npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01521-9

A combined ML and DFT strategy for the prediction of dye candidates for indoor DSSCs

Check for updates

Carmen Coppola[1,2,3,8], Anna Visibelli[4,8], Maria Laura Parisi[1,2,3,5], Annalisa Santucci[4,6,7], Lorenzo Zani[2], Ottavia Spiga[4,6,7] & Adalgisa Sinicropi[1,2,3,5]

The excellent ability of dye-sensitized solar cells (DSSCs) to capture ambient light and convert it into electric current makes them attractive power sources for indoor applications, including powering Internet of Things (IoT) devices. In this context, substantial research efforts have been devoted to the discovery of novel organic dyes able to harvest energy from a wide range of indoor light sources at different intensities. However, such activities are often based on trial-and-error procedures which are frequently expensive and time-consuming. Here, Machine Learning (ML) techniques and Density Functional Theory (DFT) methods have been combined in a two-stage approach, with the aim to accelerate the design of new, synthetically accessible organic dyes for indoor DSSC applications. By predicting the power conversion efficiency (PCE) under different indoor light sources and intensities, potentially high-performance organic dyes have been identified.

In the last decades, research in the field of dye-sensitized solar cells (DSSCs)[1] has achieved some notable breakthroughs. First introduced by Michel Grätzel and co-workers in 1991[2], DSSCs have gained the interest of the photovoltaic (PV) community thanks to simple fabrication methods based on cost-effective and scalable raw materials[3]. The versatility to tune colors, shapes, and sizes makes a large array of applications attainable, ranging from building-integrated photovoltaics (BIPV) to the incorporation into wearable/portable electronics[4–6]. DSSCs show unique performances under ambient illumination, making them one of the technologies of choice for indoor applications[7,8]. This is related on one hand to their excellent ability to capture diffused light, and on the other to the typical spectral distribution of common indoor light sources, for which the highest theoretical PV efficiencies can be attained for relatively large bandgap values (1.8–2.0 eV), corresponding to those of typical DSSC dyes[9]. Indeed, thanks to their ability to work efficiently under a wide range of indoor light conditions, DSSCs have been proposed as an ideal sustainable power source for IoT (Internet of Things) devices, such as wireless networks and sensors, helping to reduce the massive use and disposal of batteries, and the associated environmental impacts[4,7,8,10–17]. The working principle of a DSSC starts with the photoexcitation of the dye that promotes the electrons from the ground to the excited state. These electrons are then injected into the conduction band of the TiO_2 semiconductor, from which they diffuse to the photoanode, and, through an external circuit, are collected at the counter electrode. The oxidized dye is then regenerated by the redox mediator, present in the electrolyte, which passes from its reduced to its oxidized form and can finally be restored via electron transfer at the counter electrode[13,14]. The first indoor DSSC was realized by Grätzel and co-workers by using the ruthenium-based N719 dye[2]. Thanks to several technology refinements, their power conversion efficiency (PCE) has then significantly improved[18], and efficiencies of up to 38% have been recorded with the co-sensitization of organic dyes under a fluorescent lamp[12]. While the sun is the only source of light in outdoor conditions, a wide range of different indoor light sources exists, i.e. fluorescent lamps—FL (CFL, T5, T8, TL84, etc.), light-emitting diodes—LED, halogen bulbs, and incandescent bulbs. Their emission spectra are found in the 350–700 nm range and their intensity, given in illuminance units (lux), is much weaker than that of sunlight[4,10,11,14,19,20]. Therefore, an efficient dye for indoor applications must have a high molar attenuation coefficient and a good spectral match with artificial light emissions in the visible range, while absorption in the near-infrared region (NIR) is not essential. This aspect implies that long conjugation frameworks could be avoided, thus the design of new dyes for indoor DSSCs could be characterized by simple synthetic routes at lower costs[20]. Additionally, it is fundamental that the dye possesses the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) energy

[1]Department of Biotechnology, Chemistry and Pharmacy, R²ES Lab, University of Siena, Siena, Italy. [2]Institute of Chemistry of Organometallic Compounds (CNR-ICCOM), Sesto Fiorentino, Italy. [3]CSGI, Center for Colloids and Surface Science, Sesto Fiorentino, Italy. [4]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy. [5]LifeCARES s.r.l | Renewable Energy and Sustainability, Arezzo, Italy. [6]Competence Center ARTES 4.0, Siena, Italy. [7]Siena-bioACTIVE—SbA, Siena, Italy. [8]These authors contributed equally: Carmen Coppola, Anna Visibelli. e-mail: adalgisa.sinicropi@unisi.it

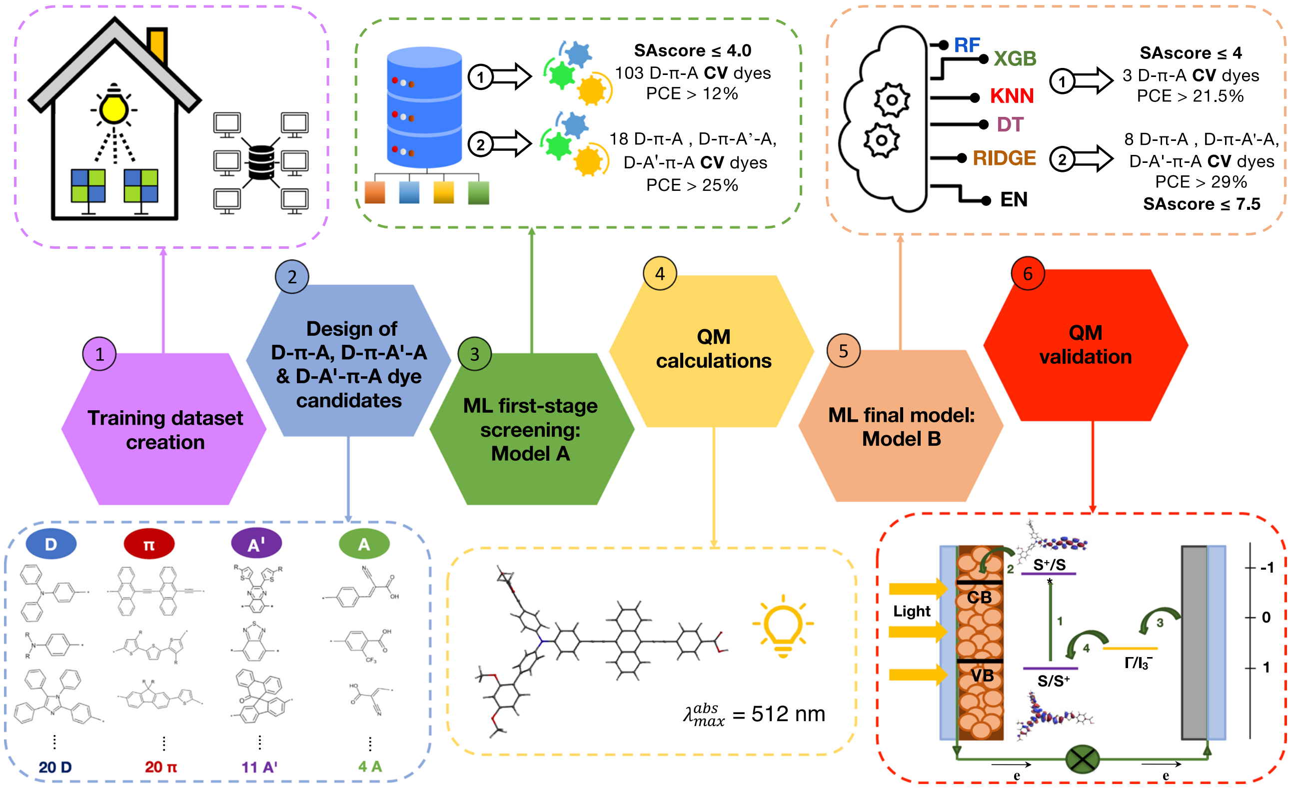


Fig. 1 | Flowchart of the strategy employed in this work.

levels well aligned with the redox potential of the redox couple and the conduction band of TiO_2, respectively, to guarantee dye regeneration and efficient charge injection into TiO_2 while minimizing electron recombination[7,10,12,14,20,21]. Considering all the above aspects, in this work, we focused our attention on metal-free organic dyes with the peculiar $D_{(Donor)}–\pi_{(spacer)}–A_{(Acceptor)}$ architecture that appear ideal candidates for indoor applications, since they could fulfill all the requirements that maximize the PCE of indoor DSSCs. In particular, they are responsible for absorption in the visible region that well-match the emission peaks of indoor lamps, and upon photoexcitation, they undergo an intramolecular charge transfer transition (ICT) from the donor to the acceptor moiety that facilitates a rapid electron injection into the semiconductor. Additionally, the optoelectronic properties of these dyes can be easily modulated by the introduction of an auxiliary acceptor unit (A') giving rise to D-π-A'-A and D-A'-π-A structures that contribute to limiting intermolecular aggregation and to ensuring good spatial separation between the HOMO and LUMO levels[7,11,14].

