Monte Carlo technique to study the adsorption affinity of azo dyes by applying new statistical criteria of the predictive potential

A.P. Toropova, A.A. Toropov, A. Roncaglioni and E. Benfenati

Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy

ABSTRACT
Azo dyes are broadly used in different industries through their chemical stability and ease of synthesis. However, these dyes are usually identified as critical environmental pollutants. Hence, a mathematical model for the adsorption affinity of azo dyes can be applied for solving tasks of medicine and ecology. Quantitative structure-property relationships for the adsorption affinity of azo dyes to a substrate (DAF, kJ/mol) were established using the Monte Carlo method by generating optimal SMILES-based descriptors. The index of ideality of correlation (IIC) and the correlation intensity index (CII) improved the model’s predictive potential, especially when they were used simultaneously. The statistical quality of the best model on the validation set was characterized by $n = 18$, $r^2 = 0.9468$, and RMSE = 1.26 kJ/mol.

Introduction
Dyes are coloured compounds with affinity for a substrate to which they are applied. Dyes interact with substrates through several mechanisms depending on the dye’s and substrate’s physical and chemical properties. Therefore the affinity and the nature of the interaction between a colourant and substrate define the quality of dyes. Based on their origin, colours can be classified as natural or synthetic. Synthetic and natural colourants can be applied on different substrates, but artificial colourings dominate the market in their availability and ease of application [1,2]. Notable that azo dyes are a large class of synthetic organic dyes that contain nitrogen as the azo group -N=N- as part of their molecular structures; more than half the commercial dyes belong to this class. Azo dyes represent the most significant production volume of dye chemistry today, and their relative importance may even increase in the future. They play a crucial role in the governance of the dye and printing market [2,3].

Azo dyes are broadly used in different industries (in the food, pharmaceutical, cosmetic, textile, and leather industries) through their chemical stability and ease of synthesis. However, these dyes are usually identified as critical environmental...
pollutants, and much attention was performed to the degradation of azo dyes using biological systems. If they are systemically absorbed, azo dyes can be metabolized via azoreductases of intestinal microflora by liver cells and skin surface bacteria. This metabolism leads to aromatic amines that can be hazardous. Hence, applying azo dyes leads to significant ecological risks [1–4].

Thus, synthetic colourants’ production and application methods pose some environmental challenges such as pollution of water bodies and occupational health issues in humans [5,6]. The affinity of azo dyes to a substrate is a significant technological and ecologic indicator [7].

Quantitative structure-property/activity relationships (QSPRs/QSARs) are a tool for developing models for different endpoints. QSPR can be applied to build models for DAF [8–13]. The model can be built up with the Monte Carlo method [14]. A QSPR model obtained from a stochastic process should be considered a random event [15]. However, the reliability of the QSPR is extremely important.

The present study aims to assess the so-called index of ideality of correlation (IIC) [16,17] and the correlation intensity index (CII) [18,19] as tools to determine the reliability of QSPR for DAF of azo dyes.

**Method**

**Data**

The experimental dye affinity to substrate data (DAF, kJ/mol) of 72 azo dyes was taken from the literature [8]. We randomly split these compounds into an active training set (≈25%), a passive training set (≈25%), a calibration set (≈25%), and a validation set (≈25%). Each of these sets has its task. The active training set serves to calculate correlation weights which give as large as possible the correlation between experimental and predicted endpoints for the active training set. The passive training set has to inspect whether these data provide a good correlation coefficient for similar compounds in the passive training set. The calibration set is a job to detect overtraining. The task for the validation set is the final estimation of the predictive potential of the model. Here we examine ten random splits.

**Optimal SMILES-based descriptor**

The optimal SMILES-based descriptor $DCW(T,N)$ is applied for a predictive model of the endpoint in this equation:

$$DAF = C_0 + C_1 \times DCW(T,N)$$  \hspace{1cm} (1)

$$DCW(T,N) = \sum CW(S_k) + \sum CW(SS_k)$$  \hspace{1cm} (2)

$T$ is an integer to separate SMILES attributes as rare or non-rare. The non-rare SMILES are employed to build up the model. The rare SMILES are not involved in building the model. $N$ is the number of epochs of the optimization of the correlation weights.
Sk is a SMILES atom, i.e., one symbol of the SMILES line (e.g. ‘=’, ‘O’) or a group of symbols that cannot be examined separately (e.g. ‘Cu’, ‘%11’). SSk is two SMILES atoms. CW(Sk) and CW(SSk) correlate with these SMILES attributes.

