comparison[67].

$$ACC = \frac{\sum_{i}^{m} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{m} \tag{2}$$

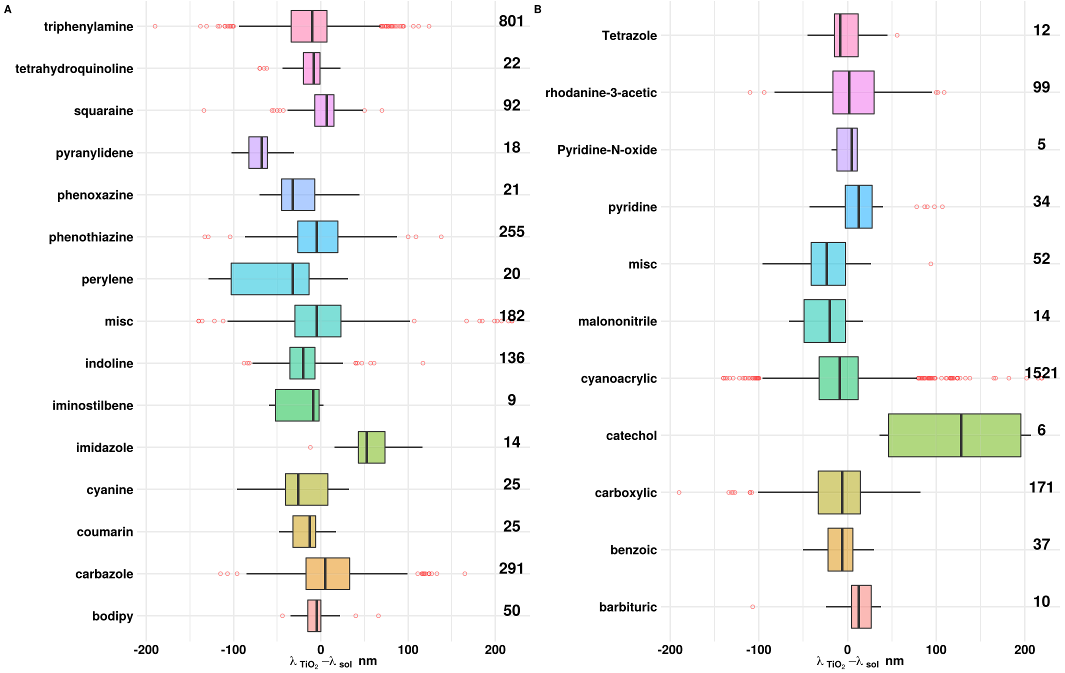


Figure 4. Box plots showing the distribution of the absorption shifts (irrespective of the solvent) based on the (A) class of the dyes and (B) the anchoring groups used. (**A**) The "misc" category includes various dyes based on pyrazoline, naphthoquinone, N,N-dialkylaniline[105], julolidine[106], bithiazole[107], cyclohexadiene[108] etc. (**B**) The "misc" category includes dyes with anchoring groups that include thiazolidine[73], aldehyde, hydantoin[109], isophorone[71], phosphonic acid and pyrimidine[110]. Numbers on the right are the counts of cases found in each category.

$$TPR = \frac{\sum_i^m \frac{tp_i}{tp_i + fn_i}}{m} \tag{3}$$

$$TNR = \frac{\sum_i^m \frac{tn_i}{tn_i + fp_i}}{m} \tag{4}$$

where, for an individual class $C_i - tp_i$, fp_i, tn_i and fn_i are the true positive, false positive, true negative and false negative counts respectively.

Results and Discussion

Manual analysis of the data. In order to ascertain if there were any noticeable patterns associated with the absorption shifts, the experimental data was analysed with respect to the class of the dye, the conjugated spacers used, and the number and types of anchoring groups. Figure 4A,B summarize the data in terms of the dye class and type of anchoring groups, respectively. For simplicity, we have ignored the solvent medium allowing for a broader analysis. Examination of Fig. 4A, shows that dyes based on imidazole[12,68] exclusively show red shifts, while those based on pyranylidene[69] exclusively show blue shifts. All other dye classes exhibit both blue and red shifts.

Analysis of the anchoring groups (Fig. 4B) suggests that, those containing catechol, pyridine, and barbituric acid are largely red-shifted. The particularly large red shifts in the catechol group are attributable to their ability to strongly adsorb on to the TiO_2 surface[70] as well as the increased dipole moment of the surface-bound metal-ligand complex[26,27]. For some of the other groups such as isophorone[71], malononitrile[72], thiazolidine[73] and benzoic acid[74], a majority of the cases are blue-shifted. The impact of multiple anchoring groups[75] was also studied (see Fig. F2 in the Supplementary Material-II). The number of dyes containing multiple anchoring groups was low (~170). Nonetheless, for the cases studied, there was no significant correlation between the number of anchoring groups and the size or direction of the spectral shift. While for some cases, little or no change was observed, others showed moderate to large shifts in either direction[76–80]. Wu *et al.*[77] observed the maximum absorption peak of three triphenylamine dyes in acetonitrile/tert-butanol (1:1, V/V) showed a red-shift with an increasing anchoring group number. Compared to the spectra in solution, the peaks did not show any change on the TiO_2 film, which was attributed to the cancelling effect of J-type aggregation and deprotonation. For squarylium dyes in particular, Connell *et al.*[81], have shown that the position of dye anchoring points can influence hydrophobicity and contact angle of dyes adsorbed to TiO_2 surfaces, which in turn can affect the absorption properties. In order

	B:N:R		NB:R		B:NR	
Method	**TRAIN**	**TEST**	**TRAIN**	**TEST**	**TRAIN**	**TEST**
GBM	0.68	0.71	0.73	0.72	0.73	0.72
RF	**0.71**	**0.76**	**0.76**	**0.80**	**0.76**	**0.80**
CART	0.61	0.61	0.65	0.65	0.65	0.65
k-NN	0.65	0.70	0.71	0.66	0.71	0.66
LDA	0.52	0.53	0.60	0.59	0.60	0.59
SVM	**0.71**	**0.73**	**0.77**	**0.79**	**0.77**	**0.79**

Table 2. Balanced accuracies obtained by the ML models for the calibration and test data.

examine the impact of mixed solvents, a total of 146 dyes in 12 different solvent mixtures were analysed. The absorption behaviour for dyes in mixed solvents is shown in Fig. F3 in the Supplementary Material-II. With the exception of solvent mixtures — methanol(MeOH)/chloroform(CHCl3), tert-butanol/acetonitrile(MeCN), ethanol(EtOH)-dichloromethane(DCM) and tetrahydrofuran(THF)-DCM, others had fewer than 10 instances. The dyes using tert-butanol/MeCN and MeOH/CHCl3 as solvents exhibited a greater tendency to blue-shift. For dyes in EtOH-DCM all three categories were equally represented while, for those in THF-DCM, a higher tendency to blue-shift was observed.

