Journal of Molecular Graphics and Modelling 28 (2010) 465–471

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Excited-state properties from ground-state DFT descriptors: A QSPR approach for dyes

Guillaume Fayet [a,b], Denis Jacquemin [c,*], Valérie Wathelet [c], Eric A. Perpète [c], Patricia Rotureau [b], Carlo Adamo [a,*]

[a] Laboratoire d'Electrochimie et Chimie Analytique, CNRS UMR-7575, Ecole Nationale Supérieure de Chimie de Paris, 11 rue P. et M. Curie, F-75231 Paris Cedex 05, France
[b] Institut National de l'Environnement Industriel et des Risques (INERIS), Parc Technologique Alata, BP2, 60550 Verneuil-en-Halatte, France
[c] Unité de Chimie Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix (FUNDP), rue de Bruxelles, 61, B-5000 Namur, Belgium

ARTICLE INFO

Article history:
Received 2 September 2009
Received in revised form 30 October 2009
Accepted 4 November 2009
Available online 13 November 2009

Keywords:
QSPR
DFT
Excited state

ABSTRACT

This work presents a quantitative structure–property relationship (QSPR)-based approach allowing an accurate prediction of the excited-state properties of organic dyes (anthraquinones and azobenzenes) from ground-state molecular descriptors, obtained within the (conceptual) density functional theory (DFT) framework. The *ab initio* computation of the descriptors was achieved at several levels of theory, so that the influence of the basis set size as well as of the modeling of environmental effects could be statistically quantified. It turns out that, for the entire data set, a statistically-robust four-variable multiple linear regression based on PCM-PBE0/6-31G calculations delivers a R^2_{adj} of 0.93 associated to predictive errors allowing for rapid and efficient dye design. All the selected descriptors are independent of the dye's family, an advantage over previously designed QSPR schemes. On top of that, the obtained accuracy is comparable to the one of the today's reference methods while exceeding the one of hardness-based fittings. QSPR relationships specific to both families of dyes have also been built up. This work paves the way towards reliable and computationally affordable color design for organic dyes.

1. Introduction

Organic chromogens play a crucial industrial role since 1880, not only as dyes or pigments, but also in more technological fields, such as thermal transfer systems, molecular switches, media storages or photovoltaic devices [1–3]. Often, industrial applications imply the design of specific dyes possessing given spectroscopic (and sometimes photochromic) properties. In that framework, beside the traditional synthesis tools belonging to the chemist cultural background, simulation approaches can provide an efficient and complementary approach for the screening of new molecules. Of course, such theoretical approaches should rely upon the development of effective tools for the prediction of the color of dyes, a great challenge in the field of theoretical chemistry. Indeed, modeling the color generated by photon absorption requires the description of low-lying excited-state(s), a demanding task. If physico-chemical properties related to the electronic distribution (such as light absorption and emission), can be accurately evaluated by quantum chemical models, the bottleneck remains the balance between accuracy and computing times. On the one hand, post-Hartree-Fock methods (EOM-CC, MR-CI or CAS-PT2) are accurate but time consuming and present a problematic scaling with the system dimension; on the other hand, semi-empirical methods (CNDO/S, INDO/S) are applicable for large systems but with a significant loss in accuracy [4,5]. In the last years, the time-dependent density functional theory (TD-DFT) [6–8] has emerged as a popular scheme able to deliver remarkable accuracies with "reasonable" computational times for both organic and inorganic species [9–16]. Nevertheless, the TD-DFT methods, even in their most modern implementations, scale as $O(N^3)$, N being the number of basis functions, the pre-factor being large. As a matter of fact, TD-DFT studies are limited to medium-size molecules (100–150 atoms, 1000–3000 basis functions), especially when atomic basis sets including diffuse/polarized orbitals are mandatory.

It is therefore interesting, to search for alternative theoretical protocols coupling speed and accuracy, especially by estimating electronic spectra from directly-available ground-state properties. In that sense, the quantitative structure activity/properties relationship (QSAR/QSPR) methods appear as ideal candidates. Nowadays such approaches are mainly used in the toxic property screening (i.e. the nitrobenzene molecule [17]) and their primary applications mainly encompass biology [18,19], toxicology [20,21] and drug design [22–24]. However, a growing number of applications have recently appeared for the prediction of physico-chemical properties [25–27]. Apart from the usual drawbacks of numerical

* Corresponding authors.
E-mail addresses: denis.jacquemin@fundp.ac.be (D. Jacquemin), carlo-adamo@enscp.fr (C. Adamo).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.11.001

methods (neural networks [28], genetic algorithms [29] and statistical regressions [30]) used for obtaining the relationship, the main limit for building up such chemical QSPR models remains the reliability of the experimental training data set.

