

Multilinear Regression and Comparative Molecular Field Analysis (CoMFA) of Azo Dye–Fiber Affinities. 2. Inclusion of Solution-Phase Molecular Orbital Descriptors

Gerrit Schüürmann*

Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research, Permoserstrasse 15,
D-04318 Leipzig, Germany

Simona Funar-Timofei

Institute of Chemistry, Romanian Academy, Bul. Mihai Viteazul 24, 1900 Timisoara, Romania

Received April 7, 2003

For a data set with 30 direct azo dyes taken from literature, quantitative structure–activity relationship (QSAR) analyses have been performed to model the affinity of the dye molecules for the cellulose fiber. The electronic structure of the compounds was characterized using quantum chemical gas-phase (AM1) and continuum-solvation molecular orbital parameters. As regards the solution phase, COSMO appears to be better suited than SM2 in quantifying relative trends of the aqueous solvation energy. For the dye–fiber affinity, the leave-one-out prediction capability of multilinear regression equations is superior to CoMFA, with predictive squared correlation coefficients ranging from 0.63 (pure CoMFA) to 0.89. At the same time, solution-phase CoMFA is superior to previously derived AM1-based CoMFA models. As a general trend, the dye–fiber affinity increases with increasing electron donor capacity that corresponds to an increasing hydrogen bond acceptor strength of the azo dyes. The discussion includes the consideration of structural features that are likely to be involved in dye–fiber and dye–dye hydrogen bonding interactions, and possible links between CoMFA electrostatic results and the atomic charge distribution of the compounds.

INTRODUCTION

Cotton consists mainly of cellulose, which is a linear glucose polymer with the sum formula $(C_6H_{10}O_5)_n$. More precisely, cellulose is a β -1,4-polyacetal of cellubiose, in which two glucose molecules are linked together via a glycosidic bond. The cellubiose units are correspondingly connected by glycosidic bonds, and the torsion around these bonds is hindered by intramolecular hydrogen bonds between 3-OH groups and ring oxygen atoms of neighboring cellubiose units, leading to a linear stiffness of the macromolecule. At the polymer scale, cellulose consists of microfibrilles that are chains of up to 5000 cellubiose units.¹ Here, the high-order sections called micelles make up ca. 60%, interrupted by amorphous sections that are more susceptible to the adsorption and fixation of dye molecules. With a density of 1.05, dry cotton has a relatively open structure, and its lateral rather than longitudinal swelling in water facilitates the entry of dye molecules.²

In contrast to the animal fiber wool, a polypeptide with strong ionic links between $-NH_3^+$ and $-CO_2^-$ groups, cellulose does not contain ionic sites, and as such is not able to fixate cationic or anionic dyes through respective Coulomb forces. As a consequence, direct dyeing of cotton requires specific structural features of the dye molecules, resulting in an overall affinity for the cellulose fiber that is called substantivity.³ For a successful uptake from aqueous solution and stable fixation in the textile fiber, the dye molecules are adsorbed in intermicellar cavities, where they are thought to form large aggregates³ that resist wash-out due to their size and suitable dye–fiber interactions.^{1,2}

Alternatively, cotton can also be dyed by developing dyes, that are formed on the fiber by chemical reaction at alkaline pH between a diazo component (e.g. diazonium salts of aromatic amines) and a coupling component like naphthol AS (anilide of 2-naphthol-3-carboxylic acid) that has no acidic groups. In contrast to the reactants, the azo product formed in situ has a particularly low water solubility, and again dye–dye aggregation and the resultant substantial molecular size support the fixation on the fiber.²

Azo dyes cover a broad range of functionalities, including the direct dyeing and developing dyeing of cellulose fibers. In the present investigation, we analyze the dye–fiber affinity of 30 sulfonated anionic dyes in terms of relevant structural features and physicochemical properties. The dye molecules possess an extended planar π -electron system and substantial molecular size, which are known prerequisites for compounds to act as direct dyes.^{1,2} Moreover, the sulfonic acid group enhances their solubility in aqueous solution, in which they are prevalent in the anionic form under neutral and alkaline conditions.

From a thermodynamic viewpoint, the transfer of the dye from the dissolved state in aqueous solution to a state fixated at the fiber can be described as consisting of mainly two steps: the adsorption of dye molecules onto the substrate surface, and the dye penetration into the fiber with an accompanying build-up of binding dye–fiber and dye–dye interaction forces. Upon adsorption onto the polysaccharide 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

* Corresponding author e-mail: gs@uoe.ufz.de.

generally observed phenomenon for a broad range of compounds.⁴

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—fiber as well as dye—dye interactions. Here, both van der Waals forces covering orientation, induction, and dispersion interaction as well as site-specific hydrogen bonding are involved. In addition to favorable dye—fiber 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 fiber matrix.^{2,3}

As mentioned above, a specific feature of cellulose fibers is their hydrogen bond network between the polymer components, involving hydroxyl groups and ether oxygen (glycoside oxygen). It follows that dye molecules with hydrogen bond donor or acceptor sites may interact more distinctly with the cellulose fiber. As a consequence, the fixation of substantive dyes such as sulfonated 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 and possibly also between cellulose ether oxygen and SO_3H groups of the dye molecule (s.b.). Moreover, the planar π -electron structure supports favorable π - π interactions between dye molecules, which together with hydrogen-bond interactions are likely to contribute to the aggregate formation of the dye molecules trapped in the fiber matrix.

The present QSAR study is based on the hypothesis that the dye—fiber affinity in terms of a suitably defined free energy term is driven either by specific dye—fiber interactions^{1,2} or by dye—dye interactions³ or both. In this respect, the discussion is broader than previous QSAR investigations that focused on dye—fiber interactions only.^{5–7} These earlier studies employed in particular CoMFA (Comparative Molecular Field Analysis)⁸ as a 3D technique that allows the analysis of intermolecular interactions in terms of site-specific steric and electrostatic fields of the substrate.

In our previous QSAR analysis of the dye—fiber interaction of 30 sulfonated azo dyes,⁷ the CoMFA method had been combined with semiempirical quantum chemical AM1⁹ calculations, comparing different CoMFA alignment rules as well as both the protonated and dissociated form of the (most acidic) sulfonic acid group of the dye molecules. According to CoMFA PLS¹⁰ models (PLS = partial least squares), the differences in dye affinity for the cellulose polysaccharide were governed mainly by polar interactions and the highest molecular orbital energy of the dye molecules. The latter suggests that the electron donor capability of the azo dyes is important for their penetration and fixation in the fiber polymer, which may also reflect a respective impact of the hydrogen bond acceptor capability of the substrates.

A unique feature of the present investigation is the inclusion of solution-phase molecular descriptors. So far, QSAR studies of the dye affinities were confined to macroscopic physicochemical properties or calculated characteristics of the gas-phase electronic and geometric structure. Since sulfonic acid dyes are taken up from aqueous solution where they are prevalent in the dissociated form at neutral and alkaline pH, both their acidity and solvation energy may

contribute to the overall affinity for the textile fiber. Moreover, both electron donor and acceptor capabilities as well as hydrogen bond donor and acceptor capabilities are affected by aqueous solvation. To address these aspects and their potential impact on the dye affinity for the textile fiber, dielectric continuum-solvation calculations^{11,12} have been performed for the series of 30 sulfonic acid monoazo dyes, employing the semiempirical models COSMO-AM1,¹³ SM2,¹⁴ and SM5.4A.¹⁵ For compounds with more than one sulfonic acid group, the calculated gas-phase and solution-phase compound acidity refers to one deprotonated site that was selected based on AM1-level dissociation energies.

MATERIALS AND METHODS

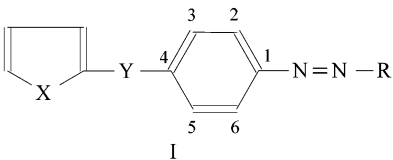
Experimental values for the dye—fiber affinity were taken from literature.¹⁶ These data are expressed as the difference between the standard chemical potential of the dye fixated in the textile fiber and the one dissolved in aqueous solution, $\Delta\mu^0$. For the sake of simplicity, the corresponding positive values, $-\Delta\mu^0$, are used for the QSAR analyses, since $-\Delta\mu^0$ increases with increasing dye—fiber affinity.

