

# Journal of Materials Chemistry A

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

*Accepted Manuscripts* are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.



[www.rsc.org/materialsA](http://www.rsc.org/materialsA)

# Evolutionary *de novo* design of phenothiazine derivatives for dye-sensitized solar cells<sup>†</sup>

Vishwesh Venkatraman,<sup>a</sup> Marco Foscato,<sup>b</sup> Vidar R. Jensen,<sup>b</sup> and Bjørn Kåre Alsberg<sup>\*a</sup>

Received Xth XXXXXXXXXXXX 20XX, Accepted Xth XXXXXXXXXXXX 20XX

First published on the web Xth XXXXXXXXXXXX 200X

DOI: 10.1039/b000000x

Traditional approaches for improving the photovoltaic performance of dye-sensitized solar cells (DSSCs) have mainly relied on judicious molecular design and device level modifications. Such schemes, however, are bound by costly and time-consuming synthesis procedures. In this paper, we demonstrate the efficacy of an alternative approach based on *in silico* evolutionary *de novo* design of novel dye structures with improved DSSC power conversion efficiency (*PCE*) values. Because the *PCE*, cannot as yet be directly computed from first principles, the evolutionary fitness function utilizes predictive structure-property relationship (QSPR) models calibrated from empirical data. Our design approach is applied to phenothiazine-based dye sensitizers. The chemical structure space is explored using a genetic algorithm that systematically assembles molecules from fragments in a synthetically tractable manner. Five novel phenothiazine dyes are proposed using our approach where all have predicted *PCE* values above 9%.

## 1 Introduction

The field of photovoltaics has seen significant developments in the recent years with a wide variety of solar photovoltaic technologies emerging that range from those based on crystalline silicon and thin films to organic and polymer cells<sup>1</sup>. Among these, dye sensitised solar cells (DSSCs) have attracted considerable attention owing to their use of cheap materials and short energy payback time<sup>2,3</sup>. The cell primarily consists of (i) a photosensitive dye that is bound to a wide band gap semiconductor photoanode (typically TiO<sub>2</sub>), (ii) a redox electrolyte (such as iodide/tri-iodide ((I<sup>-</sup>/I<sub>3</sub><sup>-</sup>))) and (iii) a platinized photocathode. The dye, upon excitation (absorption of incoming light), injects electrons into the conduction band of the TiO<sub>2</sub>, that are then transported (through diffusion) to the back collecting electrode. The oxidized dye is regenerated by electron donation from the electrolyte while the resulting triiodide is reduced back to iodide ions at the counter-electrode<sup>4,5</sup>. The overall efficiency of conversion of solar-to-electrical energy of the cell *PCE* is given by:

$$PCE(\%) = \frac{J_{sc} \cdot V_{oc} \cdot FF}{P_{inc}} \times 100 \quad (1)$$

where  $J_{sc}$  is the short circuit current (measured in mA cm<sup>-2</sup>),  $V_{oc}$  is the open circuit voltage (measured in mV),  $FF$  is the

fill factor and  $P_{inc} = 100W/cm^2$  is the intensity of the incident light (for the AM1.5 simulated solar illumination).

Although a number of ruthenium-based dye sensitizer<sup>6,7</sup> with efficiencies of up to 10% have been reported, concerns over the availability of the expensive rare earth metals has led to research on ruthenium-free alternatives. For instance, zinc-porphyrin dyes exceeding 12% conversion efficiency<sup>8</sup> have been seen as promising substitutes. Interest has also been increasing in metal-free organic dyes which have potentially lower production costs and easily tunable absorption and electrochemical properties<sup>9</sup>. Highest efficiencies reported for these organic dyes are around 10.2%<sup>10,11</sup>. However, further improvements in terms of performance are required in order to make DSSCs more commercially viable and competitive with existing technologies.

In DSSCs, the functions of electron transport, light absorption and hole transport are handled by separate materials that make it fairly modular<sup>5</sup>. Improvements in the device efficiency can be obtained through optimisation of the component materials. For example, the use of alternative metal oxides such as SnO<sub>2</sub><sup>12</sup> and ZnO<sup>13</sup> as the photoanode have been investigated while other studies have focused on iodide-free systems and counter electrode materials<sup>14</sup>. The majority of the efforts are largely centered on the dye sensitizer since it governs the photon harvesting and influences many of the key electron transfer processes that impact photovoltaic performance<sup>5,9,15</sup>. To this end, several metal-free organic dyes such as squaraines<sup>16</sup>, hemicyanines<sup>17</sup>, triphenylamines<sup>18,19</sup>, coumarins<sup>20,21</sup> and phenothiazines<sup>22-24</sup> have been analyzed<sup>9,25,26</sup>. For optimal performance, dyes are required to have a wide absorption in the visible region, high

<sup>†</sup> Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

<sup>a</sup> Department of Chemistry, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway. E-mail: alsberg@ntnu.no

<sup>b</sup> Department of Chemistry, University of Bergen, Bergen, Norway

molar extinction coefficients, photochemical stability and oxide surface anchoring, to name a few<sup>27,28</sup>. The ease of design and modification has enabled a number of molecular configurations such as  $D-\pi-A$  and  $D-A-\pi-A$  to be studied<sup>9,25</sup>.

To date, most dyes have generally been discovered through experimental means, which are often time-consuming and expensive. While there is a need for new molecules with novel and desirable attributes, current materials discovery that is largely guided by chemical intuition and serendipity, cannot simultaneously explore and optimise relevant parameters in the vast structure space. In this article, a new approach based on *in silico* evolutionary *de novo* design is presented, wherein molecular structures are systematically assembled and iteratively refined in a computer according to principles of evolution<sup>29,30</sup>. Central to this scheme is the computation of the value of the property (to be optimised) for each molecule, also referred to as the *fitness*. Given the interest in high performance devices, we focus our attention on maximizing the *PCE* (see Eqn. 1) as fitness in our evolutionary *de novo* scheme. Although physical characteristics such as charge collection efficiency<sup>31</sup>, charge recombination and device dependent internal resistances<sup>32</sup> can influence the DSSC performance, we show that improvements can be made through structural modification of the dye sensitizer alone<sup>33</sup>.

## 2 Methods and Materials

### 2.1 Evolutionary *de novo* Design

A combinatorial exploration of the space of all possible chemical compounds is impractical, and therefore, an efficient method for searching this space is needed. Here, we make use of a *de novo* design strategy based on genetic algorithms (GA)<sup>34</sup> which uses the principles of natural evolution (such as selection, mutation and crossover) to identify new structures with improved fitness. Since the stochastic search mechanism in GAs enables rapid navigation of the chemical landscape, they have been widely applied in areas such as computer-aided drug discovery<sup>35–37</sup> and materials design<sup>29,38,39</sup>.

In the framework of the evolutionary design strategy, the key concepts are the molecular representation and the evaluation of the quality i.e. the fitness of the generated structures. Here, the molecule is encoded as a metagraph containing vertices that represent small fragments (single atoms, functional groups, substructures), and edges that specify the relationships (bonds) between them<sup>40</sup>.