Several π-conjugated systems such as furan, thiophene and fused aromatic rings have been incorporated into the D-π-A architecture as π-linkers[82]. These units not only affect the light absorption regions of the DSSCs, but also influence the electron injection into the TiO_2 surface. For the dyes investigated in this study, a majority of the structures contained thiophene[83] and its derivatives (such as thienothiophene[84], indacenodithiophene[85], dithienopyrrole[86]) as the π-bridge. Figure F4 in Supplementary Material-II provides a box-plot of the absorption shifts for the various π-linkers (over 40 categories identified) used in the dyes. The conjugated spacers based on vinylene, ethynylene, furan, thiazole, thiophene and other fused aromatic segments (indole[87], fluorene[88], benzothiadiazole[89]) showed similar peak shifts in both directions. Other groups such as diphenylquinoxaline[90], 1-chlorobuta-1,3-diene[91], dithienobenzotriazole[92] and dithienobenzofurazan[93] found in a limited number of cases were largely associated with red-shifted peaks. On the other hand, those containing linkers based on fused thiophene derivatives such as dithienopyrrolobenzotriazole[94], cyclopentadithiophene[95], thienothienopyrrole[96], silolodithiophene[97] were blue shifted by more than 50 nm compare to the solution.

In conclusion, while for some choices of the dye class, anchoring groups and π-spacers we can identify clear patterns, in most cases there is no obvious pattern that can be discerned to predict the nature of the shift. In order to consider more formally, the effect of the structure on the adsorption behaviour, we employ machine learning the problem of predicting the type of the spectral shift and infer which features influence a set of observations.

Classification performance.

Table 2 summarizes the performance of the ML models across the calibration and test sets. In most cases, values for the two sets closely match one other, suggesting that the models generalize well. A comparative evaluation shows that both RF and SVM outperform other models on all classification tasks. The best performance is seen for the case *B*:*NR*, where the RF model achieves a cross-validated *BACC* = 0.76 during training and a slightly higher value of 0.80 on the test set containing 484 data points. On the same data set however, the LDA model performs only marginally better than random and achieves only a 50% accuracy on the other sets. Other models (*k*-NN, RF, SVM, GBM) are relatively more successful in separating classes by non-linear boundaries. Although the other binary classification problem *NB*:*R* has a moderate class imbalance (~2.5:1), RF classification accuracies are only slightly lower with *BACC* of 0.72 on the calibration and 0.73 on the test set.

In the case of the multiclass *B*:*N*:*R* problem, a *BACC* = 0.71 for the calibration data is obtained. Corresponding values for the test set are somewhat higher at 0.76. To better understand the classification performance, the ML predictions for the test set was examined on a class-wise basis. Values of the per class balanced accuracy (*BACC*), sensitivity (*TPR*) and specificity (*TNR*) are shown as bar plots in Fig. 5. For the blue-shifted (*B*) cases, all models show a high sensitivity, albeit with a fairly high rate of false alarms (decreased specificity). Given that there are twice as many cases of blue-shifted dyes, the classifier favours the majority class. A common practice to address the class imbalance problem is to balance them artificially where, for instance, cases from the minority class are replicated or alternatively by ignoring cases from the majority class. However, for the data sets in this study, no visible improvement in performance was observed when such schemes were used.

Descriptor analysis.

Compared to the CART scheme which uses a single tree, the tree-based ensemble models such as RF and GB are somewhat "black box" in nature given that they use multiple trees to arrive at a given outcome. We have therefore attempted to interpret the models by way of variable importance plots (shown in Fig. 6) that provide a qualitative understanding of the contribution that each input variable makes to the model[98,99]. We focus on the RF models that show the best performance. To this end, we examined the 10 most significant variables in the RF model trained for each task. The variable contributions are scaled to have a maximum value of 100 and those with higher values are expected to have a high predictive power. Given that the data contains various solvents, the relative polarity (P') is a major contributor. To assess the impact of a single descriptor, we calculated the accuracy of the classifier obtained by setting a threshold (typically the mean) on the value of the variable[98]. For the test set, the constitutional descriptor "FractionCSP3" (fraction of sp3-hybridized carbon atoms) yielded a single variable classifier with an accuracy of 54% and 58% for the entire data. The BalabanJ descriptor is indicative of a large degree of branching of the molecule. This agrees with the experimental data

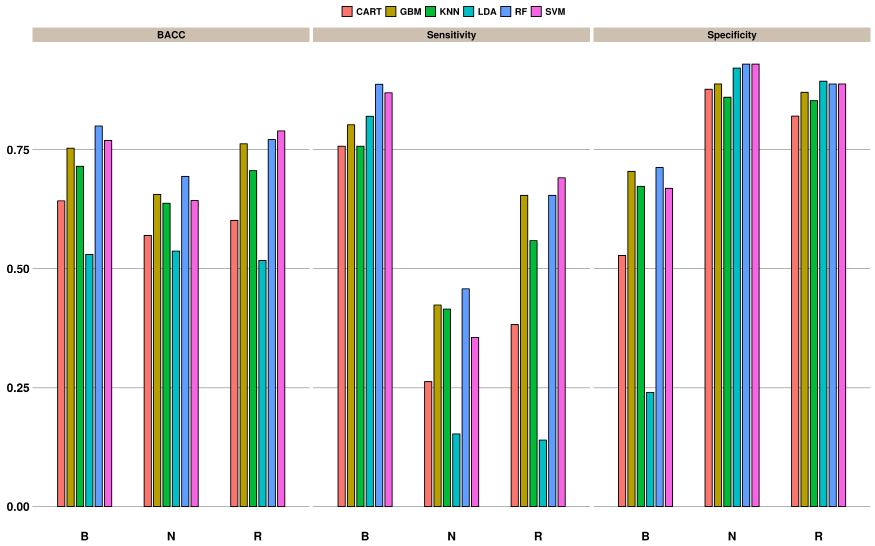


Figure 5. Bar plots showing the multiclass prediction performance on the test data. For each model, the per-class balanced accuracy, sensitivity and specificity are compared.