Molecular Modeling and CoMFA. Initial 3D molecular structures were generated using the SYBYL package¹⁷ with the TRIPOS force field, and all geometries were further optimized using the semiempirical quantum chemical AM1⁹ method as implemented in MOPAC 93.¹⁸ Aqueous solution was modeled in the continuum-solvation approach^{11,12} through application of the semiempirical methods COSMO-AM1¹³ and SM2¹⁴ as available in MOPAC and AMSOL 4.0.¹⁹ For the latter, AM1 gas-phase geometries were employed, while with COSMO a final solution-phase geometry optimization was performed due to the availability of analytical gradients. In view of some systematic differences between the COSMO and SM2 results as outlined below, the more recent continuum-solvation method SM5.4A¹⁵ including solution-phase geometry optimization was used to generate a further set of solvation energies for comparison, employing the software package AMSOL 6.5.3.²⁰

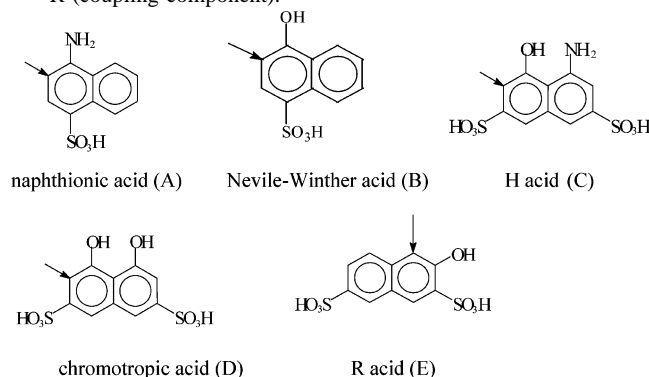
For the CoMFA alignment of the molecules (SYBYL option RMS_FIT), compound #1 as dye with the greatest affinity for the textile fiber was used as template, taking the azo nitrogen atoms as well as atoms C₁ and C₄ as anchor sites (see structural formula in Table 1 below). The CoMFA steric and electrostatic fields were computed using a sp³ carbon with charge +1, and partial atomic charges as provided by COSMO and SM2, using the standard CoMFA cutoff of 30 kcal/mol.

For the 3D grid to evaluate steric and electrostatic interactions between the probe atom and the dye molecules, boxes with dimensions 28 × 18 × 20 Å (COSMO) and 28 × 18 × 18 Å (SM2) and a grid spacing of 2 Å were used. Statistical analysis was carried out by the PLS method¹⁰ with the leave-one-out cross-validation procedure to determine the optimal number of components to be used in the final PLS models (without cross-validation). Variable selection was performed with a minimal σ (column filter) value of 2.00 kcal/mol and standard block scaling (CoMFA STD option of SYBYL).

Molecular Descriptors. Besides hydrophobicity in terms of the logarithmic octanol/water partition coefficient taken from our previous study,⁷ various characteristics of the

Table 1. Experimental Dye–Fiber Affinity¹⁶ in Terms of $-\Delta\mu^0$ [kJ/mol] and COSMO-AM1 Results of Anionic Azo Dyes (I)


no.	X	Y	R ^a	$-\Delta\mu^0$ [kJ/mol]	E_{HOMO}^b [eV]	E_{LUMO}^b [eV]	EN ^b [eV]	HARD ^b [eV]
1	-S-	-CH=CH-	A	15.80	-8.65	-1.14	4.90	3.76
2	-CH=CH-	-CH=CH-	A	14.25	-8.78	-1.06	4.92	3.86
3	-S-	-CONH-	A	13.08	-8.61	-1.13	4.87	3.74
4	-CH=CH-	-CONH-	A	12.00	-8.64	-1.11	4.88	3.76
5	-S-	-CH=CH-	B	9.66	-8.71	-1.47	5.09	3.62
6	-S-	-CH=CH-	C	9.45	-8.73	-1.78	5.26	3.48
7	-CH=CH-	-CH=CH-	B	9.20	-8.87	-1.45	5.16	3.71
8	-S-	-CONH-	C	9.03	-8.78	-1.65	5.21	3.56
9	-S-	-CO-	A	8.78	-8.83	-1.21	5.02	3.81
10	-CH=CH-	-CH=CH-	C	8.40	-8.69	-1.77	5.23	3.46
11	-CH=CH-	-CONH-	C	8.28	-8.74	-1.82	5.28	3.46
12	-S-	-CONH-	B	7.15	-8.96	-1.43	5.20	3.76
13	-S-	-CH=CH-	D	7.06	-8.73	-1.86	5.30	3.43
14	-CH=CH-	-CO-	A	7.02	-8.84	-1.22	5.03	3.81
15	-CH=CH-	-CONH-	B	6.52	-8.95	-1.44	5.19	3.76
16	-S-	-CH=CH-	E	6.27	-8.78	-1.59	5.19	3.59
17	-S-	-CONH-	D	6.23	-8.92	-1.79	5.36	3.56
18	-CH=CH-	-CH=CH-	D	6.02	-8.86	-1.87	5.36	3.49
19	-CH=CH-	-CH=CH-	E	5.81	-8.93	-1.6	5.26	3.66
20	-CH=CH-	-CONH-	D	5.18	-8.92	-1.8	5.36	3.56
21	-S-	-CONH-	E	5.10	-9.09	-1.51	5.30	3.79
22	-S-	-CO-	C	4.64	-8.85	-1.51	5.18	3.67
23	-CH=CH-	-CONH-	E	4.26	-9.05	-1.48	5.26	3.78
24	-S-	-CO-	B	4.22	-9.32	-1.49	5.40	3.91
25	-CH=CH-	-CO-	C	4.10	-8.86	-1.78	5.32	3.54
26	-CH=CH-	-CO-	B	4.05	-9.33	-1.5	5.41	3.91
27	-S-	-CO-	D	3.85	-9.09	-1.83	5.46	3.63
28	-CH=CH-	-CO-	D	3.43	-9.24	-1.87	5.56	3.68
29	-S-	-CO-	E	3.22	-9.44	-1.59	5.51	3.92
30	-CH=CH-	-CO-	E	2.84	-9.35	-1.61	5.48	3.87

^a R (coupling component):^b COSMO-AM1 results for the highest occupied molecular orbital energy, E_{HOMO} , the lowest unoccupied molecular orbital energy, E_{LUMO} , the molecular electronegativity, EN = $-1/2 (E_{\text{HOMO}} + E_{\text{LUMO}})$ and hardness, HARD = $-1/2 (E_{\text{HOMO}} - E_{\text{LUMO}})$.²¹

electronic structure of the compounds were quantified in the gas phase (AM1) or in simulated aqueous solution (COSMO, SM2).

The quantum chemical descriptors included HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) energies (E_{HOMO} , E_{LUMO}) of the compounds in the gas phase (AM1) as well as in aqueous solution (COSMO, SM2), and the associated parameters molecular electronegativity, EN = $-1/2 (E_{\text{HOMO}} + E_{\text{LUMO}})$ and hardness, HARD = $-1/2 (E_{\text{HOMO}} - E_{\text{LUMO}})$.²¹

Gas-phase acidity ΔH (AM1) as well as solution-phase acidities ΔH_{aq} (COSMO) and ΔG_{aq} (SM2) of the dye molecules were calculated as method-specific differences in enthalpy or free energy between the anion (A^-) and the neutral form (AH) of the compound, omitting the constant energies of H_2O and H_3O^+ (when considering the proton-transfer $\text{AH} + \text{H}_2\text{O} \rightarrow A^- + \text{H}_3\text{O}^+$)²² for the sake of simplicity. Moreover, solvation energies of the dye compounds in aqueous solution were quantified in terms of COSMO solvation enthalpies, ΔH_s , and SM2 free energies of solvation, ΔG_s , at 25 °C. In addition, SM5.4A free energies of solvation were calculated for comparison (s.a.).

Model Calibration and Validation. Besides PLS analyses to derive the CoMFA models using the SYBYL package, multiple linear regression (MLR) has been performed by the STATISTICA program,²³ applying standard criteria for a stepwise variable selection. The calibration quality of both MLR and PLS models has been characterized using the squared correlation coefficient

$$r^2 = 1 - \frac{\text{RSS}}{\text{SS}} \quad (1)$$

and standard error

$$\text{SE} = \sqrt{\frac{\text{RSS}}{n-p}} \quad (2)$$

with SS (sum of squares) and RSS (residual sum of squares) being defined as

$$\text{SS} = \sum_{i=1}^n (y_i - y_0)^2 \quad (3)$$

and

$$\text{RSS} = \sum_{i=1}^n (y_i - y_i^{\text{fit}})^2 \quad (4)$$

Here, n is the number of compounds (in our case: $n = 30$), p is the number of model parameters (e.g. $p = 3$ for a 2-variable MLR), y_i is the i th experimental value (in our case: dye–fiber affinity $-\Delta\mu^0$ [kJ/mol]), y_0 is the respective mean, and y_i^{fit} is the i th calculated value using an MLR or PLS model calibrated with the total set of compounds.

As a rough measure of the prediction capability, the predictive squared correlation coefficient

$$q^2 = 1 - \frac{\text{PRESS}}{\text{SS}} \quad (5)$$

and standard error of prediction

$$\text{SEP} = \sqrt{\frac{\text{PRESS}}{n-p}} \quad (6)$$

with

$$\text{PRESS} = \sum_{i=1}^n (y_i - y_i^{\text{calc}})^2 \quad (7)$$

(predictive residual sum of squares) based on the leave-one-out scheme were used. In eq 7, y_i^{calc} denotes the i th predicted value of the i th submodel calibrated without the i th experimental value. While the leave-one-out approach is not

conservative but tends to increasingly overestimate the prediction power of regression models with increasing number of compounds,²⁴ it is still a reasonable method for smaller data sets, particularly in cases where the compounds do not belong to just one congeneric set, as is also the case in the present investigation.