To date, considerable research efforts have been made to discover novel organic dyes for DSSCs with indoor applications, leading to excellent results[7,8,11,20]. However, the traditional process for the development of new sensitizers is often based on trial-and-error protocols, which frequently turn out to be expensive and time-consuming. In this framework, data-driven approaches, such as Machine Learning (ML) techniques, can be considered a valuable strategy to accelerate the discovery of novel materials for optoelectronic applications. In particular, the use of ML techniques in combination with quantum chemical approaches based on *state-of-the-art* Density Functional Theory (DFT) methods was proven to be a successful methodology for predicting the performance of novel compounds in solar energy conversion devices[22–31]. Recently, several ML models have been built by implementing DFT data on electronic structures and energetics to predict the PCE of different kinds of solar cells[23–26,32–34]. For example, in the framework of DSSCs, Ju et al.[25] reported a multiple linear regression model (MLR) combined with a hybrid genetic algorithm (GA) to predict the PCE values of N-annulated perylene (N-P) organic sensitizers. The results inspired them to design novel dyes of the same class. Still, N-P organic sensitizers with PCE > 13% have been discovered by Zhang et al.[34] by combining DFT methods with three learners. Lu et al.[26] implemented two GA-MLR quantitative structure-property relationship (QSPR) models with DFT calculations to design BODIPY-based sensitizers and predict their PCE values; Wen et al.[23] merged the approach based on ML and DFT for the identification of novel organic dyes with the estimation of their synthetic accessibility.

However, to the best of our knowledge, this approach has never been applied to the discovery of novel organic dyes for indoor DSSCs' applications. In this work, ML techniques and DFT methods have been combined for the first time in a two-stage approach to automatically design new organic dyes with D-π-A, D-π-A'-A, and D-A'-π-A architectures and predict their PCE in indoor DSSCs under different artificial lighting conditions. The flowchart of the employed strategy is reported in Fig. 1. Firstly, the structures of 61 organic dyes applied in indoor DSSCs with PCE > 3% have been collected from the literature. Their building blocks have been fragmented and recombined to generate novel dye candidates. Hence, a two-stage ML approach has been developed. More specifically, it is composed of Model A which relies on 1442 molecular descriptors and 2 experimental descriptors for preliminary screening, and Model B which includes 24 new descriptors from DFT calculations and 2 more experimental descriptors for second-stage screening. Moreover, as an important element in the design of novel dyes, the synthetic accessibility of the new candidates has been initially evaluated by calculating their SAscore parameter[35]. First, the two-stage ML model has been applied to D-π-A dye candidates with a restricted SAscore. Once the model has been satisfactorily assessed, it has been applied to all the D-π-A, D-π-A'-A, and D-A'-π-A candidates for PCE prediction, and their SAscores have been then computed. At this stage, the synthetic feasibility of the resulting top candidates has been more thoroughly evaluated by means of a dedicated retrosynthetic analysis. The approach presented here led to the prediction of three novel and potentially very efficient organic dye candidates for indoor DSSCs.

Results

Design of dye candidates

The chemical structures of the 61 collected dyes (Table S1) have been fragmented using the RDkit Python package (version 2023.09.5). Thus, 20 donor (D), 20 spacer (π), 11 auxiliary acceptor (A'), and 4 acceptor (A) building blocks have been obtained (Fig. 2). Then, they have been randomly combined via the RDkit Python package to automatically generate 2.000 D-π-A, 240.000 D-π-A'-A e 240.000 D-A'-π-A molecular structures. It should be noted that while this technique could lead to the generation of molecular structures that are permutations of existing ones, it offers a strategic method for exploring a wide chemical space and it has already successfully been used in the literature, even for the identification of novel dye candidates[23].

These molecules are characterized by donor groups mostly based on triarylamine, dialkylamine, *N,N*-diarylamine, *N*-aryl phenothiazine, diphenyl-substituted pyranylidene, carbazole, fluorenyl indoline, and triphenylimidazole moieties often decorated with *p,o*-alkoxy groups or *n*-hexyl chains.

The spacers are represented by thiophene-, thieno[3,2-*b*]thiophene-, anthracene-, terthiophene-, cyclopentadithiophene-, benzene-, fluorene-, perylene-, and acetylene-based units, often decorated by *n*-hexyl chains. The presence of bulky alkoxy and alkyl substituents on D and π groups is expected to support the formation of an insulating layer on the semiconductor surface, while at the same time reducing dye aggregation and improving hydrophobicity, which are all crucial factors for ensuring the long-term stability of the cell[5,36,37]. Moreover, the incorporation of ethynyl-based units and the extended conjugation length in the π spacer could shift the dyes' absorption in the visible region of the spectrum and lead to strong ICT transitions from the donor to the acceptor moieties[5,37]. Electron-deficient auxiliary acceptor groups are mainly represented by benzothiadiazole, benzotriazole, thieno[3,4-*b*]pyrazines, benzo[3,4-*b*]pyrazine, quinoxaline, and spiro[fluorene-9,9'-phenanthren]-10'-one groups, whose presence could enhance the light-harvesting ability of the dyes and enable an effective photoinduced charge separation, helping in the optimization of energy gaps and leading to red-shifted absorptions[38–40]. Moreover, fine-tuning the A' unit could potentially reduce intermolecular aggregation and,

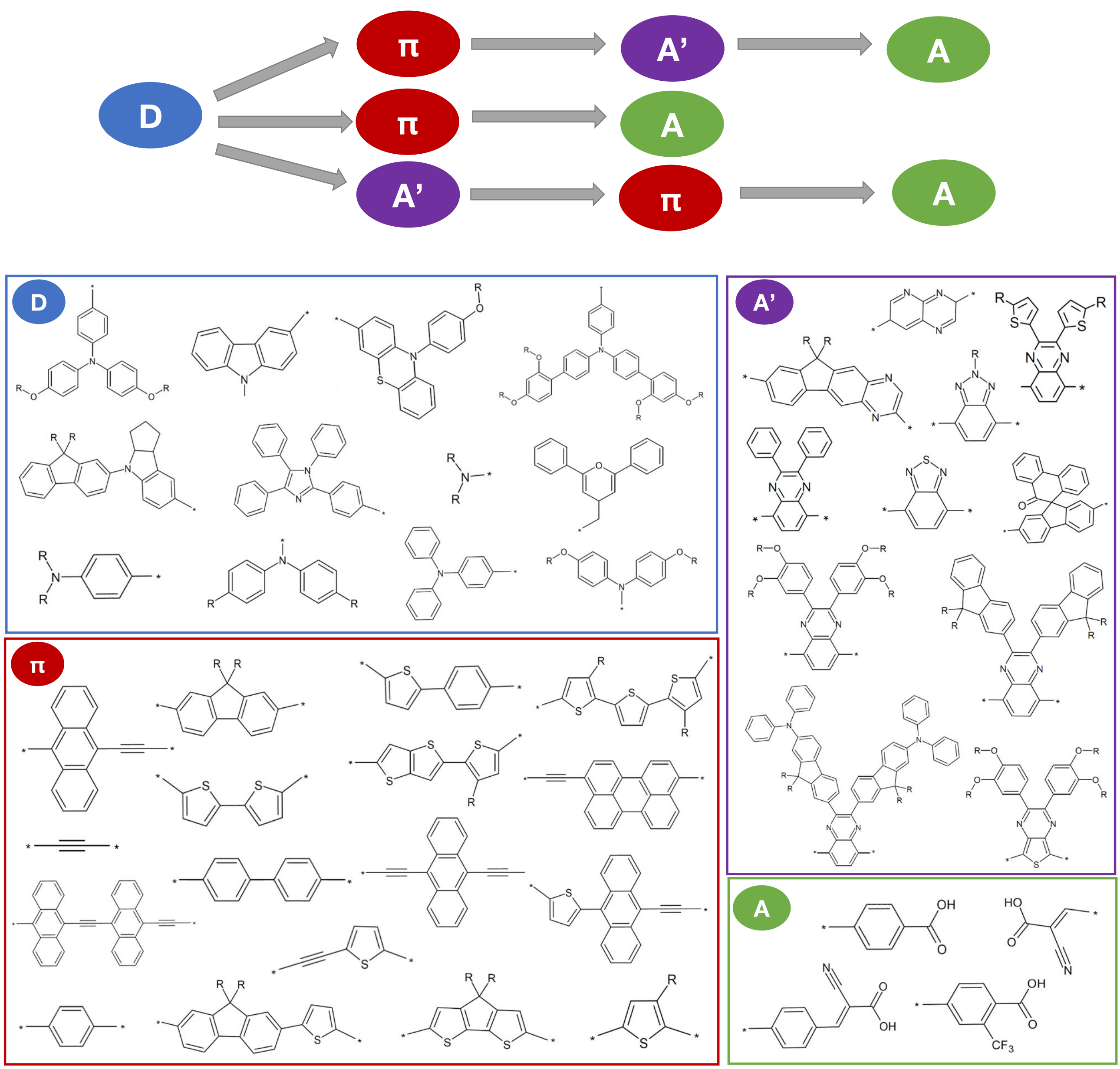


Fig. 2 | Chemical structures of donor (D), spacer (π), auxiliary acceptor (A'), and acceptor (A) building units. "R" is used to indicate variable alkyl chains (from 4 to 12 C atoms).

consequently, limit charge recombination processes[41,42]. Finally, the acceptor moieties of the dye candidates are mostly based on electron-withdrawing units such as the commonly employed cyanoacrylic acid and benzoic acid, possibly decorated with a trifluoromethyl group in the ortho position.