In this paper we propose the development of a QSPR protocol for the prediction of the main π–π^* transition of two families of industrial organic dyes: 9,10-anthraquinones (AQ) and azobenzenes (AB, see Fig. 1). Together they represent about 90% of today's world dye production [2,31]. We are aware of only a few previous works using information theory for the spectral properties of these dyes. The first two, by three of us [32,33], aims at an optimal combination of TD-DFT results obtained with different functionals, in order to reproduce the absorption wavelengths of AQ dyes, a computationally successful but demanding approach. The third, by Åstrand and coworkers [34] relates the nitrogen double bond lengths (and critical points of the electron density) to the transition energies of a large set of AB compounds. While such approach delivers accurate transition energies, it remains difficult to generalize as a family-specific geometrical parameter appears necessary to obtain valuable predictions. We have found several QSPR works carried out for dyes but these studies focus on ground-state related properties (thermal stability [35], affinity with fibres [36], adsorption [37] or acidity [38]), and excludes, to the best of our knowledge, properties related to the excited-states, such as the color. Here, the developed protocol starts from DFT calculations of the ground electronic states and includes general molecular descriptors belonging to the so-called family of "conceptual" DFT [39,40]. Correlations between these data and the experimental ones have been brought to light to obtain a predictive model for the most important excited-state property. Our results are compared to these of TD-DFT calculations.

2. Method

2.1. Data set

The choice of the training set of experimental data, a critical point in any QSPR analysis, is difficult here, as experimental conditions (solvent) might significantly influence the measured spectral properties, i.e. the wavelength of maximal absorption (λ_{max}). Therefore, to allow straightforward and consistent comparisons with previous TD-DFT benchmarks [41], we have chosen 24 anthraquinones and 22 azobenzenes. For each series, a large panel of substituents has been included, so that the experimental λ_{max} almost cover the full width of the visible spectrum. Note that, for several cases, experimental wavelengths differing by a few nanometers have been reported [42–48] and the reference data reported in Tables 2 and 3 correspond to the average value.

2.2. Computational details

The molecular structures and properties of all molecules have been modeled with the Gaussian03 package [49] by using a DFT approach relying on the parameter-free PBE0 hybrid functional [50]. The geometrical parameters of all molecules have been taken in Ref. [41], that is structural optimizations have been performed with the 6-311G(d,p) basis sets. The nature of the stationary points has been checked previously by showing the absence of imaginary frequency in these structures. Calculations of the molecular descriptors have been carried out using a rather small basis set, namely 6-31G, and a much larger one, 6-311+G(2d,p), that is required to obtain converged λ_{max} within the TD-DFT framework [5,10,41]. The determination of the descriptors has been carried out both in gas phase and in solution. In this latter case, the surrounding effects have been included by means of the well-recognized polarizable continuum model (IEF-PCM) [51], in which the solute part (the dye) is lying inside a cavity, and the solvent is represented as a structureless material, characterized by its macroscopic properties.

2.3. Molecular descriptors

To characterize the structures of our 46 compounds, eight descriptors have been extracted from the quantum chemical calculations. Five of them issue from the conceptual DFT [39,40] namely ionization potential (*IP*), electron affinity (*EA*), electronegativity (χ), hardness (η) and electrophilicity index (ω). Electronegativity characterizes the electron donor/acceptor behavior of the system and has been defined as [52]:

$$\chi = \frac{IP + EA}{2} \tag{1}$$

Therefore, χ is equal, but opposite in sign, to the chemical potential (μ). The definition of the hardness (η) was given by Parr and Pearson [53,54] and a three-point finite difference approximation leads to the following working definition:

$$\eta = IP - EA \tag{2}$$

These descriptors have already been used in QSPR models for different properties [55]. Starting from these two quantities, the electrophilicity index was defined by Parr et al. [56] as:

$$\omega = \frac{\mu^2}{2\eta} \tag{3}$$

and has already been used in the framework of a QSAR approach for the prediction of biological activity [57]. Following a standard procedure based on the Koopmans' theorem, the ionization potential (*IP*) and the electron affinity (*EA*) have been computed from the energies of the highest occupied and the lowest unoccupied molecular orbitals, respectively (ε_{HOMO} and ε_{LUMO}). Therefore, Eq. (2) corresponds directly to the HOMO–LUMO gap and we expect a strong correlation with the absorption wavelengths. The electronic structure has also been described by the norm of the dipole moment (*DM*), which yields the electronic asymmetry and the mean (α) and anisotropic polarizabilities ($\Delta\alpha$). These three latter terms are, respectively, given by:

$$DM = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2} \tag{4}$$

$$\alpha = \frac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}) \tag{5}$$

$$\Delta\alpha = \left(\frac{1}{2}\right)^{1/2} [(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{xx} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{yy})^2 + 6(\alpha_{xy}^2 + \alpha_{xz}^2 + \alpha_{yz}^2)]^{1/2} \tag{6}$$

where μ_i and α_{ii} are the components of the dipole and polarizability matrix, respectively. It should be stressed that the five conceptual DFT parameters, as well as the *DM* are obtained straightforwardly from single-point calculations, while α and $\Delta\alpha$ require a coupled-perturbed process that is computationally more demanding.