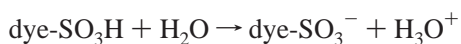
For comparison, additional q^2 and SEP tests of the prediction capability were performed based on a leave-5-out approach for all finally reported multilinear regression models, considering that the compound set consists of five structural subclasses A–E (see Table 1 below) each of which contain six compounds. To achieve a reasonable decomposition of the data set for the leave-5-out approach, the following strategy was applied: First, the compounds were ordered according to decreasing dye affinity, leading to a respective ordering within each of the five subclasses A–E. Then, the compounds from subclasses A and C were allocated to six leave-5-out subsets (group 1 to group 6) in the order of decreasing dye affinity (the two compounds of subclasses A and C with greatest dye affinity were allocated to group 1, those of second-greatest dye affinity to group 2 etc.), while for subclasses B and D the opposite ordering was applied (the two compounds of B and D with lowest dye affinity were allocated to group 1 etc.). Finally, for subclass E an ad hoc allocation procedure was applied, aiming at an overall balanced distribution of the compounds into the leave-5-out subsets with respect to dye affinity. The resultant allocation of the compounds to the six leave-5-out subsets is specified in the legend of Table 3 (see below).

RESULTS AND DISCUSSION

The experimental dye affinities¹⁶ listed in Table 1 show a variation in the free energy of fixation at the fiber, $\Delta\mu^0$, from -15.8 kJ/mol to -2.48 kJ/mol. This corresponds to equilibrium constants K of 585 to 3 quantifying the ratio of dye concentration in the fiber to the one in aqueous solution ($\Delta\mu^0 = -RT^* \ln K$) and indicates a 200-fold difference in the uptake efficiency of dyes from the dye bath to the textile.

As can be seen from Table 2, the calculated gas-phase acidity of the dye molecules, ΔH (AM1), varies by 71.7 kJ/mol (from -225.2 kJ/mol to -153.5 kJ/mol), while the calculated solution-phase acidities, ΔH_{aq} and ΔG_{aq} , cover differences of 24 kJ/mol (COSMO) and 43.6 kJ/mol (SM2), respectively. Note, however, that ΔH , ΔH_{aq} , and ΔG_{aq} are confined to the difference in energy between anion and acid, omitting the energies associated with H_2O and H_3O^+ that are needed for describing the complete energy balance of the proton transfer from the sulfonic acid to H_2O (s.a.).

Considering experimental values for the gas-phase heats of formation of H_2O and H_3O^+ , $H_f(\text{H}_3\text{O}^+) = 591$ kJ/mol and $H_f(\text{H}_2\text{O}) = -241.8$ kJ/mol,²⁵ the gas-phase energy for the proton transfer from the dye molecule to H_2O



would be higher than the ΔH (AM1) entries in Table 2 by 832.8 kJ/mol and thus range from 607.6 kJ/mol to 679 kJ/mol. Correspondingly, inclusion of both the above-mentioned gas-phase heats of formation of H_3O^+ and H_2O as well as their solvation free energies in aqueous solution, $\Delta G_s(\text{H}_3\text{O}^+) = -435.1$ kJ/mol and $\Delta G_s(\text{H}_2\text{O}) = -26.4$ kJ/mol,²⁶ would

Table 2. Calculated AM1 Heat of Formation (H_f) and Gas-Phase Dissociation Energy (ΔH), COSMO Solvation Enthalpy (ΔH_s) and Solution-Phase Dissociation Energy (ΔH_{aq}), and SM2 Solvation Free Energy (ΔG_s) and Solution-Phase Free Energy of Dissociation (ΔG_{aq})^a

no.	AM1		COSMO		SM2	
	H_f [kJ/mol]	ΔH [kJ/mol]	ΔH_s [kJ/mol]	ΔH_{aq} [kJ/mol]	ΔG_s [kJ/mol]	ΔG_{aq} [kJ/mol]
1	167.5	-157.6	-197.9	-427.6	-165.4	-361.7
2	149.5	-157.4	-194.7	-428.6	-100.3	-362.2
3	-3.7	-153.5	-224.1	-428.1	-203.5	-361.6
4	-25.8	-153.8	-221.6	-428.5	-119.3	-360.8
5	-0.7	-172.7	-174.8	-438.4	-157.7	-373.2
6	-483.8	-203.0	-301.4	-439.7	-257.7	-332.6
7	-19.0	-172.7	-173.6	-436.9	-92.5	-374.3
8	-649.6	-199.6	-347.1	-429.8	-294.7	-334.5
9	3.2	-164.4	-215.6	-428.4	-186.5	-360.4
10	-494.9	-203.4	-315.1	-440.1	-188.8	-333.6
11	-679.0	-198.7	-331.1	-439.1	-205.9	-332.6
12	-172.2	-168.9	-204.1	-434.5	-196.0	-371.3
13	-654.7	-210.2	-317.0	-429.1	-257.9	-332.6
14	-16.8	-165.0	-211.5	-429.2	-110.0	-360.5
15	-194.3	-169.1	-200.4	-437.6	-111.9	-374.0
16	-482.5	-202.3	-293.1	-444.1	-262.4	-333.5
17	-827.2	-216.4	-338.4	-441.6	-298.9	-338.4
18	-677.2	-209.1	-305.7	-442.9	-192.5	-331.9
19	-501.3	-201.8	-293.1	-439.1	-199.1	-331.7
20	-849.5	-216.4	-338.5	-434.7	-207.3	-338.8
21	-654.4	-199.5	-336.6	-426.2	-301.7	-333.4
22	-639.9	-209.1	-334.1	-444.4	-257.8	-330.7
23	-673.1	-202.1	-331.2	-443.4	-210.7	-331.9
24	-160.8	-182.8	-185.6	-450.2	-162.1	-368.9
25	-665.2	-207.7	-316.7	-437.9	-200.7	-332.9
26	-180.4	-182.1	-181.4	-448.7	-103.7	-374.3
27	-819.7	-225.2	-333.4	-440.2	-258.8	-336.8
28	-841.4	-217.7	-323.4	-437.2	-202.0	-331.7
29	-645.8	-207.1	-320.9	-430.8	-262.0	-332.9
30	-669.6	-208.1	-324.0	-431.9	-207.1	-334.1

^a COSMO-AM1 results refer to molecular geometries optimized in the solution phase, while SM2 results refer to AM1 geometries (cf. section Materials and Methods). The dissociation energies are calculated as energy differences between the acid (AH) and anion (A⁻), omitting constant energy terms for H_2O and H_3O^+ as discussed in the text.²²

shift upward the table entries of ΔH_{aq} (COSMO) and ΔG_{aq} (SM2) by $832.8 - 408.7$ kJ/mol = 424.1 kJ/mol, resulting in calculated solution-phase proton-transfer energies of -26.1 kJ/mol to -3.5 kJ/mol (COSMO) and 50.9 kJ/mol to 93.4 kJ/mol (SM2), respectively. The latter values could, in principle, be directly converted to $\text{p}K_a$ data through the formula $\Delta G_{\text{aq}} = 5.71 \text{ p}K_a + 9.94$.²² However, the substantial differences between the absolute values derived from COSMO and SM2 confirm earlier findings^{22,27–29} that with current continuum-solvation parametrizations, absolute predictions of $\text{p}K_a$ are generally not feasible. Instead, correction of absolute and relative scales of the predicted dissociation energies through regression on experimental data would be needed to arrive at realistic values for the solution-phase $\text{p}K_a$. The latter, however, is not needed for the present investigation, where relative acidity values are sufficient to address the question whether the intrinsic (gas-phase) or solution-phase acidity of the dye molecules has some impact on their affinity for the cellulose fiber.

As regards solvation energies, ΔH_s (COSMO) and ΔG_s (SM2) yield energy ranges of -347.1 to -173.6 kJ/mol and -301.7 to -92.5 kJ/mol, respectively. Apart from expected differences between COSMO enthalpies (that are in fact confined to electrostatic contributions) and SM2 free energies

Table 3. Multilinear Regression Equations for the Dye–Fiber Affinity in Terms of $-\Delta\mu^0$ [kJ/mol] with Quantum Chemical Gas-Phase and Solution-Phase Parameters^a