Application of a two-stage ML model (Model A and Model B) to D-π-A dye candidates with SAscore ≤ 4

After applying the fragmentation-recombination technique, we conceptually divided the work into two parts. In the first instance, to reduce computational costs, we focused our attention on the 2.000 D-π-A candidates and preliminarily evaluated their synthetic accessibility using the Ertl and Schuffenhauer method[35]. Known as the SAscore, it is based on the combination of molecule complexity and fragment contributions (Eq. 1) and can assume values between 1 (easy to synthesize) and 10 (difficult to synthesize).

$$\text{SAscore} = \text{fragmentScore-complexityPenalty} \tag{1}$$

where "fragmentScore" represents the contribution of each fragment in a molecule divided by the total number of fragments, and "complexityPenalty" accounts for complex structural features. As suggested in the literature[23,34], assessing synthetic accessibility is crucial to guarantee a feasible feature synthesis of the dye candidates (for a more detailed discussion, see below).

Only a restricted number of D-π-A candidates with SAscore ≤ 4 have been considered here, leading to 1520 molecules.

Model A implementation

A preliminary screening has been conducted by running Model A to predict dye candidates' PCE under different light sources and intensities (see the corresponding descriptors in Table S2). The pre-processed dataset of collected dyes has been split into two sets: the training set and the validation set. The training set constitutes 85% of the dataset and has been used to train the model to learn the hidden features. The remaining 15% of the dataset belongs to the validation set, which has been used to test the model performances after each training phase. Each network has been trained and tested on 100 runs with a different dataset split. Model A achieves a high prediction performance (Mean Absolute Error (MAE): 1.45, Root Means Squared Error (RMSE): 4.31, Standard Deviation (SD): 0.23, Coefficient of determination (R^2): 0.90) and predicts for most candidates a PCE in the range of 6–11%. Interestingly, a relevant number of promising dyes with PCE exceeding 12% or even 13% has been also found (Figure S1). As a result, 103 novel molecular structures (**CV** dyes) with PCE > 12% and SAscore ≤ 4 have been identified. Additionally, the Reaxys database (www.reaxys.com) has been queried to verify that these molecules have never been employed as indoor dyes so far.

Model B implementation

To improve the accuracy of the PCE prediction, descriptors based on the optoelectronic properties of dyes should also be included, as they could affect the working mechanism of indoor DSSCs. Therefore, DFT calculations have been performed on the 103 **CV** molecules to obtain 24 QM descriptors which have been included in Model B (Table S2). The molecular structures and the ground-state optimized geometries of the 103 **CV** dyes are reported in Table S3.

To assess the performance of Model B, it was trained and tested across 100 runs using different dataset splits, consistently maintaining the same 85%–15% train-validation ratio as Model A. Thereafter, R^2, MAE, RMSD, and SD over 100 runs have been calculated and compared in Fig. 3.

Based on the results displayed in Fig. 3, the XGB model demonstrated exceptional efficacy in accurately predicting most of the PCE values (MAE: 1.19, RMSE: 2.88, STD: 0.17, R^2: 0.94). Additionally, the performances of Model B trained on XGB are significantly improved compared to Model A (MAE: 1.45, RMSE: 4.31, STD: 0.23, R^2: 0.90). This result suggests that the implementation of additional descriptors, such as the quantum descriptors obtained from DFT calculations, can capture the key molecular characteristics relevant to the prediction of PCE. To understand the contribution of these features in the PCE prediction, a feature importance bar graph is reported in Fig. 4. The feature importance technique assigns a score to the input features based on their usefulness in predicting a target variable. In this case, it indicates the value of each attribute in the construction of the boosted decision trees within the model. Thus, the higher the importance score of an attribute, the more significant it is in the decision-making process[43].

The bar graph shows that the most relevant feature is EL-ECB (energy difference between LUMO and conduction band of TiO_2). Other important attributes come from the energy of frontier molecular orbitals (FMOs), i.e. HOMO-1, HOMO, LUMO, and LUMO + 1, and the light-harvesting ability of the dye. Considering experimental descriptors (light intensity, light source, electrolyte, and counter electrode), the PCE prediction is most affected by the light intensity and the light source of the employed indoor lamps.

Then, Model B was applied to four independent datasets (**D1, D2, D3**, and **D4**) of the 103 **CV** molecules where the quantum descriptors related to the absorption maxima values were computed using two different levels of theory and THF and DCM solvents (see the Methods section for details). For each dataset, every molecular structure has been associated with all the possible combinations of the experimental descriptors to find the one capable of maximizing the PCE of the dye. Indeed, unlike previous studies, the indoor focus required specific experimental descriptors, such as the type of light source and intensity of indoor lamps, that result in various PCE values depending on their combinations. The Test set of Model B has been constructed according to the scheme reported in Figure S2, and it results in 280.160 entries. Model B predicts 280.160 PCEs ranging from 4.30% to 22.04%, with an average PCE of 13.84%. Since no significant differences have been detected in the PCE prediction from the application of Model B to the four independent datasets, only the results of **D1** are discussed. The highest predicted PCE and the relative SAscore of the 103 **CV** molecules are reported in Table S4. The highest PCE has been obtained for all the molecules at 6000 lux as light intensity and with FTO/PEDOT as the counter electrode. The distribution of the best (>14%) predicted PCE values is shown in Figure S3. In particular, three promising organic dyes (**CV77, CV85**, and **CV86**) with predicted PCE > 21.5% under different conditions have been identified (Table S5). The two best results for these molecules are shown in Table 1. A detailed description of the molecular and electronic properties of these molecules is reported in Section B of SI.

Considering the significant predicted performances of **CV77, CV85**, and **CV86** dye candidates, besides calculating their SAscore (*see above*), their synthetic accessibility was evaluated in more detail by carrying out a detailed retrosynthetic analysis, (see Supporting Information, Section D, for details), with the aim to determine the most efficient bond disconnections to reduce their molecular complexity and identify suitable starting materials for their preparation, either easily accessible or commercially available. On that basis, possible preparation routes have been suggested for all compounds, describing the synthetic steps necessary for their assembly and discussing the potential challenges associated with the most important transformations. In general, feasible synthetic sequences have been found for all compounds, which were backed up by previous results reported in the relevant literature (see Supporting Information for references).

Application of the two-stage ML model (Model A and Model B) to D-π-A, D-A'-π-A, and D-π-A'-A dye candidates

Based on the promising results described above, the two-stage ML model has been applied to all the dye candidates resulting from the fragmentation-recombination technique. This involved the examination of the 2.000 D-π-A, 240.000 D-π-A'-A e 240.000 D-A'-π-A molecular structures without any restriction imposed by the SAscore evaluation. The aim was to discover dye candidates with the highest predicted PCE values.

Hence, Model A has been run for preliminary screening and it predicts 18 novel **CV** (**CV104** - **CV121**) dye candidates (4 D-π-A, 1 D-A'-π-A, and 13 D-π-A'-A) with PCE > 25% under different light sources and intensities.

