

Modelling heterocyclic azo dye affinities for cellulose fibres by computational approaches

Simona Funar-Timofei^{a,*}, Walter M.F. Fabian^b, Ludovic Kurunczi^c, Mohammad Goodarzi^d, Syed Tahir Ali^e, Yvan Vander Heyden^d

^a Institute of Chemistry of the Romanian Academy, Bul. Mihai Viteazu 24, 300223 Timisoara, Romania

^b Institut für Chemie, Karl-Franzens Universität Graz, Heinrichstr. 28, A-8010 Graz, Austria

^c "Victor Babes" University of Medicine and Pharmacy Timisoara, Faculty of Pharmacy, P-ta E. Murgu 2-4, 300034 Timisoara, Romania

^d Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

^e Department of Chemistry, Federal Urdu University of Arts, Science & Technology, University Road, Block 9, Gulshan-e-Iqbal, Karachi, Pakistan

ARTICLE INFO

Article history:

Received 2 August 2011

Received in revised form

16 January 2012

Accepted 17 January 2012

Available online 25 January 2012

Keywords:

Comparative molecular field analysis

(CoMFA)

Comparative molecular similarity index

analysis (CoMSIA)

Dye

Affinity

Cellulose

Density functional theory (DFT)

ABSTRACT

Textile dyeing has economical and ecological implications. Our application of QSAR techniques to dye–cellulose binding is based on the hypothesis of specific dye–fibre interactions. As an alternative to classical QSAR studies, comparative molecular field analysis was previously used to predict technical dye adsorption properties. This paper presents a structure–affinity study of heterocyclic azo dye adsorption on cellulose fibre by multiple linear regression (MLR), comparative molecular field (CoMFA) and comparative molecular similarity index (CoMSIA) analysis. Structural descriptors, derived from the minimum energy conformers, obtained by molecular mechanics and semiempirical level quantum chemical calculations, were correlated with dye affinity for cellulose by MLR. Models with predictive power were obtained. Despite these good results 3D-QSAR CoMFA and CoMSIA approaches gave a deeper insight into dye–cellulose interactions. Dye conformers obtained by the AM1 and *ab initio* approaches were aligned using atom per atom superposition of a common frame. A comparative analysis on statistic performances and predictive model power was performed for the AM1 and DFT variants of CoMFA and CoMSIA models. Statistically significant models were established for 16 molecules and validated by an external test set of 5 compounds, yielding the best predictive DFT CoMFA model [$r^2 = 0.960$, with 4 components, $q^2 = 0.707$ and $SEE = 1.12$] and several CoMSIA models for AM1 and *ab initio* cases [$r^2 = 0.935 \div 0.947$, $q^2 = 0.688 \div 0.739$ and $SEE = 1.32 \div 1.46$]. The contour maps obtained from 3D-QSAR studies were appraised for affinity trends for the investigated dye molecules. Results indicate a predominant hydrogen donor ability of the dye molecules for both AM1 and *ab initio* variants to play a significant role in dye binding to cellulose. Electrostatic interactions derived from DFT charges bring an important contribution to dye affinity, too. The information obtained from both CoMFA and CoMSIA 3D contour maps may be used in the design of new heterocyclic azo dyes.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

In operations involved in textile industry physical and chemical processes are employed until the final product, e.g. a fibre with a suitable colour, is obtained. Several by-products, e.g. from colouring are generated this way, which are in most cases slopped into wastewaters [1]. Most dyes are very stable, soluble in water, resistant to chemical reactions and poorly biodegradable. In

addition, by uncontrolled oxidation reactions very toxic secondary products can be generated. Coloured effluents can cause problems in several ways: dyes can have acute and chronic effects on exposed organisms depending on the exposure time, and dye concentration [2]. Moreover, already small quantities of dyes are visible and undesirable in water [3,4]. New approaches for the reduction of dye concentration in wastewaters are reported in Refs. [5,6].

From a thermodynamic point of view, the transfer of the dye from the aqueous dissolved state to a fixated state at the fibre can be described as consisting of mainly two steps: the adsorption of dye molecules onto the substrate surface, and the dye penetration into the fibre [7].

* Corresponding author. Tel.: +40 256 491818; fax: +40 256 491824.

E-mail address: timofei@acad-icht.tm.edu.ro (S. Funar-Timofei).

Upon adsorption onto the polysaccharide (cellulose) surface, the dye molecules lose part of their aqueous solvation shell. Accordingly, this step could be expected to be related to the hydrophobicity of the dye molecules. Note that the accumulation of solutes in the interfacial region between water and organic solvents or other organic materials is a generally observed phenomenon for a broad range of compounds [8]. Cellulose has pockets of higher and lower affinities, and water forms hydrogen-bonded clusters around the higher affinity sites, while the lower affinity sites are lacking in significant interactions [9].

The second major step in the dyeing process, the penetration and fixation in the textile matrix, should be governed by the free energy balance between intermolecular interactions among the textile components and dye–fibre as well as dye–dye interactions [7]. Here, both van der Waals forces covering orientation, induction and dispersion interactions as well as site-specific hydrogen bonds are involved. In addition to favourable dye–fibre interactions, the extended π -electron system and associated planar structure of the dye molecules facilitate their intermolecular aggregation, which supports the overall dye fixation in the fibre matrix [10,11].

A specific feature of cellulose fibres is the hydrogen bond network between the polymer components, involving hydroxyl groups and ether oxygen (glycoside oxygen) [7]. It follows that dye molecules with hydrogen bond donor or acceptor sites may interact more distinctly with the cellulose fibre. As a consequence, the fixation of substantive dyes like sulphonated azo compounds in the polymer probably involves hydrogen bond-type interactions, such as between cellulose hydroxyl groups as H donor and the azo bridge as H acceptor.

Classical quantitative structure–affinity relationships [12–17] have been reported in the literature for the dye adsorption into the cellulose fibre. Based on the hypothesis of specific dye–fibre interactions and as alternative to classical QSAR studies, comparative molecular field analysis (CoMFA) [7,18–22] and comparative molecular surface analysis (CoMSA) [23–26] were used to predict technical dye adsorption properties. The CoMSA models indicated that a pharmacophore (called ‘tinctophore’) concept was suitable for the description of the dye–fibre interactions. These models also showed that dye–cellulose interactions were well-defined resembling drug–receptor interactions and that shape determined the activity rules not by the steric repulsion but as a cofactor determining the electrostatic potential distribution. Wojciechowski and Wolska [27] studied the substantivity and spatial structure of some soluble polycyclic dyes for dyeing cellulose fibres by the semiempirical AM1 approach. They concluded that probable sites of a substitution of sulphonyl groups in polycyclic dyes were determined by the molecular energy changes and that dye affinity to the cellulose fibre may depend on the molecule dipole moment.

The MTD (Minimal Topologic Difference) [28] method was considered a precursor of the well known 3D CoMFA (Comparative Molecular Field Analysis) method [29,30] for modelling the ligand–biologic activity through intermolecular interactions. In previous studies [31,32] MTD (Minimal Topologic Difference method) and MTD-PLS (Minimal Topologic Difference method in a Projection in Latent Structures variant) calculations were applied to the adsorption on cellulose of the same series of heterocyclic azo dyes which is studied in this paper. The model predictive power was not tested by an external test set in both cases. It was found that hydrophobic interactions of the condensed heteroaromatic fragment with the fibre and H-bond interactions with donors attached to a specific naphthalenic substitution position increase the affinity. Steric interactions of two naphthalenic side groups and too polar, (negative charge), H-bond acceptor groups in the above mentioned hydrophobic “pocket” decrease the affinity.

In this paper MLR (multiple linear regression), CoMFA and CoMSIA (comparative molecular similarity index) [33] methods were used in the study of dye–cellulose interactions using a series of heterocyclic azo dyes and the results were compared. Additional structural dye features important for binding to fibre were derived, and the type of dye–fibre interactions specific for the dye adsorption emphasized. The predictive power of all models was checked by different approaches.

2. Methods

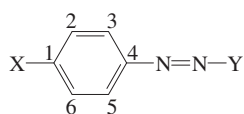
2.1. Definition of target property and molecular structures

The experimental dye affinity values (the difference of the standard chemical potential of the dye adsorbed in the fibre and the one dissolved in the dye bath, at equilibrium) of the 21 heterocyclic monoazo dyes have been taken from the literature [34–36]. The structures of these compounds are presented in Table 1.