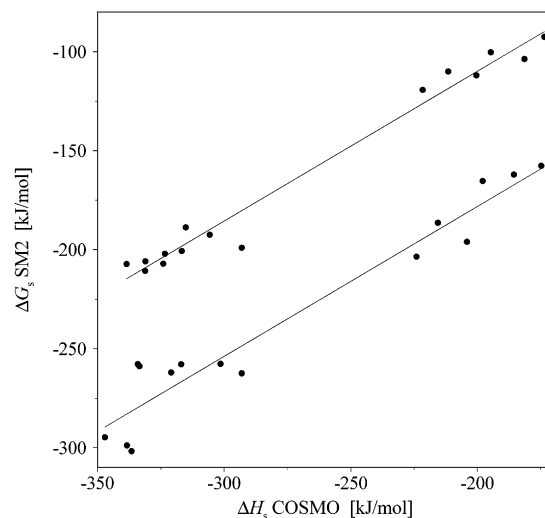
MLR model no.	regression equation $-\Delta\mu^0 =$	calibration			leave-1-out		leave-5-out ^b	
		r^2	SE	F ratio	q^2	SEP	q^2	SEP
1	$5.9 (\pm 1.1) E_{\text{HOMO}} + 4.4 (\pm 1.1) E_{\text{LUMO}} + 67.1 (\pm 9.0)$	0.77	1.66	44.9	0.71	1.87	0.73	1.80
2	$9.6 (\pm 1.4) E_{\text{HOMO}} (\text{COSMO}) + 6.2 (\pm 1.2) E_{\text{LUMO}} (\text{COSMO}) + 102.1 (\pm 11.7)$	0.78	1.61	48.9	0.73	1.79	0.74	1.78
3	$-10.3 (\pm 1.1) \text{EN} + 61.7 (\pm 5.7)$	0.76	1.65	90.7	0.72	1.80	0.73	1.77
4	$-15.6 (\pm 1.7) \text{EN} (\text{COSMO}) + 88.6 (\pm 8.7)$	0.76	1.67	87.5	0.72	1.80	0.72	1.79
5	$6.0 (\pm 1.0) E_{\text{HOMO}} + 0.065 (\pm 0.015) \Delta H + 73.1 (\pm 8.2)$	0.78	1.61	49.1	0.73	1.81	0.75	1.71
6	$8.4 (\pm 1.3) E_{\text{HOMO}} (\text{COSMO}) + 0.077 (\pm 0.014) \Delta H + 97.2 (\pm 11.3)$	0.80	1.53	55.5	0.76	1.71	0.76	1.71
7	$-9.4 (\pm 1.1) \text{EN} + 0.098 (\pm 0.047) \Delta H_{\text{aq}} (\text{COSMO}) + 100.2 (\pm 19.0)$	0.80	1.56	53.2	0.68	1.94	0.74	1.77
8	$6.5 (\pm 0.9) E_{\text{HOMO}} + 0.016 (\pm 0.005) \Delta H_{\text{s}} (\text{COSMO}) + 0.15 (\pm 0.04) \Delta H_{\text{aq}} (\text{COSMO}) + 134.1 (\pm 18.7)$	0.81	1.53	37.6	0.76	1.74	0.74	1.79
9	$9.2 (\pm 1.3) E_{\text{HOMO}} (\text{COSMO}) + 0.021 (\pm 0.005) \Delta H_{\text{s}} (\text{COSMO}) + 0.14 (\pm 0.04) \Delta H_{\text{aq}} (\text{COSMO}) + 156.2 (\pm 19.6)$	0.81	1.54	36.6	0.75	1.76	0.67	1.97
10	$234.4 (\pm 33.8) E_{\text{HOMO}} + 12.6 (\pm 1.9) E_{\text{HOMO}}^2 + 3.0 (\pm 0.7) E_{\text{LUMO}} + 1097.6 (\pm 152.4)$	0.92	1.02	94.9	0.89	1.19	0.88	1.23
11	$231.3 (\pm 34.0) E_{\text{HOMO}} + 12.6 (\pm 1.9) E_{\text{HOMO}}^2 - 6.1 (\pm 1.5) \text{EN} + 1097.4 (\pm 152.3)$	0.92	1.02	95.0	0.89	1.19	0.88	1.23

^a AM1 (if not specified) or COSMO as indicated. E_{HOMO} = highest occupied molecular orbital energy [eV], E_{LUMO} = lowest unoccupied molecular orbital energy [eV], EN = molecular electronegativity [eV], ΔH = AM1 dissociation energy [kJ/mol], ΔH_{aq} = COSMO dissociation energy [kJ/mol], ΔH_{s} = COSMO solvation enthalpy [kJ/mol] (cf. Table 2 and section Materials and Methods). Statistical parameters: r^2 = squared correlation coefficient (eq 1), SE = standard error (eq 2), $F = F_{1,28}$ (MLR model #1–2) or $F_{2,27}$ (#3–7) or $F_{3,26}$ (#8–11), q^2 = predictive squared correlation coefficient (eq 5, leave-1-out and leave-5-out, respectively), SEP = predictive standard error (eq 6). For each model parameter, the associated standard error is indicated in parentheses. ^b The leave-5-out statistics have been generated using the following 5 groups that each consist of 6 different compounds (compound numbering according to Table 1): group 1 = #1, 6, 19, 26, 28; group 2 = #2, 8, 16, 24, 27; group 3 = #3, 10, 15, 20, 30; group 4 = #4, 11, 12, 18, 23; group 5 = #7, 9, 17, 22, 29; group 6 = #5, 13, 14, 21, 25.

(that contain also parametrized contributions for dispersion interaction and cavity formation), the variation in the energy values differs also significantly (173.5 kJ/mol vs 209.2 kJ/mol). With SM5.4A (results not shown in Table 2) the predicted energy gain upon solvation is smaller (from -201.1 kJ/mol to -85.4 kJ/mol) and shows a substantially smaller range of variation (115.7 kJ/mol) for the 30 compounds. At the same time, the SM5.4A solvation energies correlate significantly better with the COSMO solvation enthalpies ($r^2 = 0.95$) than with the corresponding SM2 results ($r^2 = 0.73$). These findings indicate that also for predicting absolute solvation energies (and associated Henry's law constants), at least a regression fit would be needed to scale the calculated values, as was discussed in earlier investigations including both semiempirical and ab initio calculation schemes.^{30,31}

Solvation Energy and Dye Affinity. In aqueous solution, the free energy of solvation can be understood as balance between different and partly opposing terms such as cavity formation energy, dispersion forces, site-specific Coulomb interactions, and hydrogen bonding. With structurally similar compounds, the increase in molecular size would be expected to decrease the energy gain upon solvation due to the energy penalty associated with forming the solute cavity. For the present set of 30 azo dyes, there is indeed a highly significant correlation between solvation energy and molecular weight as indicated by r^2 values of 0.96 (ΔH_{s} COSMO), 0.67 (ΔG_{s} SM2), and 0.97 (SM5.4A), respectively. It suggests that for this compound class, the variation in aqueous solvation energy is dominated by size-related components such as dispersion forces and cavity formation energy (except for differences due to different numbers of SO_3H groups, s.b.).

As regards the electrostatic contributions, increasing the polarity is usually expected to increase the energy stabilization upon solvation, although this approximate view does not account for peculiarities of the electronic structure such

**Figure 1.** SM2 solvation free energy (ΔG_{s}) vs COSMO solvation enthalpy (ΔH_{s}) for the set of 30 azo dyes. The regression lines refer to the subsets of 15 thioether compounds (bottom) and the 15 compounds without a thioether function (top).

as higher-order moments (quadrupole etc.) and their response to external fields (polarizability).³⁰ For the present set of compounds, there is indeed almost no correlation between the solvation energy and the AM1 dipole moment (r^2 below 0.1).

The direct comparison between ΔH_{s} (COSMO) and ΔG_{s} (SM2) is shown in Figure 1. A detailed inspection of the data distribution reveals the following features: First, there is a clear separation between the 12 compounds with one sulfonic acid group on the right-hand side of Figure 1 and the 18 compounds with two such groups in their coupling component on the left-hand side. It shows the importance of the SO_3H group for the overall solvation energy and reflects the well-known fact that sulfonic acid groups usually

enhance the aqueous solubility of organic compounds. Second, the subset of 15 azo dyes with a thioether bridge (compounds 1, 3, 5, 6, etc.) shows systematically greater negative SM2 solvation free energies than the other 15 dyes (compounds 2, 4, 7, 10, etc.). It reveals an apparent artifact of SM2 that seems to overestimate the stabilizing solvation contribution of $-S-$ (thioether) units significantly. This interpretation is further supported by the fact that within each of these two complementary subgroups of 15 compounds, there is a high correlation between the COSMO and SM2 results ($r^2 = 0.94$ for thioether compounds, $r^2 = 0.98$ for other compounds). It follows that for the remainder of our investigation, we will focus on COSMO when analyzing the potential impact of aqueous solvation on the electronic structure of the dye and its affinity for the cellulose fiber. Note further that with SM5.4A¹⁵ as a more recently parametrized scheme of the SMx family of methods,¹² the 15 thioether compounds do not show any excess energy stabilization through aqueous solvation, and for all 30 compounds the correlation with COSMO is high as mentioned above.

Detailed inspection of the dependence of solvation energy upon molecular structure reveals further that besides the sulfonic acid group, the amid group as a bridge between the aromatic rings (fragment Y in Table 1) is also associated with a systematically greater energy stabilization through aqueous solvation.

Coming back to the initial hypothesis of a possible impact of solvation energy on dye affinity, the corresponding statistical analysis shows that for the present set of 30 anionic azo dyes, the energy gain upon aqueous solvation does not correlate as a single parameter with the affinity for the textile fiber. These findings suggest that the energy penalty associated with the desolvation of the dye molecules upon transfer from aqueous solution into the cellulose polymer is not a major factor for the discrimination between low and high dye–fiber affinity. As shown below, however, stepwise multilinear regression identifies ΔH_s (COSMO) as a significant parameter in 3-variable equations to model $-\Delta\mu^0$ (s.b., Table 3).

