CROATICA CHEMICA ACTA
CCACAA **77** (1–2) 377–389 (2004)

ISSN-0011-1643
*CCA*-2939
*Original Scientific Paper*

# Categorical Modeling of the Flow Pattern of Liquid Organic Compounds Between Blade Electrodes Using Semiempirical and *ab initio* Quantum Chemical Descriptors*

**Takahiro Suzuki,[a] Kohei Yoshida,[b] Hiroya Onizuka,[b] Yoshio Iwai,[b] Yasuhiko Arai,[b] Aynur Aptula,[c] Ralph Kühne,[c] Ralf-Uwe Ebert,[c] and Gerrit Schüürmann [c,**]**

[a]*Natural Science Laboratory, Toyo University, 2-11-10 Oka, Asaka, Saitama 351-8510, Japan*

[b]*Department of Chemical Engineering, Faculty of Engineering, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan*

[c]*Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research, Permoserstr. 15, D-04318 Leipzig, Germany*

RECEIVED MAY 5, 2003; REVISED OCTOBER 21, 2003; ACCEPTED OCTOBER 28, 2003

For a data set of 30 organic fluids, categorical modeling has been employed to predict the flow pattern under an external electric field. To this end, a previously generated data set was augmented by 10 compounds with new experimental results, and quantum chemical methods have been used to characterize the geometric and electronic structure of the molecules on both the semiempirical and *ab initio* levels of theory. Both linear discriminant analysis (LDA) and binary logistic regression (BLR) have been employed to model the flow rate (high *vs*. low) and flow direction (left *vs*. right). For the flow rate, good LDA and BLR calibration statistics using the dipole moment, hydrophobicity and some charged partial surface area (CPSA) descriptors is accompanied with moderate prediction statistics, as evaluated through simulated external validation, and activity scrambling shows that chance correlation is not relevant. Additional neural network analyses yielded no stable models due to constraints imposed by the data set size. For the flow direction, LDA and BLR calibration and prediction statistics show more variation among the different models generated, with an overall performance inferior to the one for the flow rate. Here, besides CPSA descriptors, two parameters characterizing the softness of the electronic structure are involved. In general, BLR is slightly superior to LDA for both properties. The results are discussed in terms of contingency table statistics and with respect to the mechanistic meaning of molecular descriptors.

*Key words*
flow pattern
organic liquids
contingency table statistics
linear discriminant analysis
binary logistic regression
quantum chemical descriptors

## INTRODUCTION

When a high electric field is applied to a solution containing a dielectric compound, convective motion of the fluid is observed. In a direct current (dc) electric field, the Coulomb force acting on a space charge may dominate the dielectrophoretic force. As a consequence, hydrodynamic instability may arise, leading to convective transport of the charge carriers or a convective current. The motion of a fluid in electric fields is known as the Sumoto effect[1] or the electro-hydrodynamic effect.[2,3]

---

* Dedicated to Professor Nenad Trinajstić on the occasion of his 65th birthday.

** Author to whom correspondence should be addressed. (E-mail: gerrit.schuurmann@ufz.de)

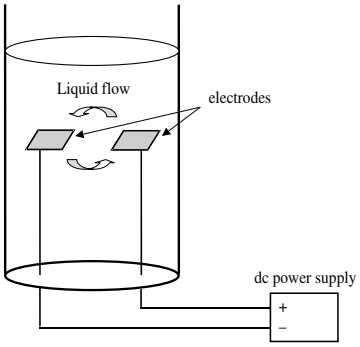


Figure 1. Apparatus for the observation of the flow patterns of liquids between the electrodes under a direct current power supply (Sumoto effect).

This phenomenon can be utilized for various technical processes. An example is the purification of 3,5-xylenol (mass fraction 95 %) + naphthalene (5 %; used as an impurity) mixtures during solidification.[4] It was shown that the turbulence in the liquid phase caused by the electric field contributed to the decrease in the concentration of naphthalene in the solid phase. Recently, another potential application has been developed based on the idea that electric energy is directly converted to kinetic energy of the fluid in terms of its convective motion, leading to micromachines or micro motors.[2,5]

Whilst there is considerable progress in exploiting the Sumoto effect for the construction of new fluid devices, only little work has been devoted to understanding the molecular mechanism of the flow of dielectric fluids subject to an external electric field. In a previous investigation,[6] the flow patterns of 20 organic liquids (containing C, H, N, O and S atoms) between blade electrodes were measured (Figure 1).

Besides the flow direction, the flow rate was determined and expressed in five categories, ranging from no flow (class 0) to very high flow (class 4), and linear discriminant analysis was applied to model the flow pattern of the compounds using experimental and calculated molecular properties. From a total of 23 descriptors including quantum chemical parameters, a combination of molecular bulk, polarity, polarizability, and H-bonding basicity appeared to influence the flow rate, whilst molecular hardness and self-polarizability normalized by molecular volume turned out to dominate the direction of the flow. However, the chemical range of compounds tested was quite limited, and the physicochemical meaning of the model for the flow direction is somewhat vague.