Each fragment contains one or more attachment points (AP), i.e. bonding positions (available valences) that may be substituted. An attachment point is characterized by a so-called attachment point class: an encoded representation, such as a kind of functional group or a retrosynthetic rule, of the chemical context from which the attachment point was gener-

ated upon fragmentation. Attachment points may be used to connect other fragments to form larger entities. The connection of fragments is governed by a set of compatibility rules that are collected in a so-called compatibility matrix. These user-defined rules may be used to restrict coupling of attachment points to connections corresponding to known and realistic chemical bonds (see Table S2 in the Electronic Supplementary Information (ESI))<sup>41,42</sup>. Using this approach diverse, yet synthetically accessible sets of compounds can be obtained. The molecular fragments were generated by applying the BRICS<sup>43</sup> fragmentation algorithm (based on 16 cleavage rules) available in the RDKit library<sup>44</sup> to a large set of synthesized dyes compiled from the literature<sup>9,25,45,46</sup>. Over 200 fragments were generated using this process (a subset of the fragments is shown in Figure F2 in the ESI).

In the *de novo* approach, molecular assembly starts with the random selection of a seed fragment (i.e., a scaffold, see Figure F1 in the ESI). The process then loops through the attachment points, randomly adding substituent groups by obeying the constraints defined by the compatibility matrix. The molecule may be further grown depending on the availability of free attachment points. In order to prevent very large size structures, the probability of substitution decreases with distance to the scaffold<sup>41</sup>. For the molecules so formed, additional filters in the form of molecular weight or number of rotatable bonds may be applied. As part of the iterative refinement, new offspring structures are generated using genetic operators such as crossover or mutation that are applied to selected compounds sampled from the population. A mutation applied to a randomly chosen fragment (graph node) can take the form of a replacement, deletion or addition (growing the molecule). A crossover between two parent molecules proceeds by swapping subgraphs connected to randomly chosen nodes bearing compatible attachment points. At the graph level, both operators generally involve removal of the edge connecting the chosen vertex (at the deletion point) and its parent and reconnecting with the new fragment.

The *fitness* of a proposed structure can be obtained from mainly three sources: 1) synthesis and experimental testing, 2) quantum chemical computations of relevant properties and 3) quantitative structure-property relationships (QSPR). Each of these fitness sources have advantages and disadvantages. Synthesis and experimental testing of new structures will give very accurate fitness values, however the cost and time required are too high to be practical in a *de novo* process where perhaps thousands of structures must be analyzed. For some cases, quantum chemical calculations may be applicable and they typically provide medium to high accuracy estimates depending on the level of theory being used. When more accurate calculations are needed, the computational cost often becomes too high. The third alternative is to use QSPR models which have medium to high accuracy, but at a very low

computational cost. The main drawback with QSPR models is that they are *local* in nature, which means that their predictions cannot always be trusted when extrapolation outside the applicability domain (AD) is performed.

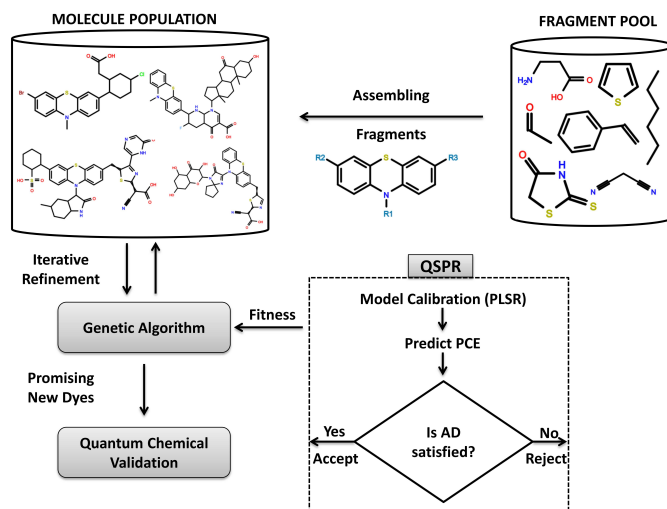
Direct computation of the *PCE* fitness for DSSCs using quantum chemical methods<sup>47–49</sup> such as density functional theory (DFT) and time-dependent DFT (TD-DFT)<sup>50,51</sup> is, as yet, not possible. These methods can only reliably model *some*, but not all, of the elementary mechanisms and electronic properties of the DSSC cell components<sup>52</sup>. Recently, Ip et al<sup>53</sup> have used quantum chemical calculations to evaluate a number of molecular properties correlating with the *PCE*. Statistical modelling based on the calculated variables was then used to obtain a degree of confidence for new dyes. Cole et al<sup>54</sup>, on the other hand, identify novel classes of dyes using data-mining procedures (based on semi-empirical computations and other molecular design criteria) followed by DFT-level calculations to assess energetic feasibility for application to DSSCs.

In this article, the third approach to fitness calculation is chosen where QSPR models based on experimental *PCE* values for several molecules are used<sup>55</sup>. We have recently demonstrated<sup>56,57</sup> that quantitative dye structure-photovoltaic performance relationship models with high predictive ability can be obtained. Building on these results, we investigate here how such QSPR models can be used for fast and accurate approximations of the *PCE* fitness in evolutionary *de novo* design<sup>29,58</sup> of new phenothiazine dyes.

Figure 1 outlines the molecular design approach adopted in this study. Starting with an initial population of molecules that are generated by a systematic assembly of fragments, improvements in the *PCE* are obtained through iterative refinement using evolutionary operations such as selection, crossover and mutation. The *PCE* for each proposed molecule is assessed using a QSPR model that provides fast quantitative estimates while adhering to applicability domain criteria established for the model. The final set of promising candidates are then further validated using density functional theory approaches<sup>59</sup>. In the following sections, we discuss the steps in more detail.

## 2.2 QSPR Modelling

A set of 117 phenothiazine-based (see Table S1 in the ESI) dye sensitizers with *PCE* ranging between 0.4% and 8.18% were compiled from the literature. During selection, it was ensured that the dyes were tested under similar conditions (same electrolyte, no coadsorber, no cosensitizer and no deaggregation agents). In previous studies<sup>56,57</sup>, we have shown that vibrational frequency based eigenvalue (EVA) descriptors<sup>60</sup> can provide robust predictions of DSSC performance. For each molecule, a semi-empirical geometry optimisation



**Fig. 1** Schematic shows an outline of the *de novo* approach to designing new phenothiazine dyes.

at the AM1 level using MOPAC<sup>61</sup> is performed and the calculated vibrational frequencies are projected onto a bounded scale ( $1 - 4000\text{cm}^{-1}$ ). This range is then sampled at fixed intervals of  $L$  and at each sample point  $x$ , a Gaussian smoothing function of fixed standard deviation  $\sigma$  is applied and the descriptor ( $D_x$ ) is calculated as:

$$D_x = \sum_{i=1}^{3N-6} \frac{1}{\sigma \sqrt{(2\pi)}} \exp\left(\frac{-(x - f_i)^2}{2\sigma^2}\right) \quad (2)$$

where,  $f_i$  is the  $i^{\text{th}}$  normal mode frequency of the compound concerned. Here, we have chosen values of  $\sigma = 2$  and  $L = 1$ .

The available experimental data was divided into separate training and test sets (50 : 50 split) using the DUPLEX algorithm<sup>62</sup>. The data were then modelled using partial least squares regression (PLSR)<sup>63</sup>. Prior to modelling, all variables were autoscaled to zero mean and unit variance. The optimal number of latent variables (LVs) was decided based on the cross-validated (10-fold) root mean square error (*RMSE*) averaged over the 50 independent runs. To guard against overfitting,  $y$ -randomization tests<sup>64</sup> (repeated 1000 times) were also applied. The predictive ability of the model was evaluated by the cross-validated correlation coefficient ( $R_{cv}^2$ ) and root mean square error (*RMSE*):

$$R_{cv}^2 = 1 - \frac{\sum (y_{obs,i} - \widehat{y}_{cv,i})^2}{\sum (y_{obs,i} - \overline{y}_{obs})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum (y_{obs} - y^{pred})^2}{n}} \quad (4)$$

where,  $\widehat{y}_{cv,i}$  is the predicted value for the excluded (cross-validated)  $i^{\text{th}}$  compound,  $y_{obs,i}$  is the corresponding ex-



perimental value and  $\overline{y_{obs}}$  is the mean of the experimental values.