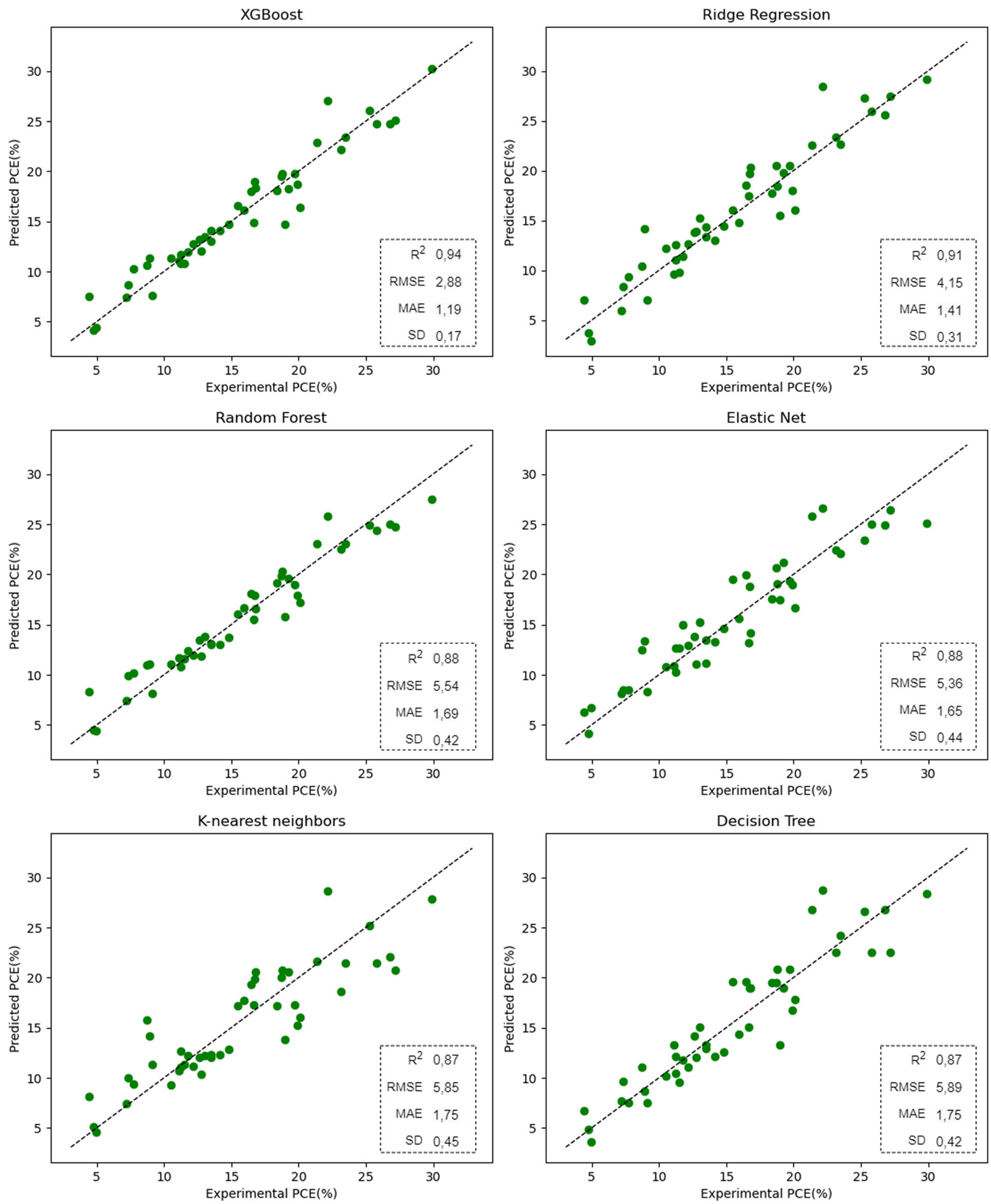
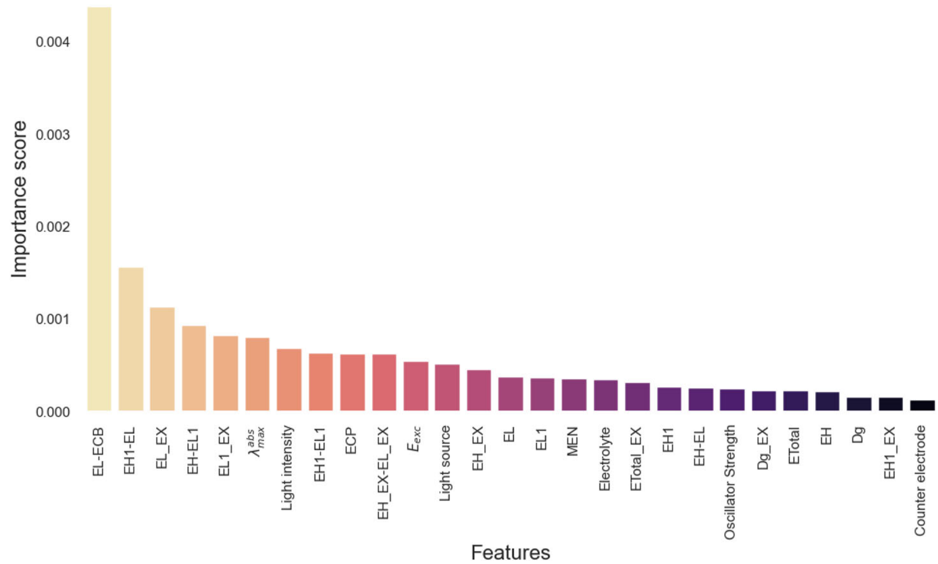


Fig. 3 | Predicted PCEs of Model B trained by XGB, Ridge, RF, EN, KNN, and DT versus experimental PCEs.

The Reaxys database (www.reaxys.com) has been queried to verify that these molecules have never been employed as indoor dyes so far.

DFT calculations have been performed to build the four independent datasets (**D1, D2, D3**, and **D4**) and Model B has been applied. Afterward, the SAscore of the novel 18 **CV** dye candidates was evaluated. The molecular structures and the ground-state optimized geometries are reported in Table S6. Also in this case, the application of Model B to the four independent datasets does not bring significant differences in the PCE prediction, and only the results of the **D1** dataset are discussed. Model B predicts PCE ranging from 27.00% to 31.66%, with an average PCE of 28.91%. The highest predicted PCE and the relative SAscore of these novel 18 **CV** dyes are reported in Table S7. It should be noted that these molecules have a SAscore ranging from 4.53 to 7.31, which aligns perfectly with the SAscore values of the training dataset, for which we got a maximum value of 7.66. In

Fig. 4 | Feature importance analysis: the bars illustrate the contribution of the features in Model B.



particular, this approach led to the identification of one promising dye candidate for each molecular architecture (**CV106, CV108**, and **CV121** for D-π-A, D-A'-π-A, and D-π-A'-A, respectively) with a PCE higher than 29% under the emissions of FLs (i.e. T5, T2, TL84, and OSRAM), regardless of the counter electrode and the electrolyte conditions, as detailed in Table S8. The two best results obtained for each dye candidate with the relative SAscore are reported in Table 2.

When comparing these data with the experimental PCE values of the dyes in the training dataset (Fig. 5), we observe that the highest reported PCE values for indoor applications are 30.24% for the D-π-A (YK8 dye), 28.95% for the D-A'-π-A (MM-6 dye), and 37.07% for the D-π-A'-A (CXC22 dye) architectures (see also Table S1). The PCE values of YK8, MM-6, and CXC22 have been obtained under the emissions of an Osram FL lamp at 1500 lux, a TL84 FL lamp at 2500 lux, and a T5 FL lamp at 6000 lux, respectively. Using Model B, we can predict these experimental PCE values at 29.19% (YK8), 28.33% (MM-6), and 34.91% (CXC22). It should be noted that extreme values are predicted less accurately by the model due to their underrepresentation in the dataset, leading to predictions that tend to be closer to the mean. Considering the particularly high PCE value of CXC22, we also compared **CV121** with the second-best experimental PCE value for the D-π-A'-A architecture which has been reported for TY6 dye at 28.50% under a T5 FL lamp at 6000 lux. Using Model B, we can predict this experimental PCE value at 26.66%. Thus, it is possible to assume that **CV106, CV108**, and **CV121** could lead to experimental PCE values in line with or higher than current literature findings.

As seen before for the dye candidates presented in Table 1, also for **CV106, CV108**, and **CV121** dye candidates a thorough retrosynthetic analysis was performed, to identify potentially privileged routes for their preparation. The results of such analysis, as well as the corresponding proposed synthetic sequences, are reported in the Supporting Information (Section D). Also in this case, the feasibility of the identified routes has been supported by citing relevant literature works describing similar synthetic operations carried out on structurally related compounds (see Supporting Information for references).

Analysis of optoelectronic properties of best-performing CV dye candidates

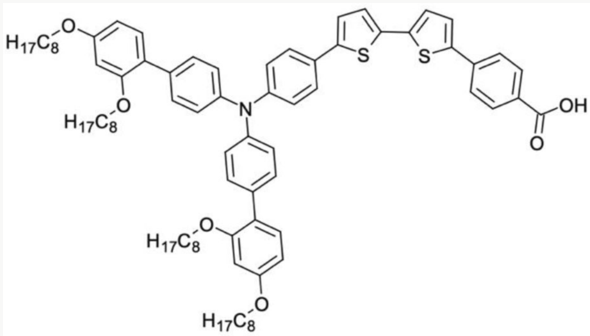
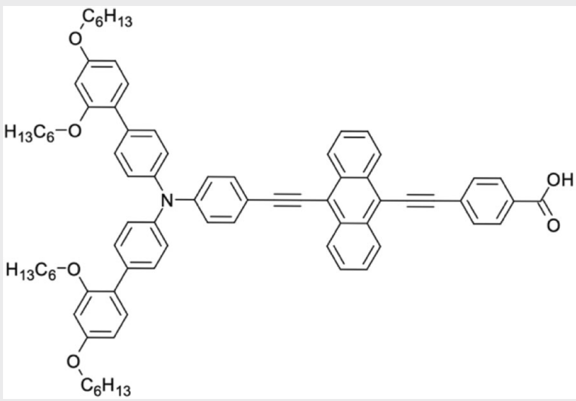
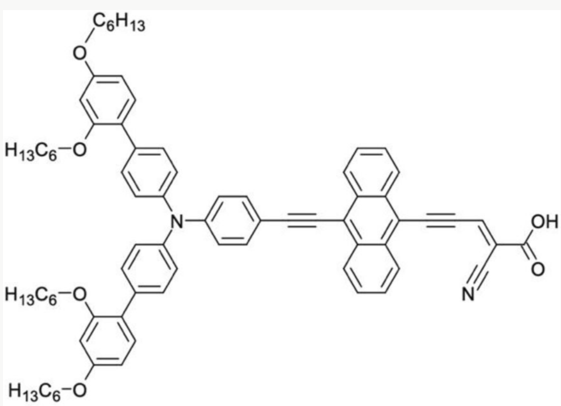
Compounds **CV106, CV108**, and **CV121** have highly conjugated π systems due to the inclusion of indacenodithiophene, ethynyl, and anthracene units. They feature arylamine derivatives and triphenylimidazole groups as donor units while electron-withdrawing cyanoacrylic acid-based groups serve as acceptors. Additionally, **CV108** and **CV121** contain a benzotriazole and a substituted quinoxaline, respectively, as their auxiliary acceptor groups. The optimized geometries of the *in-vacuo* ground state and the first excited state in THF and DCM of the three molecules are reported in Figure S4. The ground state geometries of three molecules show dihedral angles in the range of 2.3°–41.4°, while calculated geometries for S_1 show overall increased planarity of the molecules in both solvents. M06-2X/6-311 + G(2 d,p) absorption (λ_{max}^{abs}) maxima, vertical excitation (E_{exc}) energies, oscillator strengths (f), and composition (%) in terms of molecular orbitals to the lowest energy ($S_0 \rightarrow S_1$) transitions in THF of compounds **CV106, CV108**, and **CV121** are reported in Table 3. Data at CAM-B3LYP/6-311 + G(2 d,p) level of theory in THF are reported in Table S9.