For the present investigation, we have generated new experimental data for 10 additional compounds. Most of the new compounds contain N, O or S atoms in order to allow a more distinct analysis of the potential impact of electron lone pairs on the flow pattern caused by the external electric field. Both semiempirical and *ab initio* quantum chemical parameters have been calculated, and linear discriminant analysis, binary logistic regression and artificial neural network modeling were employed to analyze the flow pattern of all 30 compounds in terms of the underlying molecular properties. Model performance was evaluated using contingency table statistics[7] and activity scrambling,[8,9] and the prediction performance was characterized through simulated external validation employing complementary subsets.[7,9-11]

## MATERIAL AND METHODS

### *Test Chemicals*

The 10 additional compounds with new experimental data are listed in Table I. They augment the previous set of 20 compounds,[6] making a total set of 30 compounds. For the flow rate modeling, data are available for all 30 compounds, whilst flow direction results could be achieved only for a subset of 26 compounds (*cf.* our earlier investigation[6]).

TABLE I. Additional set of 10 organic liquids with experimental results for the flow between the electrodes

| Compound no. | Compound | Purity % | Supplier | Temperature °C | Flow rate[(a)] L | Direction (– , +) |
|---|---|---|---|---|---|---|
| 21 | 2,4-pentanedione | ≥ 99 | Tokyo[(b)] | 40 | 3 | → |
| 22 | aniline | ≥ 99 | Wako[(c)] | 40 | 2 | ← |
| 23 | chlorobenzene | ≥ 99 | Wako | 40 | 3 | ← |
| 24 | di-*t*-butyl sulfide | ≥ 99 | Wako | 40 | 2 | → |
| 25 | dodecane | ≥ 99 | Tokyo | 40 | 0 | – |
| 26 | phenol[(d)] | ≥ 98 | Wako | 50 | 2 | → |
| 27 | 1-propanol | ≥ 99.5 | Wako | 40 | 3 | → |
| 28 | pyrrole | ≥ 99 | Tokyo | 40 | 3 | ← |
| 29 | α-toluenthiol | ≥ 99 | Wako | 40 | 1 | → |
| 30 | tri-n-propylamine | ≥ 98 | Tokyo | 40 | 3 | → |

[(a)] Observed by the naked eye: class 0, no flow; class 1, low; class 2, medium; class 3, high; class 4, very high. [(b)] Tokyo Kasei Kogyo Co., Ltd., Tokyo, Japan. [(c)] Wako Pure Chem. Ind., Ltd., Osaka, Japan. [(d)] m.p. = 41 °C.

*Flow Pattern Measurements*

The flow pattern (flow rate and direction of flow) between the electrodes in a glass tube with an inner diameter of 17.6 mm was measured using the same apparatus as in the previous study.[6] Experimental measurements were performed under a dc field of 400 V at 40 °C, except for phenol where 50 °C was selected because of its higher melting point.

The initial experimental determination of the flow rate allocates each compound into one of the five classes (0 = no flow, 1 = low, 2 = medium, 3 = high, 4 = very high). Since this qualitative measurement is based on inspection by the naked eye, the data are subsequently re-allocated into two broader classes: low (covers original classes 0, 1, 2), and high (covers original classes 3 and 4). The latter two categories are used for the classification modeling as described below.

*Molecular Descriptors*

Besides hydrophobicity in terms of the decadic logarithm of the octanol/water partition coefficient, log $K_{ow}$,[12] and the experimental dipole moment, $\mu^{exp}$,[13] quantum chemical calculations have been performed to characterize potentially relevant aspects of the geometric and electronic structure of the compounds. To this end, initial 3-dimensional structures of the compounds were generated with the SYBYL software,[14] and subsequent geometry optimization was performed on two levels of theory: the MOPAC[15] package was used for semiempirical AM1 calculations,[16] and Gaussian 98[17] was employed for density functional theory (DFT) calculations using the hybrid functional B3LYP[18,19] with the polarized double-zeta split valence basis set 6-31G**.[20,21]

As regards geometric features, the molecular surface area (SA) and volume (V) were calculated using the MOLSV programme[22] and MST atomic radii listed elsewhere.[23] In order to characterize the electron density distribution of the molecules, the dipole moment was also calculated (to generate a second set in addition to the experimental values taken from literature), and Mulliken population analysis was used to specify the maximum positive and negative charge ($Q_{max}^+$ and $Q_{max}^-$) as well as the relative positive and negative charge (RPCG and RNCG).[24] The latter belong to the series of charged partial surface area (CPSA) descriptors[24] that encode various aspects of the molecular disposition to undergo electrostatic interactions. In addition to the standard set of 25 CPSA descriptors, the following additional four parameters were also calculated for all 30 compounds: the surface area of the most positive and most negative atoms in the molecule, SAPX and SANX, and the maximum value of positive or negative atomic charge multiplied by its respective surface area fraction ($SA_{atom}/SA_{total}$), QPSX and QNSX.

Besides net atomic charges, several parameters were calculated that characterize the disposition of the electronic structure to donate or accept electron charge. Among these are the energies of the highest occupied and lowest unoccupied molecular orbitals, $E_{HOMO}$ and $E_{LUMO}$, and the closely related molecular electronegativity, EN = $-\frac{1}{2}$ ($E_{HOMO}$ + $E_{LUMO}$), and molecular hardness, HD = $-\frac{1}{2}$ ($E_{HOMO}$ − $E_{LUMO}$).[25] Note that all geometric and electronic parameters mentioned so far were calculated on both the semiempirical and *ab initio* levels of theory.