Starting structures were generated using the SYBYL package [37], with the TRIPOS force field [38]. The dye structures obtained thereby were further optimized using the AM1 semiempirical quantum-mechanics procedure [39] and partial atomic charges from the semiempirical AM1 method were employed. The conformational behaviour of the molecules was described elsewhere [32]. In addition, these AM1 geometries for each dye compound from the series were used as input structures for optimization by density functional theory (B3LYP/6-31G(d)) [40] by the Gaussian 2009 software [41]. All optimized structures were characterized as true minima by frequency calculations (NImag = 0 for each compound). ZPE corrections were obtained by the standard rigid rotor – harmonic oscillator approximations with unscaled frequencies. The results thus obtained (see [Supplementary material](#)) demonstrated that all the RMSD (root mean square deviation) values between the heavy atom coordinates corresponding to the AM1 and the *ab initio* calculations were below 0.5 Å (value usually involved to detect redundant structures). Moreover, the AM1 and the *ab initio* atomic charges are approximately correlated with correlation coefficients between 0.95 and 0.98. The electronic energy, zero-point correction, thermal correction to enthalpy and thermal correction to Gibbs free energy of the energy minimized compounds are presented in the [Supplementary material](#).









Twenty-one types of descriptors were calculated by the Dragon 5.5 (Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy) software. This software automatically eliminates descriptors with constant values. Therefore, after eliminating them, a total number of 1386 descriptors remained: 31 constitutional, 82 topological and 45 walk and path counts descriptors, 28 connectivity indices, 47 information indices, 96 2D autocorrelation indices (like: ATS4p – Broto–Moreau autocorrelation of a topological structure – lag 4/weighted by atomic polarisabilities), 105 edge adjacency indices, 64 Burden eigenvalues, 15 topological charge (Galvez) indices, 44 eigenvalue-based indices, 41 Randic molecular profiles, 42 geometrical descriptors, 150 RDF (radial distribution function) descriptors, 160 3D-MoRSE descriptors (3D-molecule representation of structure based on electron diffraction), 99 WHIM descriptors, 185 Getaway (like: R6m – R autocorrelation of lag 6/weighted by atomic masses) descriptors, 14 functional groups counts, 22 atom-centred fragments, 20 molecular properties, 24 2D binary fingerprints and 72 2D frequency fingerprints.

Further 12 hydrophobicity parameters were computed by the ALOGPS 2.1 program [42,43]; further also logKow [44], AlogP [42], ClogP2 (ClogP v.4.0, Biobyte, Claremont, CA, USA), IA_logP and IA_logS (octanol/water partition coefficient water solubility, respectively calculated by Interactive Analysis, using neural networks technology), XlogP, COSMO frag [45], MLOGP1 – Moriguchi

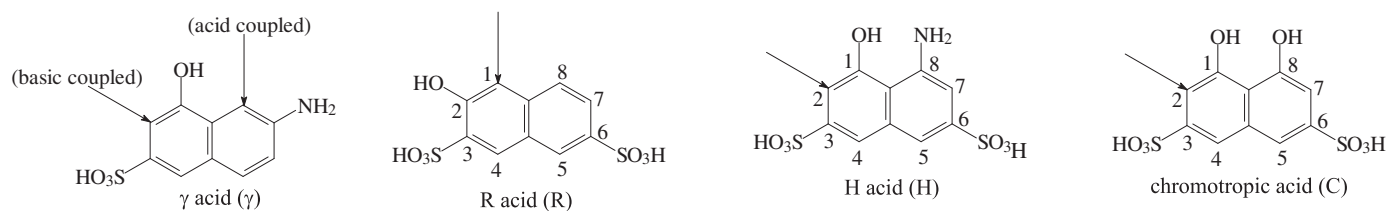
Table 1The structures of the studied heterocyclic azo dyes, their experimental affinities to cellulose fibre (A) and dye molecule descriptors used in the MLR models.^a

No.	X	Y	A (kJ mol ⁻¹)	ASA–	logD _{PHYSPROP}	Bond count	Szeged index	R6m	ATS4p
1		γ _b	22.26	273.43	5.24	56	6200	0.293	3.815
2		γ _b	15.69	266.68	1.56	54	5674	0.288	3.681
3		γ _a	14.35	288.1	0.87	54	5878	0.281	3.675
4		R	9.62	353.77	1.71	58	7759	0.399	3.912
5		R	8.79	347.84	−0.98	56	7145	0.38	3.79
6		γ _b	13.18	261.02	0.25	47	3826	0.296	3.539
7		γ _b	10.92	240.49	1.45	46	3826	0.317	3.517
8		H	14.48	318.99	3.63	60	8341	0.401	3.977
9		C	10.50	328.24	0.99	59	8341	0.414	3.96
10		C	7.70	322.09	−1.75	57	7694	0.411	3.845
11		R	5.23	331.02	−2.18	49	4969	0.369	3.664
12		γ _b	8.58	256.38	1.65	45	3826	0.331	3.494
13		H	13.56	312.7	0.09	58	7694	0.394	3.864

Table 1 (continued)

No.	X	Y	A (kJ mol ⁻¹)	ASA–	logD _{PHYSPROP}	Bond count	Szeged index	R6m	ATS4p
14		R	4.48	296.4	–1.83	48	4969	0.378	3.644
15		R	4.60	324.66	0.08	47	4969	0.386	3.625
16		C	3.59	316	–2.95	50	5394	0.423	3.726
17		C	1.92	311.51	–0.61	48	5394	0.451	3.689
18		C	2.97	295.74	–2.52	49	5394	0.439	3.708
19		H	9.49	306.7	–2.53	51	5394	0.407	3.747
20		H	7.24	286.12	0	50	5394	0.424	3.729
21		H	6.61	301.91	2.02	49	5394	0.435	3.711

^a A: experimental affinities; Y: coupling components (see below): γ : acid coupled in acidic (γ_a), respectively basic (γ_b) medium, R: R acid, H: H acid, C: chromotropic acid; ASA–: solvent accessible surface area of all atoms with negative partial charge; logD_{PHYSPROP}: logarithm of distribution coefficient at isoelectric point calculated from the PHYSPROP database; Bond count – number of bonds in the molecule including bonds of hydrogen atoms; Szeged index: sum of the number of atoms on both sides of each bond (those atoms only which are nearer to the given side of the bond than to the other); R6m: R autocorrelation of lag 6/weighted by atomic masses; ATS4p: Broto-Moreau autocorrelation of a topological structure – lag 4/weighted by atomic polarizabilities.



octanol–water partition coefficient [46] and, finally, the ClogP1 parameter (the logarithm of octanol/water partition coefficient) (ClogP v.1.0.0, Biobyte, Claremont, CA, USA).

18 Quantum chemical descriptors derived from AM1 semi-empirical calculations were, also used, i.e. dipole moment, HOMO and LUMO energies, electrophilicity index, COSMO area and volume [47], Mulliken electronegativity, absolute hardness, nucleophilic and electrophilic delocalisabilities, atom polarisability, minimum and maximum atom charge, static isotropic average alpha.

52 additional descriptors were calculated by the MarvinSketch 5.4.0.1 software (ChemAxon Ltd., Budapest, Hungary): molecular polarisability, geometric descriptors (e.g.: Dreiding energy, Minimal projection area, Maximal projection area, Minimal projection radius, Maximal projection radius, van der Waals surface area), solvent accessible area descriptors, for solvent radius = 1.4 (e.g.: ASA–: solvent accessible surface area of all atoms with negative

partial charge), atom/bond count descriptors, (e.g. bond count: number of bonds in the molecule including bonds of hydrogen atoms), distance based indices, e.g.: Szeged index – the Szeged index extends the Wiener index for cyclic graphs; it represents the sum of counting the number of atoms on both sides of each bond (those atoms only which are nearer to the given side of the bond than to the other), number of acceptor/donor sites, refractivity, logD_{PHYSPROP}: the logarithm of the distribution coefficient at isoelectric point calculated from the PHYSPROP database.

Dye structures were, also, modelled by the conformational search ability of the Omega v.2.4.3 (OpenEye Scientific Software, Santa Fe, NM 87507) software. Omega employs a rule-based algorithm [48] in combination with variants of the Merck force field 94 [49]. The following parameters were used for the conformer generation with Omega v.2.4.3: a maximum of 200 conformers per compound and an energy cut-off of 10 kcal/mol relative to the

global minimum identified from the search. SMILES notation was used as program input. The force field used was the 94s variant of the MMFF (Merck Molecular force field) with coulomb interactions and the attractive part of the van der Waals interactions. To avoid redundant conformers, an RMSD fit of 0.5 Å was used for duplicate removal, with respect to the number of flexible bond from the molecule. Atomic partial charges were calculated for each conformer by the AM1-BCC model [50,51].

2.2. MLR calculations

MLR analysis [52] has been applied after variable selection carried out by the genetic algorithm included in the MobyDigs program [53], using the RQK fitness function [54], with leave-one-out cross-validation correlation coefficient as constrained function to be optimized, a crossover/mutation trade-off parameter $T = 0.5$ and a model population size $P = 50$. In a genetic algorithm procedure for variable selection, a population of strings is randomly created. The process of variable selection is described in the [Supplementary material](#).