2.4. Development and evaluation of models

The statistical models were developed and evaluated by means, on the one hand, of the best multi-linear regression (BMLR) analysis as integrated in Codessa software [58], and, on the other hand, the MLR approach implemented in Statgraphics [59]. In

Codessa, the set of descriptors is cleaned from insignificant descriptors ($R^2 < 0.1$), then the best 2 parameter-regressions is constructed upon the statistical significance and non-colinearity criteria ($R^2 < 0.6$) of the selected descriptors. The final model with the "best" representation of the property in the given descriptor set is obtained by adding descriptors to the original 2 terms equation. On the contrary, in Statgraphics, the treatment is to allow for step-by-step elimination the less significant independent variables, on the basis of the statistical reliability of each parameter. Results of both approaches have been used and we propose here the most reliable model in each case.

The reliability of the models can be measured through several parameters, such as the resulting R^2 and the corrected R^2_{cv} (cross-validated value [60]) or R^2_{adj} (adjusted to the number of variables used [30]), but also the obtained mean absolute deviations (MAE) and standard deviation (d_R), the latter yielding the predictive power of the obtained MLR [30,32].

3. Results and discussion

The QSPR methodology was applied to predict the wavelength of maximum absorption (λ_{max}). The eight descriptors have been calculated for the 24 AQ with the 6-311+G(2d,p) and 6-31G basis sets with and without ethanol solvent (see Supporting Information, Tables S1–S3). The corresponding information for the AB set are located at Table S4–S6. Note that we have not carried out gas-phase calculations with the most extended atomic basis set as including the solvent effects does not provoke a significant loss of *cpu* time when large basis sets are used [51].

3.1. Linear correlations

Initially, we have determined the linear correlation coefficients, R^2, between each descriptor and the experimental λ_{max} in order to estimate the influence of the basis set and the usefulness of solvent effects. The results are reported in Table 1. It is obvious that the modification of the basis set has only a marginal influence on the R^2 for the full set. Indeed, typical variations are limited to 0.02 and do not systematically represent an improvement. Such statement also holds for the AB set, while the correlations of AQ seem more basis set dependent, especially for the average and anisotropic polarizabilities. On the contrary, including environmental effects during the computation of the descriptors offers a significant gain in accuracy. This is striking in the full set for which an improvement in R^2 is found for all descriptors, but η. A similar behavior is found for both AQ and AB, though in the former case, one notable exception exists (α). Therefore, we advocate that a PCM-PBE0/6-31G model is sufficient in order to determine the molecular descriptors of dyes. In that sense QSPR approach already represents a net efficiency gain over TD-DFT for which larger basis sets are mandatory to obtain converged transition energies [5,10,16,41]. To illustrate this point, the computational effort required to compute the eight PCM-PBE0/6-31G descriptors of the full AQ set is more than one order of magnitude smaller than for the corresponding PCM-TD-DFT/6-311+G(2d,p) calculations. By discarding the bulk solvent effects from the calculation (PBE0/6-31G), an additional speed-up factor of three is attained.

Amongst all parameters, the chemical hardness, η, systematically provides the largest R^2 close to 0.8-0.9. This is not a surprise as it is equivalent to the gap between frontier orbitals in the conceptual DFT framework. As the majority (but not all) of the λ_{max} reported in this contribution formally corresponds to the promotion of one electron from the HOMO to the LUMO, a strong correlation was foreseeable. Such correlation is still not yet perfect in part due to the lack of self-interaction corrections in our DFT model. Extra corrections could be included, for instance through the ADSIC scheme [61], but at the expense of a much larger effort. Therefore, one can consider the MLR mentioned below as inexpensive ways to correct the crude wavelength estimates given by the chemical hardness. Indeed, using a simple linear correction on the PCM-PBE0/6-31G, η provides a MAE of 28.2 nm and a standard deviation of 34.7 nm, that is, accuracies possibly sufficient for a first screening but not sufficient for dyes design. Interestingly, several other parameters nicely correlate with the λ_{max}, especially if the two series of dyes are considered separately. For instance, the anisotropic polarizability presents a R^2 of 0.75 for AQ and 0.85 for AB (both with the PCM-PBE0/6-31G). This is also quite unsurprising as the polarizabilty depends upon of the differences between the energies of occupied and virtual orbitals in the sum-over-states approximation.

For the records, we have also tested the adequacy of simple structural descriptors. We have chosen the central diazo bond length for AB and the average distance of carbonyl bond for AQ, as these units are widely considered as chromophoric centers. The R^2 obtained with these descriptors are 0.77 and 0.67, for AB and AQ, respectively. Obviously such correlation coefficients, while sizable, remain smaller than the one obtained with the hardness in Table 1. We have chosen not to use these geometrical descriptors in the following, as our goal is to design a procedure able to correctly reproduce the electronic properties using generic descriptors applicable to a large panel of compounds. Indeed, while the N=N distance is an efficient descriptor for AB dyes such input could obviously not be obtained for AQ derivatives.

3.2. Multi-linear regressions for the entire data set

As our main focus is to develop a general model, we have started with simulation on the entire data set, with the selected PCM-

Table 1
Correlation coefficients (R^2) of computed descriptors with the experimental λ_{max} of the dye sets considered in this study.