In addition, the following wavefunction-based descriptors were calculated only for the AM1 scheme: acceptor (nucleophilic) and donor (electrophilic) delocalizability ($D^N$ and $D^E$; for mathematical formula see reviews),[25,26] self-polarizability (SP),[26] and self-polarizability normalized by molecular volume (SP/V).

Finally, a simple count of electron lone pairs at oxygen, nitrogen and sulfur atoms, NLP, was included in order to check their potential impact on the flow pattern. In total, 83 descriptors were considered for the statistical analysis, including 42 AM1 parameters and 38 DFT parameters.

*Classification Modeling*

For modeling the flow direction (left, right) and the flow rate (low, high), the following statistical methods were employed: linear discriminant analysis (LDA) as implemented in STATISTICA,[27] and binary logistic regression (BLR) as available in SPSS.[28] Moreover, backpropagation neural network (NN)[29] modeling using in-house software was applied to the 30-compound set with flow rate data.

In LDA, the class membership of the compounds is fitted into the equation

$$d = a_0 + \sum_{k=1}^{m} a_k x_k \qquad (1)$$

where $d$ denotes the canonical discrimination function, $x_k$ is the $k$-th molecular property taken into account with its coefficient $a_k$, $a_0$ is a constant, and $m$ is the final number of descriptors included. The corresponding BLR equation reads

$$\ln\left(\frac{P}{1-P}\right) = a_0 + \sum_{k=1}^{m} a_k x_k \qquad (2)$$

where $P$ is the probability of a compound to belong to a certain class (*e.g.* low flow rate), and $1–P$ is the probability of the opposite event (*e.g.* high flow rate), which combine to the so-called odds ratio $P/(1–P)$.

With both LDA and BLR, the number of model parameters is $m$+1 (# variables plus constant). LDA model building was performed in a stepwise manner, using Fisher's $F$ test, Wilks $\lambda$ (see below) and the $p$ level of significance as criteria for determining the optimal number of descriptors during calibration of the classification

functions. The resultant descriptor combinations were then also used for the BLR and NN modeling.

The NN classification model has no explicit equation, but is defined through a learning algorithm, architecture and some technical parameters. In our case, the backpropagation learning rule was applied to a 3-layer setting, with 3+1 input-layer nodes (3 descriptors + bias), 2+1 hidden-layer nodes, and one output-layer node, using a sigmoidal transfer function (1/(1+*x*)), a learning rate of 0.95 and a momentum of 0.35. Since the NN results may depend on the initial weights, every NN calibration (and associated prediction) was repeated three times with three different randomly selected sets of starting weights, and the final statistics were calculated as average values of the three individual runs.[30]

The input bias is connected to the hidden-layer bias, which in turn is connected to the output node; all other nodes are fully connected, resulting in a total of 10 NN weights as model parameters. Here, only the descriptor combinations selected through LDA were included, and for the calibration five (arbitrarily selected) compounds were separated as an internal prediction set in order to evaluate the impact of the number of iteration cycles on the prediction performance.[30]

### *Contingency Table Statistics*

For the case of two classes (in our case: two flow-rate classes or two flow-direction classes), the contingency table has the following general form:

TABLE II. Two-dimensional 2x2 contingency table

| Category type | Experimental category | | Total |
|---|---|---|---|
| Predicted | $n_{11}$ | $n_{12}$ | $n_{1\bullet}$ |
| category | $n_{21}$ | $n_{22}$ | $n_{2\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet\bullet} = N$ |

In this table, the (predicted and experimental) categories are numbered (*e.g.* 1 = low flow rate, 2 = high flow rate), and the cell entry $n_{ij}$ denotes the number of compounds that belong to the predicted category $i$ and the experimental category $j$. The total number of compounds,

$$N = \sum_i \sum_j n_{ij} \equiv n_{\bullet\bullet} \tag{3}$$

is 30 for the flow rate modeling, and 26 for the flow direction modeling (see above), and the marginal totals

$$n_{i\bullet} = \sum_j n_{ij} \tag{4}$$

and

$$n_{\bullet j} = \sum_i n_{ij} \tag{5}$$

quantify the number of compounds belonging to the *i*th predicted class (Eq. 4) and *j*th experimental class (Eq. 5), respectively.

The concordance is defined as the proportion of compounds where the predicted and experimental categories agree, which in terms of the contingency table entries can be written as

$$\text{Concordance} = \frac{1}{N} \sum_i n_{ii} \tag{6}$$

A more demanding parameter is the so-called kappa index,[31]

$$\kappa = \frac{(1/N) \sum_i n_{ii} - (1/N^2) \sum_i n_{i\bullet} n_{\bullet i}}{1 - (1/N^2) \sum_i n_{i\bullet} n_{\bullet i}} \tag{7}$$

where the fraction of agreement that could have been obtained by chance is eliminated. Another more sophisticated measure of association is the $\lambda_B$ parameter,[32]

$$\lambda_B = \frac{\sum_i \max_j (n_{ij}) - \max_j (n_{\bullet j})}{N - \max_j (n_{\bullet j})} \tag{8}$$

This parameter compares the prediction based on marginal totals with the prediction based on the classification model, and its values represent the reduction in prediction error of the latter (true) model as compared to the former (simple) model. All three parameters (Eqs. 6–8) range from 0 (no agreement or no prediction capability) to 1 (full agreement or full prediction capability).