2.3. CoMFA/COMSIA calculations

2.3.1. Structural alignment procedures

In cellulose dyeing there is few reported data on the dye orientation inside the dyed fibre. Thus, a predominantly parallel orientation of the dye molecules to the fibre axis was found by electro-optical methods [55]. The dye linearity is also, considered to bring the dye molecule in a close proximity to the cellulose fibre and to be responsible for efficient dye–fibre interactions [56]. The number of azo groups [57,58], as well as the presence of conjugated double bonds [59] or the length of planar dye molecules [60] can be factors that influence the dye substantivity. Coplanarity of the dye molecule is also considered an important dye parameter which influences the dyeing capacity [56,57,61]. Better coplanarity of azo dyes derived from benzene sulphonamide intermediates increased the exhaustion on cellulose fibre [62]. Therefore the alignment rules used in this study were related to the dye molecular axis.

For each compound the conformation of lowest energy derived from the different applied approaches was used in CoMFA/CoMSIA calculations. The molecules were aligned according to the RMS_FIT option within SYBYL (compound **1** with the highest affinity was considered as template).

Five different structure alignments were considered for the 21 heterocyclic azo dyes. In alignment model **1**, structures were built and optimized using the Tripos 5.2 force field [38] and partial charges were determined using the Gasteiger–Hückel (G–H) electronegativity-based method, as implemented in Sybyl 8.0. In alignment models **2** and **3**, the AM1 optimized structures were applied. The AM1 semiempirical quantum-mechanics partial atomic charges were used. The dye structures in alignment model **4** were minimized by the MMFF94s force field and the partial charges by the AM1-BCC model were employed. In alignment model **5** the structures optimized by the DFT (*ab initio*) approach and the corresponding Mulliken charges were used. In the alignment models **1**, **2**, **4** and **5** the following atoms were considered as template: the nitrogen atoms of the azo group plus the carbon atoms of the central phenyl fragment (see [Table 1](#)). The alignment rule **3** used the nitrogen atoms of the azo group plus the atoms C¹ and C⁴ (see [Table 1](#)).

2.3.2. CoMFA and CoMSIA parameters

Steric and electrostatic CoMFA fields were computed using a sp³ carbon with a charge of +1 as probe atom and a grid spacing of 2 Å. Electrostatics was suppressed within the 30 kcal/mol steric energy cut-off.

The CoMSIA similarity indices fields (steric, electrostatic, hydrophobic and hydrogen bond-donor and -acceptor) were calculated at the grid lattice points using a common probe atom of 1 Å radius, a charge of +1, hydrophobicity of +1, a value of +1 for hydrogen bond donating and accepting properties, and an attenuation factor of 0.3.

A box with the dimensions 30 × 18 × 14 Å was used. The PLS method [63] with the leave-one-out (LOO) [64] cross-validation procedure was carried out to determine the optimal number of components to be used in the final PLS analysis (that without cross-validation). All cross-validation calculations were performed with a minimal σ (column filter) value of 2.00 kcal/mol. “CoMFA STD” (block) scaling of independent variables within SYBYL was used for the steric and electrostatic fields.

2.4. Model validation

To test the external predictive power, the following statistical measures were used [65]: (1) squared correlation coefficient R^2 between the predicted and observed activities as well as squared correlation coefficient by cross-validation (q^2); (2) coefficient of determination for linear regressions with intercepts set to zero, i.e. R_0^2 (predicted versus observed activities), and $R_0'^2$ (observed versus predicted activities); (3) slopes k and k' of the above mentioned two regression lines. All these measures were applied over the test set compounds. The following conditions should be satisfied for a model with acceptable predictive power:

$$q^2 > 0.5 \quad (1)$$

$$R^2 > 0.6 \quad (2)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \leq k \leq 1.15 \quad (3)$$

or

$$\frac{(R^2 - R_0'^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \leq k' \leq 1.15 \quad (4)$$

$$|R_0^2 - R_0'^2| < 0.3 \quad (5)$$

Additional statistical parameters: root mean square error of prediction (RMSEP), relative standard error of prediction (RSEP%) and mean absolute error (MAE%) were employed to test the predictive model ability [66]:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^n (y_{\text{pred}} - y_{\text{obs}})^2}{n}} \quad (6)$$

$$\text{RSEP}(\%) = 100 \sqrt{\frac{\sum_{i=1}^n (y_{\text{pred}} - y_{\text{obs}})^2}{\sum_{i=1}^n (y_{\text{obs}})^2}} \quad (7)$$

$$\text{MAE}(\%) = \frac{100}{n} \sqrt{\sum_{i=1}^n |y_{\text{pred}} - y_{\text{obs}}|} \quad (8)$$

where y_{obs} is the observed dye affinity of the compound in the sample, y_{pred} the predicted affinity (either as internal, cross-validated, or external, test set prediction) and n the number of samples in the test set.

3. Results and discussion

3.1. Design of training and test sets

A training set of 16 compounds and a test set of five: **2, 5, 6, 11** and **18** (see Table 1) were considered. It is desirable that the two sets contain similar molecules. The test set compounds were selected consulting the scores scatter plots of the first three factors in the model constructed using as descriptor matrix the set of CoMFA (steric and electrostatic) variables included in the final PLS model for the 21 analyzed compounds. We have included in the test set one of two similar compounds (grouped together) positioned on the opposite sides of the plot origin in the four quadrants of the respective plots.

3.2. MLR

In case of a widely distributed dataset in QSAR studies, the data are normalized based on the autoscaling method [67], which can be described as:

$$XT_{mj} = \frac{X_{mj} - \bar{X}_m}{S_m} \quad (9)$$

where for each variable m , XT_{mj} and X_{mj} are the values j for the variable m after and before scaling respectively, \bar{X}_m is the mean and S_m the standard deviation of the variable.

All 1468 descriptors calculated for the dye molecules, derived from AM1 semiempirical calculations, were normalized based on the autoscaling method. Starting from the normalized descriptors, variable selection was carried out by a genetic algorithm.

In the MobyDigs software 50% of the population individuals is iteratively recreated after 2000 iterations; the eliminated individuals are the worst. This process does provide not a single model but a population of acceptable models. A flowchart of variable selection by genetic algorithm is presented in the Supplementary material.

All statistical tests were performed at a significance level of 5%. In MLR calculations outliers were tested by the standardized residuals of less than -2.5 or more than $+2.5$ [68] and by the value of residual greater than three times the value of standard error in calculation [69], as implemented in the MobyDigs program [53]. The Kubinyi fitness function [54] was also used to check the goodness of fit of the obtained MLR models.

The goodness of prediction of the MLR models was checked by the Akaike Information Criterion (AIC) [70], the multivariate K correlation index [71], Y-scrambling [72] and bootstrapping [73]. All these calculations were performed by the MobyDigs software. The leave-one out cross-validation procedure [74] was, also, employed for internal validation. Several MLR models were thus obtained. Good correlations with dye affinity were noticed in case of the following equations:

where r^2 represents the correlation coefficient, q^2 – leave-one-out cross-validation parameter, q_{boot}^2 – bootstrapping parameter, $a(r^2)$ and $a(q^2)$ – Y-scrambling variables, r_{adj}^2 – adjusted r^2 , SDEP – standard deviation error in prediction, F – Fischer test, SEE – standard error of estimate, AIC – Akaike Information Criterion, the multivariate K correlation indices (Kx – the multivariate correlation index of the matrix of X descriptors and Kxy – the multivariate correlation index of the matrix of X descriptors and Y response variable), FIT – the Kubinyi fitness function.

The descriptors included in equations (10) and (11) before normalization are presented in Table 1.

The predictive power of equations (10) and (11) was checked (see Supplementary material). In fact, for judging the performance of a multivariate calibration model, the root mean squared error of prediction (RMSEP) is extensively used as a criterion; often it is even the sole criterion. But in this study the predictive ability of both 2D- and 3D-QSAR models are not only expressed in terms of only RMSEP, RSEP%, MAE% but they were also evaluated based on the statistical measures presented by equations (1)–(5). These models have good predictive power according to the Tropsha's criteria.

2D autocorrelation descriptors are calculated from the molecular graph with a series of weighting schemes, including weighting by atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarisabilities. The spatial autocorrelation is then evaluated by considering separately all the contributions of each different path length (lag) in the molecular graph, as collected in the topological distance matrix. The ATS4p descriptor characterizes the polarisability of the molecules and the interactions between each pair of atom in the molecules. The presence of this descriptor in equation (11) suggests the favourable influence of dye polarisability on the affinity for fibre.