**The Monte Carlo optimization**

Equation (2) needs the numerical data for the correlation weights, and Monte Carlo optimization is a tool to calculate them. Here four target functions for the Monte Carlo optimization are examined:

\[
TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1
\]

\[
TF_1 = TF_0 + \text{IIIC}_C \times 0.5
\]

\[
TF_2 = TF_0 + \text{CII}_C \times 0.3
\]

\[
TF_3 = TF_0 + \text{IIIC}_C \times 0.3 + \text{CII}_C \times 0.5
\]

\(r_{AT}\) and \(r_{PT}\) are correlation coefficients between observed and predicted endpoints for the active and passive training sets, respectively.

\(\text{IIIC}_C\) is the index of ideality of correlation [16,17]. \(\text{IIIC}_C\) is calculated with data on the calibration set as follows:

\[
\text{IIIC}_C = r_C \frac{\min(-\text{MAEC}, +\text{MAEC})}{\max(-\text{MAEC}, +\text{MAEC})}
\]

\[
\min(x, y) = \begin{cases} x, & \text{if } x < y \\ y, & \text{otherwise} \end{cases}
\]

\[
\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases}
\]

\[
-\text{MAEC} = \frac{1}{-N} \sum |\Delta_k|, \; -N \text{ is the number of } \Delta_k < 0
\]

\[
+\text{MAEC} = \frac{1}{+N} \sum |\Delta_k|, \; +N \text{ is the number of } \Delta_k \geq 0
\]

\[
\Delta_k = \text{observed}_k - \text{calculated}_k
\]

The \(\Delta_k\) is the difference between the observed and predicted parameter values for the \(k\)-th dye.

The correlation intensity index (CII), similarly to the above \(\text{IIIC}_C\), was developed to improve the quality of the Monte Carlo optimization used to build up QSPR/QSAR models. \(\text{CII}_C\) is calculated as follows [18,19]:

\[
\text{CII}_C = 1 - \sum \text{Protest}_k
\]
\[
Protest_k = \begin{cases} 
    r_k^2 - r^2, & \text{if } r_k^2 - r^2 > 0 \\
    0, & \text{otherwise}
\end{cases}
\]

\(r^2\) is the coefficient of determination for a set that contains \(n\) substances. \(r_k^2\) is the coefficient of determination for \(n-1\) substances of a group after removing the \(k\)-th substance. Thus, if \((r_k^2 - r^2)\) is more than zero, the \(k\)-th meaning is an ‘oppositionist’ for the correlation between experimental and predicted values of the set. A small sum of ‘protests’ means a more ‘intensive’ correlation.

Results and discussion

Comparison of target functions

A comparison of the determination coefficients on the validation set showed that the best target function is \(TF_3\), calculated with Equation (6).

A possible reason for the effectiveness of the objective function calculated, taking into account the IIIC together with the CII lies in the significant difference in the estimates of the correlation quality by means of these two values (Figure 2). The CII recognizes correlations even if they are non-linear or fuzzy. In contrast, the IIIC reacts very strongly to the different dissymmetric departures of points (i.e. to the difference of values ‘experiment – calculation’).

The best model

The best model according to the determination coefficient on the validation set is the following:

\[
DAF = -20.40(\pm 1.72) + 0.7427(\pm 0.0439) \times DCW(1, 15)
\]

Table 1 lists the statistical characteristics of the model. Figure 3 contains the graphical representation of the best model.

Mechanistic interpretation

Table 2 lists promoters for an increase or decrease of DAF. The presence of nitrogen atoms increases DAF; the number of rings from 1 to 4 increases DAF, but four rings with nitrogen and five with Sulphur reduce it. However, it is clearer to single out molecular features and compare their total correlation contribution to

| Table 1. Statistical characteristics of the best model calculated with Equation (15). |
|---|---|---|---|---|---|---|---|
| Set | \(n^*\) | \(r^2\) | CCC | IIIC | CII | \(Q^2\) | RMSE | \(F\) |
| Active training | 18 | 0.7480 | 0.8558 | 0.5504 | 0.8333 | 0.6409 | 2.18 | 47 |
| Passive training | 18 | 0.7897 | 0.8834 | 0.7324 | 0.9196 | 0.7148 | 1.93 | 60 |
| Calibration | 18 | 0.9440 | 0.9358 | 0.9707 | 0.9624 | 0.9283 | 1.15 | 270 |
| Validation | 18 | 0.9468 | 0.9403 | 0.6028 | 0.9628 | 0.9312 | 1.26 |
| All | 72 | 0.8418 | 0.9051 | 0.7716 | 0.9082 | 0.8295 | 1.65 |