### *Model Validation*

The predictive performance of the LDA and BLR classification models was evaluated applying two techniques: simulated external validation using complementary subsets,[7,9-11] and activity scrambling.[8,9] For the former, a total set of 30 compounds (including 20 compounds from the previous investigation)[6] was ordered by increasing molecular weight, and then two subgroups were formed by allocating all odd-numbered compounds to group1, and all even-numbered compounds to group2, as specified in Table III. Through this construction, both subgroups contain 15 compounds (flow rate) and 13 compounds (flow direction), and cover almost the same range of molecular size. The LDA and BLR models were then calibrated for group1 and group2 separately, and the resultant submodels were used to predict the flow pattern of the complementary subsets (group1 flow pattern predicted from group2 model, and *vice versa*).

Whilst simulated external validation provides information about the prediction capability of a given model, activity scrambling allows characterizing the degree of chance correlation.[8,9] For both the flow rate and flow direction, half of the compounds of category 1 (low flow rate or left flow direction) were (wrongly) allocated to category 2 (high flow rate, or right flow direction), and

TABLE III. Set of 30 compounds with all molecular descriptors used in the final LDA and BLR models [a]

| Compound no. | Compound name | $\mu^{exp}$ D | log $K_{ow}$ | HD eV | $D^N$ $(eV)^{-1}$ | SP/V $(eV \cdot Å^3)^{-1}$ | WPSA-1 Å$^4$ | WPSA-1$^{B3LYP}$ Å$^4$ | WPSA-3 Å$^4$ | WPSA-3$^{B3LYP}$ Å$^4$ | DPSA-2 Å$^2$ | PNSA-3 Å$^2$ | QPSX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | acetone[c] | 2.88 | –0.24 | 5.76 | –2.30 | 0.0684 | 41.0 | 40.5 | 6.3 | 5.5 | 262.5 | –15.4 | 0.047 |
| 2 | benzene[b], [d] | 0 | 2.13 | 5.10 | –3.60 | 0.0840 | 49.7 | 49.0 | 9.5 | 4.1 | 290.1 | –10.5 | 0.030 |
| 3 | benzoic acid[c] | 1.73 | 1.87 | 4.81 | –4.43 | 0.0796 | 54.8 | 56.0 | 12.7 | 8.7 | 513.9 | –31.2 | 0.116 |
| 4 | bromobenzene[b] | 1.55 | 2.99 | 4.83 | –3.81 | 0.0658 | 70.5 | 43.6 | 10.3 | 4.3 | 310.9 | –7.8 | 0.035 |
| 5 | butyronitrile[b] | 3.60 | 0.60 | 6.66 | –2.97 | 0.0717 | 47.5 | 49.7 | 6.7 | 6.7 | 257.4 | –7.5 | 0.040 |
| 6 | cyclohexane[b], [d] | 0 | 3.44 | 7.30 | –4.33 | 0.0742 | 79.8 | 80.0 | 10.0 | 7.6 | 425.6 | 0.0 | 0.025 |
| 7 | cyclopentanone[b] | 3.00 | 0.31 | 5.69 | –3.45 | 0.0738 | 57.7 | 57.4 | 8.9 | 7.0 | 393.8 | –15.1 | 0.041 |
| 8 | dibutylamine[c] | 1.04 | 2.83 | 6.21 | –7.00 | 0.0710 | 167.1 | 168.1 | 21.0 | 0.1 | 996.3 | –2.3 | 0.037 |
| 9 | dibutyl ether[b] | 1.22 | 3.21 | 6.64 | –6.26 | 0.0701 | 166.0 | 167.3 | 20.6 | 16.0 | 922.6 | –2.3 | 0.035 |
| 10 | 3,5-dimethylphenol[b] | 1.76 | 2.35 | 4.68 | –5.37 | 0.0786 | 87.8 | 90.9 | 14.1 | 12.0 | 567.8 | –14.8 | 0.097 |
| 11 | ethanol[b] | 1.69 | –0.30 | 7.22 | –1.96 | 0.0394 | 36.8 | 35.9 | 5.4 | 4.8 | 183.3 | –11.1 | 0.094 |
| 12 | hexane[c], [d] | 0.08 | 4.00 | 7.41 | –4.54 | 0.0703 | 90.8 | 106.4 | 11.8 | 10.1 | 516.9 | –0.2 | 0.034 |
| 13 | isobutyraldehyde[c] | 2.58 | 0.36 | 5.69 | –3.00 | 0.0694 | 52.2 | 52.1 | 7.1 | 5.7 | 318.0 | –15.6 | 0.038 |
| 14 | naphthalene[c] | 0 | 3.35 | 4.22 | –5.91 | 0.0887 | 74.9 | 76.8 | 14.5 | 6.6 | 489.5 | –12.9 | 0.031 |
| 15 | nitrobenzene[c] | 4.22 | 1.85 | 4.75 | –4.25 | 0.0810 | 43.8 | 44.8 | 10.4 | 6.0 | 468.5 | –36.9 | 0.038 |
| 16 | *o*-xylene[b] | 0.62 | 3.12 | 4.85 | –5.04 | 0.0810 | 79.2 | 80.7 | 5.0 | 8.0 | 466.0 | –7.3 | 0.031 |
| 17 | phenyl acetate[b] | 1.91 | 1.49 | 4.83 | –5.45 | 0.0798 | 85.8 | 85.0 | 16.4 | 9.8 | 660.3 | –22.9 | 0.050 |
| 18 | quinoline[b] | 2.18 | 2.03 | 4.36 | –5.60 | 0.0884 | 69.4 | 72.7 | 15.7 | 7.1 | 456.3 | –15.1 | 0.036 |
| 19 | sulfolane[c] | 4.81 | 1.23 | 5.65 | –3.25 | 0.0663 | 55.6 | 56.1 | 13.3 | 9.2 | 1273.7 | –75.1 | 0.047 |
| 20 | thiophene[c] | 0.55 | 1.81 | 4.73 | –2.96 | 0.0743 | 47.9 | 47.9 | 16.4 | 7.4 | 358.5 | –11.4 | 0.113 |
| 21 | 2,4-pentanedione[c] | 1.80[e] | 0.14 | 5.53 | –3.71 | 0.0710 | 61.5 | 60.9 | 10.4 | 8.9 | 540.0 | –24.4 | 0.048 |
| 22 | aniline[b] | 1.53 | 0.90 | 4.49 | –4.38 | 0.0799 | 57.2 | 57.0 | 10.5 | 7.2 | 442.1 | –17.1 | 0.077 |
| 23 | chlorobenzene[b] | 1.69 | 2.84 | 4.86 | –3.79 | 0.0700 | 42.0 | 42.0 | 8.6 | 4.2 | 281.4 | –9.9 | 0.063 |
| 24 | di-*tert*-butyl sulfide[c] | 1.50 | 4.61[f] | 4.55 | –6.50 | 0.0677 | 131.4 | 133.3 | 16.6 | 14 | 859.2 | –1.0 | 0.030 |
| 25 | dodecane[c], [d] | 0 | 6.80 | 7.30 | –8.77 | 0.0720 | 262.4 | 278.5 | 24.6 | 44.6 | 1659.7 | –0.2 | 0.035 |
| 26 | phenol[c] | 1.45 | 1.50 | 4.76 | –4.00 | 0.0807 | 49.7 | 49.7 | 10.1 | 6.3 | 346.8 | –17.7 | 0.098 |
| 27 | 1-propanol[b] | 1.68 | 0.25 | 7.17 | –2.67 | 0.0649 | 52.2 | 52.4 | 7.6 | 6.8 | 264.8 | –9.5 | 0.113 |
| 28 | pyrrole[c] | 1.80 | 0.75 | 5.02 | –3.09 | 0.0793 | 41.5 | 45.6 | 9.9 | 5.2 | 275.2 | –11.7 | 0.096 |
| 29 | α-toluenethiol[b] | 1.70[e] | 2.45[f] | 4.47 | –4.92 | 0.0744 | 68.6 | 69.2 | 11.6 | 6.5 | 424.4 | –8.9 | 0.034 |
| 30 | tri-n-propylamine[c] | 0.90[e] | 2.79 | 5.97 | –7.84 | 0.0711 | 173.9 | 179.4 | 21.5 | 17.6 | 1088.3 | –1.0 | 0.035 |