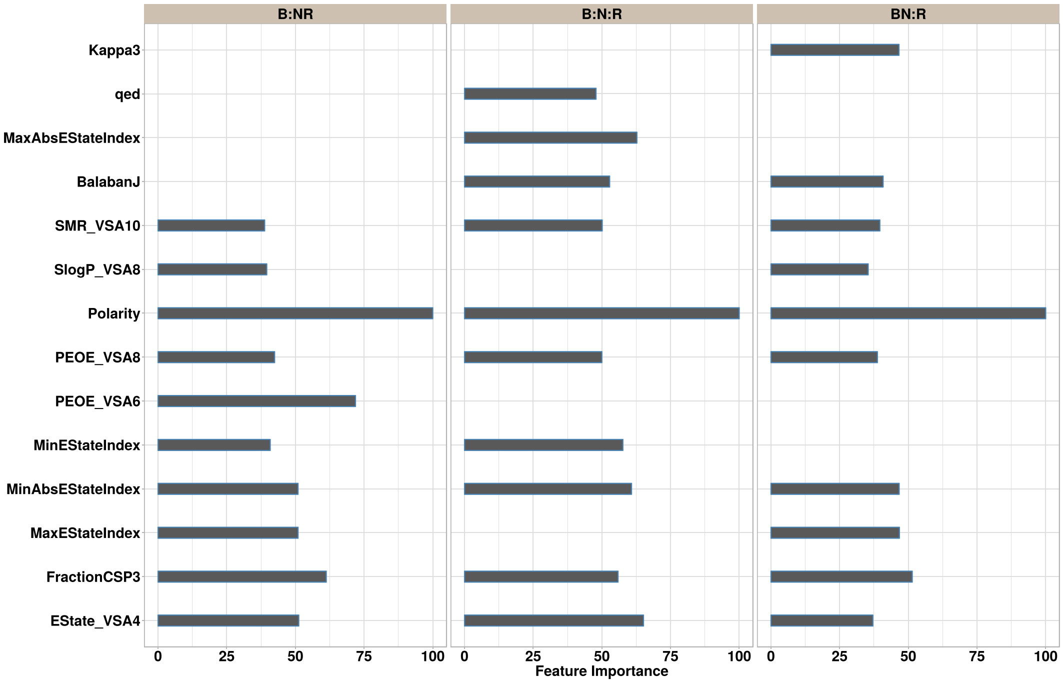


Figure 6. Variable importance plots for the RF model computed for each task: *B*/*NR*, *B*/*N*/*R* and *NB*/*R*. Only the top 10 most important variables are shown for each task. The bars show the contribution of the matching feature to the prediction. A missing bar for a given variable indicates that the said variable was ranked lower.

which shows that the large size of branched dyes can lead to a poor dye loading on the TiO_2 surface[28,100]. The E-state descriptors (MinEStateIndex, MaxEStateIndex, MinAbsEStateIndex, MaxAbsEStateIndex) for each atom in a given molecule reflect the steric and electronic effects of the surrounding atoms.

External validation. In order to test the performance of the ML models on unseen data, we examined the absorption behaviour for 3 dyes quercetin, 2,5-dihydroxytetraphthalic acid and carminic acid (purchased from Sigma-Aldrich) in ethanol and THF (see Fig. 7). Experimental details are provided in the Supplementary Material. While two of the dyes (T01, T02) show negligible change on adsorption, carminic acid (T03) shows a very small blue shift. However, based on the selected criteria ($|\Delta\lambda| \leq 10$) they are categorized as *NR*. The RF predictions for the dyes are listed in Table 3, which shows that all instances are correctly classified. We also investigated electronic absorption spectra of the isolated dye as well as those adsorbed on titania using a $(TiO_2)_9$ cluster[101]. The Gaussian 09[102] calculations were carried out using the B3LYP functional and the 6-31G(d,p) basis

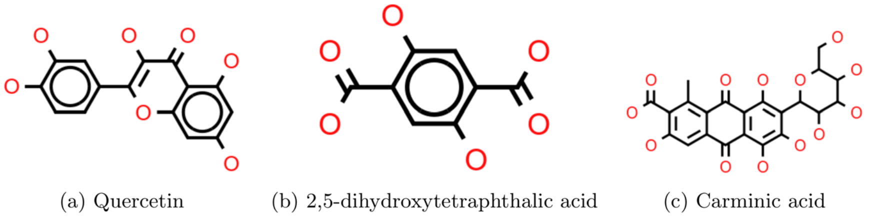

(a) Quercetin (b) 2,5-dihydroxytetraphthalic acid (c) Carminic acid

Figure 7. Structures of 3 dyes quercetin, 2,5-dihydroxytetraphthalic acid and carminic acid (purchased from Sigma-Aldrich).

Dye	Solvent	λ^{max}_{soln}	$\lambda^{max}_{TiO_2}$	Shift	ML	DFT
T01	Ethanol	373	376	NR	NR	NR
	THF	370	376	NR	NR	NR
T02	Ethanol	372	370	NR	NR	NR
	THF	376	370	NR	NR	NR
T03	Ethanol	495	487	NR	NR	NR
	THF	497	487	NR	NR	NR

Table 3. Comparison of the experimental and machine learning (RF) predictions for dyes (T01: quercetin, T02: 2,5-dihydroxytetraphthalic acid and T03: carminic acid shown in Fig. 7) in different solvents. The UV-VIS absorption spectra for the dyes in ethanol and THF and on TiO2 are shown in Figs F5–F7 in the Supplementary Material-II.