The three **CV** dyes present absorption maxima in the range of 486–549 nm (2.55–2.26 eV), which are associated with intramolecular charge transfer transitions mainly involving HOMO→LUMO orbitals. Other minor contributions to the lowest energy transitions involve HOMO-1→LUMO orbitals for **CV106** and **CV108** and HOMO→LUMO + 1 orbitals for **CV121**. Essentially, all the dye candidates match well with artificial light emissions in the visible range. In particular, they can be employed in the presence of fluorescent T5 lamps and other commonly employed FLs, such as TL84, T2, and T8, whose emission peaks range from 450 nm (2.76 eV) to 610 nm (2.03 eV). Moreover, the **CV106, CV108**, and **CV121** dyes can also be combined with the emission peaks of blue, green, and orange-red LED lamps (from 450 nm to 620 nm)[11,44,45]. The Kohn-Sham FMOs energies and their electron density distributions at the B3LYP/6-31 G* level of theory in THF are reported in Fig. 6.

From Table 3 and Fig. 6, it is possible to notice a red-shift going from **CV106** to **CV121**, which displays the most red-shifted computed absorption maximum, in agreement with the smallest HOMO-LUMO gap. On the other hand, **CV108** displays a slight bathochromic shift compared to **CV106**, despite its larger frontier orbital gap. However, it should be considered that a larger fraction of the HOMO-1→LUMO transition contributed to the main absorption band of **CV106** compared to that of **CV108**, which, instead, is dominated by the HOMO→LUMO transition.

The inspection of the wavefunction plots reveals that the electron density distribution in HOMOs is mainly located on the donor and the conjugated scaffold of the dye candidates, while the LUMOs are mostly located on the acceptor moieties, again with a considerable contribution of the conjugated scaffold. Such superposition of HOMOs and LUMOs on the central scaffold of the **CV** dyes supports a good degree of intramolecular charge transfer upon photoexcitation, suggesting a high intensity of the related transitions, as shown also by the large oscillator strengths values (Table 3)[46].

Table 1 | High-performance organic CV molecules with predicted PCE > 21.5% resulting from Model B

Dye Name	2D Structure	Predicted PCE (%)
CV77		**22.04** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^- and $[Co(bpy)_3]^{2+/3+}$) **22.02** (TL84 and T8 at 6000 lux, FTO/PEDOT, I^-/I_3^- and $[Co(bpy)_3]^{2+/3+}$)
CV85		**21.90** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^- and $[Co(bpy)_3]^{2+/3+}$) **21.82** (Osram 14 W T2 at 6000 lux, FTO/PEDOT, I^-/I_3^- and $[Co(bpy)_3]^{2+/3+}$)
CV86		**21.63** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^- and $[Co(bpy)_3]^{2+/3+}$) **21.54** (Osram 14 W T2 at 6000 lux, FTO/PEDOT, $[Co(bpy)_3]^{2+/3+}$)

Additionally, it is possible to notice that all the LUMO energies are higher than the conduction band of TiO_2 (−4.00 eV), predicting a favorable electron injection. Even if the B3LYP functional has a small fraction of Hartree-Fock exchange which limits the accuracy in the prediction of HOMO energies[47], it is possible to affirm that all the HOMO energies of the **CV** dyes are aligned with the redox potential of the most commonly employed redox couple I^-/I_3^- electrolyte (−4.80 eV), ensuring dye regeneration[48]. Additionally, it is possible to assume a feasible dye regeneration after photoexcitation in the presence of Co-based (from ca. −4.85 to ca. −5.04 eV) redox couple systems[49,50].

Discussion

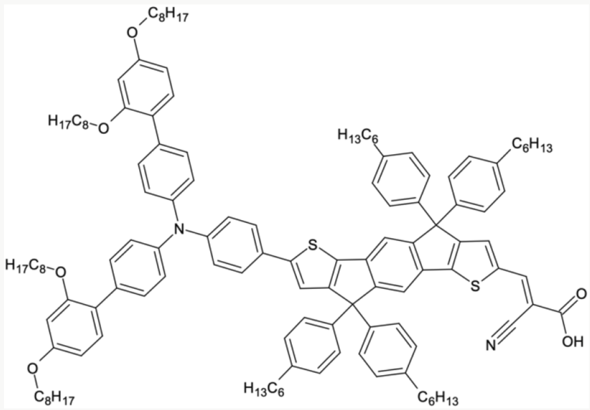
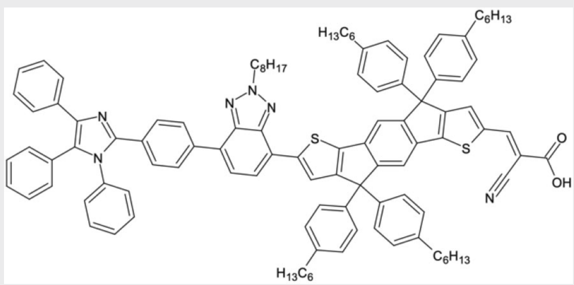
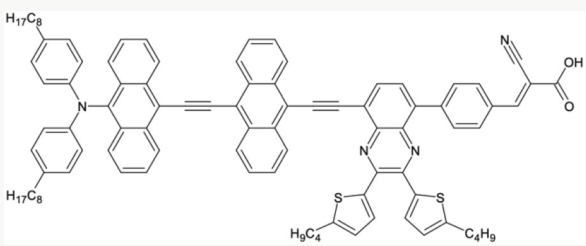
In this work, a two-stage approach combining ML and DFT methods has been applied for the first time to identify new potential organic dye candidates for indoor DSSCs' application by predicting PCE values under different light sources and intensities. The developed approach led to the identification of three promising dye candidates (**CV106, CV108**, and **CV121** with D-π-A, D-A'-π-A, and D-π-A'-A architectures respectively) for indoor DSSCs with PCE > 29% under different artificial illumination conditions.

Additionally, the synthetic accessibility of the dye candidates has been evaluated and the Reaxys database has been queried to verify that these molecules have never been employed as indoor dyes so far.

The developed ML-DFT protocol represents a powerful approach to accelerate the discovery of novel organic dyes for indoor DSSCs' applications. However, a current limitation could be the size of the dataset for training the ML model. Nevertheless, we constructed our dataset to adequately capture crucial patterns in indoor DSSCs' behavior and used literature data on indoor DSSCs that are available up to now. A future direction for this work could involve expanding the training dataset as more data for dyes for indoor applications becomes available.

In this work, a set of molecular descriptors that effectively define critical aspects of the dyes has been incorporated. Nevertheless, adding more

Table 2 | High-performance organic CV molecules with predicted PCE > 29% resulting from Model B

Architecture	Dye name	2D Structure	Predicted PCE (%)	SAscore
D-π-A	**CV106**		**29.26** (T5 at 6000 lux, FTO/PEDOT, $[Co(bpy)_3]^{2+/3+}$) **29.12** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^-)	6.13
D-A'-π-A	**CV108**		**30.46** (T5 at 6000 lux, FTO/PEDOT, $[Co(bpy)_3]^{2+/3+}$) **30.44** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^-)	5.62
D-π-A'-A	**CV121**		**29.83** (T5 at 6000 lux, FTO/PEDOT, $[Co(bpy)_3]^{2+/3+}$) **29.81** (T5 at 6000 lux, FTO/PEDOT, I^-/I_3^-)	4.95

chemical and structural features could improve the model's accuracy but significantly increase the computational costs for computing DFT descriptors.

Another aspect to consider for future research is the experimental validation of the predicted dye candidates. Indeed, finding confirmation of the predictive claims presented in this study would be key to understanding the true potential and limitations of this methodology and building confidence in its wider application to different problems in the field of energy materials research. With this goal, as mentioned above, the synthetic accessibility of the dye candidates reported in Tables 1, 2 has been analyzed, and possible pathways for their synthesis have been proposed. The feasibility of the suggested synthetic routes will be tested by attempting the preparation of at least some of these compounds, exploiting the authors' previous experience in the synthesis of complex organic molecules for energy applications (see, for example, refs. 51–54). After successful completion of the synthesis, the dyes will be used to sensitize nc-TiO_2 photoanodes, which would in turn be employed for the fabrication of DSSCs on a small laboratory scale (0.25 cm^2), on which the authors have accumulated significant expertise over the years. Cells will be built with different electrolyte mixtures and counter electrode materials, as suggested in the study, to test the relative compatibility of the dye candidates. Finally, their power conversion efficiency will be assessed by measurements conducted under various kinds of indoor light sources and compared to the computational predictions.

In this context, the goal for future efforts is to adopt an iterative framework where the model is continuously updated with validated data, which could contribute to enhancing its efficiency. Therefore, expanding the methodology to create a larger and mixed dataset that encompasses experimental and computed data would lead to further advancing the research in indoor DSSC technology. Additionally, for future works, it should be taken into account that in the upcoming years, it is highly advisable to conduct indoor PV measurements according to the technical specification IEC TS 62607–7–2 and to the European standard EN 12464, which defines the appropriate artificial lighting conditions in work environments[55].