[a] Explanation of descriptor abbreviations (see also the text): $\mu^{exp}$ = experimental dipole moment; log $K_{ow}$ = decadic logarithm of the octanol/water partition coefficient; HD = molecular hardness = – ½ ($E_{HOMO} - E_{LUMO}$) with $E_{HOMO}$ and $E_{LUMO}$ denoting the energies of the highest occupied and lowest unoccupied molecular orbitals, respectively; $D^N$ = nucleophilic delocalizability = acceptor delocalizability;[25,26] SP/V = self-polarizability[26] normalized by molecular volume; WPSA-1 = total surface weighted partial positive surface area;[24] WPSA-3 = atomic charge weighted partial positive surface area, weighted by total surface area;[24] DPSA-2 = difference between total charge weighted positive and negative surface areas;[24] PNSA-3 = atomic charge weighted partial negative surface area;[24] QPSX = maximum value of the product of positive net atomic charge and its associated surface area fraction.

[b], [c] Superscript indicates the allocation of the compound to one the two complementary subsets group1[b] and group2[c] in the context of simulated external validation.

[d] Chemicals not in data set for flow direction.

[e] Value calculated by AM1[16] (due to lacking experimental value).

[f] Value calculated by CHEMICALC-2[33] (due to lacking experimental value).

correspondingly half of the compounds belonging originally to category 2 were allocated to category 1. In this setting, 50 % of the compounds are allocated wrongly, and a randomly selected compound would have a 50 % chance to belong to either of the two categories (for both the flow rate and flow direction). Consequently, model calibration would be expected to yield 50 % error, and significantly better calibration results would indicate that the model can be trained to predict noise, and thus would be based, at least partly, on chance correlations.