### 2.3 Applicability Domain Analysis (ADAN)

However well-trained, there is no guarantee that the obtained model will be predictive on future datasets<sup>65</sup>. Every model therefore, has an applicability domain (AD) which is determined by the chemical and mechanistic variability associated with the molecules in the calibration set. A query molecule can be considered to be inside or outside the AD, and to this purpose a number of measures have been proposed<sup>66</sup>. A popular measure for linear models is the *leverage* which for a given observation  $i$  is calculated as  $h_i = \mathbf{s}_i^T (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{s}_i$ <sup>67</sup> where the superscript  $T$  refers to the transpose of the matrix/vector,  $\mathbf{S}$  is the matrix containing the PLSR scores and  $\mathbf{s}_i$  is the scores vector for compound  $i$  over  $A$  PLSR components. Compounds for which  $h_i > \frac{3p}{n}$  are generally considered as being outside the AD. Here,  $p$  is the number of model parameters and  $n$  is the number of compounds in the training set.

In a recent article, Carrio et al.,<sup>68</sup> have listed six quantitative criteria for assessing the reliability of QSPR predictions. The values obtained are transformed into qualitative binary rules (0-inside, 1-outside) by comparing it with the 95% percentile of the training set metrics. Reliability of the predictions can thus be evaluated in terms of the rules being violated. These metrics use Euclidean distances and differences calculated in the PLSR-score space where very large values (distances/differences) are indicative of poor predictive ability. The metrics are:

$D_1$ : Euclidean distance (in the latent variable space of the PLSR model) between the query molecule and the centroid of the training set.

$D_2$ : Euclidean distance (in the latent variable space of the PLSR model) between the query molecule and its nearest neighbour in the training set.

$D_3$ : Distance to model ( $D_{ModX}$ )<sup>69</sup> obtained from the  $X$  residuals of the projection of the new object on to the PLSR model.

$D_4$ : Difference between the predicted  $PCE$  and the mean  $PCE$  value of the training set.

$D_5$ : Difference between the predicted  $PCE$  and the experimental value for the closest compound in the training set.

$D_6$ : Prediction accuracy based on the standard error of prediction ( $SDEP$ ) for 5% of the nearest neighbours (in the training set) of the query compound.

In addition to the leverage and other ADAN measures, we also calculate the prediction uncertainty<sup>70</sup> using a model-based bootstrapping procedure. Given a training set of size  $N$  (here  $N = 59$ ), multiple random samples (with replacement) of the same size are generated. For each such sample a PLSR model is calculated and applied to the query molecule. In this study, a total of  $M = 500$  bootstrap PLSR models were generated and the uncertainty  $\tilde{y}_i$  associated with the prediction for molecule  $i$  is calculated as the standard deviation of the distribution:

$$\tilde{y}_i = \sqrt{\frac{\sum_{j=1}^M (\hat{y}_j - \hat{y})^2}{M-1}} \quad (5)$$

where  $M$  is the number of bootstrap models,  $\hat{y}_j$  is the prediction for  $j^{th}$  model and  $\hat{y}$  obtained using the full training set model is taken as the mean of the distribution.

### 2.4 Quantum Chemical Validation

DFT and TDDFT have become the standard choice for electronic structure calculations on molecules and solids and have been increasingly used for theoretical validation of DSSC properties<sup>49,59,71,72</sup>. As mentioned earlier, although it is currently not possible to directly estimate the  $PCE$  from first principles, one can still assess qualitatively or even to some degree quantitatively, relevant aspects of the proposed dyes.

With reference to Equation 1, the short circuit current  $J_{sc}$ <sup>73</sup> is given by:

$$J_{sc} = \int LHE(\lambda) \phi_{inj} \eta_{col} d\lambda \quad (6)$$

where  $LHE(\lambda)$  is the light harvesting efficiency at wavelength  $\lambda$ ,  $\phi_{inj}$  is the electron injection efficiency and  $\eta_{col}$  is the charge collection efficiency. Since the DSSCs only differ in the dyes being used, it is reasonable to assume that  $\eta_{col}$  is constant. Thus, the two main factors influencing  $J_{sc}$  are the  $LHE(\lambda)$  and  $\phi_{inj}$ , where the former is given by:

$$LHE(\lambda) = 1 - 10^{-f_{max}} \quad (7)$$

where  $f_{max}$  is the oscillator strength of the dye corresponding to the maximum absorption wavelength ( $\lambda_{max}$ ) of the dye. The charge injection efficiency ( $\phi_{inj}$ ) is closely related to the driving force of electron injection ( $\Delta G_{inj}$ ) i.e. the free energy change (in eV) given by:

$$\Delta G_{inj} = E^{dye*} - E_{CB}^{TiO_2} \quad (8)$$

where  $E^{dye*}$  is the excited state oxidation potential of the dye and  $E_{CB}$  is the conduction band edge of the semiconductor (experimental value of -4.0 eV vs. vacuum is used<sup>74,75</sup>). Both  $\phi_{inj}$  and  $LHE(\lambda)$  have to be as high as possible in order to maximize the photocurrent.

In addition to the spectral absorption properties, another important requirement for donor-acceptor dyes is that of an efficient intramolecular charge separation which facilitates electron injection from the excited dye into the semiconductor. To this end, we have quantified these charge-transfer (CT) excitations using the approach proposed by Le Bahers et al.<sup>76</sup>. The method starts by identifying barycenters of regions (upon excitation) of electron density depletion ( $R_-$ ) and increment ( $R_+$ ) given by:

$$R_+ = \frac{\int r \rho_+(r) dr}{\int \rho_+(r) dr} \quad (9)$$

and

$$R_- = \frac{\int r \rho_-(r) dr}{\int \rho_-(r) dr} \quad (10)$$

Here,  $\rho_+(r)$  and  $\rho_-(r)$  define the points in space where an increase or decrease in density is produced and are calculated as:

$$\rho_+(r) = \begin{cases} \Delta\rho(r), & \text{if } \Delta\rho(r) > 0 \\ 0, & \text{if } \Delta\rho(r) < 0 \end{cases} \quad (11)$$

and

$$\rho_-(r) = \begin{cases} \Delta\rho(r), & \text{if } \Delta\rho(r) < 0 \\ 0, & \text{if } \Delta\rho(r) > 0 \end{cases} \quad (12)$$

where  $\Delta\rho(r)$  is defined as the difference between the total density of the considered excited ( $\rho^{ES}(r)$ ) and ground states ( $\rho^{GS}(r)$ ):

$$\Delta\rho(r) = \rho^{ES}(r) - \rho^{GS}(r) \quad (13)$$

Based on Equations 9-13, the computed charge-transfer parameters include:

$d_{CT}$  The charge transfer distance calculated as the difference between the barycenters  $d_{CT} = |R_+ - R_-|$ .

$q_{CT}$  The amount of charge transferred  $q_{CT} = \int \rho_+(r) dr$ .