Overall, we demonstrated that the presented ML-DFT protocol has the potential to generate new insights compared to the traditional, trial-and-error procedures routinely employed for developing organic light-harvesting materials.

Methods

Data collection and pre-processing

Despite the large number of organic dyes reported up to now, to create the training dataset of this work, we had to focus on dyes developed for indoor conditions. To this end, we collected 61 molecules individually tested as dyes in indoor DSSCs with PCE > 3% under different light sources and intensities from the literature published until the end of 2023 (Table S1). To ensure the consistency of the training dataset, data from co-sensitized molecules and metal-based sensitizers have been excluded. The 61 dyes are representative of the three most employed architectures

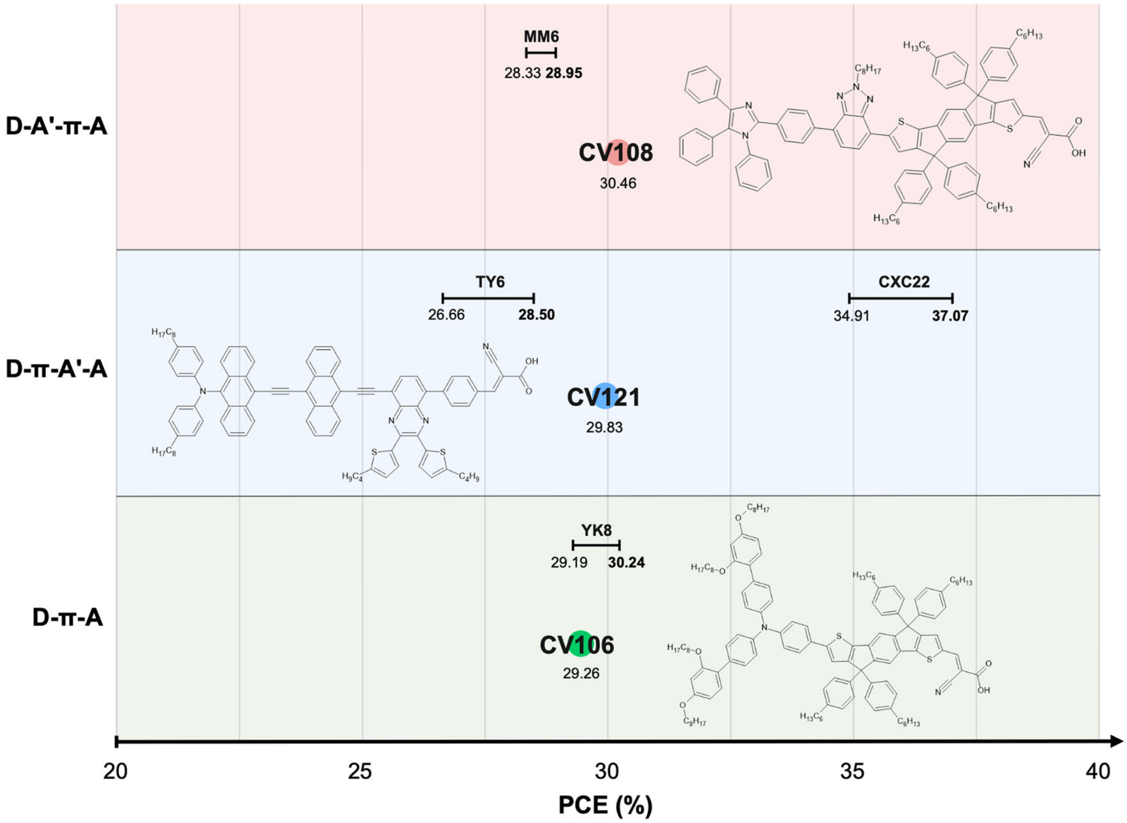


Fig. 5 | Best predicted dye candidates (**CV106, CV108**, and **CV121**) vs top-performing dyes of the training dataset (MM6, CXC22, TY6, and YK8) with predicted and experimental (in bold) PCE values.

Table 3 | M06-2X/6-311 + G(2 d,p) absorption (λ_{max}^{abs}) maxima, vertical excitation (E_{exc}) energies, oscillator strengths (f), and composition (%) in terms of molecular orbitals to the lowest energy ($S_0 \rightarrow S_1$) transitions in THF of compounds CV106, CV108, and CV121

Molecule	λ_{max}^{abs}	E_{exc}	f	Contribution (%)
CV106	486	2.55	2.02	53% H→L 35% H-1→L
CV108	490	2.53	2.48	74% H→L 10% H-1→L
CV121	549	2.26	1.78	66% H→L 20% H→L + 1

(D-π-A, D-π-A'-A, and D-A'-π-A) of organic dyes applied in indoor conditions. Moreover, they belong to different dye families, featuring a variety of functional groups that result in chemical structures with different sizes, complexity, and connectivity. This diversity allowed us to include various structural and electronic properties in our dataset that are crucial for ensuring the predictivity and the generalizability of the model. Additionally, the generalizability of the model is further supported by the variety of experimental conditions of light sources and intensities considered in the dataset. The 61 dyes' chemical structures have been identified by Simplified Molecular-Input Line-Entry Systems (SMILES)[56] and used to calculate 1442 molecular descriptors with Mordred[57] Python library which has free accessibility and provides high computational speed. In particular, the Mordred Python library includes over 1800 descriptors related to structural, topological, and physico-chemical properties. From this collection, 1442 descriptors have been able to capture the wide range of molecular features that could influence dye performance, from basic atom counts to complex three-dimensional spatial data. The choice of using Mordred was also supported by its efficiency in computing descriptors for complex molecules which is fundamental in the first-stage screening of our protocol to sort out molecules without incurring heavy computational costs. In particular, we included topological, geometrical, spatial, physico-chemical, fragment-based, and autocorrelation molecular descriptors. This large selection of descriptors gives insights into molecules' structural properties that need to be addressed in the design process of novel dye candidates and could also influence how dyes interact with the semiconductor surface. Moreover, they consider structure-property relationships of distinct fragments, as well as the connectivity and arrangement of atoms in the dye candidates. The inclusion of such molecular descriptors allows us to account for the key features that may correlate with the light absorption and charge transport behavior of the dye candidates, thus contributing to the PCE prediction in indoor DSSCs. In particular, the first-stage screening (Model A) has been carried out using the 1442 molecular descriptors and 2 experimental descriptors, *i.e.* light source (i.e., LED, T5, etc.) and light intensity (200–6000 lux). In the second-stage screening (Model B), 24 new quantum descriptors obtained from DFT calculations and 2 more experimental descriptors (the counter electrode and the electrolyte composition) have been also included since dataset molecules have been tested with different electrolytes (I^-/I_3^-, EL-201, $[Co(bpy)_3]^{2+/3+}$, $[Cu(dmp)_2]^{2+/1+}$ $[Cu(tmby)_2]^{2+/1+}$) and counter electrodes (Pt, PEDOT, PVP-Pt, Pt/C). Descriptors from DFT calculations have been selected considering the basic operation principles of indoor DSSCs and following previous works[23,25,26]. In particular, they have been computed or derived from dye geometric and electronic structures and they are highly related to the UV-Vis properties of the dye candidates. Table S2 reports all the descriptors' information. In both Model A and Model B, the training dataset has been pre-processed, encoding each categorical value between 0 and n_classes-1, while numerical columns have been standardized by removing the mean and scaling to unit variance.

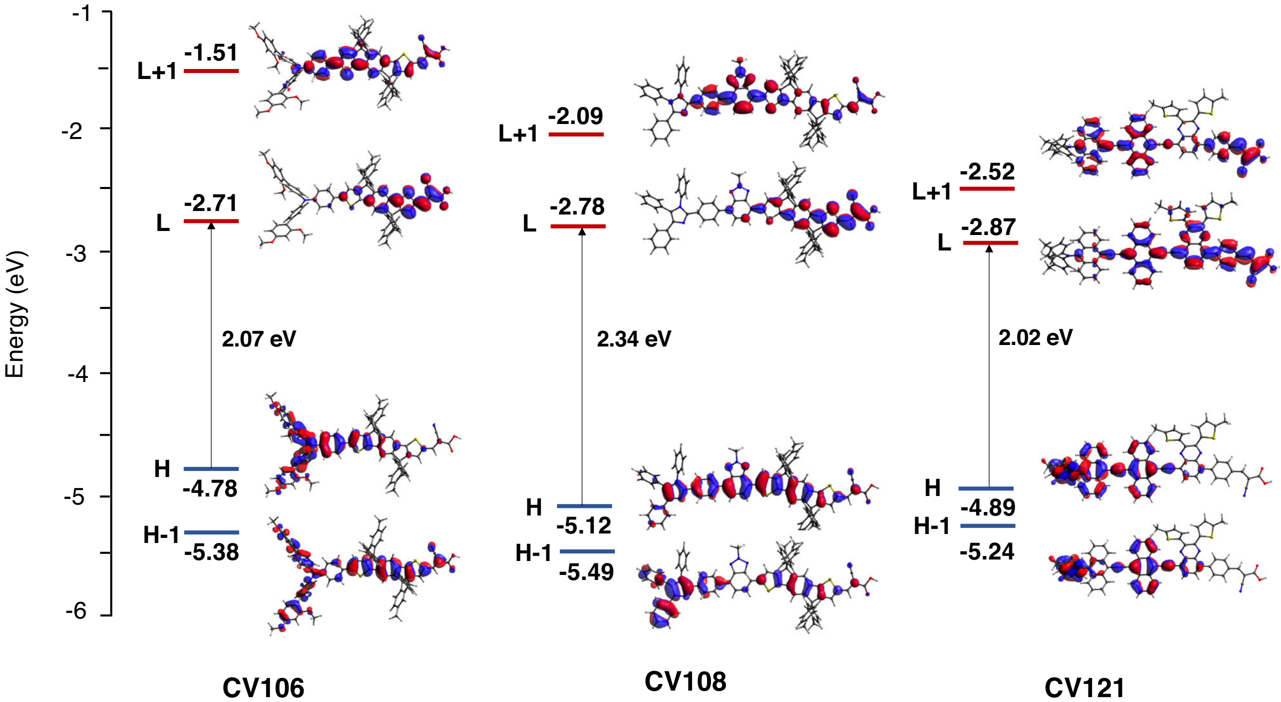


Fig. 6 | Energy levels and electron density distribution of FMOs of compounds **CV106, CV108**, and **CV121** in THF.