The Getaway (Geometry, Topology, and Atom-Weights Assembly) descriptors are derived from the leverage matrix which is deduced by centring all atomic coordinates. They are geometric descriptors encoding information on the effective position of substituents and fragments in the molecular space. They are independent of molecule alignment and, to some extent, account also for information on molecular size and shape and for specific atomic properties. R indices are molecular descriptors which encode information on structural fragments and therefore seem to be particularly suitable for describing differences in congeneric series of molecules. Descriptor R6m gives information about the presence of significant substituents in the molecule. More information on the influence of this descriptor on the dye affinity is presented in Section 3.3.

Dye polarity is favourable for binding to cellulose. The affinity increases with increased dye distribution coefficient measured at the isoelectric point, and with increased molecular dimension.

$$A = 9.46(\pm 0.34) + 1.79(\pm 0.37)\log D_{\text{PHYSPROP}} + 12.21(\pm 1.32)\text{bond count} - 10.64(\pm 1.31)\text{Szeged index}$$

$$N_{\text{training}} = 16 \quad r^2 = 0.947 \quad q^2 = 0.914 \quad q_{\text{boot}}^2 = 0.899 \quad r_{\text{adj}}^2 = 0.934$$

$$Kx = 53.74 \quad Kxy = 57.8 \quad \text{SDEP} = 1.451 \quad \text{SEE} = 1.31 \quad F = 71.59 \quad \text{AIC} = 2.863$$

$$a(r^2) = 0.493 \quad a(q^2) = 0.009 \quad \text{FIT} = 8.431 \quad (10)$$

$$A = 9.32(\pm 0.41) + 4.61(\pm 0.55)\text{ATS4p} - 3.68(\pm 0.56)\text{R6m} - 2.74(\pm 0.74)(\text{ASA}-)$$

$$N_{\text{training}} = 16 \quad r^2 = 0.923 \quad q^2 = 0.869 \quad q_{\text{boot}}^2 = 0.858 \quad r_{\text{adj}}^2 = 0.930$$

$$Kx = 57.39 \quad Kxy = 57.65 \quad \text{SDEP} = 1.783 \quad \text{SEE} = 1.58 \quad F = 48.07 \quad \text{AIC} = 4.156$$

$$a(r^2) = 0.494 \quad a(q^2) = -0.045 \quad \text{FIT} = 5.695 \quad (11)$$

The molecular surface area measures the extent to which a molecule is exposed to the external environment. This descriptor is related to binding, transport, and solubility. The highest negative values of the solvent accessible surface area, favourable for dye binding to cellulose, correspond to dye molecules having γ_b acid as coupling component. The solvent accessible surface area of atoms with negative partial charge can be related to the polar interactions of the dye molecules in the process of binding to cellulose. Dyeing carried out in a basic environment is favourable for binding to cellulose.

As the information encoded in the descriptors included in equations (10) and (11) is not always very clear, additional 3D-QSAR calculations were employed to have a deeper insight into dye features important for binding to cellulose.

3.3. CoMFA and CoMSIA

In order to choose the structure alignment, CoMFA (steric and electrostatic) and CoMSIA (steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor) were used for the set of 21 compounds.

In the alignment model 3, computed by CoMFA for the entire series of 21 dyes the electrostatic field dominates the steric one, opposite to models 1, 2, 4 and 5 (see Table 2). Better statistical results were noticed for alignment model 5, in comparison to the others.

In CoMSIA models (see Table 3) the hydrogen bond donor field dominates the steric, electrostatic, hydrophobic and hydrogen bond acceptor ones in all alignments 1–4. Good statistical results were noticed for alignment model 2. Both alignment models 2 and 5 include same atoms as template.

To rank the models obtained by the CoMFA and CoMSIA approaches tackled using structures obtained by AM1 or DFT (*ab initio*) quantum chemical calculations, the statistical parameters should be as high as possible. In the mean time higher r^2 values must be accompanied by q^2 values as near as possible to the r^2 ones (to avoid overfitting) [75]. In the model construction phase the danger of overfitting was avoided as much as possible by using the

cut-off value for the number of significant latent variables when the statistical performances are not augmented more than 5%. Still the best models must be those with a minimal gap between the r^2 and q^2 values. Also an important characteristic of a good model is their high predictive power. Usually this capacity is certified by using an external test set, i.e. predicting the activity for molecules which are not used in the model construction. The analysis of these features must be performed comparing the AM1 and DFT variants of the models. For this purpose CoMFA and CoMSIA models were constructed for the whole set of 21 compounds, but also splitting the series in training and external test sets. The results are presented in Tables 4 and 5.

Seeing the intrinsic characteristics of the CoMFA and CoMSIA methods, one can expect differences between the AM1 and the *ab initio* calculations based models in the statistic performances, and in the steric and electrostatic contributions. Thus the comparative analysis must be focused on these aspects. To facilitate the insight into the major differences between the AM1 and the *ab initio* results, the small details were disregarded, in favour of the coherent, larger patterns.

Thus, first the differences between the 3D geometrical structures resulted from the two quantum chemical methods were searched. For this purpose the RMSD (root mean square deviation) values between the heavy atom coordinates corresponding to the AM1 and the *ab initio* calculations for every molecule were determined. All the values fall below 0.5 Å usually involved in detecting redundant conformations (see the Supplementary material). So it can be considered that from the two methods essentially the same geometrical structures were delivered. In the steric field of CoMFA or CoMSIA no significant differences are expected between the AM1 and *ab initio* versions. This fact is illustrated in the contour maps discussed later. These conclusions are retrieved inspecting the data from Tables 4 and 5: the models CoMFA(S) and CoMSIA(S) present approximately the same AM1 and DFT statistical performances either for 21, or for 16 compounds.

The charge distribution calculation method generally affects the 3D-QSAR procedure performances. The effects of the empirical and semiempirical calculation methods were tested recently [76]. The

Table 2
CoMFA results with steric and electrostatic fields for the series of 21 dyes.^a

No.	Alignment models	q^2	SDEP	r^2	SEE	F	Relative contributions	
							Steric	Electrostatic
1	Tripos, G–H ^b	0.700	2.896	0.898 (2)	1.685	79.51	0.514	0.486
2	AM1, AM1	0.693	2.928	0.835 (2)	2.143	45.701	0.559	0.441
3	AM1, AM1	0.648	3.226	0.896 (3)	1.753	48.799	0.433	0.567
4	MMFF94s, AM1-BCC	0.672	3.024	0.817 (2)	2.258	40.255	0.569	0.431
5	B3LYP 6-31d, Mulliken charges	0.789	2.443	0.933(3)	1.410	78.525	0.534	0.466

^a q^2 – leave-one-out ‘cross-validated r^2 ’; SDEP – standard errors of predictions; r^2 – conventional correlation coefficient (the optimum PLS-component number used in the non-crossvalidated PLS analysis is given in parentheses); F – Fischer test; SEE – standard errors of estimates.

^b G–H – Gasteiger–Hückel charges.

Table 3
CoMSIA results with steric, electrostatic, hydrophobic, donor and hydrogen bond acceptor fields for the series of 21 azo dyes.^a

No.	Alignment models	q^2	SDEP	r^2	SEE	F	Relative contributions				
							Steric	Electrostatic	Hydrophobic	Donor	Acceptor
1	Tripos, G–H ^b	0.759	2.667	0.928 (3)	1.457	73.184	0.117	0.207	0.148	0.292	0.236
2	AM1, AM1	0.784	2.524	0.932 (3)	1.421	77.209	0.123	0.159	0.168	0.305	0.245
3	AM1, AM1	0.679	3.078	0.926 (3)	1.475	71.312	0.080	0.204	0.128	0.306	0.282
4	MMFF94s, AM1-BCC	0.780	2.553	0.924 (3)	1.499	68.885	0.143	0.161	0.178	0.337	0.181
5	B3LYP 6-31d, Mulliken charges	0.759	2.594	0.898(2)	1.684	79.62	0.124	0.171	0.163	0.303	0.239

^a q^2 – leave-one-out ‘cross-validated r^2 ’; SDEP – standard errors of predictions; r^2 – conventional correlation coefficient (the optimum PLS-component number used in the non-crossvalidated PLS analysis is given in parentheses); F – Fischer test; SEE – standard errors of estimates.

^b G–H – Gasteiger–Hückel charges.