According to the Marcus electron transfer theory<sup>77</sup>, the total reorganization energy  $\lambda_{tot}$  can also affect the kinetics of electron injection. Therefore, in order to enhance the  $J_{sc}$  and to minimize energy losses, a small reorganization energy is favoured. In the calculation of the reorganization energy which has an inner-sphere contribution  $\lambda_i$  and an outer-sphere contribution  $\lambda_s$  (due to the surrounding dielectric medium), the latter is often ignored<sup>78,79</sup>. The internal reorganization energy  $\lambda_i$ <sup>80</sup> is therefore calculated as:

$$\lambda_i = [E_0(G_+) + E_+(G_0)] - [E_+(G_+) + E_0(G_0)] \quad (14)$$

where  $E_j$  is the energy of the system in charge state  $j$  and is dependent on  $G_k$  which is the geometry of the system in charge state  $k$ . Here, 0 and + correspond to the neutral and cationic charge state respectively.

## 2.5 Computational Details

The PLSR modelling was carried out using the *pls*<sup>81</sup> package available in the *R*<sup>82</sup>. The structures emerging from the *de novo* were subjected to conformational search using *cxcalc* (based on the Dreiding force field<sup>83</sup>). The lowest energy conformer was then geometry optimised, using the AM1 Hamiltonian in MOPAC<sup>61</sup>. The EVA descriptors were then calculated using these optimised geometries. The *de novo* algorithm and molecular descriptor calculation routines were implemented in an in-house software written in Java using the CDK toolkit<sup>84</sup>.

For further quantum chemical analysis, the molecular geometries were optimised with DFT using the Beckes three-parameter and Lee-Yang-Parr hybrid (B3LYP) functional<sup>85</sup> and a 6-31G(d) basis set. For the simulation of the absorption spectra, TD-DFT calculations were carried out in dichloromethane solution using the CAM-B3LYP functional with the same basis set. The 30 lowest excitation energies were considered in these calculations. Solvation effects were included by means of the conductor-like polarizable continuum model (CPCM)<sup>86</sup>. Charge transfer parameters were calculated using software provided by Le Bahers et al.<sup>76</sup>. All calculations were carried out using the *Gaussian 09* program<sup>87</sup>.

## 3 Results and Discussion

### 3.1 QSPR Models

For the calibration set containing 59 phenothiazine dyes, a 3-component PLSR model with  $R_{cv}^2 = 0.72$  and  $RMSE = 1.06$  was obtained (see Table S1 in the ESI for additional details). The true test for any model is to check its predictions against molecules that were not part of the training. The performance of the proposed model was therefore further verified using a test set of 58 dyes that span an experimental *PCE* range (0.4% to 8%) similar to that of the training set. Table 1 summarizes the test set performance of the model. Overall, the model exhibits good predictive characteristics with  $R_{test}^2 = 0.68$  and  $RMSE_{test} = 1.13$  which mirrors the values obtained with the training data. The table also shows additional model diagnostics that include the leverage and the number of ADAN rules violated. With respect to the leverage, none of the predictions exceed the cutoff value of  $h^* = 0.15$ . For predictions with no violations, the maximum error in the *PCE* is around 2% and increases to 3.5% for up to 3 violations. Overall, only 9 dyes have a absolute prediction error greater than 2%. A

**Table 1** Independent external validation set performance for the experimental values of the *PCE*. Test set molecules have been sorted (ascending) according to the number of ADAN violations.

Molecule	<i>PCE</i>	QSPR	$h_i$	ADAN	Molecule	<i>PCE</i>	QSPR	$h_i$	ADAN
T01	5.60	4.40 ± 0.458	0.047	0	T30	8.07	6.88 ± 0.404	0.022	0
T02	6.02	5.75 ± 0.439	0.046	0	T31	5.23	5.87 ± 0.530	0.032	0
T03	0.80	1.68 ± 0.945	0.093	0	T32	6.29	6.03 ± 0.312	0.035	0
T04	4.56	3.80 ± 0.637	0.034	0	T33	6.02	5.56 ± 0.385	0.026	0
T05	0.60	0.97 ± 0.564	0.103	0	T34	5.60	4.98 ± 0.364	0.013	0
T06	6.85	6.27 ± 0.799	0.047	0	T35	4.53	5.44 ± 0.434	0.040	0
T07	6.70	5.94 ± 0.480	0.084	0	T36	5.36	5.66 ± 0.463	0.065	0
T08	5.73	5.50 ± 0.381	0.014	0	T37	7.98	5.66 ± 0.371	0.004	0
T09	4.80	3.68 ± 0.774	0.052	0	T38	5.39	6.08 ± 0.337	0.039	0
T10	7.30	6.63 ± 0.454	0.021	0	T39	3.94	6.40 ± 0.510	0.043	1
T11	3.56	4.28 ± 0.456	0.031	0	T40	3.60	3.44 ± 0.539	0.021	1
T12	7.38	7.47 ± 0.452	0.042	0	T41	2.10	3.90 ± 0.588	0.028	1
T13	1.90	2.06 ± 0.640	0.037	0	T42	3.54	4.83 ± 0.349	0.012	1
T14	2.48	4.29 ± 0.433	0.004	0	T43	6.40	6.14 ± 0.855	0.026	1
T15	5.12	5.15 ± 0.429	0.013	0	T44	2.24	5.27 ± 0.375	0.005	1
T16	3.91	4.01 ± 0.352	0.031	0	T45	6.82	3.86 ± 0.687	0.019	1
T17	8.08	7.37 ± 0.414	0.035	0	T46	5.22	4.20 ± 0.548	0.033	1
T18	7.13	5.82 ± 0.329	0.005	0	T47	1.30	0.87 ± 0.657	0.078	1
T19	6.32	6.32 ± 0.396	0.018	0	T48	6.13	5.26 ± 0.437	0.017	1
T20	4.40	4.40 ± 0.961	0.032	0	T49	4.07	4.25 ± 0.360	0.024	1
T21	5.53	5.52 ± 0.441	0.022	0	T50	4.39	5.41 ± 0.408	0.020	1
T22	6.80	6.21 ± 0.490	0.040	0	T51	4.80	4.61 ± 0.402	0.032	1
T23	6.37	6.12 ± 0.347	0.011	0	T52	0.40	2.67 ± 0.613	0.066	1
T24	5.12	4.56 ± 0.317	0.012	0	T53	4.43	4.49 ± 0.367	0.022	1
T25	6.80	6.06 ± 0.469	0.010	0	T54	0.50	1.90 ± 1.088	0.093	1
T26	1.30	1.85 ± 0.523	0.097	0	T55	6.44	7.26 ± 0.744	0.108	2
T27	4.79	5.17 ± 0.463	0.016	0	T56	5.40	4.49 ± 0.590	0.013	2
T28	7.44	5.32 ± 0.515	0.036	0	T57	6.72	4.22 ± 0.412	0.035	2
T29	4.41	5.35 ± 0.519	0.003	0	T58	3.88	4.12 ± 0.595	0.043	2

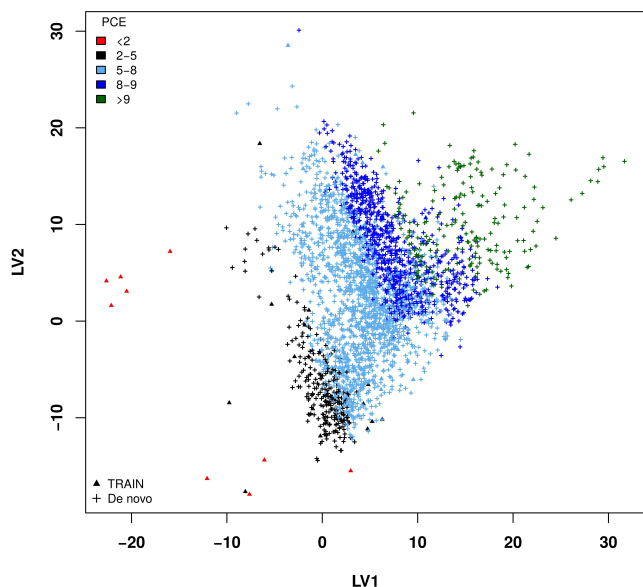
commonly violated rule is that of the distance to the model *DModX* which provides an indication of the distance from the observation to the model plane. Unlike the other AD measures, the uncertainty relies purely on the variability of the predictions that results from using different training sets drawn from the same population. Thus, compounds with very small uncertainties are expected to have relatively small prediction errors. High uncertainties are generally associated with dyes containing features that are very sensitive to the training set composition. However, for the test set observations in Table 1, the calculated uncertainties do not show any significant variations and are largely below 1.