Category	BACC	ACC	Sensitivity	Specificity
B:NR	0.81	0.83	0.77	0.85
NB:R	0.81	0.75	0.94	0.67
B:N:R	0.71	0.73	0.56	0.86

Table 4. Classification performance for the RF model on a second independent test set.

set for the C, H, O and N atoms and the effective core potential LANL2DZ basis set for the Ti atoms. Solvent effects were considered using the using the conductor-like polarizable continuum model[103] (CPCM) along with the CAM-B3LYP functional. Computation times varied between 6–10 hours per structure. Although TD-DFT was not able to accurately predict the absorption peaks, its performance with respect to identifying the nature of the shift is comparable with that of the ML approach, albeit at a much higher computational cost.

The predictive performance of the RF model was also tested on an additional unseen data set. A second round of literature search was undertaken that yielded an additional set of 60 data points corresponding to 34 diverse dyes that included triphenylamine, indoline, bodipy, julolidine, and pyrenoimidazole based donors. Solvents in this list included dichloromethane (14 cases), THF (19), acetonitrile (6), toluene (9), DMF (6) and methanol (6). Table 4 summarizes the performance of the RF model on the second test set. The evaluation metrics are similar to those seen for the initial test set and reinforce the initial assessment of generalizability of the models.

Overall, the ML models, trained using only "two-dimensional" information consisting of atom types and connectivity, are capable of identifying dyes that have a propensity to blue- or red-shift on adsorption. A similar theoretical assessment using TD-DFT approaches requires the calculations to be carried out on both the isolated dye molecule isolated and adsorbed on TiO_2 clusters[32]. While the descriptor calculations are completed in less than a second, evaluations using DFT/TD-DFT approaches took more than 6–8 hours per structure. On the other hand, despite not being provided with the details about contributing factors such as the adsorption mode and strength of dye-cluster coupling, the ML models are able to deduce the nature of the absorption shift with reasonable accuracy (~70–80%).

In this work, we have outlined a data-driven approach that we believe could serve as a useful tool to exclude dyes adsorbed on TiO_2 that are likely to exhibit undesirable photosensitization behaviour. We have shown that the approach can indicate the exact nature of the spectral shift in 70–80% of the dyes inspected. The predictive models afford a higher reliability than any experienced human expert. In addition, the advantages in terms of speed and versatility will certainly outweigh any possible gains that may be achieved with the more time-consuming DFT-based approaches. The models can be easily integrated into material screening frameworks that allow for the rapid computational assessment of candidate structures.

Data availability

The data used in the article were manually extracted from literature. The molecular structures (in SMILES format), values of the spectral shift and corresponding references have been made available in the Supplementary Information. Scripts used in the calculation of the molecular descriptors are available upon request.

Received: 31 July 2019; Accepted: 4 November 2019;
Published online: 18 November 2019