	Full set			AQ set			AB set		
	6-31G		6-311+G(2d,p)	6-31G		6-311+G(2d,p)	6-31G		6-311+G(2d,p)
	Gas	Solvent	Solvent	Gas	Solvent	Solvent	Gas	Solvent	Solvent
ε_{HOMO}	0.18	0.30	0.31	0.79	0.84	0.88	0.32	0.63	0.64
ε_{LUMO}	0.08	0.10	0.09	<0.01	0.01	0.04	0.05	0.11	0.11
DM	0.22	0.23	0.23	0.17	0.24	0.28	0.74	0.76	0.77
η	0.83	0.79	0.77	0.91	0.92	0.93	0.84	0.90	0.91
χ	0.01	0.05	0.06	0.45	0.63	0.73	0.04	0.14	0.16
ω	0.14	0.20	0.22	0.04	0.05	0.08	0.09	0.23	0.25
α	0.15	0.23	0.20	0.52	0.30	0.69	0.79	0.89	0.86
$\Delta\alpha$	0.04	0.10	0.10	0.76	0.75	0.84	0.79	0.85	0.84

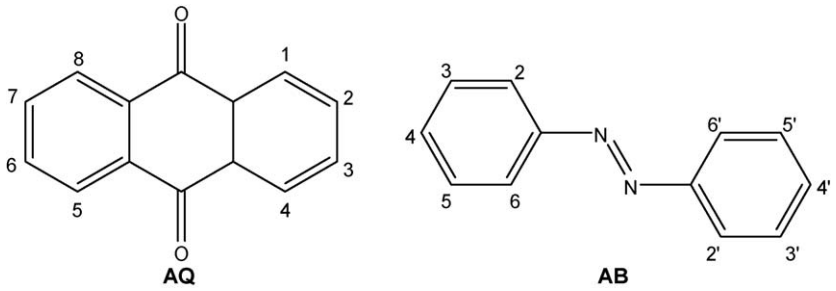

Fig. 1. Sketch of the chromophores investigated in this study: 9,10-anthraquinones (AQ) and azobenzenes (AB).

PBE0/6-31G descriptors. On this basis, we obtain the following MLR equation

$$\lambda_{max} = 947.37 + 0.76\alpha - 0.56\Delta\alpha - 6490.74\chi + 5266.55\omega \quad (7)$$

with all parameters significant at a 99% confidence level. Eq. (7) delivers a R^2_{adj} of 0.93, a MAE limited to 14.5 nm and a d_R of 20.4 nm, much better than any of the R^2 of Table 1. Such standard error is small, as it means that the longest wavelength of maximal absorption of (similar) dyes not included in our training set would be semi-quantitatively predicted (±20 nm). This is illustrated in Fig. 2a. We underline that the four descriptors included in Eq. (7) can be straightforwardly computed, without any data specific to a particular structure, which is an obvious advantage of the proposed approach. Indeed, typically selected variables (bond lengths, number of hydrogen bonds, etc.) in the field of dye design are highly family-dependent [34,62]. It is also worth to point out that Eq. (7) provides a much better accuracy than the hardness taken alone (R^2 of 0.79, see Table 1). In fact, imposing the hardness in a MLR cannot outperform Eq. (7) though it slightly improves the hardness-alone equation. For instance,

$$\lambda_{max} = 1009.25 - 5.46DM - 4251.60\eta \quad (8)$$

yields a R^2_{adj} of 0.82. The associated MAE and d_R are 24.2 nm and 32.2 nm, respectively, about 50% larger than the one of Eq. (7), though Eq. (8) favorably avoids the calculation of the molecular polarizability.

If the calculations with the large basis set are at hand, one can select the very similar,

$$\lambda_{max} = 947.32 + 0.54\alpha - 0.49\Delta\alpha - 6898.91\chi + 5568.94\omega \quad (9)$$

though the resulting statistical parameters are only slightly improved compared to the 6-31G calculations: R^2_{adj} of 0.94, a MAE limited to 13.5 nm and a d_R of 19.1 nm. This confirms that, for the purpose of obtaining conceptual DFT descriptors, using an extended basis set, such as 6-311+G(2d,p), is not worth the computational effort. If calculations with the solvent (here PCM) model are unavailable (which is unfortunately the case in several computational chemistry codes), we recommend the following five-parameter PBE0/6-31G model,

$$\lambda_{max} = 908.38 + 1.03\alpha - 0.83\Delta\alpha + 7.32DM - 5974.99\chi + 4728.30\omega \quad (10)$$

with all parameters significant at a 99% confidence level. Such correlation presents a R^2 of 0.90 and a R^2_{adj} of 0.88, significantly better than the one obtained with the η taken alone that yields $R^2 = R^2_{adj}$ of 0.83 (see Table 1). Eq. (10) provides a MAE of 18.3 nm and a d_R of 26.0 nm. Such deviations might look a bit unsatisfying compared to the one obtained in PCM. However, the state-of-the-art PCM-TD-PBE0/6-311+G(2d,p) yields a MAE of 20.5 nm on the same set, that is provides a similar accuracy (without need of any statistical help), but for a computational effort about 30 times larger. To decrease this effort, one might rely on the TD-DFT/6-31G values reported in Tables 2 and 3. These data provide a MAE of 32.8 nm, significantly larger than with the converged TD-DFT approach. By performing a SLR on the TD-DFT/6-31G results, one decreases the MAE to a 27.8 nm, value, for a computational cost that remains about twice larger than the one used to compute all descriptors of Eq. (10). In that sense, it seems that the MLR equations might indeed be quite powerful and efficient schemes for computing the absorption wavelengths.