Table 4

CoMFA and CoMSIA statistical results for the training set of 21 dyes for alignment model 2 ('cross-validated r^2 (q^2)) for the optimum number of PLS-components; conventional r^2 , standard errors of estimates (SEE).^a

Method	AM1 minimisation					DFT minimisation				
	r^2	SEE	RMSEP	MAE	q^2	r^2	SEE	RMSEP	MAE	q^2
CoMFA(S,E)	0.835	2.143	1.98	28.72	0.693	0.933	1.41	1.27	20.59	0.798
CoMFA(S)	0.821	2.235	2.07	29.4	0.637	0.851	2.099	1.89	27.54	0.598
CoMFA(E)	0.952	1.229	1.07	21.17	0.704	0.921	1.525	1.37	22.64	0.712
CoMSIA(S,E)	0.801	2.356	2.18	30.35	0.674	0.924	1.458	1.35	22.5	0.821
CoMSIA(S)	0.786	2.442	2.26	31.19	0.577	0.789	2.429	2.25	30.39	0.579
CoMSIA(E)	0.802	2.421	2.18	29.05	0.618	0.847	2.065	1.91	27.45	0.643
CoMSIA(H)	0.915	1.584	1.42	22.46	0.726	0.935	1.388	1.25	21.55	0.768
CoMSIA(D)	0.813	2.282	2.11	28.53	0.496	0.671	2.948	2.8	31.82	0.46
CoMSIA(A)	0.77	2.607	2.35	29.43	0.595	0.683	2.975	2.75	32.58	0.499
CoMSIA(D,A)	0.885	1.846	1.66	25.99	0.662	0.79	2.421	2.24	29.53	0.549
CoMSIA(H,A,D)	0.931	1.43	1.29	22.42	0.771	0.93	1.437	1.29	21.88	0.765
CoMSIA(S,E,H)	0.856	2.064	1.86	28.1	0.693	0.926	1.436	1.33	21.75	0.787
CoMSIA(S,E,A)	0.823	2.222	2.06	28.53	0.721	0.873	1.88	1.74	24.94	0.743
CoMSIA(S,E,D)	0.924	1.499	1.35	22.59	0.746	0.921	1.487	1.38	21.74	0.753
CoMSIA(S,E,D,A)	0.913	1.6	1.44	24.34	0.762	0.877	1.852	1.71	24.51	0.714
CoMSIA(S,E,H,D,A)	0.932	1.421	1.28	22.02	0.784	0.898	1.684	1.56	23.49	0.759
CoMSIA(S,H,A)	0.854	2.022	1.87	27.76	0.75	0.852	2.032	1.88	27.55	0.74
CoMSIA(S,H,D,A)	0.94	1.331	1.2	20.73	0.8	0.947	1.248	1.12	20.49	0.812
CoMSIA(E,H,A)	0.828	2.191	2.03	28.29	0.712	0.876	1.858	1.72	24.67	0.744
CoMSIA(E,H,D,A)	0.92	1.54	1.39	23.66	0.762	0.88	1.831	1.69	24.32	0.713
CoMSIA(S,H,D)	0.945	1.279	1.15	19.8	0.758	0.934	1.357	1.26	20.8	0.763
CoMSIA(S,E,H,D)	0.938	1.357	1.22	20.35	0.768	0.931	1.386	1.28	20.25	0.781
CoMSIA(S,E,H,A)	0.844	2.086	1.93	28.3	0.734	0.892	1.74	1.61	24.75	0.768
CoMSIA(E,H,D)	0.929	1.448	1.3	21.89	0.736	0.921	1.48	1.37	21.77	0.747

^a S – steric field, E – electrostatic field, H – hydrophobic field, D – donor field, A – acceptor field; RMSEP – root mean square error of prediction; MAE – mean absolute error.

ab initio charges were also implied in CoMFA studies, the results being better than for the semiempirical methods [77]. Thus the changes in *ab initio* charges against the AM1 charges were examined. The linear dependence between the two charge populations for every molecule (all the Pearson product-moment correlation coefficients were above 0.95) presents outliers and high leverage points. The different charges on the atoms corresponding to these points represent the significant differences between the AM1 and *ab initio* approaches. These differences can be summarized as follows. For the molecules with high “activity” (mainly benzothiazole derivatives, but also others) in the *ab initio* variant the positive charge has been augmented in the opposite side region of the sulphur atom belonging to the heterocycles (or the negative charge has diminished). In contrast, the negative charge on the S atom (or

on the corresponding N atoms in other molecules) is greater. On the other hand, for the molecules with weak “activity”, on the N atoms corresponding to S in the benzothiazole environment are no significant charge changes (eventually a small diminution of the negative charge), as pointed out by the linear relationships. Also the *ab initio* variant has presented greater negative charges on the OH and NH₂ side groups of the coupling components. These findings might be retrieved in differences in the electrostatic field contributions of the AM1 and *ab initio* based models.

Examining the CoMFA(E) models from Tables 4 and 5 the statistical performances do not differ noticeable between the AM1 and DFT variants. On the other hand adding the steric field to the electrostatic one, the CoMFA(S,E) case, for the 21 compounds set the DFT version presents a better model, while for the training set

Table 5

CoMFA and CoMSIA statistical results for the training set of 16 dyes for alignment model 2 ('cross-validated r^2 (q^2)) for the optimum number of PLS-components; conventional r^2 , standard errors of estimates (SEE).^a

Method	AM1 minimisation					DFT minimisation				
	r^2	SEE	RMSEP		q^2	r^2	SEE	RMSEP		q^2
			Training	Test				Training	Test	
I CoMFA(S,E)	0.987	0.702	0.555	1.652	0.723	0.96	1.188	0.985	1.011	0.707
II CoMFA(S)	0.806	2.409	2.17	2.14	0.415	0.656	3.092	2.89	3.35	0.341
III CoMFA(E)	0.965	1.114	0.92	1.4	0.588	0.963	1.139	0.94	2.18	0.665
IV CoMSIA(S,E)	0.754	2.713	2.445	2.518	0.466	0.905	1.687	1.521	0.977	0.689
V CoMSIA(S)	0.604	3.321	3.11	3.48	0.35	0.588	3.387	3.17	3.57	0.316
VI CoMSIA(E)	0.692	3.04	2.74	2.54	0.375	0.826	2.282	2.06	1.77	0.488
VII CoMSIA(H)	0.929	1.52	1.32	2.3	0.587	0.945	1.338	1.16	1.55	0.678
VIII CoMSIA(D)	0.892	1.802	1.62	2.61	0.394	0.859	2.142	1.85	3.02	0.444
IX CoMSIA(A)	0.688	3.058	2.76	2.19	0.392	0.677	3.113	2.81	2.69	0.373
X CoMSIA(D,A)	0.91	1.711	1.482	2.299	0.594	0.866	2.083	1.804	2.737	0.554
XI CoMSIA(H,A,D)	0.941	1.383	1.198	1.49	0.681	0.984	0.742	0.615	1.982	0.769
XII CoMSIA(S,E,H)	0.85	2.207	1.911	1.928	0.559	0.931	1.497	1.297	0.917	0.72
XIII CoMSIA(S,E,D)	0.934	1.465	1.269	1.463	0.647	0.936	1.389	1.252	1.548	0.668
XIV CoMSIA(S,E,A)	0.8	2.45	2.208	1.454	0.577	0.882	1.882	1.697	1.006	0.623
XV CoMSIA(S,E,H,D,A)	0.935	1.458	1.263	1.271	0.688	0.945	1.338	1.158	1.403	0.738
XVI CoMSIA(S,H,D,A)	0.939	1.406	1.22	1.32	0.698	0.947	1.315	1.14	1.54	0.739

^a S – steric field, E – electrostatic field, H – hydrophobic field, D – donor field, A – acceptor field; RMSEP – root mean square error of prediction.

the two versions are sensibly the same. The CoMSIA(E) *ab initio* model for 21 compounds is slightly better than the AM1 equivalent, but with overall weak performances. For the training set the corresponding models are also unconvincing. In the CoMSIA(S,E) case the models for the whole set of molecules reproduce the CoMFA(S,E) behaviour. The training set DFT CoMSIA(S,E) model is better, but their external predictive power is questionable (see below). Briefly the function of the electrostatic field in the CoMFA and CoMSIA modelling tend to favour the use of DFT charges. If a greater reliance is admitted for the *ab initio* charges (and principally for their variation in the series), this is probably reflected in somewhat better statistical performances.

Inspecting the external predictive performances of the models (Table 5, RMSEP values) one attains the conclusion of Hawkins [78]: the small holdout samples (test sets) are not reliably able to recognize a good model, they display large random variability. Thus for example five models – CoMSIA(S,E), CoMSIA(E), CoMSIA(A), CoMSIA(S,E,H) and CoMSIA(S,E,A) – from sixteen in Table 5 present smaller external (test set) RMSEP values than the corresponding internal cross-validated RMSEP's. This fact is suspicious and appears probably by chance. Hawkins concludes and demonstrates that in case of limited number of data one can assess more trustworthily a model deducing it from the whole dataset and relying on the internal cross-validation prediction. This is why in Table 4 the internal RMSEP values are included and used in the comparison of the predictive power of the models.