References

1. Hagfeldt, A., Boschloo, G., Sun, L., Kloo, L. & Pettersson, H. Dye-sensitized solar cells. *Chem. Rev.* **110**, 6595–6663, https://doi.org/10.1021/cr900356p (2010).
2. Brennaman, M. K. *et al.* Finding the way to solar fuels with dye-sensitized photoelectrosynthesis cells. *J Am. Chem. Soc.* **138**, 13085–13102, https://doi.org/10.1021/jacs.6b06466 (2016).
3. Dandin, M., Abshire, P. & Smela, E. Optical filtering technologies for integrated fluorescence sensors. *Lab Chip* **7**, 955, https://doi.org/10.1039/b704008c (2007).
4. Yamazaki, M. *et al.* Non-emissive colour filters for fluorescence detection. *Lab Chip* **11**, 1228, https://doi.org/10.1039/c0lc00642d (2011).
5. Yamazaki, M., Krishnadasan, S., deMello, A. J. & deMello, J. C. Non-emissive plastic colour filters for fluorescence detection. *Lab Chip* **12**, 4313, https://doi.org/10.1039/c2lc40718c (2012).
6. Lee, C.-P. *et al.* Recent progress in organic sensitizers for dye-sensitized solar cells. *RSC Adv.* **5**, 23810–23825, https://doi.org/10.1039/c4ra16493h (2015).
7. Urbani, M., Grätzel, M., Nazeeruddin, M. K. & Torres, T. Meso-substituted porphyrins for dye-sensitized solar cells. *Chem. Rev.* **114**, 12330–12396, https://doi.org/10.1021/cr5001964 (2014).
8. Venkatraman, V., Raju, R., Oikonomopoulos, S. P. & Alsberg, B. K. The dye-sensitized solar cell database. *J. Cheminf.* **10**, https://doi.org/10.1186/s13321-018-0272-0 (2018).
9. Li, W., Wu, Y., Zhang, Q., Tian, H. & Zhu, W. D-a-p-a featured sensitizers bearing phthalimide and benzotriazole as auxiliary acceptor: Effect on absorption and charge recombination dynamics in dye-sensitized solar cells. *ACS Appl. Mater. Inter.* **4**, 1822–1830, https://doi.org/10.1021/am3001049 (2012).
10. Namuangruk, S. *et al.* D-d-p-a-type organic dyes for dye-sensitized solar cells with a potential for direct electron injection and a high extinction coefficient: Synthesis, characterization, and theoretical investigation. *J. Phys. Chem. C* **116**, 25653–25663, https://doi.org/10.1021/jp304489t (2012).
11. Zilberberg, K., Meyer, J. & Riedl, T. Solution processed metal-oxides for organic electronic devices. *J Mater. Chem. C* **1**, 4796, https://doi.org/10.1039/c3tc30930d (2013).
12. Kumar, D., Thomas, K. R. J., Lee, C.-P. & Ho, K.-C. Novel pyrenoimidazole-based organic dyes for dye-sensitized solar cells. *Org. Lett.* **13**, 2622–2625, https://doi.org/10.1021/ol2006874 (2011).
13. Hua, Y. *et al.* New phenothiazine-based dyes for efficient dye-sensitized solar cells: Positioning effect of a donor group on the cell performance. *J Power Sources* **243**, 253–259, https://doi.org/10.1016/j.jpowsour.2013.05.157 (2013).
14. Cherepy, N. J., Smestad, G. P., Grätzel, M. & Zhang, J. Z. Ultrafast electron injection: implications for a photoelectrochemical cell utilizing an anthocyanin dye-sensitized TiO_2 nanocrystalline electrode. *J Phys. Chem. B* **101**, 9342–9351, https://doi.org/10.1021/jp972197w (1997).
15. Cheng, M. *et al.* Efficient organic dye-sensitized solar cells: Molecular engineering of donor-acceptor-acceptor cationic dyes. *ChemSusChem* **6**, 2322–2329, https://doi.org/10.1002/cssc.201300481 (2013).
16. Cheng, M. *et al.* Effect of the acceptor on the performance of dye-sensitized solar cells. *Phys. Chem. Chem. Phys.* **15**, 17452, https://doi.org/10.1039/c3cp52314d (2013).
17. Fakis, M. *et al.* Excited state and injection dynamics of triphenylamine sensitizers containing a benzothiazole electronaccepting group on TiO_2 and al2o3 thin films. *J Phys. Chem. C* **118**, 28509–28519, https://doi.org/10.1021/jp509971q (2014).
18. Margalias, A. *et al.* The effect of additional electron donating group on the photophysics and photovoltaic performance of two new metal free d-p-a sensitizers. *Dye. Pigment.* **121**, 316–327, https://doi.org/10.1016/j.dyepig.2015.05.028 (2015).
19. Bahng, H.-W., Hagfeldt, A. & Moser, J.-E. Donor effect on the photoinduced interfacial charge transfer dynamics of d-p-a diketopyrrolopyrrole dye sensitizers adsorbed on titanium dioxide. *J Phys. Chem. C* **122**, 19359–19369, https://doi.org/10.1021/acs.jpcc.8b04819 (2018).
20. Hagberg, D. P. *et al.* A novel organic chromophore for dye-sensitized nanostructured solar cells. *Chem. Commun.* **2245**, https://doi.org/10.1039/b603002e (2006).
21. Lin, L.-Y. *et al.* Organic dyes containing coplanar diphenyl-substituted dithienosilole core for efficient dye-sensitized solar cells. *The J. Org. Chem.* **75**, 4778–4785, https://doi.org/10.1021/jo100762t (2010).
22. Zhang, L. & Cole, J. M. Dye aggregation in dye-sensitized solar cells. *J Mater. Chem. A* **5**, 19541–19559, https://doi.org/10.1039/c7ta05632j (2017).
23. Bayer, E., Egeter, H., Fink, A., Nether, K. & Wegmann, K. Complex formation and flower colors. *Angewandte Chemie Int. Ed. Engl.* **5**, 791–798, https://doi.org/10.1002/anie.196607911 (1966).
24. Goto, T. & Kondo, T. Structure and molecular stacking of anthocyanins—flower color variation. *Angewandte Chemie Int. Ed. Engl.* **30**, 17–33, https://doi.org/10.1002/anie.199100171 (1991).
25. Dangles, O., Elhabiri, M. & Brouillard, R. Kinetic and thermodynamic investigation of the aluminium-anthocyanin complexation in aqueous solution. *J. Chem. Soc., Perkin Trans.* **2**, 2587–2596, https://doi.org/10.1039/p29940002587 (1994).
26. An, B.-K., Hu, W., Burn, P. L. & Meredith, P. New type II catechol-thiophene sensitizers for dye-sensitized solar cells. *J. Phys. Chem. C* **114**, 17964–17974, https://doi.org/10.1021/jp105687z (2010).
27. Rajh, T. *et al.* Surface restructuring of nanoparticles: an efficient route for ligand-metal oxide crosstalk. *J. Phys. Chem. B* **106**, 10543–10552, https://doi.org/10.1021/jp021235v (2002).
28. Fischer, M. K. R. *et al.* D-π-a sensitizers for dye-sensitized solar cells: Linear vs branched oligothiophenes. *Chem. Mater.* **22**, 1836–1845, https://doi.org/10.1021/cm903542v (2010).
29. Zhang, L., Cole, J. M. & Dai, C. Variation in optoelectronic properties of azo dye-sensitized TiO2 semiconductor interfaces with different adsorption anchors: Carboxylate, sulfonate, hydroxyl and pyridyl groups. *ACS Appl. Mater. Inter.* **6**, 7535–7546, https://doi.org/10.1021/am502186k (2014).
30. Fang, H. *et al.* Effects of molecular structure and solvent polarity on adsorption of carboxylic anchoring dyes onto TiO_2 particles in aprotic solvents. *Langmuir* **33**, 7036–7042, https://doi.org/10.1021/acs.langmuir.7b01442 (2017).
31. Fantacci, S., Angelis, F. D. & Selloni, A. Absorption spectrum and solvatochromism of the [ru(4,4′-COOH-2,2′-bpy)2(NCS)2] molecular dye by time dependent density functional theory. *J Am. Chem. Soc.* **125**, 4381–4387, https://doi.org/10.1021/ja0207910 (2003).
32. de Armas, R. S. *et al.* Real-time TD-DFT simulations in dye sensitized solar cells: The electronic absorption spectrum of alizarin supported on TiO2nanoclusters. *J. Chem. Theory Comput.* **6**, 2856–2865, https://doi.org/10.1021/ct100289t (2010).