## RESULTS AND DISCUSSION

Experimental data of the additional set of 10 compounds are listed in Table I. In this set, dodecane is the only compound with no Sumoto effect. At the same time, dodecane is the only compound without any heteroatom, and consequently has no dipole moment. This suggests that at least some net polarity in the molecule is required for a movement induced by an external electric field. A similar situation was observed in the previous study with 20 compounds,[6] where the three compounds without net flow were benzene, cyclohexane and hexane, all of which have no permanent dipole moment.

As regards molecular descriptors, AM1 and B3LYP show similar trends. Greater deviations are observed for some CPSA parameters such as the partial positive and partial negative surface areas (PPSA-1, PNSA-1) and the total positive charge weighted surface area (PPSA-2) of bromobenzene. These differences are due to the fact that the net atomic charge of bromine is positive with AM1 (0.0528 a.u.), but negative according to the B3LYP/6-31G** Mulliken population analysis (–0.1327 a.u.). The former is in accord with the π electron donor capacity (mesomeric substituent effect), whilst the latter reflects the σ electron acceptor capacity (inductive substituent effect). Note further that a recent comparative analysis of semiempirical and *ab initio* calculations for a set of 607 organic compounds revealed systematic differences with regard to descriptors that depend on net atomic charges.[25] In Table III, all descriptor values used in the final classification models are listed for all 30 compounds.

### *Flow Rate Modeling*

In the upper part of Table IV, one-variable LDA classification statistics are shown for all the variables that are used in the final multi-variable LDA equations. Among these parameters, the experimental dipole moment, $\mu^{exp}$, is the best single variable to discriminate between the high and low flow rates, achieving an agreement between experimental and calculated categories of 80 % (concordance = 0.80; see Eq. 6). The respective $\kappa$ and $\lambda_B$ values are close to 60 % (0.594 and 0.571, respectively; see Eqs. 7 and 8), indicating that the simple concordance is probably a too optimistic measure of the actual discrimination power.

The next best variable to discriminate between the high and low flow rates is log $K_{ow}$, and all other parame-

TABLE IV. One-variable LDA statistics for the flow rate (high *vs.* low, 30 compounds) and flow direction (left *vs.* right, 26 compounds)[(a)]

| Endpoint<br>Descriptor | Wilks $\lambda$ | $F$ value | $p$ level | Concordance | $\kappa$ | $\lambda_B$ |
|---|---|---|---|---|---|---|
| Flow rate | | | | | | |
| $\mu^{exp}$ | 0.602 | 18.5 | 0 | 0.800 | 0.594 | 0.571 |
| log $K_{ow}$ | 0.661 | 14.3 | 0 | 0.700 | 0.395 | 0.357 |
| WPSA-1$^{B3LYP}$ | 0.959 | 1.2 | 0.282 | 0.600 | 0.269 | 0.143 |
| WPSA-1 | 0.953 | 1.4 | 0.249 | 0.567 | 0.094 | 0.071 |
| DPSA-2 | 0.993 | 0.2 | 0.666 | 0.533 | 0.146 | 0 |
| $D^N$ | 0.929 | 2.1 | 0.156 | 0.533 | 0.045 | 0 |
| Flow direction | | | | | | |
| QPSX$^{B3LYP}$ | 0.830 | 4.9 | 0.036 | 0.692 | 0.365 | 0.333 |
| WPSA-3$^{B3LYP}$ | 0.920 | 2.1 | 0.161 | 0.654 | 0.282 | 0.250 |
| SP/V | 0.855 | 4.1 | 0.055 | 0.615 | 0.225 | 0.167 |
| PNSA-3 | 0.980 | 0.5 | 0.493 | 0.615 | 0.186 | 0.167 |
| DPSA-2 | 0.978 | 0.5 | 0.470 | 0.577 | 0.217 | 0.083 |
| HD | 0.998 | 0.1 | 0.830 | 0.615 | 0.176 | 0.167 |
| WPSA-3 | 0.989 | 0.3 | 0.617 | 0.538 | 0.035 | 0 |

[(a)] Units – dipole moment: D; delocalizability: $eV^{-1}$; hardness: eV; charged partial surface area: Å or Å au. The quantum chemical descriptors have been calculated with AM1 (no specification) or B3LYP/6-31G** (specified through superscript B3LYP). For explanations of descriptor abbreviations see the legend to Table III.

TABLE V. Contingency table statistics of four LDA classification models (CMs) for the flow rate (high *vs.* low, 30 compounds)[a],[b]