Dye Acidity and Affinity for the Textile Fiber. In Figure 2, gas-phase and solution-phase acidity is compared for all 30 dyes, using AM1 and COSMO-AM1 as calculation schemes. The left part of the plot contains all dyes with two sulfonic acid groups in their coupling component. For this subset of 18 compounds with ΔH values between -230 and -195 kJ/mol, aqueous solution introduces significant changes in the intrinsic acidity trend, such that there is indeed no correlation ($r^2 = 0$) between ΔH and ΔH_{aq} (COSMO).

By contrast, the gas-phase and solution-phase acidity correlate pretty well for the 12 dyes with only one sulfonic acid group (right-hand part of Figure 2 with ΔH range of -190 kJ/mol to -150 kJ/mol, $r^2 = 0.87$), and here omission of the first four compounds in the top right corner of the plot increases the r^2 to 0.97. The analysis shows that with sulfonic acid dyes, the trend in intrinsic acidity may be significantly affected by aqueous solvation.

Since dye fixation in the fiber is accompanied by successive desolvation, this process may in principle be influenced by both intrinsic and solution-phase acidity. To maintain electrical neutrality, the SO_3^- groups of the dye molecules become either shielded through according counterions (e.g.

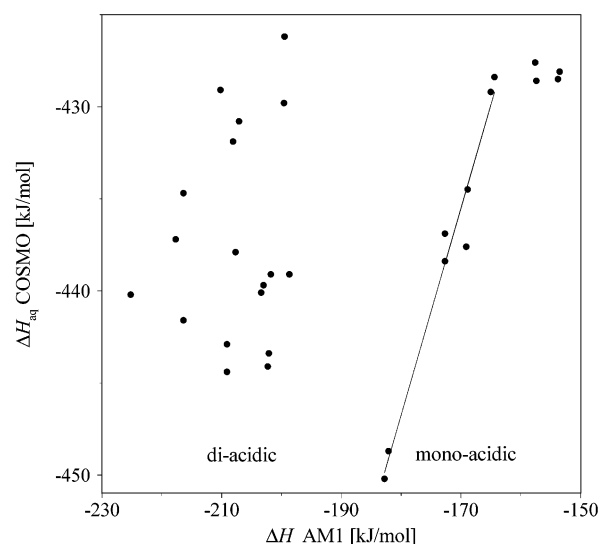


Figure 2. COSMO solution-phase acidity (ΔH_{aq}) vs AM1 gas-phase acidity (ΔH) for the set of 30 azo dyes. Among the 15 compounds with one sulfonic acid group on the right-hand side of the plot, omission of compounds no. 1–4 (see Table 1) in the top right yields the regression line as indicated with an r^2 of 0.97.

Na^+) or they finally become protonated, participating in hydrogen bonds to the cellulose ether oxygen or hydroxyl oxygen according to



linkages. Here, the sulfonic O–H bond strength is primarily related to the dye's intrinsic acidity, but peculiarities of the aqueous solvation may well affect the probability to form such coordinative $O \cdots H-O$ bridges. Note further that the formation of dye aggregates may also involve coordinative linkages between protonated sulfonic acid groups and hydrogen bond acceptor groups such as the azo bridge, implying that the azo dyes would act both as hydrogen bond donors and acceptors.

Nonetheless, comparison of $-\Delta\mu^0$ with ΔH (AM1) and ΔH_{aq} (COSMO) reveals that there is no overall correlation between dye–fiber affinity and compound acidity. Only for the subset of 12 dye molecules with one sulfonic acid function, $-\Delta\mu^0$ generally increases with decreasing compound acidity in the gas phase ($r^2 = 0.73$) and in aqueous solution ($r^2 = 0.61$). Here, the two most acidic compounds #24 and #26 have the lowest affinities to the cellulose fiber ($-\Delta\mu^0$ values of 4.22 and 4.05 kJ/mol, respectively), while compounds #1 and #2 with greatest $-\Delta\mu^0$ values (15.80 kJ/mol, 14.25 kJ/mol) have the lowest acidities both in the gas phase and in solution.

Coming back to the total set of 30 dyes with different numbers of acidic groups, compound acidity clearly has no dominating influence on the dye–fiber affinity. However, multilinear regression of $-\Delta\mu^0$ as shown below reveals that the intrinsic molecular acidity becomes a significant descriptor in combination with the solution-phase HOMO energy.

Gas-Phase vs Solution-Phase Electronic Structure and Dye Affinity. Aqueous solution affects not only the acidity of the azo dyes but also subtle electronic characteristics such as polarity, polarizability, and electron donor and acceptor as well as hydrogen bond donor and acceptor capabilities. This is demonstrated by comparing gas-phase values for the

dipole moment, molecular electronegativity EN, and hardness HARD as well as for E_{HOMO} and E_{LUMO} with the solution-phase counterparts as calculated with COSMO and SM2. The respective r^2 values are 0.14 and 0.22 (dipole moment), 0.75 and 0.54 (HARD), 0.87 and 0.59 (EN), 0.86 and 0.44 (E_{HOMO}), and 0.77 and 0.91 (E_{LUMO}). Similar to the situation with the solvation energy as discussed above, SM2 yields systematically lower EN values for the 15 thioether derivatives as compared to the other 15 compounds without such $-S-$ units.

The differences in r^2 for the frontier orbitals show that the impact of aqueous solvation upon the electron donor capability of the azo dyes is slightly different to the one on their capability to interact as electron acceptor. Note, however, that in the SM2 scheme, frontier orbital energies have a somewhat different meaning, because in this method Koopmans' theorem (according to which E_{HOMO} and E_{LUMO} are directly related to the ionization potential and electron affinity, respectively) is no longer valid.¹²

Comparison of the dye–fiber affinity, $-\Delta\mu^0$, with E_{HOMO} and E_{LUMO} reveals that for both the gas-phase and solution-phase parameters, the overall data distribution is nonlinear and contains some clustering. However, regression of $-\Delta\mu^0$ on both E_{HOMO} and E_{LUMO} yields reasonable statistics ($r^2 = 0.77$, q^2 (leave-one-out) = 0.71), with a slight preference for the COSMO model ($r^2 = 0.78$, q^2 (leave-one-out) = 0.73; see Table 3). At the same time, the intercorrelation between both parameters is quite low (AM1: $r^2 = 0.28$; COSMO: $r^2 = 0$). In case of the COSMO-based MLR model #2 (Table 3), the difference between the E_{HOMO} and E_{LUMO} coefficient is just significant with respect to the associated standard errors, indicating a respective difference to the simple sum $E_{\text{HOMO}} + E_{\text{LUMO}}$ that is directly proportional to the molecular electronegativity EN.

Since E_{HOMO} increases with increasing electron donor capacity and E_{LUMO} decreases with increasing electron acceptor capacity, MLR models #1 and 2 imply that the dye–fiber affinity increases with increasing electron donor and decreasing electron acceptor strength of the dye molecules. Note that this finding does not fit well to the hypothesis that the fixation of substantive dyes is exclusively driven by a proper dye–dye aggregation in the cavities of the cellulose micelles.³ In such a case, intermolecular association would be expected to be favored by both strong donor and acceptor capabilities of the molecular components, which would result in E_{HOMO} and E_{LUMO} regression coefficients of similar size but opposite sign. Alternatively, the association tendency of the dye molecules would also increase with either increasing electron donor capacity (keeping acceptor strength essentially constant) or increasing electron acceptor capacity (keeping donor strength more or less constant), which is, however, also not in accord with MLR models #1 and 2. As a consequence, the present findings support the view that for the dye fixation in the cellulose matrix, stabilizing dye–fiber interactions are important.

At the same time, however, it seems also unclear why a lowered readiness to accept electrons should strengthen favorable dye–fiber interactions. A possible explanation could be that despite the above-mentioned difference in size between the E_{HOMO} and E_{LUMO} regression coefficients, the actually relevant property might be the closely related molecular electronegativity EN.

When using EN as single variable, the regression statistics are of similar quality (MLR models #3 and 4 in Table 3). While the regression fit suggests a very small preference for the gas-phase parameter (SE values of 1.65 vs 1.67), the leave-one-out prediction results in essentially identical statistics (SEP = 1.80 in both cases). For both models, the three greatest calibration outliers are compounds #1, #2, and #15, and further three compounds still have fit errors above 2 kJ/mol (#12, #14, #23).

The negative EN regression coefficient indicates that the dye affinity for the textile fiber increases with decreasing electronegativity. It means that the dye fixation in the cellulose polysaccharide is favored by lowering the molecule's tendency to attract electron charge or by increasing its electron donor capacity as was already seen with E_{HOMO} .