Machine Learning techniques

Model A. The first-stage screening (Model A), which is based on structural and physicochemical properties, has been implemented using the Extreme Gradient Boosting model- XGBoost (XGB 1.7.6)[43], an ML technique that generates a predictive learner by combining a collection of weak predictive models, allowing the optimization of an arbitrary differentiable cost function. Decision trees have been selected as weak predictors and they outperform all the other algorithms in manipulating small and structured data.

Model B. The second-stage screening (Model B), which incorporates quantum descriptors obtained from DFT calculations, has been built by comparing the performance of six regressors in predicting the PCE of dyes: XGBoost (XGB)[43], Random Forest (RF)[58], Ridge regression (Ridge)[59], Elastic Net (EN)[60], K-nearest neighbors (KNN)[61], Decision Tree (DT)[62]. The implementation of these models enabled a comprehensive investigation of both linear and non-linear relationships within the dataset. XGB and RF were selected based on their interpretability and robustness in handling complex and high-dimensional data. Furthermore, these ensemble techniques can estimate the contributions of features in predicting PCE values by effectively reducing overfitting and leveraging feature importance. In contrast, as linear models, EN and Ridge provide a foundation for understanding the linear contributions of the descriptors in predicting PCE values. KNN and DT, on the other hand, provide insights into non-linear data distribution and potential interaction effects among features.

The models' hyperparameters have been selected through a grid search, testing every possible configuration to achieve the parameter set that guarantees the best results. The ML model performances have been evaluated in terms of Coefficient of determination (R^2), Root Means Squared Error (RMSE), Mean Absolute Error (MAE), and Standard Deviation (SD).

Quantum Mechanical (QM) methods

All QM calculations have been performed using Gaussian 16, Revision C.01 suite of programs[63]. The geometries of the 61 dye molecules from the training dataset and the molecules from the virtual screening (here referred to as **CV** dyes) have been built replacing alkyl chains with methyl groups to reduce the computational cost. The ground-state (S_0) geometries have been optimized using DFT[64,65] at the B3LYP[66,67]/6-31 G* level of theory in *vacuo*. Vibrational frequency calculations have been performed at the same level of theory to check that the stationary points were true energy minima. Lowest energy excited state (S_1)-optimized geometries have been computed at TD-CAM-B3LYP[68]/6-31 G* level of theory, including solvent effects by using the polarizable continuum model (PCM)[69]. The S_0 energies and the electron density distribution of frontier molecular orbitals (FMOs) were assessed at the B3LYP/6-31 G* level of theory including the solvent effects by PCM. Absorption maxima (λ_{max}^{abs}) vertical excitation energies (E_{exc}), oscillator strength (*f*), and compositions in terms of molecular orbitals for the lowest singlet-singlet excitations ($S_0 \rightarrow S_1$) of the **CV** molecules have been calculated at CAM-B3LYP/6-311 + G(2 d,p) and M06-2X[70]/6-311 + G(2 d,p) levels of theory as these two functionals exhibit the best fitting with the experimental absorptions of the 61 molecules from the training dataset ($\Delta E_{comp-exp} = 0.11$ and $\Delta E_{comp-exp} = 0.10$ eV, respectively). Concerning solvents, the 61 molecules from the training dataset have been characterized by using their experimental solvents. Dichloromethane (DCM) and tetrahydrofuran (THF) were the most employed, thus they have been used to characterize the electronic and UV-Vis properties of the **CV** molecules. Therefore, four independent datasets have been created based on the level of theory employed in the absorption maxima calculations and the relative solvent: **D1**. M06-2X/6-311 + G(2 d,p)-THF; **D2**. M06-2X/6-311 + G(2 d,p)-DCM; **D3**. CAM-B3LYP/6-311 + G(2 d,p)-THF, and **D4**. CAM-B3LYP/6-311 + G(2 d,p)-DCM.

To further confirm the results from Model B application, the redox properties of **CV77, CV85**, and **CV86** have been calculated at the MPW1K[71]/6-31 + G* level of theory in THF, following a computational protocol previously developed in our research group[49].

Data availability

The datasets and the data analyzed during the current study are included in Supplementary Information files. Additional data will be available from the corresponding author on reasonable request.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 24 July 2024; Accepted: 7 January 2025;
Published online: 12 February 2025