| Association parameter | CM1 | CM2 | CM3 | CM4 |
|---|---|---|---|---|
| All chemicals | | | | |
| Concordance | 0.833 | 0.900 | 0.867 | 0.900 |
| $\kappa$ | 0.660 | 0.798 | 0.730 | 0.798 |
| $\lambda_B$ | 0.643 | 0.786 | 0.714 | 0.786 |
| Group1 calibration | | | | |
| Concordance | 0.867 | 0.800 | 0.867 | 0.867 |
| $\kappa$ | 0.733 | 0.598 | 0.733 | 0.733 |
| $\lambda_B$ | 0.714 | 0.571 | 0.714 | 0.714 |
| Group1 prediction | | | | |
| Concordance | 0.733 | 0.733 | 0.733 | 0.733 |
| $\kappa$ | 0.454 | 0.463 | 0.473 | 0.463 |
| $\lambda_B$ | 0.429 | 0.429 | 0.429 | 0.429 |
| Group2 calibration | | | | |
| Concordance | 0.867 | 1 | 0.933 | 1 |
| $\kappa$ | 0.728 | 1 | 0.864 | 1 |
| $\lambda_B$ | 0.714 | 1 | 0.857 | 1 |
| Group2 prediction | | | | |
| Concordance | 0.733 | 0.733 | 0.733 | 0.733 |
| $\kappa$ | 0.464 | 0.464 | 0.464 | 0.464 |
| $\lambda_B$ | 0.429 | 0.429 | 0.429 | 0.429 |

[a] The subsets group1 and group2 contain 15 compounds each and are specified in Table III.

[b] CM1 = $0.697\,\mu^{exp} - 0.904 \log K_{ow} - 0.585\,D^N - 1.93$
CM2 = $-0.717\,\mu^{exp} + 0.866 \log K_{ow} - 0.0184$ WPSA-1 + 0.879
CM3 = $0.533\,\mu^{exp} - 0.715 \log K_{ow} + 0.0106$ DPSA-2 – 0.330
CM4 = $-0.716\,\mu^{exp} + 0.891 \log K_{ow} - 0.0184$ WPSA-$1^{B3LYP}$ + + 0.860

[c] Descriptor abbreviations are explained in the legend to Table III.

TABLE VI. Contingency table statistics of four BLR classification models (CMs) for the flow rate (high *vs.* low, 30 compounds)[a],[b],[c]

| Association parameter | CM1 | CM2 | CM3 | CM4 |
|---|---|---|---|---|
| All chemicals | | | | |
| Concordance | 0.867 | 0.900 | 0.900 | 0.900 |
| $\kappa$ | 0.730 | 0.798 | 0.798 | 0.798 |
| $\lambda_B$ | 0.714 | 0.786 | 0.786 | 0.786 |
| Group1 calibration | | | | |
| Concordance | 1 | 1 | 1 | 1 |
| $\kappa$ | 1 | 1 | 1 | 1 |
| $\lambda_B$ | 1 | 1 | 1 | 1 |
| Group1 prediction | | | | |
| Concordance | 0.800 | 0.800 | 0.733 | 0.733 |
| $\kappa$ | 0.602 | 0.602 | 0.473 | 0.464 |
| $\lambda_B$ | 0.571 | 0.571 | 0.429 | 0.429 |
| Group2 calibration | | | | |
| Concordance | 1 | 1 | 1 | 1 |
| $\kappa$ | 1 | 1 | 1 | 1 |
| $\lambda_B$ | 1 | 1 | 1 | 1 |
| Group2 prediction | | | | |
| Concordance | 0.800 | 0.800 | 0.867 | 0.800 |
| $\kappa$ | 0.602 | 0.602 | 0.738 | 0.602 |
| $\lambda_B$ | 0.571 | 0.571 | 0.714 | 0.571 |

[a] The subsets group1 and group2 contain 15 compounds each and are specified in Table III.

[b] CM1 = $2.71(\pm 1.97)\,\mu^{exp} - 1.50(\pm 0.796) \log K_{ow} - 1.07(\pm 0.644) D^N - 6.05(\pm 4.72)$
CM2 = $-4.67(\pm 3.36)\,\mu^{exp} - 1.47(\pm 0.763) \log K_{ow} + 0.046(\pm 0.027)$ WPSA-1 – $7.96(\pm 6.70)$
CM3 = $3.43(\pm 2.43)\,\mu^{exp} - 1.27(\pm 0.692) \log K_{ow} + 0.010(\pm 0.004)$ DPSA-2 – $5.75(\pm 5.02)$
CM4 = $4.73(\pm 3.36)\,\mu^{exp} - 1.44(\pm 0.759) \log K_{ow} + 0.045(\pm 0.026)$ WPSA-$1^{B3LYP}$ – $8.09(\pm 6.67)$

[c] Descriptor abbreviations are explained in the legend to Table III.

ters show a much poorer discrimination performance as single variables. This is also reflected by Wilks $\lambda$, which evaluates the total variance in terms of the between-category and within-category variance, and is defined as the portion of the total variance that is due to the within-group variance (ranging from 0 for perfect discrimination to 1 for no discrimination). Taking $\mu^{exp}$ as an example, the value of Wilks $\lambda$ indicates that *ca.* 60 % of the total variance is due the within-category variance, and thus only 40 % due to the between-group variance.

Tables V and VI summarize the calibration and prediction statistics of the four best LDA and BLR classification models (CMs). For the total set of 30 compounds, the concordance ranges from 0.833 to 0.900, and the best $\kappa$ and $\lambda_B$ values (0.798 and 0.786, respectively) are achieved for CM2 (LDA, BLR), CM3 (BLR) and CM4 (LDA, BLR). Note that all four models include $\mu^{exp}$ and $\log K_{ow}$ and differ only in their third variable, which is either the AM1 acceptor delocalizability $D^N$ (CM1) or some CPSA parameter (CM2: WPSA-1; CM3: DPSA-2; CM4: WPSA-$1^{B3LYP}$). The regression coefficients of the final LDA and BLR equations are given in the legends to Tables V and VI.