Thus analyzing together the two tables for the CoMFA case the best proposed model is the CoMFA(S,E) DFT variant. The CoMSIA method provide some good models: CoMSIA(S,H,D,A) in the DFT and AM1 variants, but the AM1 CoMSIA(S,E,H,D,A) can, also, be used. Plots of dependence of experimental versus predicted affinity values are presented for the DFT CoMSIA(S,H,D,A) training set model (Fig. 1) and in Supplementary materials for other good models.

Y-randomization was used to test the robustness of the CoMFA and CoMSIA models for 21 compounds. The dependent variable (Y) was randomly shuffled and a new model was developed using the original independent-variables matrix. The dye affinity values were randomly shuffled five times (two cases being presented in the Supplementary material) and each time CoMFA/CoMSIA models were built. The results indicate that all models obtained in all these five cases were poor and without predictive power.

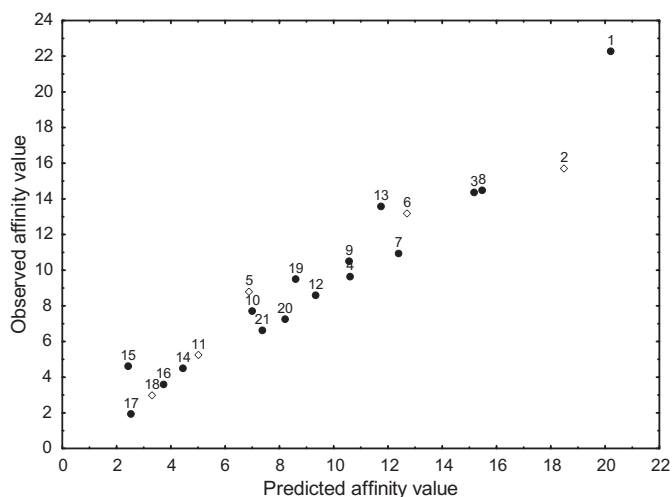


Fig. 1. Dependence of observed versus predicted affinity values for the DFT CoMSIA XVI model (black circles – training set compounds; white rhombus – test set compounds).

From the analysis of the CoMSIA contour maps (model XVI) for the AM1 and *ab initio* versions, as has been expected, the hydrophobic and the H bond (acceptor and donor) contours are nearly the same using the results from the two quantum chemical variants. Also the steric contour maps appear to present the same pattern in the analyzed cases, in a nice agreement with the small RMSD values presented above.

The DFT CoMFA(S,E) contour plot (Fig. 2) indicates positive charge favourable for high affinities on the OH and NH₂ groups of the coupling components (e.g. H and chromotropic acids) and negative charge on the N atoms in opposite side of the sulphur atom in the heterocyclic moiety and in position 6 of the naphthalene fragment (e.g. H and chromotropic acids).

The inspection of the values of the (ASA–) detrimental descriptor in MLR model described by equation (11) indicates that the main contribution to this parameter is assured by the sulphonic acid groups of the R and chromotropic acids. The same information is retrieved in the electrostatic unfavourable plot (Fig. 2): the highest (ASA–) value was noticed in case of compound 4, where unfavourable regions for increased affinity correspond to sulphonic acid groups.

Similar trends were observed in the steric contour plots of AM1 CoMSIA XV (Fig. 3) and AM1 and *ab initio* XVI models (Supplementary material). The contributions to the steric field emphasize the presence of bulkier moieties at the free end region of heterocyclic fragment of the dye molecules as favourable for increased affinity for cellulose. This is supplemented with the information described below: the preferred hydrophobic character of this region. On the other hand the presence of bulky groups at the other end of the dye molecules, like the sulphonic acid group attached in the position 6 of the naphthalene nucleus of the coupling component (e.g. H, chromotropic and R acids), decrease the dye affinity.

The R6m descriptor in MLR model 11 seems to account for this detrimental influence on the dye affinity. Mantle and heavy atoms in the molecule result in higher values for R6m. The peripheral sulphonic groups (S atom) are good candidates to give significant R6m values. The inspection of the values of this parameter, indicate compounds 17 and 18 as presenting the highest R6m. This can be corroborated nicely with the detrimental presence of the sulphonic acid group of compound 17 (Fig. 3).

A hydrophobic site is located in the heterocyclic moiety and may be attributed to the attached methyl substituent, as derived from the hydrophobic contour maps for the training set of AM1 CoMSIA(S,E,H,D,A) (Fig. 4) and both AM1 and *ab initio* CoMSIA(S,H,D,A) models (Supplementary material). Having in mind that this is also

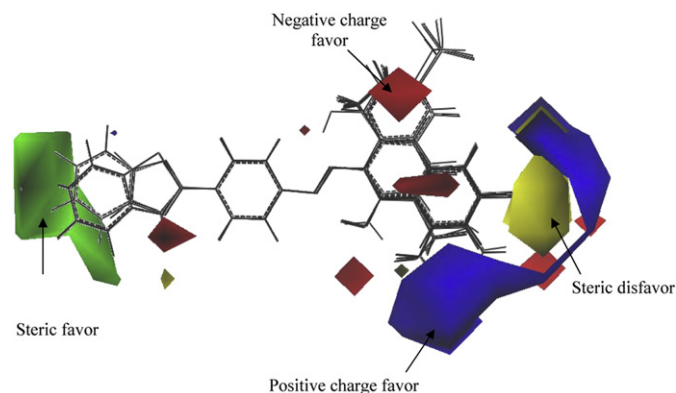


Fig. 2. CoMFA contour maps for the steric and electrostatic fields (stdev*coeff) on all superposed dye molecules from the training set for DFT case.

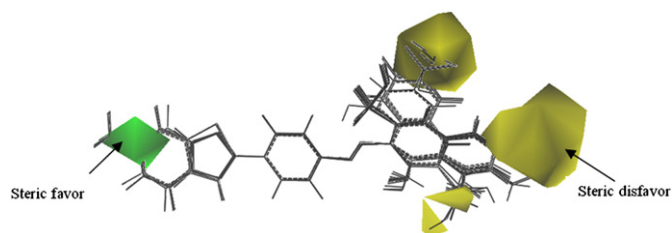


Fig. 3. CoMSIA steric contour map (stdev*coeff) on all superposed dye molecules from the training set for the CoMSIA(S,E,H,D,A) AM1 model.

a sterically tolerant region one would expect that increased alkyl substituents placed here would lead to higher affinities. Increased polar contributions, attributed to increased number of nitrogen atoms in the heterocyclic moiety diminished the affinity for fibre. The presence of hydrophilic polar groups (like: hydroxy or sulphonic acid groups), attached to the coupling component (e.g. the chromotropic acid) in positions 7, respectively 3 of the naphthalene moiety decreased the affinity.

The donor hydrogen field brings an important contribution in comparison to the hydrogen acceptor field in AM1 CoMSIA models X, XI, XV and XVI (Table 5) or the hydrophobic field (AM1 models XI, XV and XVI), even if the statistical parameters did not improve for the *ab initio* case (see Supplementary material). In AM1 models XIII, XV and XVI it even outruns the electrostatic and steric contributions. Therefore the hydrogen bond contribution for dye–cellulose interactions could be considered important.

The hydrogen bond donor field in AM1 CoMSIA XV model (Fig. 5) and both AM1 and *ab initio* CoMSIA(S,H,D,A) models (Supplementary material) confirms the favourable contribution to the affinity of the amino group attached in position 8 of the naphthalene moiety of the coupling component (e.g. H acid). The presence of the sulphonic acid group in position 3 of the naphthalene moiety of the coupling component (e.g. in the H acid) contributes to increased the affinity by its hydrogen bond acceptor capacity. The presence of polar (like hydroxy) groups attached in position 1 to the coupling component (like γ_b and γ_a) are not favourable for dye binding, being probably implied in intramolecular hydrogen bonds with the azo group.

The hydrogen atom attached to the nitrogen atom in the heterocyclic moiety decreased the dye affinity. The heterocyclic moiety, having hydrogen bond acceptor ability can increase the dye affinity. The sulphonic acid group in the position 6 on the naphthalene moiety of the R-acid coupling component is detrimental for the dye affinity, as observed in the steric field. This information can be interpreted either as the location of this group in the inner solution from the matrix fibre [32], or being not involved in the dye binding to the fibre, but probably in the dye solubilisation; or to the lack of structural diversity in the coupling component in this position. As expected, an amino group attached to the coupling

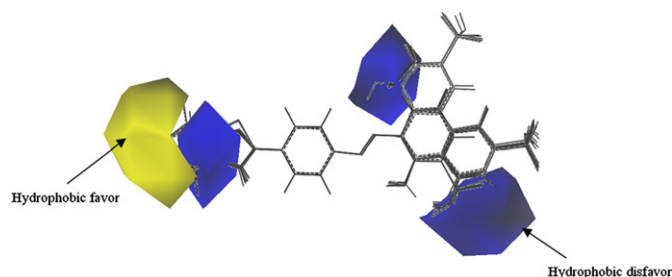


Fig. 4. CoMSIA hydrophobic contour map (stdev*coeff) on all superposed dye molecules from the training set for the CoMSIA(S,E,H,D,A) AM1 model.