### 3.2 QSPR-Guided Design of Dyes

Evolutionary algorithms have shown great promise in the search for solutions to molecular design problems that can be highly non-linear<sup>88</sup>. A central part of this approach consists of evaluating the fitness of the proposed structures. Although limited by their domain of applicability, QSPR models typ-

ically obtain quantitative estimates of the *PCE* at the order of < 10 seconds (on a Intel Core i5-2400 CPU @ 3.10GHz) and therefore can be easily integrated into a virtual high-throughput screening framework<sup>89,90</sup>. Other aspects relevant to the design are the fragment database and the genetic algorithm parameters.

For this study, we used a population size of 200 molecules and the genetic algorithm ran for up to 200 generations with tournament selection. Crossover and mutation probabilities were set to 0.50 with five offspring produced in each generation. In order to limit the extent of extrapolation, model predictions which had more than three ADAN violations were not allowed to enter the population. An additional constraint of maximum 1000 Daltons was also imposed to restrict the size of the molecules. Over ten independent runs were executed which yielded around 9000 unique structures with predicted *PCE* in the range 5-10%. The score plot based on the first two latent variables in Figure 2 summarizes the *PCE* values of the final populations of the repeated runs. As can be seen from the plot, the increasing scores along both LVs cor-



**Fig. 2** PLSR scores plot of the first two LVs shows the positions of the training set observations (shown as filled triangles) and the molecules produced from multiple *de novo* runs (shown as “+” symbols). The points have been coloured according to the value of the *PCE*.

respond to increasing *PCE*. A clear separation is also seen between molecules with *PCE* < 2% (shown as red triangles) and those that exceed 5%. Although molecules with *PCE* > 10% were also identified, the number of ADAN violations were also found to increase. Setting the ADAN violation threshold to three allows us to exclude molecules that are too dissimilar or too far from the training set to make reliable predictions of the *PCE*.

### 3.3 Quality Assessment

In order to validate the predictions, five structures (see Table 2) were subjected to additional calculations at the DFT level of theory. All selected molecules have predicted *PCE* above 9% and associated uncertainties around  $\pm 2\%$  (see Table 3). As part of the DFT-based evaluation of the dyes, we focus on aspects that can be used as a guide to setting performance expectations. In particular, we examine key features such as the UV-vis spectra, charge-transfer character, reorganization energy for oxidation and the free energy of dye oxidation in solution<sup>53</sup>.

As can be seen from the ground state density plots in Table 2, the HOMO is largely localized over the phenothiazine moiety (donor) while the LUMO is distributed electron def-

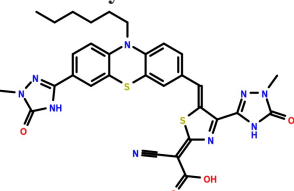
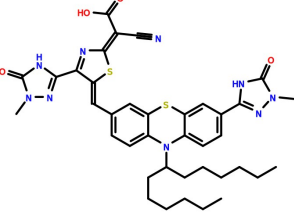
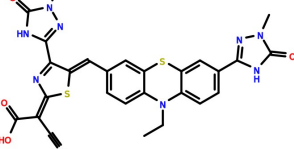
icient groups (acceptor and anchoring groups) that are indicative of an efficient charge separation<sup>5</sup>. The presence of these electron-withdrawing groups (such as dihydrothiazole) have been shown to improve the spectral response and charge injection<sup>92</sup>. The computed maximum absorption wavelengths  $\lambda_{max}$ , oscillator strengths ( $f_{max}$ ) and the light harvesting efficiency  $LHE(\lambda)$  are listed in Table 3. For all the investigated dyes, the first significant transition is calculated at around 450-530 nm, with oscillator strengths ( $f_{max}$ ) marginally greater than 1, thus predicting a strong absorption in the visible region (see simulated absorption spectra in Table 2). The dyes PTZ1-PTZ3 additionally contain electron-deficient triazole groups which have been shown to have excellent electron transport<sup>93,94</sup> properties. Among the proposed structures, dye PTZ-1 which also contains a strong electron donating group (julolidine), exhibits the maximum absorption wavelength at 537 nm that is dominated by a HOMO→LUMO transition. PTZ-1 also shows the strongest charge transfer character with a computed  $q_{CT}$  value of 0.72 and a  $d_{CT}$  of 3.76Å. The dye PTZ-4 which is predicted to have the highest *PCE* = 9.52 and  $LHE(\lambda) = 0.92$ , contains a nitrophenyl moiety as an additional anchoring group. Cong et al.<sup>95</sup>, have suggested that such substituents can act as robust anchors by increasing the electron injection ability of the dye. The largest value of the calculated injection driving force ( $\Delta G_{inj}$ ) are associated with dyes PTZ-4 and PTZ-5. Closer inspection of the reorganization energies,  $LHE_{\lambda}$  and  $\Delta G_{inj}$  in Table 3 suggests that these dyes can be expected to have large values of the  $J_{sc}$ .

Compared with the highest *PCE* of 8.18% (C09 in the ESI) as reported in the literature, an improvement of more than 1 percentage point has been predicted for the proposed dyes. This increase may be attributed to the additional anchoring groups present in these dyes. For instance, as shown by Zervaki et al<sup>96</sup>, the piperidine moiety in PTZ2 (see Table 2) can act as an additional anchoring group. It has been observed that incorporating bulky alkyl chains into the dye structure can prevent aggregation<sup>97,98</sup>. Furthermore, bulky groups also facilitate the suppressing of interfacial charge transfer recombination<sup>33</sup>. For example, in PTZ-1, the secondary electron donor unit i.e. julolidine not only extends the absorption region but may also limit aggregation<sup>99</sup>. In C09, this function is typically performed by the hexyl and hexyloxy groups. To investigate this effect further, we replaced (manually) the julolidine group in PTZ-1 with alkyl chains. The results of these substitutions are summarized in Table 4. Interestingly, for the bulky alkyl substituents, only a marginal decrease of up to 0.2% in the predicted *PCE* is seen. Although a much smaller ethyl substituent significantly increases the *PCE* to over 11%, the predictive uncertainties also increase as indicated by the number of ADAN violations.

It should be pointed out that the whole concept of an applicability domain is heavily influenced by the idea of trusting



**Table 4** Modifications to the dye PTZ-1 listed in Table 2. The julolidine moiety was replaced with different alkyl chains that have been used in most studies.

Dye	PCE	ADAN
	9.30±2.26	2
	9.10±2.23	4
	11.18±2.67	5

the predictive ability of a model when the query structures are similar to those that were used in the calibration data set. In our case, our objective is to perform an *extrapolation* (beyond the region covered by the calibration data) by searching for new structures that are by definition different and hopefully better than those seen before. The ADAN measures that we have adopted in this study are used to provide warnings when substantial extrapolation by the model occurs.