Interestingly, the above-mentioned hydrogen-bonding interactions provide a further possibility to interpret the regression results: Decreasing EN is in accord with increasing the capability of the azo bridge to act as H bond acceptor according to



linkages as well as according to respective dye–dye H-bond linkages. By contrast, increasing the H bond donor capacity at respective sites of the dye molecules such as $-\text{SO}_3\text{H}$, $-\text{OH}$, and $-\text{NH}_2$ would imply an increased electronegativity, because donation of positively charged H results in excess negative charge at the heavy atom connected to that H (hydrogen bond donor sites accept electronic charge, and hydrogen bond acceptor sites donate electronic charge). As a consequence, the negative correlation between $-\Delta\mu^0$ and EN also suggests that the variation in the H bond capability of the azo group is a crucial parameter for the fixation of the azo dyes in the cellulose polymer.

Molecular hardness does not correlate at all with $-\Delta\mu^0$, which indicates that for the fixation of sulfonated azo dyes in the polysaccharide, the global polarizability of the dye molecules is not an important factor.

Multilinear regression of $-\Delta\mu^0$ on both E_{HOMO} (AM1 or COSMO) and the gas-phase acidity ΔH (AM1) results in leave-one-out q^2 values of 0.73 and 0.76, respectively (MLR models #5 and 6 in Table 3, with squared intercorrelation coefficients between E_{HOMO} and ΔH as well as between E_{HOMO} (COSMO) and ΔH of 0.24 and 0.13, respectively). As before, the positive E_{HOMO} coefficients indicate an increased dye–fiber affinity with increasing electron donor capacity or increasing H-bond acceptor strength, and the positive sign of the ΔH regression coefficient shows that the dye fixation in the cellulose matrix is favored by a decreased acidity of the dye molecule.

Note further that MLR model #6 is one example for a regression equation that includes both intrinsic (gas-phase) and solution-phase properties of the azo dyes and is the statistically best two-variable regression relationship for $-\Delta\mu^0$ when employing AM1 and COSMO parameters (leave-one-out: $q^2 = 0.76$, SEP = 1.71). Similar calibration statistics but significantly inferior leave-one-out prediction statistics are achieved with EN (AM1) and ΔH_{aq} (COSMO) as two MLR variables ($q^2 = 0.68$, SEP = 1.94). Here, the positive sign of the ΔH_{aq} regression coefficient shows again

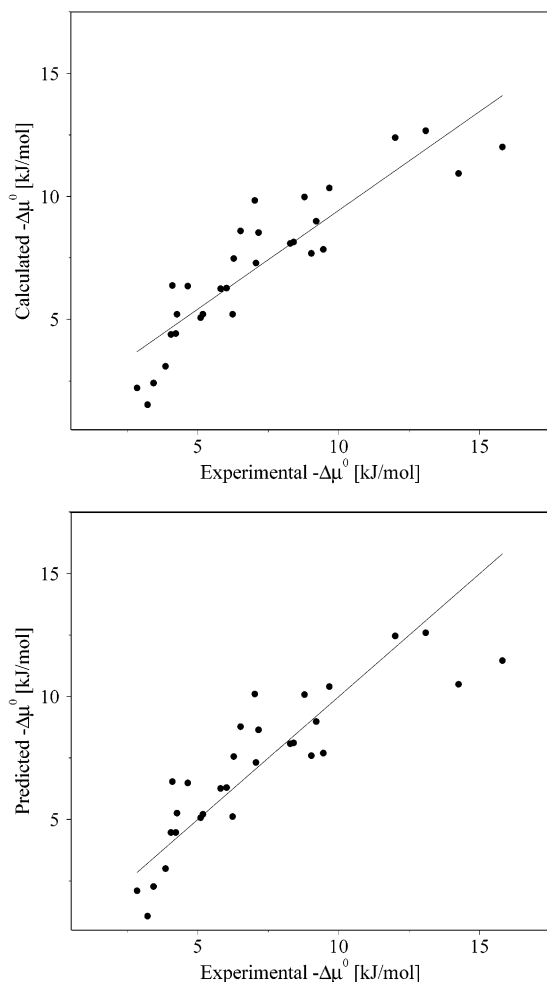


Figure 3. Calculated (top) and leave-one-out-predicted (bottom) vs experimental dye–fiber affinity in terms of $-\Delta\mu^0$ [kJ/mol] for the 30 azo dyes according to MLR model #6 (Table 3). The calibration plot includes the MLR regression line (top), while in the prediction plot, the $y = x$ line is shown (bottom).

that a large acidity of the dye tends to weaken its affinity for the textile fiber.

Among the three-variable regression equations, the combination of E_{HOMO} , ΔH_{aq} (COSMO), and ΔH_s (COSMO) yields the best statistics (leave-one-out: $q^2 = 0.76$, SEP = 1.74), and a similar result is achieved when replacing E_{HOMO} by its solution-phase counterpart (MLR models #8 and 9 in Table 3; for both equations, all squared intercorrelations are below 0.09). According to the positive regression coefficient of ΔH_s , $-\Delta\mu^0$ tends to be greater if the energy gain through aqueous solvation is relatively small. Interestingly, for the subset of 12 dyes with one sulfonic acid group, the two-variable regression of $-\Delta\mu^0$ on ΔH and ΔH_s (COSMO) yields highly significant statistics ($r^2 = 0.92$, SE = 1.06) and a corresponding interpretation of the combined impact of acidity and solvation energy on the dye–fiber affinity.

In Figure 3, the data distribution of MLR model #6 is shown for both the calibration (top) and leave-one-out prediction (bottom) mode. As can be seen from the figure, both plots contain some nonlinear trend that is not captured by the present multilinear regression model. This deviation from linearity is also observed with the other MLR models, indicating that there might be room for improvement through inclusion of quadratic terms or of other molecular descriptors.

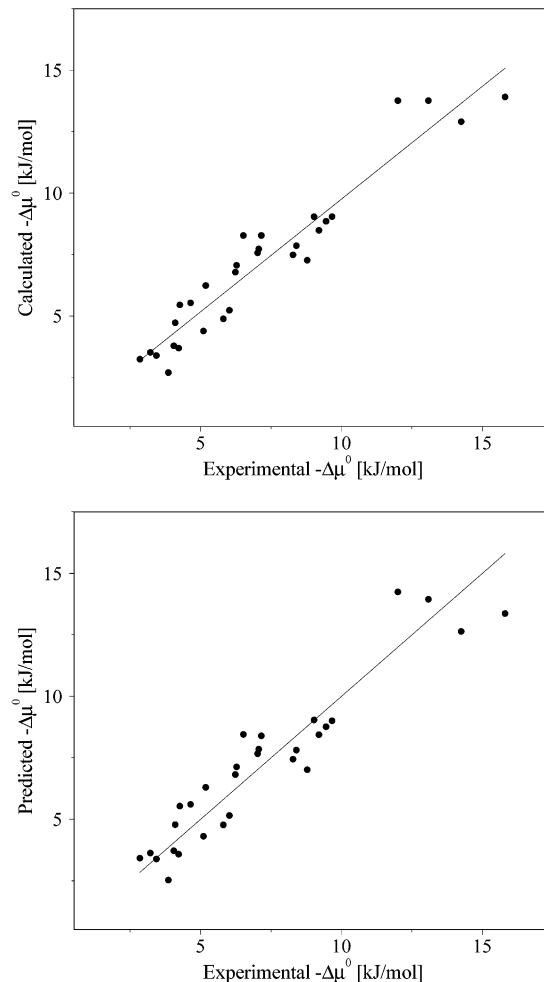


Figure 4. Calculated (top) and leave-one-out-predicted (bottom) vs experimental dye–fiber affinity in terms of $-\Delta\mu^0$ [kJ/mol] for the 30 azo dyes according to MLR model #10 (Table 3). The calibration plot includes the MLR regression line (top), while in the prediction plot, the $y = x$ line is shown (bottom).

The former strategy was tested, resulting in MLR models #10 and 11 as statistically best respective regression equations, employing E_{HOMO}^2 as quadratic term (leave-one-out: $q^2 = 0.89$, SEP = 1.19; cf. Table 3; note, however, the substantial squared intercorrelation between E_{HOMO} and EN of 0.77). With these models, the distribution of calculated vs experimental data has a linear shape, as is demonstrated for MLR model #10 in Figure 4. However, the physical meaning of including E_{HOMO}^2 is not clear, since it implies (considering the positive sign of the regression coefficient) that below a certain electron donor strength (that decreases with decreasing E_{HOMO}), the readiness of the dye molecule for undergoing favorable interactions with the fiber components or with other dye molecules increases with further decreasing E_{HOMO} . Note also that with COSMO, the respective statistics are significantly inferior (although still improved as compared to the two-variable regression equations; results not shown).