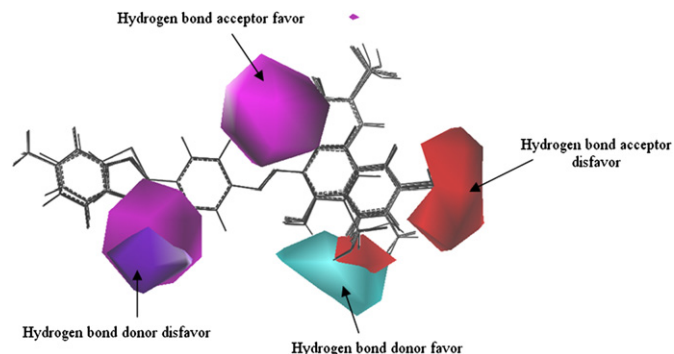


Fig. 5. CoMSIA hydrogen bond donor and acceptor contour maps (stdev*coeff) on all superposed dye molecules from the training set for the CoMSIA(S,E,H,D,A) AM1 model.

component in position 8 (Fig. 5) is not favourable for acceptor–hydrogen bonding, being favourable as hydrogen bond donor group.

The ATS4p descriptor of MLR model 11 can be related to how the polarisability is distributed along the topological structure of the molecule. The highest values were observed for compounds **8** and **9**. These molecules possess three very polarisable S atoms, which can form many pairs with other relatively polarisable atoms at topological distance 4.

4. Conclusion

A comparative structure–affinity study of heterocyclic azo dye adsorption on cellulose fibre by multiple linear regression (MLR), comparative molecular field analysis (CoMFA) and comparative molecular similarity index (CoMSIA) analysis is presented. Good correlations and models with predictive power were obtained. Five different alignments were used in the molecule superposition, both in CoMFA and CoMSIA calculations. The characteristics of the CoMFA and CoMSIA methods were compared by AM1 and the *ab initio* calculations based models in the statistic performances, and in the steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor contributions for both the entire and splitted sets of compounds (training and test sets). The CoMSIA models including the steric, hydrophobic, hydrogen bond donor and acceptor contributions in both DFT and AM1 variants, and the AM1 CoMSIA (steric, electrostatic, hydrophobic and donor and acceptor) model were found to be good candidates for predicting the dye affinity. The information obtained from contour maps obtained from 3D-QSAR studies was compared to the influence of different descriptors included in the best MLR models. Similar trends were observed with those derived from the electrostatic (influence of negative charges) and steric fields. The presence of the sulphonic acid group attached to the coupling component decreases the dye affinity. The contributions to the steric and hydrophobic fields emphasize the presence of bulkier moieties in the region of the dye heterocyclic fragment as favourable for increased affinity for cellulose. Increased polar groups, attributed to increased number of nitrogen atoms in the heterocyclic moiety decreased the affinity for fibre.

Acknowledgements

This project was financially supported by Ministerul Educatiei, Cercetarii si Tineretului – Autoritatea Nationala pentru Cercetare Stiintifica (MedC-ANCS), contract grant number: 71GR/2006 and Project 1.1 of the Institute of Chemistry Timisoara of the Romanian Academy. The authors are indebted to Prof. Tudor Oprea from the University of New Mexico (Albuquerque, U.S.A.), for giving access to

the Sybyl software. Gaussian acquisition was funded by CNCSIS-UEFISCSU project PN II-RU PD_502/2010.

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dyepig.2012.01.015.