## 4 Conclusions

In this article, we have outlined a computationally efficient procedure to identify novel phenothiazine dyes with improved properties. At the heart of the approach is a QSPR-driven methodology combined with an evolutionary strategy for dye design. The nature-inspired approach has been able to identify a diverse collection of promising alternatives that have been further verified using theoretical studies. A logical extension of this work would be to apply the scheme to a multi-objective problem where, for instance, both current, voltage and other important parameters can be optimized. The approach proposed here is not limited to DSSCs and can be easily expanded to a wider array of applications.

## Acknowledgements

The Norwegian Research Council (NFR) is acknowledged for financial support from the eVITA (Grant No. 205273) and the CLIMIT (Grant No. 233776) and for CPU resources granted through the NOTUR supercomputing programme. We also thank ChemAxon (<http://www.chemaxon.com>) for free academic use of the Marvin package.

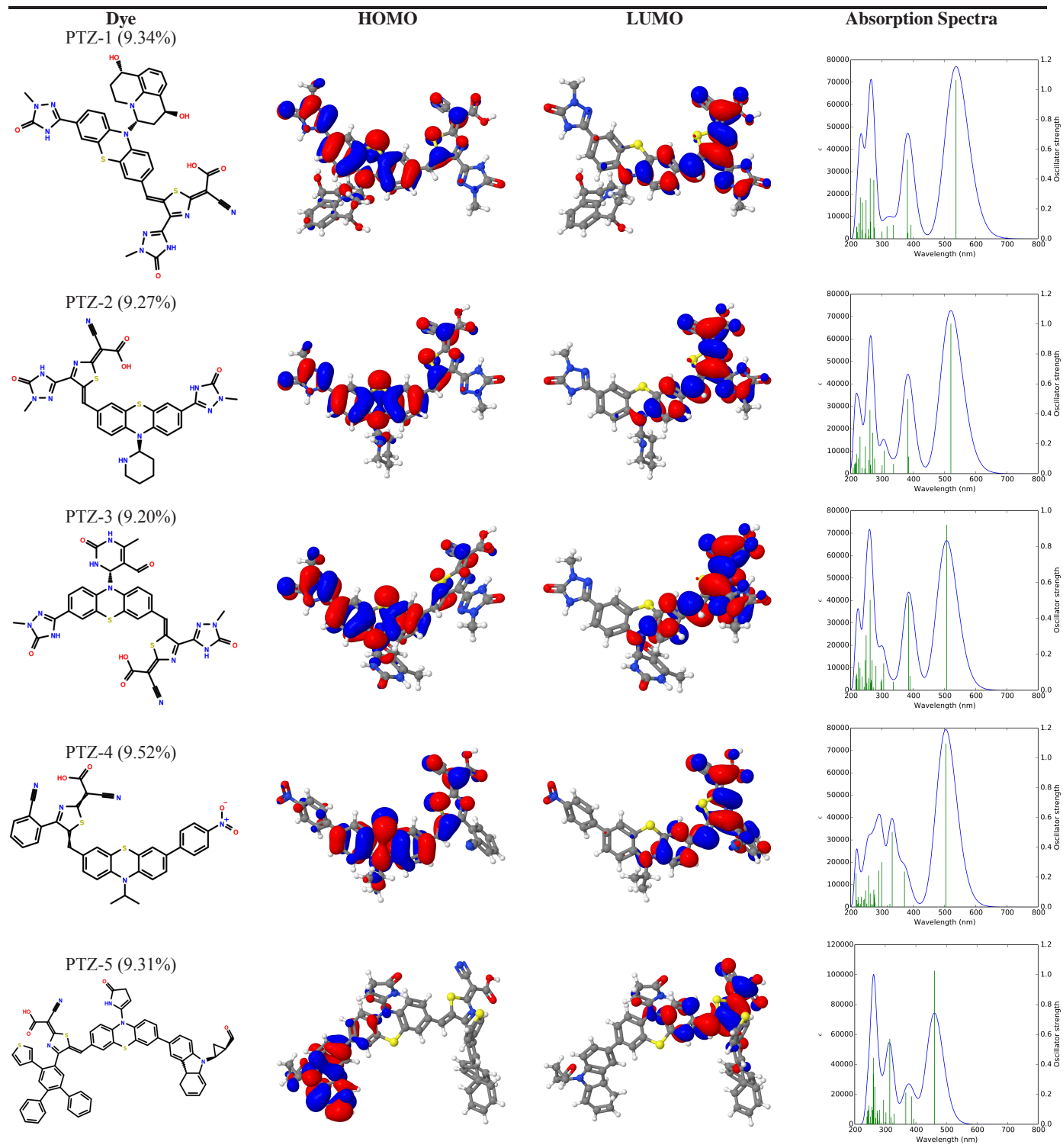
## References

- B. Parida, S. Iniyar and R. Goic, *Renew. Sust. Energ. Rev.*, 2011, **15**, 1625–1636.
- B. Hardin, H. Snaith and M. McGehee, *Nat. Photon.*, 2012, **6**, 162–169.
- H. M. Upadhyaya, S. Senthilarasu, M.-H. Hsu and D. K. Kumar, *Sol. Energy Mater. Sol. Cells*, 2013, **119**, 291–295.
- M. Grätzel, *J. Photochem. Photobiol. C: Photochem. Rev.*, 2003, **4**, 145–153.
- A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo and H. Pettersson, *Chem. Rev.*, 2010, **110**, 6595–6663.
- T. Bessho, E. Yoneda, J.-H. Yum, M. Guglielmi, I. Tavernelli, H. Imai, U. Rothlisberger, M. K. Nazeeruddin and M. Grätzel, *J. Am. Chem. Soc.*, 2009, **131**, 5930–5934.
- C.-C. Chou, K.-L. Wu, Y. Chi, W.-P. Hu, S. J. Yu, G.-H. Lee, C.-L. Lin and P.-T. Chou, *Angew. Chem. Int. Edit.*, 2011, **50**, 2054–2058.
- A. Yella, H.-W. Lee, H. N. Tsao, C. Yi, A. K. Chandiran, M. Nazeeruddin, E. W.-G. Diau, C.-Y. Yeh, S. M. Zakeeruddin and M. Grätzel, *Science*, 2011, **334**, 629–634.
- A. Mishra, M. Fischer and P. Bäuerle, *Angew. Chem. Int. Edit.*, 2009, **48**, 2474–2499.
- D. Joly, L. Pellejá, S. Narbey, F. Oswald, J. Chiron, J. N. Clifford, E. Palomares and R. Demadrille, *Sci. Rep.*, 2014, **4**.
- W. Zeng, Y. Cao, Y. Bai, Y. Wang, Y. Shi, M. Zhang, F. Wang, C. Pan and P. Wang, *Chem. Mater.*, 2010, **22**, 1915–1925.
- E. Ramasamy and J. Lee, *J. Phys. Chem. C*, 2010, **114**, 22032–22037.
- T. Yoshida, J. Zhang, D. Komatsu, S. Sawatani, H. Minoura, T. Pauporté, D. Lincot, T. Oekermann, D. Schlettwein, H. Tada, D. Whrle, K. Funabiki, M. Matsui, H. Miura and H. Yanagi, *Adv. Funct. Mater.*, 2009, **19**, 17–43.
- M. Wang, C. Grätzel, S. M. Zakeeruddin and M. Grätzel, *Energy Environ. Sci.*, 2012, **5**, 9394–9405.
- C.-H. Siu, L. T. Lin Lee, P.-Y. Ho, P. Majumdar, C.-L. Ho, T. Chen, J. Zhao, H. Li and W.-Y. Wong, *J. Mater. Chem. C*, 2014, **2**, 7086–7095.
- C. Qin, W.-Y. Wong and L. Han, *Chem. Asian J.*, 2013, **8**, 1706–1719.
- Y.-S. Chen, C. Li, Z.-H. Zeng, W.-B. Wang, X.-S. Wang and B.-W. Zhang, *J. Mater. Chem.*, 2005, **15**, 1654–1661.
- G. Wu, F. Kong, Y. Zhang, X. Zhang, J. Li, W. Chen, W. Liu, Y. Ding, C. Zhang, B. Zhang, J. Yao and S. Dai, *J. Phys. Chem. C*, 2014, **118**, 8756–8765.
- C. Sakong, H. J. Kim, S. H. Kim, J. W. Namgoong, J. H. Park, J.-H. Ryu, B. Kim, M. J. Ko and J. P. Kim, *New J. Chem.*, 2012, **36**, 2025–2032.
- K. D. Seo, H. M. Song, M. J. Lee, M. Pastore, C. Anselmi, F. D. Angelis, M. K. Nazeeruddin, M. Grätzel and H. K. Kim, *Dyes Pigments*, 2011, **90**, 304–310.
- K. D. Seo, I. T. Choi, Y. G. Park, S. Kang, J. Y. Lee and H. K. Kim, *Dyes Pigments*, 2012, **94**, 469–474.
- X. Yang, J. Zhao, L. Wang, J. Tian and L. Sun, *RSC Adv.*, 2014, **4**, 24377–24383.
- S. H. Bae, K. D. Seo, W. S. Choi, J. Y. Hong and H. K. Kim, *Dyes Pigments*, 2015, **113**, 18–26.