In the following, we investigate the two families of dyes separately, as building up specific approaches for these two series of dyes might be very useful for practical applications. We expect to obtain larger correlation coefficients for most descriptors, once only a single family of chemical molecule is considered, as the similarity of the compounds often helps in improving the statistical parameters.

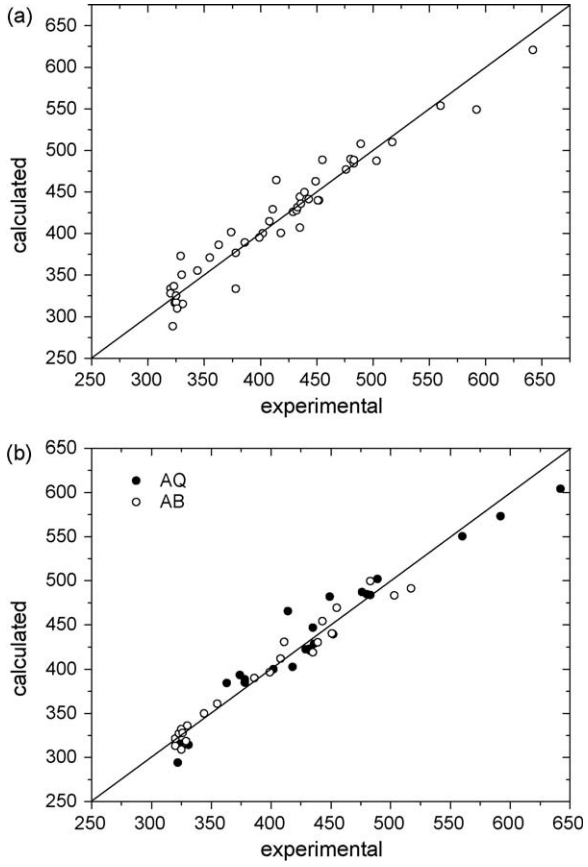

Fig. 2. Plot of the λ_{max} (nm) calculated from QSPR models (Eq. (7) in (a), Eq. (13) (AQ) and 17 (AB) in (b)) and the experimental values of 46 anthraquinone (AQ) and azobenzene (AB) molecules.

Table 2
λ_{max} (nm) of 9,10-anthraquinones (AQ) obtained from QSPR models.

Substituent	Eq. (7)	Eq. (13)	TD-DFT		Exp	Ref.
			Gas/6-31G [a]	Ref [41][b]		
None	288	305	302	321	322	[41]
2-Me	317	318	305	324	324	[41]
1-Me	315	320	319	339	331	[41]
2-OMe	386	387	367	391	363	[41]
1,2-OMe	402	405	362	386	374	[41]
1-OMe	333	382	361	387	378	[41]
2-OH	377	365	358	389	378	[41]
1-OH	400	399	386	398	402	[41]
1,2,3-OH	464	467	373	395	414	[41]
1,3-OH	401	402	388	408	418	[41,43]
1,8-OH	426	420	405	419	429	[41]
1,5-OH	428	420	404	417	432	[41]
1,5-OH-3-Me	432	429	403	415	433	[44]
1,2-OH	444	445	414	426	435	[41,43]
1,3,8-OH-6-Me	436	441	413	433	436	[41]
2-NH_2	463	455	405	459	449	[45]
1,3,6,8-OH	440	441	426	451	452	[41]
1-NH_2	477	469	426	463	476	[45]
1,4-OH	490	482	442	456	480	[41]
1,2,4-OH	484	486	441	456	483	[41]
1,4,5-OH	508	499	453	469	489	[41]
1,4,5,8-OH	554	550	485	504	560	[41]
1,4-NH_2	549	577	486	522	592	[43]
1,4-NHEt	621	626	529	568	642	[46]

[a] Using the geometry of Ref. [41].
[b] IEF-PCM-TD-PBE0/6-311+G(2d,p)//IEF-PCM-PBE0/6-311G(d,p) approach.