33. Sousa, C., Tosoni, S. & Illas, F. Theoretical approaches to excited-state-related phenomena in oxide surfaces. *Chem. Rev.* **113**, 4456–4495, https://doi.org/10.1021/cr300228z (2012).
34. Zhang, L., Liu, X., Rao, W. & Li, J. Multilayer dye aggregation at dye/TiO_2 interface via π-π stacking and hydrogen bond and its impact on solar cell performance: A DFT analysis. *Sci. Rep.* **6**, https://doi.org/10.1038/srep35893 (2016).
35. Mendizabal, F., Mera-Adasme, R., Xu, W.-H. & Sundholm, D. Electronic and optical properties of metalloporphyrins of zinc on TiO2 cluster in dye-sensitized solar-cells (DSSC). a quantum chemistry study. *RSC Adv.* **7**, 42677–42684, https://doi.org/10.1039/c7ra08648b (2017).
36. Pastore, M. & Angelis, F. D. Aggregation of organic dyes on TiO2 in dye-sensitized solar cells models: An ab initio investigation. *ACS Nano* **4**, 556–562, https://doi.org/10.1021/nn901518s (2009).
37. de Armas, R. S., San-Miguel, M. A., Oviedo, J., Márquez, A. & Sanz, J. F. Electronic structure and optical spectra of catechol on TiO2nanoparticles from real time TD-DFT simulations. *Phys. Chem. Chem. Phys.* **13**, 1506–1514, https://doi.org/10.1039/c0cp00906g (2011).
38. Agrawal, S. *et al*. Optical properties and aggregation of phenothiazine-based dye-sensitizers for solar cells applications: A combined experimental and computational investigation. *J Phys. Chem. C* **117**, 9613–9622, https://doi.org/10.1021/jp4026305 (2013).
39. Jacquemin, D., Wathelet, V., Perpète, E. A. & Adamo, C. Extensive TD-DFT benchmark: Singlet-excited states of organic molecules. *J Chem. Theory Comp.* **5**, 2420–2435, https://doi.org/10.1021/ct900298e (2009).
40. Adamo, C. & Jacquemin, D. The calculations of excited-state properties with time-dependent density functional theory. *Chem. Soc. Rev.* **42**, 845–856, https://doi.org/10.1039/c2cs35394f (2013).
41. Anselmi, C., Mosconi, E., Pastore, M., Ronca, E. & Angelis, F. D. Adsorption of organic dyes on TiO2 surfaces in dyesensitized solar cells: interplay of theory and experiment. *Phys. Chem. Chem. Phys.* **14**, 15963, https://doi.org/10.1039/c2cp43006a (2012).
42. Calogero, G., Bartolotta, A., Marco, G. D., Carlo, A. D. & Bonaccorso, F. Vegetable-based dye-sensitized solar cells. *Chem. Soc. Rev.* **44**, 3244–3294, https://doi.org/10.1039/c4cs00309h (2015).
43. Rajan, K. (ed.) *Informatics for Materials Science and Engineering* (Elsevier, 2013).
44. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 053208, https://doi.org/10.1063/1.4946894 (2016).
45. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comp. Mater*. **3**, https://doi.org/10.1038/s41524-017-0056-5 (2017).
46. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nat.* **559**, 547–555, https://doi.org/10.1038/s41586-018-0337-2 (2018).
47. Gubernatis, J. E. & Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2**, 120301, https://doi.org/10.1103/PhysRevMaterials.2.120301 (2018).
48. Zhao, J., Yang, X., Cheng, M., Li, S. & Sun, L. Molecular design and performance of hydroxylpyridium sensitizers for dye-sensitized solar cells. *ACS Appl. Mater. Inter.* **5**, 5227–5231, https://doi.org/10.1021/am4010545 (2013).
49. Reichardt, C. & Welton, T. *Solvents and Solvent Effects in Organic Chemistry* (Wiley-VCH Verlag GmbH & Co. KGaA, 2010).
50. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study1. *Intell. Data Anal.* **6**, 429–449, https://doi.org/10.3233/IDA-2002-6504 (2002).
51. Todeschini, R. & Consonni, V. (eds) *Molecular Descriptors for Chemoinformatics* (Wiley-VCH Verlag GmbH & Co. KGaA, 2009).
52. Baskin, I. Chapter 1. fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening. In Varnek, A. (ed.) *Chemoinformatics Approaches to Virtual Screening*, 1–43, https://doi.org/10.1039/9781847558879-00001 (Royal Society of Chemistry, 2008).
53. Ruggiu, F., Marcou, G., Solov'ev, V., Horvath, D. & Varnek, A. Isida fragmentor (2017).
54. Hall, L. H. & Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J Chem. Inf. Model.* **35**, 1039–1045, https://doi.org/10.1021/ci00028a014 (1995).
55. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **18**, 464–477, https://doi.org/10.1016/s1093-3263(00)00068-1 (2000).
56. Landrum, G. Rdkit: Open-source cheminformatics software, https://www.rdkit.org/ (2018).
57. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition* (John Wiley & Sons, Inc., 1992).
58. Luts, J. *et al*. A tutorial on support vector machine-based methods for classification problems in chemometrics. *Anal. Chim. Acta* **665**, 129–145, https://doi.org/10.1016/j.aca.2010.03.030 (2010).
59. Loh, W.-Y. Classification and regression trees. *WIRES Data Min Knowl.* **1**, 14–23, https://doi.org/10.1002/widm.8 (2011).
60. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, https://doi.org/10.1023/a:1010933404324 (2001).
61. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals Stat.* **29**, 1189–1232 (2001).
62. Dearden, J., Cronin, M. & Kaiser, K. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Env. Res.* **20**, 241–266, https://doi.org/10.1080/10629360902949567 (2009).
63. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **29**, 476–488, https://doi.org/10.1002/minf.201000061 (2010).
64. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
65. Carrillo, H., Brodersen, K. H. & Castellanos, J. A. Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy. In *ROBOT2013: First Iberian Robotics Conference*, 347–361, https://doi.org/10.1007/978-3-319-03413-3_25 (Springer International Publishing, 2014).
66. Tharwat, A. Classification assessment methods. *Appl. Comp. Inf.* **0**, 0, https://doi.org/10.1016/j.aci.2018.08.003 (2018).
67. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45**, 427–437, https://doi.org/10.1016/j.ipm.2009.03.002 (2009).
68. Karthik, D., Kumar, V., Thomas, K. J., Li, C.-T. & Ho, K.-C. Synthesis and characterization of thieno[3,4-d]imidazolebased organic sensitizers for photoelectrochemical cells. *Dye. Pigment.* **129**, 60–70, https://doi.org/10.1016/j.dyepig.2016.02.009 (2016).
69. Marco, A. B. *et al*. Pyranylidene/thienothiophene-based organic sensitizers for dye-sensitized solar cells. *Dye. Pigment.* **161**, 205–213, https://doi.org/10.1016/j.dyepig.2018.09.035 (2019).
70. Liu, Y., Dadap, J. I., Zimdars, D. & Eisenthal, K. B. Study of interfacial charge-transfer complex on TiO2particles in aqueous suspension by second-harmonic generation. *J Phys. Chem. B* **103**, 2480–2486, https://doi.org/10.1021/jp984288e (1999).
71. Chen, Y.-C., Lin, Y.-Z., Cheng, Y.-T. & Chang, Y. J. Synthesis of novel isophorone-based dyes for dye-sensitized solar cells. *RSC Adv.* **5**, 96428–96436, https://doi.org/10.1039/c5ra17898c (2015).
72. Qian, X. *et al*. Indeno[1,2- b]indole-based organic dyes with different acceptor groups for dye-sensitized solar cells. *Dye. Pigment.* **139**, 274–282, https://doi.org/10.1016/j.dyepig.2016.12.028 (2017).
73. Naik, P., Su, R., Elmorsy, M. R., El-Shafei, A. & Adhikari, A. V. New di-anchoring a–π-d–π-a configured organic chromophores for DSSC application: sensitization and co-sensitization studies. *Photochem. Photobiol. Sci.* **17**, 302–314, https://doi.org/10.1039/c7pp00351j (2018).
74. Song, X. *et al*. Influence of ethynyl position on benzothiadiazole based d–a–π–a dye-sensitized solar cells: spectral response and photovoltage performance. *J Mater. Chem. C* **4**, 9203–9211, https://doi.org/10.1039/c6tc03418g (2016).
75. Kumar, D. & Wong, K.-T. Organic dianchor dyes for dye-sensitized solar cells. *Mater. Today Energy* **5**, 243–279, https://doi.org/10.1016/j.mtener.2017.05.007 (2017).