References

- [1] Chacón JM, Leal MaT, Sánchez M, Bandala ER. Solar photocatalytic degradation of azo-dyes by photo-Fenton process. *Dyes Pigm* 2006;69:144–50.
- [2] Pavan FA, Dias SLP, Lima EC, Benvenutti EV. Removal of Congo red from aqueous solution by anilinepropylsilica xerogel. *Dyes Pigm* 2008;76:64–9.
- [3] Pearce CI, Lloyd JR, Guthrie JT. The removal of colour from textile wastewater using whole bacterial cells: a review. *Dyes Pigm* 2003;58(3):179–96.
- [4] Metivier-Pignon H, Faur C, Le Cloirec P. Adsorption of dyes onto activated carbon cloth: using QSPRs as tools to approach adsorption mechanisms. *Chemosphere* 2007;66:887–93.
- [5] Li P, Zhao G, Zhao K, Gao J, Wu T. An efficient and energy saving approach to photocatalytic degradation of opaque high-chroma methylene blue wastewater by electrocatalytic pre-oxidation. *Dyes Pigm* 2012;92(3):923–8.
- [6] Burkinshaw SM, Kabambe O. Attempts to reduce water and chemical usage in the removal of bifunctional reactive dyes from cotton: part 2 bis(vinyl sulfone), aminochlorotriazine/vinyl sulfone and bis(aminochlorotriazine/vinyl sulfone) dyes. *Dyes Pigm* 2011;88:220–9.
- [7] Schüürmann G, Funar-Timofei S. Multilinear regression and comparative molecular field analysis (CoMFA) of Azo dye-fiber affinities. 2. Inclusion of solution-phase molecular orbital descriptors. *J Chem Inf Comput Sci* 2003;43:1502–12.
- [8] Pratt LR, Pohorille A. Hydrophobic effects and modeling of biophysical aqueous solution interfaces. *Chem Rev* 2002;102:2671–92.
- [9] Baird MS, Hamlin JD, O'Sullivan A, Whiting A. An insight into the mechanism of the cellulose dyeing process: molecular modelling and simulations of cellulose and its interactions with water, urea, aromatic azo-dyes and aryl ammonium compounds. *Dyes Pigm* 2008;76:406–16.
- [10] Gordon PF, Gregory P. Organic chemistry in colour. Heidelberg: Springer-Verlag; 1982.
- [11] Bach H, Pfeil E, Phillippar W, Reich M. Molekülbau und Haftung substantiver Farbstoffe auf Cellulose. *Angew Chem* 1963;75:407–16.
- [12] Timofei S, Schmidt W, Kurunczi L, Simon Z, Sallo A. A QSAR study for cellulose fiber adsorption of anthraquinone vat dyes. *Dyes Pigm* 1994;24:267–79.
- [13] Timofei S, Kurunczi L, Schmidt W, Fabian WMF, Simon Z. Structure-affinity binding relationships by principal-component-regression analysis of anthraquinone dyes. *Quant Struct-Act Relat* 1995;14:444–9.
- [14] Timofei S, Kurunczi L, Schmidt W, Simon Z. Structure-affinity binding relationships of some 4-aminoazobenzene derivatives for cellulose fiber. *Dyes Pigm* 1995;29:251–8.
- [15] Timofei S, Kurunczi L, Schmidt W, Simon Z. Lipophilicity in dye-cellulose fiber binding. *Dyes Pigm* 1996;32:25–42.
- [16] Timofei S, Kurunczi L, Schmidt W, Simon Z. Dye structure-affinity relationships by the MTD method. *Rev Roum Chim* 1997;42:687–92.
- [17] Timofei S, Kurunczi L, Suzuki T, Fabian WMF, Muresan S. Multiple linear regression (MLR) and neural network (NN) calculations of some disazo dye adsorption on cellulose. *Dyes Pigm* 1997;34:181–93.
- [18] Timofei S, Kurunczi L, Schmidt W, Simon Z. Steric and electrostatic effects in dye-cellulose interactions by the MTD and CoMFA approaches. *SAR QSAR Environ Res* 2002;13:219–26.
- [19] Fabian WMF, Timofei S, Kurunczi L. Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and multiconformational minimal steric difference (MTD) calculation of anthraquinone dye-fiber affinities. *J Mol Struct-THEOCHEM* 1995;340:73–81.
- [20] Fabian WMF, Timofei S. Comparative molecular field analysis (CoMFA) of dye-fiber affinities II: symmetrical bisazo dyes. *J Mol Struct-THEOCHEM* 1996;362:155–62.
- [21] Oprea TI, Kurunczi L, Timofei S. Quantitative structure-activity relationship studies of disperse azo dyes. Toward the negation of the pharmacophore theory of dye-fiber interaction? *Dyes Pigm* 1997;33:41–64.
- [22] Timofei S, Fabian WMF. Comparative molecular field analysis (CoMFA) of heterocyclic monoazo dye-fiber affinities. *J Chem Inf Comput Sci* 1998;38:1218–22.
- [23] Polanski J, Gieleciak R, Wyszomirski M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and anthraquinone dyes. *J Chem Inf Comput Sci* 2003;43:1754–62.
- [24] Polanski J, Gieleciak R, Wyszomirski M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm* 2004;62:61–76.
- [25] Polanski J, Gieleciak R, Magdziarz T, Bak A. GRID formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, Azo dyes, and HEPT derivatives. *J Chem Inf Comput Sci* 2004;44:1423–35.
- [26] Polanski J, Gieleciak R, Bak A. Probability issues in molecular design: predictive and modeling ability in 3D-QSAR schemes. *Comb Chem High T Scr* 2004;7:793–807.
- [27] Wojciechowski K, Wolska A. Substantivity and spatial structure of soluble polycyclic dyes for dyeing cellulose fibres. *Dyes Pigm* 2005;65:111–6.
- [28] Simon Z, Chiriac A, Holban S, Ciubotariu D, Mihalas GI. Minimum steric difference. The MTD method for QSAR studies. Letchworth: Research Studies Press; 1984.
- [29] Cramer III RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 1988;110:5959–67.
- [30] Kubinyi H. The third dimension. In: Kubinyi H, editor. QSAR: an introduction, in 3D – QSAR in drug design. Theory, methods and applications. Leiden: ESCOM; 1993. p. 3–10.
- [31] Timofei S, Kurunczi L, Simon Z. Structure-affinity relationships by the MTD method for binding to cellulose fibre of some heterocyclic monoazo dyes. *MATCH-Commun Math Co* 2001;44:349–60.
- [32] Kurunczi L, Funar-Timofei S, Bora A, Seclaman E. Application of the MTD-PLS method to heterocyclic dye-cellulose interactions. *Int J Quantum Chem* 2007;107:2057–65.
- [33] Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 1994;37:4130–46.
- [34] Alberti G, Cerniani A, De Giorgi MR, Seu G. Affinità, calore ed entropia standard di tinture di alcuni azocoloranti benzotiazolici. *Red Sem Fac Sci* 1978;48:267–73.
- [35] Alberti G, Cerniani A, De Giorgi MR, Seu G. Dyeing thermodynamics of direct azo-dyes derived from 3-(p-aminophenyl)-1,2,4-triazole and 5-(p-aminophenyl)-1,2,3,4-tetrazole. *Ann Chim (Rome)* 1981;295–8.
- [36] Alberti G, Cerniani A, De Giorgi MR, Seu G. Thermodynamique de teinture sur rayonne et coton, solidité et mesure de la couleur sur fibres de colorants azoïques anioniques dérivés de l'imidazole. *Teintex* 1981;1–2:17–26.
- [37] SYBYL 8.0, Tripos Associates, St. Louis, MO
- [38] Clark M, Cramer III RD, van Opdenbosch N. Validation of the general purpose Tripos 5.2 force field. *J Comput Chem* 1989;10:982–1012.
- [39] Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc* 1985;107:3902–9.
- [40] Parr RG, Yang W. Density functional theory of atoms and molecules. New York: Oxford University Press; 1989.
- [41] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 09, Revision A.1. Wallingford CT: Gaussian, Inc.; 2009.
- [42] Tetko IV, Tanchuk VY. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J Chem Inf Comp Sci* 2002;42:1136–45.
- [43] Tetko IV, Tanchuk VY, Villa AEP. Prediction of n-octanol/water partition coefficients from physprop database using artificial neural networks and E-state indices. *J Chem Inf Comp Sci* 2001;41:1407–21.
- [44] Meylan WM, Howard PH. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J Pharm Sci* 1995;84:83–92.
- [45] Hornig M, Klamt A. COSMO frag: a novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. *J Chem Inf Model* 2005;45(5):1169–77.
- [46] Moriguchi I, Hirono S, Liu Q, Nakagome I, Matsushita Y. Simple method of calculating octanol/water partition coefficient. *Chem Pharm Bull* 1992;40:127–30.
- [47] Klamt A, Schüürmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc Perkin Trans* 1993;2:799–805.
- [48] Tresadern G, Bemporad T, Howe A. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *J Mol Graph Modell* 2009;27:860–70.
- [49] Halgren TA. MMFF VI.MMFF94s option for energy minimization studies. *J Comput Chem* 1999;20:720–9.
- [50] Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J Comput Chem* 2000;21(2):132–46.
- [51] Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem* 2002;23(16):1623–41.
- [52] Wold S, Dunn III WJ. Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. *J Chem Inf Comp Sci* 1983;23:6–13.
- [53] Todeschini R, Consonni V, Mauri A, Pavan M. In: Leardi R, editor. Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks. Amsterdam: Elsevier; 2004. p. 141–67 [Chapter 5].
- [54] Todeschini R, Consonni V, Mauri A, Pavan M. Detecting 'bad' regression models: multicriteria fitness functions in regression analysis. *Anal Chim Acta* 2004;515:199–208.
- [55] Trotman ER. Dyeing and chemical technology of textile fibres. 4th ed. London: Griffin; 1970. p. 410.
- [56] Vickerstaff T. The physical chemistry of dyeing. London, Edinburgh: Imperial Chemical Industries Limited; 1954. pp. 175–176.
- [57] Venkataraman K. The chemistry of synthetic dyes, vol. II. New York: Academic Press; 1952.

- [58] Moryganov PV, Mel'nikov BN. O svyazi mezhdru srodstvom prijamyh crasitelei i ih stroenim (Relationship between the affinity of direct dyes and their structure). *Colloid J (USSR)* 1957;19:100–3.
- [59] Schaeffer A. Warum sind substantive Farbstoffe substantiv? *Melliand* 1958; 39:68–74.
- [60] Giles CH, Hassan ASA. Adsorption of organic surfaces. V. A study of the adsorption of dyes and other organic solutes by cellulose and chitin. *J Soc Dyers Colour* 1958;74:846–57.
- [61] Peters RH. Textile chemistry. In: *The physical chemistry of dyeing*, vol. III. Amsterdam: Elsevier Scientific Publishing; 1975.
- [62] Wang L, Pan X, Wang F, Yang L, Liu L. Structure-properties relationships investigation on the azo dyes derived from benzene sulfonamide intermediates. *Dyes Pigm* 2008;76:636–45.
- [63] Wold S, Ruhe A, Wold H, Dunn III WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Siam J Sci Stat Comput* 1984;5:735–43.
- [64] Cramer RD, Bunce JD, Patterson DE, Frank IE. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant Struct-Act Relat* 1988;7:18–25.
- [65] Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aid Mol Des* 2003;17:241–53.
- [66] Goodarzi M, Deshpande S, Murugesan V, Katti SB, Prabhakar YS. Is feature selection essential for ANN modeling? *QSAR Comb Sci* 2009;28:1487–99.
- [67] Latorre MJ, Pena R, Pita C, Botana A, Garca S, Herrero C. Chemometric classification of honeys according to their type. II. Metal content data. *Food Chem* 1999;66:263–8.
- [68] Frank H, Althoen SC. Outliers. In: *Statistics: concepts and applications*. Cambridge: Cambridge University Press; 1994. p. 142–3.
- [69] Todeschini R, Consonni V. *Handbook of molecular descriptors*. Weinheim: Wiley; 2000. p. 369.
- [70] Gentleman JF, Wilk MB. Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics* 1975;31:387–410.
- [71] Todeschini R, Consonni V, Maiocchi A. The K correlation index: theory development and its applications in chemometrics. *Chemometr Intell Lab* 1998;46:13–29.
- [72] Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: applications to variable selection. *J Chemometr* 1996;10: 521–32.
- [73] Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987;82: 171–200.
- [74] Wold S. Cross validity estimation of the number of components in factor and principal components models. *Technometrics* 1978;20:397–405.
- [75] Gramatica P. Principles of QSAR model validation: internal and external. *QSAR Comb Sci* 2007;26(5):694–701.
- [76] Mittal RR, Harris L, McKinnon RA, Sorich MJ. Partial charge calculation method affects CoMFA QSAR prediction accuracy. *J Chem Inf Model* 2009; 49:704–9.
- [77] Jung D, Floyd J, Gund TM. A comparative molecular field analysis (CoMFA) study using semiempirical, density functional, Ab initio methods and pharmacophore derivation using DISCOtech on Sigma 1 ligands. *J Comput Chem* 2004;25:1385–99.
- [78] Hawkins DH. The problem of overfitting. *J Chem Inf Comput Sci* 2004;44: 1–12.