3.3. 9,10-Anthraquinones

The 8 descriptors, calculated for the 24 AQ set, are listed in Supporting Information (see Tables S1–S3) and the correlation efficiency of the hardness is again impressive ($R^2 \sim 0.92$), while the electrophilicity index correlates especially poorly with the absorption wavelengths (see Table 1 and Section 3.1). Using our method of choice, PCM-PBE0/6-31G, and performing a simple linear regression on the hardness yields:

$$\lambda_{max} = 986.10 - 4052.80\eta \quad (11)$$

that already allows accurate evaluations with a MAE of 16.5 nm and a standard deviation of 22.2 nm, only. Using MLR, one can slightly improve Eq. (11),

$$\lambda_{max} = 716.85 + 0.71\alpha - 3281.61\eta \quad (12)$$

Eq. (12) provides a MAE of 14.2 nm and a standard deviation of 18.8 nm. Likewise, the equations

$$\lambda_{max} = 1076.54 - 6540.48\chi + 4428.56\omega \quad (13)$$

$$\lambda_{max} = 837.4 + 3784\omega + 0.57\alpha - 5513\chi \quad (14)$$

similarly provide MAE limited to 15.3 nm and 12.2 nm, respectively. Due to its simplicity (no polarizability) we cope for model (13) here (see Fig. 2b). As for the full set, these values are in the line of the one obtained by straightforward (but computationally

Table 3
λ_{max} (nm) of azobenzenes (AB) obtained from QSPR models.

Substituent	Eq. (7)	Eq. (16)	TD-DFT		Exp	Ref.
			Gas/6-31G[a]	Ref [41][b]		
None	334	312	310	342	320	[47]
4-F	328	320	314	345	320	[47]
4-Me	337	330	317	350	323	[47]
4-Br	317	330	323	355	325	[47]
4,4′-F	325	299	318	347	325	[47]
4,4′-Br	310	316	336	367	326	[47]
4-Phenylazomaleinanil	373	350	325	357	329	[42]
4,4′-Me	350	335	323	357	330	[47]
4-OMe	355	353	331	366	344	[43,47]
4,4′-OMe	371	364	344	380	355	[43,47]
4-NH_2	389	389	350	399	386	[42]
4,4′-NH_2	395	392	366	422	399	[43]
4-NMe_2	415	424	368	418	408	[42]
4-NHPh	429	425	382	438	411	[42]
6′-OBu-2,6-NH_2-3,3′-azopdipyridine	407	421	369	401	435	[42]
2′-NH_2-azobenzenenaphtalene	449	429	398	451	439	[42]
4-NO_2-4′-NH_2	441	454	403	483	443	[42]
2,4-NH_2-azobenzenenphtalene	440	442	385	441	451	[42]
4-NO_2-4′-N(Et)(CH_2CH_2CN)	489	471	417	492	455	[42]
4-NO_2-4′-NHPh	488	496	440	527	483	[42]
4-NO_2-4′-N(Et)(CH_2CH_2OH)	487	485	429	513	503	[42]
4-NO_2-2-Cl-4′-N(Et)(CH_2CH_2OH)	510	490	439	525	517	[42]

[a] Using the geometry of Ref. [41].
[b] IEF-PCM-TD-PBE0/6-311+G(2d,p)//IEF-PCM-PBE0/6-311G(d,p) approach.

intensive) TD-DFT calculation (R^2 = 0.96, MAE of 18.8 nm [41]). On the basis of the PCM-PBE0/6-311+G(2d,p) descriptors, one can obtain equations similar to Eqs. (12)–(14) but they do not significantly improve nor worsen the above-mentioned values. In short, for anthraquinones, the hardness taken alone already provides a semi-quantitative evaluation of the λ_{max}, that can be slightly improved, at a zero computational cost, by the selection of the two-variables MLR (13).

3.4. Azobenzenes

A data set of 22 azobenzene molecules was also investigated with the same levels of calculations (see descriptors values in Supporting Information). As discussed in Section 3.1, linear correlation coefficients with the experimental λ_{max} in Table 1 are in the line of previous observations, with the hardness being the most correlated descriptor independently of the selected theoretical model. For AB, using the three same descriptors (average polarizability, dipole moment and electronegativity):

$$\lambda_{max} = 511.84 + 0.30\alpha + 5.78DM - 1645.20\chi \quad (15)$$

$$\lambda_{max} = 440.23 + 0.43\alpha + 4.84DM - 1337.66\chi \quad (16)$$

$$\lambda_{max} = 422.45 + 0.61\alpha + 9.69DM - 1332.93\chi \quad (17)$$

on the basis of PCM-PBE0/6-311+G(2d,p), PCM-PBE0/6-31G and PBE0/6-31G calculations, respectively. In each case, each term is statistically significant at the 99% confidence level. These equations (respectively) provide R^2 of 0.97, 0.97, 0.95, R^2_{adj} of 0.96, 0.96, 0.94, MAE of 10.3 nm, 10.3 nm, 12.5 nm and d_R of 13.1 nm, 13.2 nm, 15.5 nm, illustrating again that including environmental effects has more impact than increasing the basis set size. The efficiency of (16) is demonstrated in Fig. 2b. For the PCM-PBE0/6-31G calculation we have also determined a simple polarizability-free MLR,

$$\lambda_{max} = 877.82 - 5336.16\chi + 3661.73\omega \quad (18)$$

that is easier to use than (16), but at the expense of a decreased accuracy (R^2_{adj} of 0.93 and d_R of 17.4 nm). For the 6-311+G(2d,p) calculation, Codessa also proposes an alternative to (15) that does not include the polarizability

$$\lambda_{max} = 893.2 + 10401\varepsilon_{HOMO} + 7291\omega + 8505\eta \quad (19)$$

though proposing a similar accuracy (R^2 = 0.95 and $R^2_{cv} = 0.92$). Obviously, all models designed for AB present encouraging correlations, particularly when comparing to the correlation between the TD-DFT calculated and experimental λ_{max} (R^2 = 0.93) [41]. If the sets of descriptors in the three models are not identical, they are close to the one used for the anthraquinone data set, χ and ω being present in both cases.