\(^*n = \text{number of compounds in a set}; \ r^2 = \text{coefficient of determination}; \ Q^2 = \text{cross-validated } r^2; \ RMSE = \text{root mean squared error}; \ F = \text{Fischer } F\text{-ratio}.\)
Table 2. Mechanistic interpretation – promoters of an increase and a decrease of DAF.

| SMILES attributes SAk | CWs run 1 | CWs run 2 | CWs run 3 | NA | NP | NC | dk |
|----------------------|-----------|-----------|-----------|----|----|----|----|
| Promoters of an increase |           |           |           |    |    |    |    |
| N.................. | 1.45058   | 1.24085   | 3.68733   | 18 | 18 | 18 | 0.0000 |
| \..N............ | 0.86749   | 2.68951   | 0.85243   | 18 | 18 | 18 | 0.0000 |
| c..............   | 0.91855   | 0.39087   | 0.02313   | 18 | 18 | 18 | 0.0000 |
| c..............   | 0.59292   | 0.39571   | 0.22744   | 18 | 18 | 18 | 0.0000 |
| c..............   | 1.08281   | 0.36335   | 0.35112   | 18 | 18 | 18 | 0.0000 |
| c..............   | 1.40532   | 1.21865   | 0.19009   | 18 | 18 | 18 | 0.0000 |
| c..............   | 0.86514   | 0.12869   | 0.99491   | 18 | 18 | 18 | 0.0000 |
| \..c............ | 1.17295   | 1.05545   | 1.01533   | 17 | 11 | 11 | 0.0119 |
| C..............   | 0.74225   | 1.35039   | 1.07951   | 11 | 12 | 4  | 0.0259 |
| s..............   | 0.83436   | 0.85554   | 0.73424   | 11 | 11 | 8  | 0.0088 |
| c..N............ | 1.10022   | 1.72242   | 1.98520   | 9  | 8  | 7  | 0.0069 |
| s..c............ | 0.26422   | 0.00880   | 1.19791   | 9  | 7  | 3  | 0.0278 |
| c..A............ | 0.56773   | 0.49658   | 1.09108   | 7  | 5  | 4  | 0.0152 |
| s..c............ | 1.13818   | 1.11876   | 1.15964   | 6  | 3  | 8  | 0.0079 |
| s..s............ | 1.08951   | 0.63707   | 1.44703   | 6  | 3  | 8  | 0.0079 |
| Promoters of a decrease |           |           |           |    |    |    |    |
| \..\..........   | -0.25405  | -0.46478  | -0.60453  | 18 | 18 | 18 | 0.0000 |
| O..=............ | -0.14589  | -0.16774  | -0.69359  | 18 | 18 | 18 | 0.0000 |
| S...c...........| -0.35864  | -0.78923  | -0.33920  | 18 | 18 | 18 | 0.0000 |
| c..O............| -0.14306  | -0.47928  | -0.31381  | 8  | 9  | 13 | 0.0132 |
| C..c............| -0.09640  | -0.01961  | -0.11468  | 4  | 6  | 3  | 0.0079 |
| n..A............| -0.93981  | -0.16267  | -0.03242  | 3  | 3  | 3  | 0.0000 |
| s..S............| -0.66599  | -0.01608  | -1.16084  | 2  | 3  | 2  | 0.0000 |

*CWs = correlation weights; NA, NP, and NC are frequencies of a SMILES attribute in the active training set, passive training set, and calibration set, respectively. The dk is the statistical defect of SMILES attribute, calculated via probability in active training set P(SA_k), probability in calibration set P'(SA_k); frequencies in active training set N(SA_k), in calibration set N'(SA_k), as the following:

\[
d_k = \frac{|P(SA_k) - P'(SA_k)|}{N(SA_k) + N'(SA_k)}
\]

compare molecular structures. Figure 4 represents a comparison of two molecules that are more than 90% the same (as well as their SMILES). Such comparisons can be helpful in the molecular design of effective dyes.