In the last two columns of Table 3, the leave-5-out q^2 and SEP values are list for all 11 MLR models. Surprisingly, for models #1–7 the leave-5-out prediction statistics are even slightly superior to the ones of the leave-1-out approach. It probably indicates some kind of fortuitous error compensation associated with the selection procedure of the six subsets

Table 4. PLS Models for the Dye–Fiber Affinity in Terms of $-\Delta\mu^0$ [KJ/Mol] Based on Solution-Phase CoMFA and Additional Gas-Phase and Solution-Phase Quantum Chemical Parameters^a

PLS model no.		no. of latent variables ^c	calibration			leave-1-out	
			r^2	SE	F ratio	q^2	SEP
1	CoMFA (S: 0.330, E: 0.670)	3	0.95	0.77	174.0	0.63	2.14
2	CoMFA (S: 0.214, E: 0.482), E_{HOMO} (0.305)	2	0.86	1.29	84.1	0.66	2.02
3	CoMFA (S: 0.052, E: 0.111), E_{HOMO} (0.439), E_{LUMO} (0.398)	1	0.78	1.59	99.8	0.73	1.76
4	CoMFA (S: 0.081, E: 0.173), EN (0.784)	1	0.78	1.59	99.7	0.73	1.76
5	CoMFA (S: 0.193, E: 0.462), E_{HOMO} (COSMO, 0.345)	2	0.84	1.37	72.1	0.70	1.89
6	CoMFA (S: 0.044, E: 0.155), E_{HOMO} (COSMO, 0.443), E_{LUMO} (COSMO, 0.359)	1	0.80	1.52	111.0	0.74	1.72
7	CoMFA (S: 0.311, E: 0.410), EN (COSMO, 0.279)	3	0.97	0.68	240.5	0.75	1.77

^a For the molecular descriptors and statistical parameters see legend of Table 3. CoMFA refers to COSMO, while the additional molecular descriptors have been calculated using AM1 (if not specified) or COSMO as indicated. ^b The values in parentheses specify the fraction of the respective descriptor in the PLS model. For CoMFA, the fractions of the steric (S) and electrostatic (E) field are given separately. ^c The optimal number of latent variables according to stepwise PRESS statistics (minimal q^2) as implemented in SYBYL.¹⁷

as described above (see section Materials and Methods and legend of Table 3). MLR models #10 and #11 yield with both leave-1-out and leave-5-out the best prediction performance as compared to all other MLR equations, and here the leave-5-out predictions are slightly inferior to the ones based on leave-1-out (q^2 : 0.88 vs 0.89, SEP: 1.23 vs 1.19). Overall, the comparison between the leave-1-out and leave-5-out results suggest that for the present data set size and composition, leave-1-out statistics provide indeed a reasonable characterization of the prediction performance.

CoMFA based on Solution-Phase Quantum Chemistry.

In contrast to multilinear regression using global molecular orbital parameters, CoMFA allows a site-specific 3D analysis of the dye–fiber affinity in terms of the steric and electrostatic fields exerted by the dye molecules. In our previous investigation⁷ that was confined to gas-phase molecular descriptors such as AM1 frontier orbital energies and net atomic charges, the best CoMFA model employing two latent variables showed a clear dominance of polar interactions over steric interactions (72% vs 28%), with $r^2 = 0.81$ and SE = 1.50. Interestingly, positively charged regions in the coupling components as well as in the X moiety (structural formula in Table 1) turned out to be important for favorable dye–fiber interactions.

The presently derived CoMFA regression models employing solution-phase geometries and atomic charges without or with additional intrinsic and solution-phase molecular descriptors are summarized in Table 4. For the pure COSMO-based CoMFA model (PLS model #1 in Table 4), cross-validation now yields an optimum of three latent variables. Interestingly, the leave-one-out predictive correlation and standard error are significantly superior to the previous gas-phase CoMFA model⁷ (q^2 : 0.63 vs 0.44, SEP: 2.14 vs 2.58) but at the same time clearly inferior to all present MLR models of Table 3 (leave-1-out: q^2 range of 0.68–0.89, SEP range of 1.94–1.19).

When including AM1 E_{HOMO} as additional parameter, two latent variables provide the best leave-one-out performance (PLS model #2 in Table 4), yielding an inferior calibration quality but superior prediction capability as compared to the pure CoMFA model (r^2 : 0.86 vs 0.95; q^2 : 0.66 vs 0.63). Comparison between PLS models #3 (CoMFA, E_{HOMO} , E_{LUMO}) and #4 (CoMFA, EN) of Table 4 with MLR models #1 (E_{HOMO} , E_{LUMO}) and #3 (EN) of Table 3 shows that the inclusion of CoMFA results in a slightly improved prediction performance (ignoring, for the sake of simplicity, the

additional difference between the PLS and MLR methods). At the same time, the use of only one latent variable reduces the calibration statistics to $r^2 = 0.78$, indicating that the much greater r^2 value of PLS models #1 and 2 simply reflects the greater number of model parameters. Note also that in PLS models #3 and 4, the CoMFA contribution is relatively small (steric field: 5.2% and 8.1%; electrostatic field: 11.1% and 17.3%).

The inclusion of solution-phase molecular orbital parameters is shown in PLS models #5–7. As a general trend, these models are superior to the ones employing intrinsic (gas-phase) molecular descriptors in combination with CoMFA: With solution-phase E_{HOMO} and two latent variables, q^2 increases to 0.70 (PLS model #5), and additional inclusion of solution-phase E_{LUMO} results in $q^2 = 0.74$ with only one latent variable (PLS model #6). The latter model shows the smallest standard error of prediction (SEP = 1.72) of all models employing CoMFA.

Finally, combination of CoMFA and solution-phase EN leads to also relatively good prediction statistics (PLS model #7 in Table 4). Again, the unusually high r^2 (0.97) and low se (0.68) are due to the use of three latent variables. As compared to the pure CoMFA model as only other PLS model with three latent variables, however, both q^2 and SEP are significantly improved (q^2 : 0.75 vs 0.63; SEP: 1.77 vs 2.14).

Like with the previous gas-phase CoMFA study,⁷ inspection of the site-specific steric and electrostatic field reveals positively charged regions of the molecular structures of the dyes as relevant features of the CoMFA models. In terms of the atomic constituents of the dye molecules, these positively charged molecular regions are probably related to sulfur atoms as well as to acidic hydrogen (H connected to O or N), both of which carry positive partial charges due to their smaller electronegativity as compared to oxygen, carbon, and nitrogen. From this viewpoint, the CoMFA result may possibly reflect H bond donor capabilities of the dye. Interestingly, however, the 3D electrostatic analysis does not identify polar interactions of the azo bridge and in particular its hydrogen bond acceptor strength as a crucial factor for modeling $-\Delta\mu^0$.

When comparing Tables 3 and 4, it is apparent that the simpler MLR models can successfully compete with the more complex CoMFA approach. In particular, the prediction power of the pure CoMFA model (PLS #1 in Table 4) is significantly inferior to MLR models employing frontier

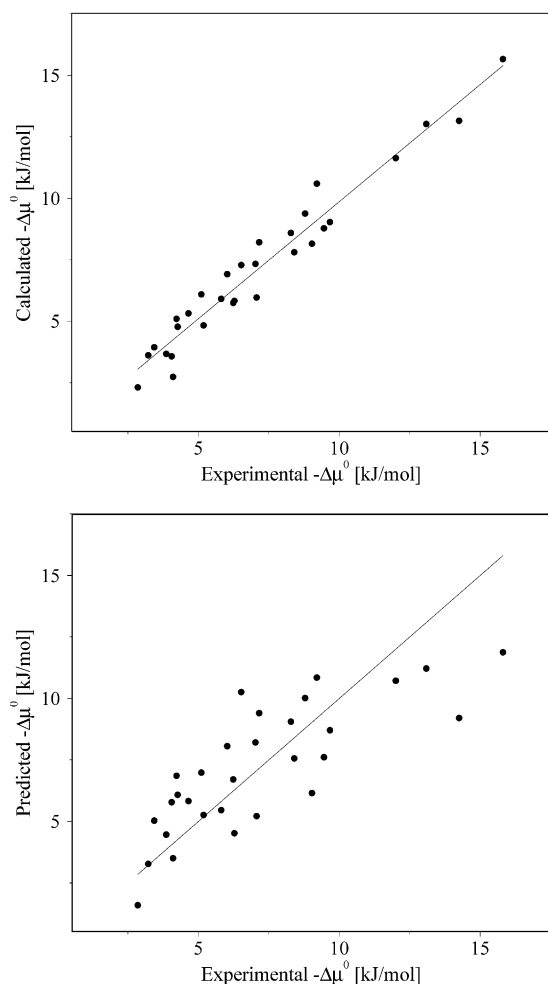


Figure 5. Calculated (top) and leave-one-out-predicted (bottom) vs experimental dye–fiber affinity in terms of $-\Delta\mu^0$ [kJ/mol] for the 30 azo dyes according to the pure CoMFA model (PLS model #1 in Table 4). The calibration plot includes the PLS regression line (top), while in the prediction plot, the $y = x$ line is shown (bottom).

orbital energies. The latter aspect is further illustrated in Figure 5, where (as with Figures 3 and 4) the data distributions for both calibration (top) and prediction (bottom) are shown. While the calibration plot looks fine ($r^2 = 0.95$, $SE = 0.77$), the leave-one-out prediction is much more scattered, with underestimations for compounds #1–4 that show the greatest dye–fiber affinities. Only PLS-CoMFA models augmented by E_{HOMO} and E_{LUMO} or EN provide prediction statistics of a similar quality as the pure MLR models.