4. Conclusions

Using a QSPR methodology relying on ground-state "conceptual DFT" descriptors, models have been designed to predict an essential excited-state property: the wavelength of maximal absorption of organic dyes. Our training set contained 24 anthraquinones and 22 azobenzenes, solvated in polar media, as these two families present the largest industrial interest. First the influence of the theoretical level used to compute the descriptors was tested and it turned out that the selection of a large basis set is useless, while the explicit consideration of bulk solvatation effects improves the correlation between the computed descriptors and the measured values. In other words, the computationally-light PCM-PBE0/6-31G scheme is sufficient to compute our descriptors. On this basis, a generic statistically-meaningful model, Eq. (7), has been set up. It yields a R^2_{adj} of 0.93, a mean absolute deviation of 14.5 nm, and a standard deviation limited to 20.4 nm. In fact, such accuracy outperforms both a simple hardness-based correction and the best theoretical tool (TD-DFT) available for modeling the transition energies of medium-sized compounds. Let us highlight that, contrary to most QSPR approaches designed previously for estimating absorption wavelength, no molecule-specific descriptor (bond length of a particular chromogens, presence of donor groups, number and position of the auxochroms, etc.) has been used. In that way, the variables included in our model are easy to determine for all dyes: no human "chemical analysis" is required.

Azobenzene and anthraquinone sets have also been considered separately, and family-specific regressions have been obtained, though the accuracy improvement compared to the full set remains rather limited, especially for the quinones.

We believe that this works opens new opportunities for reliable and quick prediction of the color of molecules. For sure, these results have to be confirmed for other classes of dyes (naphtoquinones, indigoids, etc.) but are certainly very encouraging.

Acknowledgements

DJ and EAP thank the Belgian National Fund for their research associate and senior research associate positions, respectively. DJ, EAP, and CA thank the Commissariat Général aux Relations Internationale and the Egide agency for supporting this work within the framework of the Tournesol Scientific cooperation between France and the Communauté Française de Belgique. DFT calculations have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the "Loterie Nationale" for the convention number 2.4578.02 and of the FUNDP.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.11.001.

References

[1] A. Natansohn, P. Rochon, Chem. Rev. 102 (2002) 4139.
[2] H. Zollinger, Color Chemistry, Syntheses Properties and Applications of Organic Dyes and Pigments, 3rd ed., Wiley–VCH, Weinheim, 2003.
[3] V. Balzani, A. Credi, M. Venturi, Molecular Devices and Machines, 1st ed., Wiley–VCH, Weinheim, 2004.
[4] J. Fabian, Theor. Chem. Acc. 106 (2001) 199.
[5] D. Jacquemin, E.A. Perpete, Chem. Phys. Lett. 429 (2006) 147.
[6] J.P. Perdew, A. Ruzsinsky, J. Tao, V.N. Staroverov, G.E. Scuseria, G.I. Csonka, J. Chem. Phys. 123 (2005) 062001.
[7] E. Runge, E.K.U. Gross, Phys. Rev. Lett. 52 (1984) 997.
[8] K. Burke, J. Werschnik, E.K.U. Gross, J. Chem. Phys. 123 (2005) 062206.
[9] D. Jacquemin, J. Preat, V. Wathelet, J.M. Andre, E.A. Perpete, Chem. Phys. Lett. 405 (2005) 429.
[10] D. Jacquemin, J. Preat, V. Wathelet, M. Fontaine, E.A. Perpète, J. Am. Chem. Soc. 128 (2006) 2072.
[11] L. Petit, A. Quartarolo, C. Adamo, N. Russo, J. Phys. Chem. B 110 (2006) 2398.
[12] A.D. Quartarolo, N. Russo, E. Sicilia, Chem. Eur. J. 12 (2006) 6797.
[13] L. Petit, C. Adamo, N. Russo, J. Phys. Chem. B 109 (2005) 12214.
[14] R. Improta, V. Barone, F. Santoro, Angew. Chem. Int. Ed. Engl. 46 (2007) 405.
[15] M. Dierksen, S. Grimme, J. Phys. Chem. A 108 (2004) 10225.
[16] M.J.G. Peach, P. Beneld, T. Helgaker, D.J. Tozer, J. Chem. Phys. 128 (2008) 044118.
[17] V.K. Agrawal, P.V. Khadikar, Bioorg. Med. Chem. 9 (2001) 3035.
[18] H. Gao, J.A. Katzenellenbogen, R. Garg, C. Hansch, Chem. Rev. 99 (1999) 723.
[19] D.A. Winkler, Brief Bioinform. 3 (2002) 73.
[20] C.D. Selassie, R. Garg, S. Kapur, A. Kurup, R.P. Verma, S.B. Mekapati, C. Hansch, Chem. Rev. 102 (2002) 2585.
[21] S.P. Bradbury, Toxicol. Lett. 79 (1995) 229.
[22] R. Garg, S.P. Gupta, H. Gao, M.S. Babu, A.K. Debnath, C. Hansch, Chem. Rev. 99 (1999) 3525.