- 24 C. Chen, J.-Y. Liao, Z. Chi, B. Xu, X. Zhang, D.-B. Kuang, Y. Zhang, S. Liu and J. Xu, *J. Mater. Chem.*, 2012, **22**, 8994–9005.
- 25 M. Liang and J. Chen, *Chem. Soc. Rev.*, 2013, **42**, 3453–3488.
- 26 J. Zhao, X. Yang, M. Cheng, S. Li and L. Sun, *J. Phys. Chem. C*, 2013, **117**, 12936–12941.
- 27 J.-H. Yum, E. Baranoff, S. Wenger, M. K. Nazeeruddin and M. Grätzel, *Energy Environ. Sci.*, 2011, **4**, 842–857.
- 28 B.-G. Kim, K. Chung and J. Kim, *Chem. Eur. J.*, 2013, **19**, 5220–5230.
- 29 Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen and B. K. Alsborg, *J. Am. Chem. Soc.*, 2012, **134**, 8885–8895.
- 30 V. Mlinar, *J. Mater. Chem.*, 2012, **22**, 1724–1732.
- 31 K. Zhu, S.-R. Jang and A. J. Frank, *J. Phys. Chem. Lett.*, 2011, **2**, 1070–1076.
- 32 N. Koide, A. Islam, Y. Chiba and L. Han, *J. Photochem. Photobiol., A*, 2006, **182**, 296–305.
- 33 J. N. Clifford, E. Martinez-Ferrero, A. Viterisi and E. Palomares, *Chem. Soc. Rev.*, 2011, **40**, 1635–1646.
- 34 J. H. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 1992.
- 35 G. Schneider and U. Fechner, *Nat. Rev. Drug Discov.*, 2005, **4**, 649–663.
- 36 M. Segall, E. Champness, C. Leeding, R. Lilien, R. Mettu and B. Stevens, *J. Chem. Inf. Model.*, 2011, **51**, 2967–2976.
- 37 C. A. Nicolaou and N. Brown, *Drug Discov. Today Technol.*, 2013, **10**, e427–e435.
- 38 N. Chakraborti, *Int. Mater. Rev.*, 2004, **49**, 246–260.
- 39 D. J. Scott, S. Manos and P. V. Coveney, *J. Chem. Inf. Model.*, 2008, **48**, 262–273.
- 40 N. Brown, B. McKay, F. Gilardoni and J. Gasteiger, *J. Chem. Inf. Model.*, 2004, **44**, 1079–1087.
- 41 M. Foscatto, G. Occhipinti, V. Venkatraman, B. K. Alsborg and V. R. Jensen, *J. Chem. Inf. Model.*, 2014, **54**, 767–780.
- 42 M. Foscatto, V. Venkatraman, G. Occhipinti, B. K. Alsborg and V. R. Jensen, *J. Chem. Inf. Model.*, 2014, **54**, 1919–1931.
- 43 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, *Chem. Med. Chem.*, 2008, **3**, 1503–1507.
- 44 *RDKit Release 2014.09.1*, 2014, RDKit (<http://www.rdkit.org>).
- 45 Y. Lin, Y. Li and X. Zhan, *Chem. Soc. Rev.*, 2012, **41**, 4245–4272.
- 46 Y. Wu and W. Zhu, *Chem. Soc. Rev.*, 2013, **42**, 2039–2058.
- 47 T. Le Bahers, F. Labat, T. Pauporté, P. P. Lainé and I. Ciofi ni, *J. Am. Chem. Soc.*, 2011, **133**, 8005–8013.
- 48 F. Labat, T. Le Bahers, I. Ciofi ni and C. Adamo, *Acc. Chem. Res.*, 2012, **45**, 1268–1277.
- 49 T. Le Bahers, T. Pauporté, P. P. Lainé, F. Labat, C. Adamo and I. Ciofi ni, *J. Phys. Chem. Lett.*, 2013, **4**, 1044–1050.
- 50 A. D. Laurent, C. Adamo and D. Jacquemin, *Phys. Chem. Chem. Phys.*, 2014, **16**, 14334–14356.
- 51 N. Agnihotri, *J. Photochem. Photobiol. C: Photochem. Rev.*, 2014, **18**, 18–31.
- 52 N. Martsinovich and A. Troisi, *J. Phys. Chem. C*, 2011, **115**, 11781–11792.
- 53 C. M. Ip, A. Eleuteri and A. Troisi, *Phys. Chem. Chem. Phys.*, 2014, **16**, 19106–19110.
- 54 J. M. Cole, K. S. Low, H. Ozoe, P. Stathi, C. Kitamura, H. Kurata, P. Rudolf and T. Kawase, *Phys. Chem. Chem. Phys.*, 2014, **16**, 26684–26690.
- 55 T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chem. Rev.*, 2012, **112**, 2889–2919.
- 56 V. Venkatraman, P.-O. Åstrand and B. K. Alsborg, *J. Comput. Chem.*, 2014, **35**, 214–226.
- 57 V. Venkatraman and B. K. Alsborg, *Dyes Pigments*, 2015, **114**, 69–77.
- 58 N. Brown, B. McKay and J. Gasteiger, *J. Comput-Aided. Mol. Des.*, 2006, **20**, 333–341.
- 59 M. Pastore, E. Mosconi, F. De Angelis and M. Grätzel, *J. Phys. Chem. C*, 2010, **114**, 7205–7212.
- 60 T. Heritage, A. Ferguson, D. Turner and P. Willett, *Perspect. Drug Discov.*, 1998, **9-11**, 381–398.
- 61 J. J. P. Stewart, *MOPAC2012*, 2012, Stewart Computational Chemistry, Colorado Springs, CO, USA, (<http://OpenMOPAC.net>).
- 62 R. D. Snee, *Technometrics*, 1977, **19**, 415–428.
- 63 H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, New York, 1989.
- 64 S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold and K. Faber, *J. Chemom.*, 2007, **21**, 427–439.
- 65 T. Stouch, J. Kenyon, S. Johnson, X.-Q. Chen, A. Doweyko and Y. Li, *J. Comput-Aided. Mol. Des.*, 2003, **17**, 83–92.
- 66 T. I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliakzova, G. Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D. W. Stanton, J. J. M. van de Sandt, W. Tong, G. Veith and C. Yang, *ATLA: Altern Lab Anim*, 2005, **33**, 155173.
- 67 J. Jaworska, N. Nikolova-Jeliakzova and T. Aldenberg, *ATLA: Altern Lab Anim*, 2005, **33**, 445459.
- 68 P. Carrió, M. Pinto, G. Ecker, F. Sanz and M. Pastor, *J. Chem. Inf. Model.*, 2014, **54**, 1500–1511.
- 69 L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361.
- 70 Y. An, W. Sherman and S. L. Dixon, *J. Chem. Inf. Model.*, 2013, **53**, 2312–2321.
- 71 F. De Angelis, S. Fantacci, A. Selloni, M. K. Nazeeruddin and M. Grätzel, *J. Phys. Chem. C*, 2010, **114**, 6054–6061.
- 72 J. Wang, M. Li, D. Qi, W. Shen, R. He and S. H. Lin, *RSC Adv.*, 2014, **4**, 53927–53938.
- 73 J. Zhang, H.-B. Li, S.-L. Sun, Y. Geng, Y. Wu and Z.-M. Su, *J. Mater. Chem.*, 2012, **22**, 568–576.
- 74 J. Asbury, Y.-Q. Wang, E. Hao, H. Ghosh and T. Lian, *Res. Chem. Intermed.*, 2001, **27**, 393–406.
- 75 R. Katoh, A. Furube, T. Yoshihara, K. Hara, G. Fujihashi, S. Takano, S. Murata, H. Arakawa and M. Tachiya, *J. Phys. Chem. B*, 2004, **108**, 4818–4822.
- 76 T. Le Bahers, C. Adamo and I. Ciofi ni, *J. Chem. Theory Comput.*, 2011, **7**, 2498–2506.
- 77 R. A. Marcus, *Rev. Mod. Phys.*, 1993, **65**, 599–610.
- 78 V. Coropceanu, J. Cornil, D. A. da Silva Filho, Y. Olivier, R. Silbey and J.-L. Brédas, *Chem. Rev.*, 2007, **107**, 926–952.
- 79 E. Maggio, N. Martsinovich and A. Troisi, *J. Phys. Chem. C*, 2012, **116**, 7638–7649.
- 80 V. Vaissier, P. Barnes, J. Kirkpatrick and J. Nelson, *Phys. Chem. Chem. Phys.*, 2013, **15**, 4804–4814.
- 81 B.-H. Mevik and R. Wehrens, *J. Stat. Softw.*, 2007, **18**, 1–24.
- 82 R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012.
- 83 S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.
- 84 C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E. Willighagen, *Curr. Pharm. Des.*, 2006, **12**, 2111–2120.
- 85 C. Lee, W. Yang and R. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
- 86 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
- 87 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro,