76. Liu, J. *et al*. Mesoscopic titania solar cells with the tris(1,10-phenanthroline)cobalt redox shuttle: uniped versus biped organic dyes. *Energy Env. Sci.* **4**, 3021, https://doi.org/10.1039/c1ee01633d (2011).

77. Wu, G. *et al*. Multiple-anchoring triphenylamine dyes for dye-sensitized solar cell application. *J Phys. Chem. C* **118**, 8756–8765, https://doi.org/10.1021/jp4124265 (2014).

78. Jiang, S., Fan, S., Lu, X., Zhou, G. & Wang, Z.-S. Double d–π–a branched organic dye isomers for dye-sensitized solar cells. *J. Mater. Chem. A* **2**, 17153–17164, https://doi.org/10.1039/c4ta03451a (2014).

79. Connell, A. *et al*. Multiple linker half-squarylium dyes for dye-sensitized solar cells; are two linkers better than one? *J. Mater. Chem. A* **3**, 2883–2894, https://doi.org/10.1039/c4ta06896c (2015).

80. Raju, T. B., Vaghasiya, J. V., Afroz, M. A., Soni, S. S. & Iyer, P. K. Twisted donor substituted simple thiophene dyes retard the dye aggregation and charge recombination in dye-sensitized solar cells. *Org. Elec.* **50**, 25–32, https://doi.org/10.1016/j.orgel.2017.07.019 (2017).

81. Connell, A. *et al*. A study of dye anchoring points in half-squarylium dyes for dye-sensitized solar cells. *J. Mater. Chem. A* **2**, 4055–4066, https://doi.org/10.1039/c3ta15278b (2014).

82. Prachumrak, N. *et al*. Improvement of d–π–a organic dye-based dye-sensitized solar cell performance by simple triphenylamine donor substitutions on the π-linker of the dye. *Mater. Chem. Front.* **1**, 1059–1072, https://doi.org/10.1039/c6qm00271d (2017).

83. Shen, P. *et al*. Efficient triphenylamine dyes for solar cells: Effects of alkyl-substituents and π-conjugated thiophene unit. *Dye. Pigment.* **83**, 187–197, https://doi.org/10.1016/j.dyepig.2009.04.005 (2009).

84. Zhang, G. *et al*. Employ a bisthienothiophene linker to construct an organic chromophore for efficient and stable dye-sensitized solar cells. *Energy Environ. Sci.* **2**, 92–95, https://doi.org/10.1039/b817990e (2009).

85. Chen, J.-H. *et al*. Organic dyes containing a coplanar indacenodithiophene bridge for high-performance dye-sensitized solar cells. *J. Org. Chem.* **76**, 8977–8985, https://doi.org/10.1021/jo201730a (2011).

86. Dong, H. *et al*. Twisted fused-ring thiophene organic dye-sensitized solar cells. *J. Phys. Chem. C* **120**, 22822–22830, https://doi.org/10.1021/acs.jpcc.6b06604 (2016).

87. Wang, Y., Wan, Z., Jia, C. & Yao, X. Indole-based organic dyes with different electron donors for dye-sensitized solar cells. *Synth. Met.* **211**, 40–48, https://doi.org/10.1016/j.synthmet.2015.10.024 (2016).

88. Marzari, G. *et al*. Fluorous molecules for dye-sensitized solar cells: Synthesis and characterization of fluorene-bridged donor/acceptor dyes with bulky perfluoroalkoxy substituents. *J. Phys. Chem. C* **116**, 21190–21200, https://doi.org/10.1021/jp305884u (2012).

89. Bolisetty, M. P., Li, C.-T., Thomas, K. J., Bodedla, G. B. & Ho, K.-C. Benzothiadiazole-based organic dyes with pyridine anchors for dye-sensitized solar cells: effect of donor on optical properties. *Tetrahedron* **71**, 4203–4212, https://doi.org/10.1016/j.tet.2015.04.089 (2015).