**Domain of applicability**

The domain of applicability for the described model calculated with Equation (15) defines the so-called statistical defects of SMILES attributes. These defects can be calculated as:

\[
d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N'(A_k)} + \frac{|P(A_k) - P''(A_k)|}{N(A_k) + N''(A_k)} + \frac{|P'(A_k) - P''(A_k)|}{N(A_k) + N''(A_k)}
\]

where \(P(A_k), P'(A_k)P''(A_k)\) are the probability of \(A_k\) in the active training set, passive training set, and calibration set, respectively; \(N(A_k), N'(A_k), N''(A_k)\) are frequencies of \(A_k\) in the active training set, passive training set, and calibration set, respectively. The statistical SMILES-defects \(D_j\) are calculated as:

\[
D_j = \sum_{k=1}^{NA} d_k
\]

where NA is the number of non-blocked SMILES attributes in the SMILES.
A SMILES falls in the domain of applicability if

$D_j < 2 + D$

In the model calculated with Equation (15), six SMILES are outside the domain of applicability. However, these do not fall in the validation set.

**Comparison with models suggested in the literature**

The average statistical characteristics for the validation set for eight random splits studied in Kumar and Kumar [8] are $\overline{R^2_v} = 0.8678$, and $\overline{RMSE}_v = 0.391$. In other words, the average statistical characteristics for the model calculated with Equation (15) are better (Table 1 and Figure 1).

*Supplementary materials* set out the technical details related to the best model (split in active training, passive training, calibration, and validation sets, values of $DCW(1,15)$, experimental and calculated values of DAF, and domain of applicability).
Figure 2. Different configurations of correlations and their assessment by means of various statistical criteria: coefficient of determination ($R^2$), correlation concordance coefficient (CCC), average metric $<r_m^2>$ [19], the index of ideality of correlation (IIC), and the correlation intensity index (IIC).

Figure 3. Graphical representation of the model calculated with Equation (15).
Conclusions

The index of ideality of correlation ($TF_1$) and the correlation intensity index ($TF_2$) improve the predictive potential of QSPR for DAF. The simultaneous use of these indices ($TF_3$) is especially effective. The advantage of the $TF_3$ is demonstrated by the considerable absolute average of the determination coefficient on ten random splits and the small dispersion of the value on the ten random splits.

Acknowledgements

The authors are grateful for the contribution of the CONCERT REACH (LIFE17 GIE/IT/000461) for support.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding

This work was supported by the LIFE-CONCERT [LIFE17 GIE/IT/000461].
ORCID

A.P. Toropova https://orcid.org/0000-0002-4194-9963
A.A. Toropov https://orcid.org/0000-0001-6864-6340
A. Roncaglioni https://orcid.org/0000-0002-1734-0939
E. Benfenati https://orcid.org/0000-0002-3976-5989

Data availability statement

The data used in this work and the models developed are freely available in the Supplementary materials section and at: http://www.insilico.eu/coral.

References

[1] L. Nambela, L.V. Haule, and Q. Mgani, A review on source, chemistry, green synthesis and application of textile colorants, J. Clean. Prod. 246 (2020), pp. 119036. doi:10.1016/j.jclepro.2019.119036.

[2] S. Benkhayra, S. M’rabet, and A. El Harfi, Classifications, properties, recent synthesis and applications of azo dyes, Helion 6 (2020), pp. e03271. doi:10.1016/j.helion.2020.e03271.

[3] H.S. Freeman, Aromatic amines: Use in azo dye chemistry, Front. Biosci. (Landmark Ed.) 18 (2013), pp. 145–164. doi:10.2741/4093

[4] F. Dehghanian, M. Kay, and D. Kahrizi, A novel recombinant AzrC protein proposed by molecular docking and in silico analyses to improve azo dye’s binding affinity, Gene 569 (2015), pp. 233–238. doi:10.1016/j.gene.2015.05.063.