References

1. Muñoz-García, A. B. et al. Dye-sensitized solar cells strike back. *Chem. Soc. Rev.* **50**, 12450–12550 (2021).
2. Burnside, S. et al. Deposition and characterization of screen-printed porous multi-layer thick film structures from semiconducting and conducting nanomaterials for use in photovoltaic devices. *J. Mater. Sci.: Mater. Electron.* **11**, 355–362 (2000).
3. Coppola, C., Parisi, M. L. & Sinicropi, A. The role of organic compounds in dye-sensitized and perovskite solar cells. *Energies (Basel)* **16**, 573 (2023).
4. Zheng, H. et al. Emerging organic/hybrid photovoltaic cells for indoor applications: recent advances and perspectives. *Sol. RRL* **5**, 1–17 (2021).
5. Royo, R., Domínguez-Celorrio, A., Franco, S., Andreu, R. & Orduna, J. Pyranylidene/trifluoromethylbenzoic acid-based chromophores for dye-sensitized solar cells. *Dyes Pigments* **206**, 110566 (2022).
6. Siva Gangadhar, P. et al. An investigation into the origin of variations in photovoltaic performance using D-D-π-A and D-A-π-A triphenylimidazole dyes with a copper electrolyte. *Mol. Syst. Des. Eng.* **6**, 779–789 (2021).
7. Devadiga, D., Selvakumar, M., Shetty, P. & Santosh, M. S. Dye-sensitized solar cell for indoor applications: a mini-review. *J. Electron Mater.* **50**, 3187–3206 (2021).
8. D'Amico, F. et al. Recent advances in organic dyes for application in dye-sensitized solar cells under indoor lighting conditions. *Materials* **16**, 7338 (2023).
9. Aftabuzzaman, M., Sarker, S., Lu, C. & Kim, H. K. In-depth understanding of the energy loss and efficiency limit of dye-sensitized solar cells under outdoor and indoor. *J. Mater. Chem. A Mater.* **9**, 24830–24848 (2021).
10. Michaels, H. et al. Dye-sensitized solar cells under ambient light powering machine learning: Towards autonomous smart sensors for the internet of things. *Chem. Sci.* **11**, 2895–2906 (2020).
11. Michaels, H., Benesperi, I. & Freitag, M. Challenges and prospects of ambient hybrid solar cell applications. *Chem. Sci.* **12**, 5002–5015 (2021).
12. Michaels, H. et al. Emerging indoor photovoltaics for self-powered and self-aware IoT towards sustainable energy management. *Chem. Sci.* **14**, 5350–5360 (2023).
13. Aslam, A. et al. Dye-sensitized solar cells (DSSCs) as a potential photovoltaic technology for the self-powered internet of things (IoTs) applications. *Sol. Energy* **207**, 874–892 (2020).
14. Venkateswararao, A., Ho, J. K. W., So, S. K., Liu, S. W. & Wong, K. T. Device characteristics and material developments of indoor photovoltaic devices. *Mater. Sci. Eng. R: Rep.* **139**, 100517 (2020).
15. Pecunia, V., Occhipinti, L. G. & Hoye, R. L. Z. Emerging indoor photovoltaic technologies for sustainable internet of things. *Adv. Energy Mater.* **11**, 2100698 (2021).
16. Prajapat, K. et al. The evolution of organic materials for efficient dye-sensitized solar cells. *J. Photochem. Photobiol. C: Photochem. Rev.* **55**, 100586 (2023).
17. Li, B., Hou, B. & Amaratunga, G. A. J. Indoor photovoltaics, The Next Big Trend in solution-processed solar cells. *InfoMat* **3**, 445–459 (2021).
18. Zhang, D. et al. A molecular photosensitizer achieves a Voc of 1.24 V enabling highly efficient and stable dye-sensitized solar cells with copper(II/I)-based electrolyte. *Nat. Commun.* **12**, 2–11 (2021).
19. Biswas, S. & Kim, H. Solar cells for indoor applications: progress and development. *Polym. (Basel)* **12**, 1338 (2020).
20. Saeed, M. A., Yoo, K., Kang, H. C., Shim, J. W. & Lee, J. J. Recent developments in dye-sensitized photovoltaic cells under ambient illumination. *Dyes Pigments* **194**, 109626 (2021).
21. Chen, C. H. et al. Rational design of cost-effective dyes for high performance dye-sensitized cells in indoor light environments. *Org. Electron* **59**, 69–76 (2018).
22. Mubashir, T. et al. Designing of symmetric and asymmetric small molecule acceptors for organic solar cells: A farmwork based on Machine learning, virtual screening and structural analysis. *J. Photochem. Photobiol. A Chem.* **444**, 114977 (2023).
23. Wen, Y., Fu, L., Li, G., Ma, J. & Ma, H. Accelerated discovery of potential organic dyes for dye-sensitized solar cells by interpretable machine learning models and virtual screening. *Sol. RRL* **4**, 1–11 (2020).
24. Liu, X. et al. Accelerating the discovery of high-performance donor/acceptor pairs in photovoltaic materials via machine learning and density functional theory. *Mater. Des.* **216**, 110561 (2022).
25. Ju, L., Li, M., Tian, L., Xu, P. & Lu, W. Accelerated discovery of high-efficient N-annulated perylene organic sensitizers for solar cells via machine learning and quantum chemistry. *Mater. Today Commun.* **25**, 101604 (2020).
26. Lu, T., Li, M., Yao, Z. & Lu, W. Accelerated discovery of boron-dipyrromethene sensitizer for solar cells by integrating data mining and first principle. *J. Materiomics* **7**, 790–801 (2021).
27. Gao, Z. et al. Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and DFT calculation. *Appl. Surf. Sci.* **568**, 150916 (2021).
28. Saßnick, H. D. & Cocchi, C. Automated analysis of surface facets: the example of cesium telluride. *NPJ Comput. Mater.* **10**, 1–9 (2024).
29. Choudhary, K. et al. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. *Chem. Mater.* **31**, 5900–5908 (2019).
30. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
31. Sahu, H. et al. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J. Mater. Chem. A Mater.* **7**, 17480–17488 (2019).
32. Meftahi, N. et al. Machine learning property prediction for organic photovoltaic devices. *NPJ Comput. Mater.* **166**, 1–8 (2020).
33. Kar, S., Roy, J. K. & Leszczynski, J. In silico designing of power conversion efficient organic lead dyes for solar cells using todays innovative approaches to assure renewable energy for future. *NPJ Comput. Mater.* **22**, 1–11 (2017).
34. Zhang, Y. et al. Accelerating the discovery of N-annulated perylene organic sensitizers via an interpretable machine learning model. *J. Mol. Struct.* **1296**, 136855 (2023).
35. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 1–8 (2009).
36. Jiang, M. L. et al. High-performance organic dyes with electron-deficient quinoxalinoid heterocycles for dye-sensitized solar cells under one sun and indoor light. *ChemSusChem* **12**, 3654–3665 (2019).
37. Li, C. T., Kuo, Y. L., Kumar, C. P., Huang, P. T. & Lin, J. T. Tetraphenylethylene tethered phenothiazine-based double-anchored sensitizers for high performance dye-sensitized solar cells. *J. Mater. Chem. A Mater.* **7**, 23225–23233 (2019).
38. Tsai, M. C. et al. Efficient anthryl dye enhanced by an additional ethynyl bridge for dye-sensitized module with large active area to drive indoor appliances. *ACS Appl Energy Mater.* **3**, 2744–2754 (2020).
39. Tsai, M. C. et al. A large, ultra-black, efficient and cost-effective dye-sensitized solar module approaching 12% overall efficiency under 1000 lux indoor light. *J. Mater. Chem. A Mater.* **6**, 1995–2003 (2018).

40. Desta, M. B. et al. Pyrazine-incorporating panchromatic sensitizers for dye sensitized solar cells under one sun and dim light. *J. Mater. Chem. A Mater.* **6**, 13778–13789 (2018).
41. Tingare, Y. S. et al. New acetylene-bridged 9,10-conjugated anthracene sensitizers: application in outdoor and indoor dye-sensitized solar cells. *Adv. Energy Mater.* **7**, 1700032 (2017).
42. Huang, R. Y., Tsai, W. H., Wen, J. J., Chang, Y. J. & Chow, T. J. Spiro[fluorene-9,9′-phenanthren]-10′-one as auxiliary acceptor of D-A-π-A dyes for dye-sensitized solar cells under one sun and indoor light. *J. Power Sources* **458**, 228063 (2020).
43. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13–17**, 785–794 (2016).
44. Elvidge, C. D., Keith, D. M., Tuttle, B. T. & Baugh, K. E. Spectral identification of lighting type and character. *Sensors* **10**, 3961–3988 (2010).
45. Baleja, R. et al. Comparison of LED properties, compact fluorescent bulbs and bulbs in residential areas. in *Proceedings of the 2015 16th International Scientific Conference on Electric Power Engineering, EPE 2015* 566–571 (2015).
46. Coppola, C. et al. DFT and TDDFT investigation of four triphenylamine/phenothiazine-based molecules as potential novel organic hole transport materials for perovskite solar cells. *Mater. Chem. Phys*. **278**, 125603 (2022).
47. Zhang, G. & Musgrave, C. B. Comparison of DFT methods for molecular orbital eigenvalue calculations. *J. Phys. Chem. A* **111**, 1554–1561 (2007).
48. Boschloo, G. & Hagfeldt, A. Characteristics of the iodide/triiodide redox mediator in dye-sensitized solar cells. *Acc. Chem. Res.* **42**, 1819–1826 (2009).
49. Mohammadpourasl, S. et al. Ground-state redox potentials calculations of D-π-A and D-A-π-A organic dyes for DSSC and visible-light-driven hydrogen production. *Energ. (Basel)* **13**, 1–10 (2020).
50. Feldt, S. M. et al. Design of organic dyes and cobalt polypyridine redox mediators for high-efficiency dye-sensitized solar cells. *J. Am. Chem. Soc.* **132**, 16714–16724 (2010).
51. Yzeiri, X. et al. Synthesis, characterization and application of quinoxaline-based organic dyes as anodic sensitizers in photoelectrochemical cells. *Dyes Pigments* **232**, 112455 (2024).
52. Goti, G. et al. Orange/Red Benzo[1,2-b:4,5-b′]dithiophene 1,1,5,5-tetraoxide-based emitters for luminescent solar concentrators: effect of structures on fluorescence properties and device performances. *Eur. J. Org. Chem.* **2024**, e202400112 (2024).
53. Bartolini, M. et al. Orange/Red Benzo[1,2-b:4,5-b′]dithiophene 1,1,5,5-tetraoxide-based emitters for luminescent solar concentrators: effect of structures on fluorescence properties and device performances. *ACS Appl. Energy Mater.* **6**, 4862–4880 (2023).
54. Castriotta, L. A. et al. Stable methylammonium-Free p-i-n perovskite solar cells and mini-modules with phenothiazine dimers as hole-transporting materials. *Energy Environ. Mater.* **6**, e12455 (2023).
55. Chakraborty, A. et al. Photovoltaics for indoor energy harvesting. *Nano Energy* **128**, 109932 (2024).
56. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
57. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform* **10**, 1–14 (2018).
58. Ho, T. K. Random Decision Forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* 278–282 (1995).
59. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
60. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
61. Cover, T. M. & Hart, P. E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
62. Quinlan, J. R. Induction of decision trees. *Mach. Learn* **1**, 81–106 (1986).
63. Frisch, M. J. et al. Gaussian 16, Revision C.01. Gaussian, Inc., Wallingford CT.
64. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev. B* **7**, 1912–1919 (1973).
65. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
66. Becke, A. D. Density-functional thermochemistry. III. The role of the exact exchange. *J. Chem. Phys.* **98**, 5648 (1993).
67. Lee, C., Yang, W. & Parr, R. G. Development of Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
68. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).
69. Tomasi, J., Mennucci, B. & Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **105**, 2999–3093 (2005).
70. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other function. *Theor. Chem. Acc.* **120**, 215–241 (2008).
71. Lynch, B. J., Fast, P. L., Harris, M. & Truhlar, D. G. Adiabatic connection for kinetics. *J. Phys. Chem. A* **104**, 4813–4815 (2000).

Acknowledgements

We acknowledge the hpc@dbcf for providing computational resources and Regione Toscana for granting the Project INSIEME (Approcci di INtelligenza artificiale, Sintesi Innovative e valutazione di sostenibilità Economico-ambientale per lo sviluppo di nuovi Materiali per la conversione e stoccaggio dell'Energia solare) - Progetti di alta formazione - Fondo Sociale Europeo + 2021-207 (FSE + 2021-2027). This research received no external funding.

Author contributions

C.C. and A.Sinicropi conceived and designed the study. C.C. and A.V. performed calculations, investigation, data analysis, and wrote the original draft. L.Z. and A.Sinicropi took care of funding acquisition. A.Sinicropi supervised the work. M.L.P, A. Santucci, L.Z., O.S. and A. Sinicropi contributed to the writing—review and editing. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01521-9.

Correspondence and requests for materials should be addressed to Adalgisa Sinicropi.