90. Li, S.-R., Lee, C.-P., Kuo, H.-T., Ho, K.-C. & Sun, S.-S. High-performance dipolar organic dyes with an electrondeficient diphenylquinoxaline moiety in the π-conjugation framework for dye-sensitized solar cells. *Chem.: Eur. J* **18**, 12085–12095, https://doi.org/10.1002/chem.201201000 (2012).

91. Lokhande, P. K. M. *et al*. Multi-dentate carbazole based schiff base dyes with chlorovinylene group in spacer for dyesensitized solar cells: A combined theoretical and experimental study. *Chem.* **4**, 4044–4056, https://doi.org/10.1002/slct.201803940 (2019).

92. Ni, J.-S., Yen, Y.-C. & Lin, J. T. Organic sensitizers with a rigid dithienobenzotriazole-based spacer for high-performance dye-sensitized solar cells. *J Mater. Chem. A* **4**, 6553–6560, https://doi.org/10.1039/c6ta02275h (2016).

93. Ni, J.-S. *et al*. Organic dyes incorporating the dithieno[30,20:3,4;200,300:5,6]benzo[1,2-c]furazan moiety for dye-sensitized solar cells. *ACS Appl. Mater. Inter.* **6**, 22612–22621, https://doi.org/10.1021/am5067145 (2014).

94. Huang, Z.-S. *et al*. Dithienopyrrolobenzotriazole-based organic dyes with high molar extinction coefficient for efficient dye-sensitized solar cells. *Dye. Pigment.* **125**, 229–240, https://doi.org/10.1016/j.dyepig.2015.10.022 (2016).

95. Li, X. *et al*. Insight into quinoxaline containing d–π–a dyes for dye-sensitized solar cells with cobalt and iodine based electrolytes: the effect of π-bridge on the HOMO energy level and photovoltaic performance. *J Mater. Chem. A* **3**, 21733–21743, https://doi.org/10.1039/c5ta07254a (2015).

96. Wang, Z. *et al*. Asymmetric 8h-thieno[20,30:4,5]thieno[3,2-b]thieno[2,3-d]pyrrole-based sensitizers: Synthesis and application in dye-sensitized solar cells. *Org. Lett.* **19**, 3711–3714, https://doi.org/10.1021/acs.orglett.7b01465 (2017).

97. Gao, Y. *et al*. Effect of an auxiliary acceptor on d–a–π–a sensitizers for highly efficient and stable dye-sensitized solar cells. *J Mater. Chem. A* **4**, 12865–12877, https://doi.org/10.1039/c6ta05588e (2016).

98. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).

99. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J Anim. Ecol.* **77**, 802–813, https://doi.org/10.1111/j.1365-2656.2008.01390.x (2008).

100. Meier, H., Huang, Z.-S. & Cao, D. Double d–π–a branched dyes – a new class of metal-free organic dyes for efficient dye-sensitized solar cells. *J Mater. Chem. C* **5**, 9828–9837, https://doi.org/10.1039/c7tc03406g (2017).

101. de Armas, R. S., Miguel, M. Á. S., Oviedo, J. & Sanz, J. F. Coumarin derivatives for dye sensitized solar cells: a TD-DFT study. *Phys. Chem. Chem. Phys.* **14**, 225–233, https://doi.org/10.1039/c1cp22058f (2012).

102. Frisch, M. J. *et al*. Gaussian 09, revision b.01. Gaussian, Inc., Wallingford CT, (2010).

103. Takano, Y. & Houk, K. N. Benchmarking the conductor-like polarizable continuum model (CPCM) for aqueous solvation free energies of neutral and ionic organic molecules. *J. Chem. Theory Comput.* **1**, 70–77, https://doi.org/10.1021/ct049977a (2004).

104. Snyder, L. R. Classification off the solvent properties of common liquids. *J Chromatogr. Sci.* **16**, 223–234, https://doi.org/10.1093/chromsci/16.6.223 (1978).

105. Duerto, I. *et al*. DSSCs based on aniline derivatives functionalized with a tert -butyldimethylsilyl group and the effect of the π-spacer. *Dye. Pigment.* **148**, 61–71, https://doi.org/10.1016/j.dyepig.2017.07.063 (2018).

106. Wu, G. *et al*. Julolidine dyes with different acceptors and thiophene-conjugation bridge: Design, synthesis and their application in dye-sensitized solar cells. *Synth. Met.* **180**, 9–15, https://doi.org/10.1016/j.synthmet.2013.07.028 (2013).

107. Lai, L.-F. *et al*. New bithiazole-functionalized organic photosensitizers for dye-sensitized solar cells. *Dye. Pigment.* **96**, 516–524, https://doi.org/10.1016/j.dyepig.2012.10.002 (2013).

108. Chen, K.-F. *et al*. Photophysical studies of dipolar organic dyes that feature a 1,3-cyclohexadiene conjugated linkage: The implication of a twisted intramolecular charge-transfer state on the efficiency of dye-sensitized solar cells. *Chem. Eur. J.* **16**, 12873–12882, https://doi.org/10.1002/chem.201001294 (2010).

109. Guo, F.-L. *et al*. Metal-free sensitizers containing hydantoin acceptor as high performance anchoring group for dyesensitized solar cells. *Adv. Func. Mater.* **26**, 5733–5740, https://doi.org/10.1002/adfm.201601305 (2016).

110. Wu, Z., Li, X., Ågren, H., Hua, J. & Tian, H. Pyrimidine-2-carboxylic acid as an electron-accepting and anchoring group for dye-sensitized solar cells. *ACS Appl. Mater. Inter.* **7**, 26355–26359, https://doi.org/10.1021/acsami.5b07690 (2015).

Acknowledgements

This work was supported through a grant (Grant No. 262152) from the Research Council of Norway (RCN).

Author contributions

V.V. and J.M. conceived the study. V.V. collected the data, performed the calculations, analysed the results and wrote the manuscript. A.E.Y. performed the experimental validation for the dyes. J.M. helped revise the manuscript. All authors discussed and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-53534-2.

Correspondence and requests for materials should be addressed to V.V.