A possible drawback of the current CoMFA analysis could be that the azo dyes were employed as neutral molecules, although with sulfonic acids, the deprotonated form is prevalent in aqueous solution under neutral and alkaline conditions. Note, however, that upon fixation of the dye in the fiber matrix, electric neutrality requires that the anionic site is either shielded by a positive counterion or protonated as discussed above. Accordingly, the actual speciation of the dye molecules is likely to change upon transfer from the water to the cellulose polymer and thus would also not correspond to the deprotonated molecular structure (apart from the question how to treat the compounds with more than one sulfonic acid group). In this context, a further

consideration is that the electronic structure of the azo dye in form of an ion pair such as $\text{R-SO}_3^-\text{Na}^+$ may be better approximated through calculations of its neutral (protonated) structure, $\text{R-SO}_3\text{H}$, than of its anionic (deprotonated) counterpart, R-SO_3^- .

We note finally that in the present regression analyses, molecular hydrophobicity in terms of the decadic logarithm of the calculated octanol/water partition coefficient, $\log K_{\text{ow}}$, did not show up as significant parameter. Possible explanations include the reasoning that upon fixation at the fiber matrix, desolvation might be either not crucial or not rate-determining, and the consideration that the specific dye–fiber interactions including hydrogen bonding cannot be accounted for properly with octanol as (actually too simple) organic phase.

CONCLUSIONS

The dye–fiber affinity of direct dyes has been explained by conflicting hypotheses: On one hand, the dyeing process was traced back to the formation of dye–dye aggregates in intermicellar cavities of the cellulose polymer, and, on the other hand, site-specific dye–fiber interactions have been claimed as a driving force for the fixation of dye molecules in the fiber matrix. Our present findings support the view that hydrogen bonding is a crucial feature in the dye–fiber interaction, which however may also be involved in the intermolecular dye–dye association. As a general trend, the dye–fiber affinity increases with increasing electron donor capacity and thus with increasing hydrogen bond acceptor strength of the dye molecule. With the azo dyes, a likely hydrogen bonding acceptor site is the azo nitrogen, interacting with OH groups of the cellulose polysaccharide.

The observed increase in dye–fiber affinity with decreasing electron acceptor strength is unexpected and deserves further attention, since it would be in conflict with building favorable dye–dye and dye–fiber interactions. For the time being, we suggest that in the respective regression relationships, the LUMO energy appears mainly as component of the molecular electronegativity (which latter is defined as half the sum of the HOMO and LUMO energy). Surprisingly, the best calibration and prediction statistics are achieved when including the HOMO energy in both linear and quadratic form. At present, no reasonable explanation can be given for this finding.

Interestingly, simple MLR models employing frontier orbital energies compete successfully with more advanced CoMFA models. The latter, however, point to positively charged molecular regions as crucial features for the dye–fiber affinity, which were not apparent from the MLR models. Moreover, CoMFA based on solution-phase atomic charges is superior to previously derived AM1-based CoMFA models for modeling the dye–fiber affinity.

A final question concerns the molecular form of the sulfonic acid groups of the dyes accumulated in the fiber matrix. Due to the constraint of electrical neutrality, the SO_3^- groups require a respective number of counterions such as Na^+ , or they become protonated and may interact as hydrogen bond donors with oxygen lone pairs of the polysaccharide, which is another issue to clarify in future investigations.

ACKNOWLEDGMENT

Financial support for a research stay of Simona Funar-Timofei at Leipzig from the Saxon Ministry of Science and Arts, reference no. 7531.50-04-840-98/2, is gratefully acknowledged.

REFERENCES AND NOTES

- (1) Kratzert, W.; Peichert, R. *Farbstoffe*; Quelle & Meier: Heidelberg, Germany, 1981; 261pp.
- (2) Gordon, P. F.; Gregory, P. *Organic chemistry in colour*; Springer-Verlag: Heidelberg, Germany, 1982; 322pp.
- (3) Bach, H.; Pfeil, E.; Philipp, W.; Reich, M. Molekülbau und Haftung substantiver Farbstoffe auf Cellulose. *Angew. Chem.* **1963**, 75, 407–416.
- (4) Pratt, L. R.; Pohorille, A. Hydrophobic effects and modeling of biophysical aqueous solution interfaces. *Chem. Rev.* **2002**, 102, 2671–2692.
- (5) Timofei, S.; Fabian, W. M. F. Comparative molecular field analysis (CoMFA) of heterocyclic monoazo dye-fibre affinities. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 1218–1222.
- (6) Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z. A review of QSAR for dye affinity for cellulose fibres. *Dyes Pigm.* **2000**, 47, 5–16.
- (7) Funar-Timofei, S.; Schüürmann, G. Comparative molecular field analysis (CoMFA) of anionic azo dye-fibre affinities. I. Gas-phase molecular orbital descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 788–795.
- (8) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959–5967.
- (9) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, 107, 3902–3909.
- (10) Wold, S. PLS for multivariate modelling. In *Chemometric methods in molecular design*; van de Waterbeemd, Ed.; VCH: Weinheim, Germany, 1995; pp 195–218.
- (11) Tomasi, J.; Persico, M. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **1994**, 94, 2027–2094.
- (12) Cramer, C. J.; Truhlar, D. G. Continuum solvation models: Classical and quantum mechanical implementations. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, U.S.A. 1995; Vol. VI, pp 1–73.
- (13) 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. 2* **1993**, 799–815.
- (14) Cramer, C. J.; Truhlar, D. G. An SCF solvation model for the hydrophobic effect and absolute free energies of solvation. *Science* **1992**, 256, 213–217.
- (15) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.* **1996**, 100, 16385–16398.
- (16) Fisichella, S.; Scarlata, G.; Torre, M. Correlation between R_m and standard affinity of some anionic azo dyes on cellulose. *J. Soc. Dyers Colour.* **1978**, 94, 521–523.
- (17) SYBYL 6.5, Tripos Associates, St. Louis, MO, 1998.
- (18) MOPAC 93. Fujitsu Limited, 9-3, Nagase 1-Chome, Mihama-ku, Chiba-city, Chiba 261, Japan, and Stewart Computational Chemistry, 15210 Paddington Circle, Colorado Springs, CO 80921, U.S.A., 1993.
- (19) AMSOL 4.0. 1993. Cramer, C. J.; Lynch, G. C.; Hawkins, G. D.; Truhlar, D. G.; Liotard, D. A. An SCF program for free energies of solvation. Quantum Chemistry Program Exchange program 606, QCPE, Indiana University, Bloomington, IN, 1993. Based on AMPAC version 2.1. by Liotard D. A.; Healy, E. F.; Ruiz, J. M.; Dewar, M. J. S. *Quantum Chemistry Program Exchange Bulletin* **1993**, 13, 78.
- (20) AMSOL 6.5.3 1998. Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. University of Minnesota, based in part on AMPAC 2.1 by Liotard, D. A., Healy, E. F., Ruiz, J. M., Dewar, M. J. S., and on the EF routines by Frank Jensen.
- (21) Schüürmann, G. Ecotoxic modes of action of chemical substances. In *Ecotoxicology*; Schüürmann, G., Markert, B., Eds.; John Wiley and Spektrum Akademischer Verlag: New York, U.S.A., 1998; pp 665–749.
- (22) Schüürmann, G. Modelling pK_a of carboxylic acids and chlorinated phenols. *Quant. Struct.-Act. Relat.* **1996**, 15, 121–132.
- (23) STATISTICA 5.5, StatSoft, Inc., Tulsa, OK, U.S.A., 2000.
- (24) Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, 88, 486–494.
- (25) Lias, S. G.; Bartmess, J. E.; Liebman, J. F.; Holmes, J. L.; Levin, R. D.; Mallard, W. G. Gas-phase ion and neutral thermochemistry. *J. Phys. Chem. Ref. Data* **1988**, 17, Suppl. 1, 1–861.
- (26) Barone, V.; Cossi, M.; Tomasi, J. A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. *J. Chem. Phys.* **1997**, 107, 3210–3221.
- (27) Schüürmann, G. Assessment of semiempirical quantum chemical continuum-solvation models to estimate pK_a of organic compounds. In *Quantitative Structure–Activity Relationships in Environmental Sciences – VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 225–242.
- (28) Schüürmann, G.; Cossi, M.; Barone, V.; Tomasi, J. Prediction of the pK_a of carboxylic acids using the ab initio continuum-solvation model PCM-UAHF. *J. Phys. Chem. A* **1998**, 102, 6706–6712.
- (29) Schüürmann, G. Quantum chemical analysis of the energy of proton transfer from phenol and chlorophenols to H_2O in the gas phase and in aqueous solution. *J. Chem. Phys.* **1998**, 109, 9523–9528.
- (30) Schüürmann, G. Prediction of Henry's law constant of benzene derivatives using quantum chemical continuum-solvation models. *J. Comput. Chem.* **2000**, 21, 17–34.
- (31) Dearden, J.; Schüürmann, G. Quantitative structure–property relationships for predicting Henry's law constant from molecular structure. *Environ. Toxicol. Chem.* **2003**, 22, 1755–1770.

CI034064F