- M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian09 Revision D.01*, Gaussian Inc. Wallingford CT 2009.
- 88 A. Lehmann and C. D. Maranas, *Ind. Eng. Chem. Res.*, 2004, **43**, 3419–3432.
- 89 R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sanchez-Carrera, L. Vogt and A. Aspuru-Guzik, *Energy Environ. Sci.*, 2011, **4**, 4849–4861.
- 90 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
- 91 N. M. O’boyle, A. L. Tenderholt and K. M. Langner, *J. Comput. Chem.*, 2008, **29**, 839–845.
- 92 C. Chen, X. Yang, M. Cheng, F. Zhang, J. Zhao and L. Sun, *ACS Appl. Mater. Inter.*, 2013, **5**, 10960–10965.
- 93 G. de Miguel, M. Wielopolski, D. I. Schuster, M. A. Fazio, O. P. Lee, C. K. Haley, A. L. Ortiz, L. Echegoyen, T. Clark and D. M. Guldi, *J. Am. Chem. Soc.*, 2011, **133**, 13036–13054.
- 94 S. Raja, C. Satheshkumar, P. Rajakumar, S. Ganesan and P. Maruthamuthu, *J. Mater. Chem.*, 2011, **21**, 7700–7704.
- 95 J. Cong, X. Yang, J. Liu, J. Zhao, Y. Hao, Y. Wang and L. Sun, *Chem. Commun.*, 2012, **48**, 6663–6665.
- 96 G. E. Zervaki, P. A. Angaridis, E. N. Koukaras, G. D. Sharma and A. G. Coutsolelos, *Inorg. Chem. Front.*, 2014, **1**, 256–270.
- 97 Y. Zhang, Y. Zhang, Z. Wang, M. Liang, D. Jia, Q. Wu and S. Xue, *J. Power Sources*, 2014, **253**, 167 – 176.
- 98 M. Garcia-Iglesias, L. Pelleja, J.-H. Yum, D. Gonzalez-Rodriguez, M. K. Nazeeruddin, M. Gratzel, J. N. Clifford, E. Palomares, P. Vazquez and T. Torres, *Chem. Sci.*, 2012, **3**, 1177–1184.
- 99 G. Wu, F. Kong, J. Li, X. Fang, Y. Li, S. Dai, Q. Chen and X. Zhang, *J. Power Sources*, 2013, **243**, 131 – 137.

**Table 2** Chemical structures of the investigated dyes. Computed isodensity surfaces (0.02 a.u.) of HOMO and LUMO orbitals of dyes. TDDFT spectra are calculated using Gausssum<sup>91</sup>.





**Table 3** Calculated TDDFT-CAM-B3LYP/6 – 31G(d) excitation energies (in dichloromethane) for the lowest transitions, oscillator strengths ( $f_{max}$ ), composition in terms of molecular orbital contributions and the light harvesting efficiency  $LHE(\lambda)$  for the free dyes. Also shown are the charge transfer distance ( $d_{CT}$  in Å) and amount of charge ( $q_{CT}$ ) transferred. The letters "H" and "L" in column 7 (Major transitions) correspond to the HOMO and LUMO orbitals. The last column  $\Delta G_{inj}$  is the driving force of charge injection (see Equation 8).

Dye	PCE(%)	$\lambda_{max}$ (nm/eV)	f	$LHE_{\lambda}$	$q_{CT}$	$d_{CT}$ Å	Major Transitions	$\lambda_i$ (eV)	$\Delta G_{inj}$ (eV)
PTZ-1	9.34±1.92	537/2.31	1.06	0.913	0.72	3.76	H-2→L (11%), H→L (77%)	0.14	-1.00
PTZ-2	9.27±2.32	521/2.38	1.00	0.900	0.68	3.54	H-1→L (15%), H→L (77%)	0.13	-0.96
PTZ-3	9.20±1.80	507/2.44	0.92	0.880	0.68	3.43	H-1→L (18%), H→L (72%)	0.20	-0.88
PTZ-4	9.52±2.07	505/2.46	1.10	0.919	0.69	3.47	H→L (86%), H-1→L (8%)	0.27	-1.18
PTZ-5	9.31±2.24	461/2.69	1.02	0.905	0.70	3.57	H-3→L (19%), H-1→L (64%)	0.21	-1.47