[5] M. Dolatabadi, M. Mehrabpour, M. Esfandyari, H. Alidadi, and M. Davoudi, Modeling of simultaneous adsorption of dye and metal ion by sawdust from aqueous solution using of ANN and ANFIS, Chemom. Intell. Lab. Syst. 181 (2018), pp. 72–78. doi:10.1016/j.chemolab.2018.07.012.

[6] N.D. Lourenço, J.M. Novais, and H.M. Pinheiro, Analysis of secondary metabolite fate during anaerobic-aerobic azo dye biodegradation in a sequential batch reactor, Environ. Technol. 24 (2003), pp. 679–686. doi:10.1080/09593330309385603.

[7] D. Beiknejad, M.J. Chaichi, and M.H. Fatemi, Prediction of ozonolytic decolorization half-lives of azo dyes in a continuous-flow system using QSAR, Dyes Pigms. 185 (2021), pp. 108915. doi:10.1016/j.dyepig.2020.108915.

[8] P. Kumar and A. Kumar, In silico enhancement of azo dye adsorption affinity for cellulose fibre through mechanistic interpretation under guidance of QSAR models using Monte Carlo method with index of ideality correlation, SAR QSAR Environ. Res. 31 (2020), pp. 697–715. doi:10.1080/1062936X.2020.1806105.

[9] K.L. Bhat, A. Garg, M. Trachtman, and C.W. Bock, Metabolites of monomethoxy-4-aminoazobenzene dyes: A computational study, Dyes Pigms 50 (2001), pp. 133–150. doi:10.1016/S0143-7208(01)00032-8.

[10] F. Luan, X. Xu, H. Liu, and M.N.D.S. Cordeiro, Review of quantitative structure-activity/property relationship studies of dyes: Recent advances and perspectives, Color. Technol. 129 (2013), pp. 173–186. doi:10.1111/cote.12027.

[11] Y. Okada and Z. Morita, Quantitative structure–property relationship modelling for photoreduction-fast azo dyes on a nylon substrate: A methodology for thermochemical analysis of the photocleavage of azo dyes using the RM1 semi-empirical molecular orbital method, Color. Technol. 133 (2017), pp. 98–115. doi:10.1111/cote.12265.

[12] J.H. Churchley, A.J. Greaves, M.G. Hutchings, D.A.S. Phillips, and J.A. Taylor, A chemometric approach to understanding the bioelimination of anionic, water-soluble dyes by a biomass—Part 3: Direct dyes, Color. Technol. 116 (2000), pp. 279–284. doi:10.1111/j.1478-4408.2000.tb00047.x.
[13] S. Timofei, W. Schmidt, L. Kurunczi, and Z. Simon, A review of QSAR for dye affinity for cellulose fibres, Dyes Pigm. 47 (2000), pp. 5–16. doi:10.1016/S0143-7208(00)00058-9.

[14] A.P. Toropova and A.A. Toropov, CORAL: Monte Carlo method to predict endpoints for medical chemistry, Mini-Rev. Med. Chem. 18 (2018), pp. 382–391. doi:10.2174/1389557517666170927154931.

[15] A.A. Toropov and A.P. Toropova, QSAR as a random event: Criteria of predictive potential for a chance model, Struct. Chem. 30 (2019), pp. 1677–1683. doi:10.1007/s11224-019-01361-6.

[16] A.P. Toropova, A.A. Toropov, E.L. Viganò, E. Colombo, A. Roncaglioni, and E. Benfenati, Carcinogenicity prediction using the index of ideality of correlation, SAR QSAR Environ. Res. 33 (2022), pp. 419–428. doi:10.1080/1062936X.2022.2076736.

[17] A.P. Toropova and A.A. Toropov, The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability? Sci. Total Environ. 586 (2017), pp. 466–472. doi:10.1016/j.scitotenv.2017.01.198.

[18] A.A. Toropov and A.P. Toropova, Correlation intensity index: Building up models for mutagenicity of silver nanoparticles, Sci. Total Environ. 737 (2020), pp. 139720. doi:10.1016/j.scitotenv.2020.139720.

[19] P. Kumar and A. Kumar, Correlation Intensity Index (CII) as a benchmark of predictive potential: Construction of quantitative structure activity relationship models for anti-influenza single-stranded DNA aptamers using Monte Carlo optimization, J. Mol. Struct. 1246 (2021), pp. 131205. doi:10.1016/j.molstruc.2021.131205.