[23] M. Grover, B. Singh, M. Bakshi, S. Singh, Pharm. Sci. Technol. Today 3 (2000) 28.
[24] M. Grover, B. Singh, M. Bakshi, S. Singh, Pharm. Sci. Technol. Today 3 (2000) 50.
[25] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279.
[26] A.R. Katritzky, V.S. Lobanov, M. Karelson, Pure Appl. Chem. 69 (1997) 245.
[27] J. Taskinen, J. Yliruusi, Adv. Drug Deliv. Rev. 55 (2003) 1163.
[28] J. Gasteiger, J. Zupan, Angew. Chem. Int. Ed. Engl. 32 (1993) 503.
[29] R. Leardi, J. Chemometr. 15 (2001) 559.
[30] P. Dagnelie, Statistique théorique et appliquée, Tomes 1 & 2, De Boeck and Larcier, Bruxelles and Paris, 1998.
[31] J. Griffiths, Colour and Constitution of Organic Molecules, Academic Press, London, 1976.
[32] E.A. Perpète, V. Wathelet, J. Preat, C. Lambert, D. Jacquemin, J. Chem. Theory Comput. 2 (2006) 434.
[33] D. Jacquemin, V. Wathelet, J. Preat, E.A. Perpète, Spectrochim. Acta A 67 (2007) 334.
[34] B. Buttingsrud, B.K. Alsberg, P.O. Åstrand, Phys. Chem. Chem. Phys. 9 (2007) 2226.
[35] J. Xu, B. Guo, B. Chen, Q. Zhang, J. Mol. Model. 12 (2005) 65.
[36] S. Timofei, W. Schmidt, L. Kurunczi, Z. Simon, Dyes Pigments 47 (2000) 5.
[37] H. Metivier-Pignon, C. Faur, P. Le Cloirec, Chemosphere 66 (2006) 887.
[38] S. Shamsipur, B. Hemmateenejad, M. Akhond, H. Sharghi, Talanta 54 (2001) 1113.
[39] H. Chermette, J. Comput. Chem. 20 (1999) 129.
[40] P. Geerlings, F. De Proft, W. Langenaeker, Chem. Rev. 103 (2003) 1793.
[41] D. Jacquemin, E.A. Perpète, G.E. Scuseria, I. Ciofini, C. Adamo, J. Chem. Theory Comput. 4 (2008) 123.
[42] F.J. Green, The Sigma–Aldrich Handbook of Stains Dyes and Indicators, Aldrich Chemical Company Inc., Milwaukee, WI, 1990.
[43] R.H. Thomson, Naturally Occurring Quinones, 2nd ed., Academic Press, London, U.K., 1971.
[44] R.C. Weast, Handbook of Chemistry and Physics, 51st ed., The Chemical Rubber Company, Cleveland, OH, 1970.
[45] K. Günaydin, G. Topcu, R.M. Ion, Nat. Prod. Lett. 16 (2002) 65.
[46] J. Fabian, M. Nepras, Collect. Czech. Chem. Commun. 45 (1980) 2605.
[47] R.M. Christie, Colour Chemistry, The Royal Society of Chemistry, Cambridge, U.K., 1991.
[48] P.H. Gore, O.H. Wheeler, J. Org. Chem. 26 (1961) 3295.
[49] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, 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, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03 Revision C.02, Gaussian Inc., Wallingford, CT, 2004.
[50] C. Adamo, V. Barone, J. Chem. Phys. 110 (1999) 6158.
[51] J. Tomasi, B. Mennucci, R. Cammi, Chem. Rev. 105 (2005) 2999.
[52] R.S. Mulliken, J. Chem. Phys. 2 (1934) 782.
[53] R.G. Parr, R.G. Pearson, J. Am. Chem. Soc. 105 (1983) 7512.
[54] R.G. Pearson, Chemical Hardness, Wiley–VCH, Weinheim, 1997.
[55] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, Chem. Phys. Lett. 323 (2000) 59.
[56] R.G. Parr, L. Szentpaly, S. Liu, J. Am. Chem. Soc. 121 (1999) 1922.
[57] J. Padmanabhan, R. Parthasarathi, V. Subramanian, P.K. Chattaraj, Bioorg. Med. Chem. 14 (2006) 1021.
[58] CodessaPro, http://www.codessa-pro.com.
[59] Statgraphics Plus 5.1, Manugistics Inc., Herndon, VA, U.S.A., 2000.
[60] M. Karelson, Molecular Descriptors in QSAR/QSPR, John Wiley & Sons, New York, 2000.
[61] I. Ciofini, C. Adamo, H. Chermette, Chem. Phys. 309 (2005) 67.
[62] P.C. Chen, Y.C. Chieh, J.C. Wu, J. Mol. Struct. (Theochem) 715 (2005) 183.