Interestingly, the group1 and group2 calibrations show significant differences between LDA and BLR. For both subsets, BLR achieves a perfect discrimination between the high and low flow rates with all four CMs (concordance = $\kappa = \lambda_B = 1$), whilst the LDA performance differs between the subsets and CMs. Here, CM2 yields perfect discrimination for group2, but is inferior to the other three CMs when being calibrated with group1.

TABLE VII. LDA and BLR calibration for the flow rate using scrambled activity categories[a]

| Descriptor set | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 | Set 7 | Set 8 | Set 9 | Set 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LDA | | | | | | | | | | |
| CM1 ($\mu^{exp}$, log $K_{ow}$, $D^N$) | 56.7 | 53.3 | 60.0 | 63.3 | 66.7 | 60.0 | 56.7 | 53.3 | 46.7 | 63.3 |
| CM2 ($\mu^{exp}$, log $K_{ow}$, WPSA-1) | 50.0 | 50.0 | 50.0 | 63.3 | 70.0 | 56.7 | 53.3 | 43.3 | 53.3 | 66.7 |
| CM3 ($\mu^{exp}$, log $K_{ow}$, DPSA-2) | 53.3 | 60.0 | 63.3 | 53.3 | 63.3 | 53.3 | 60.0 | 63.3 | 56.7 | 63.3 |
| CM4 ($\mu^{exp}$, log $K_{ow}$, WPSA-1$^{B3LYP}$) | 50.0 | 50.0 | 53.3 | 56.7 | 66.7 | 53.3 | 56.7 | 46.7 | 50.0 | 66.7 |
| BLR | | | | | | | | | | |
| CM1 ($\mu^{exp}$, log $K_{ow}$, $D^N$) | 56.7 | 53.3 | 60.0 | 70.0 | 66.7 | 60.0 | 56.7 | 53.3 | 46.7 | 63.3 |
| CM2 ($\mu^{exp}$, log $K_{ow}$, WPSA-1) | 50.0 | 50.0 | 50.0 | 63.3 | 70.0 | 56.7 | 60.0 | 40.0 | 53.3 | 63.3 |
| CM3 ($\mu^{exp}$, log $K_{ow}$, DPSA-2) | 53.3 | 56.7 | 63.3 | 56.7 | 60.0 | 53.3 | 60.0 | 63.3 | 60.0 | 63.3 |
| CM4 ($\mu^{exp}$, log $K_{ow}$, WPSA-1$^{B3LYP}$) | 50.0 | 50.0 | 53.3 | 56.7 | 66.7 | 53.3 | 53.3 | 46.7 | 50.0 | 63.3 |

[a] For each of the 10 sets, a randomly selected half of the compounds belonging (truly) to the category of low flow rate were allocated wrongly to the high flow rate category, and *vice versa* for the compounds of the (true) high flow rate category. Descriptor abbreviations are explained in the legend to Table III.

With both LDA and BLR, however, the prediction performance characterized through the simulated external validation approach (group1 flow rate predicted from group2-calibrated model, and group2 flow rate predicted from group1-calibrated model) is significantly inferior to the calibration statistics. Taking group1 as an example, the LDA calibration $\lambda_B$ value ranges from 0.571 to 0.714 (with $\lambda_B = 1$ for all four BLR models, s.a.), whilst the LDA prediction $\lambda_B$ is 0.429 for all four CMs (Table V), and 0.429 to 0.571 for BLR (Table VI). Interestingly, the BLR models CM1 and CM2 are clearly superior to CM3 and CM4 in predicting the flow rate of group1 (*e.g.* $\kappa$ 0.602 *vs.* 0.473 and 0.464, Table VI), and CM3 achieves the best prediction statistics for group2 (concordance = 0.867, $\kappa = 0.738$, $\lambda_B = 0.714$).

Taking calibration and prediction statistics together, the present analysis shows that a relatively good calibration is accompanied with only moderate prediction power. This difference between the calibration success and the actual prediction capability (as estimated through simulated external validation) suggests that a larger set of compounds will be needed to derive classification models with an improved power for external prediction.

Coming back to the descriptors, all four classification models include the dipole moment and the molecular hydrophobicity. The former appears to be mechanistically clear, since the molecular charge distribution and in particular the net polarity is expected to play a crucial role in the response of the molecule to an external electric field. At first sight, the mechanistic meaning of log $K_{ow}$ is less obvious. However, hydrophobicity in terms of $K_{ow}$ results from a complex interaction of the compound with the solvents, water and octanol, including electrostatic and dispersion forces and hydrogen bonding interactions as well as an entropy component. From this viewpoint, it appears that the solute susceptibility to electrostatic interactions is the component of log $K_{ow}$ that affects its flow rate under the influence of an external electric field.

In CM2, the AM1 acceptor delocalizability $D^N$ is used as the third descriptor. $D^N$ characterizes the molecule's disposition to accept excess charge offered by nucleophiles,[25] and as such is a mechanistically reasonable parameter in the context of the Sumoto effect. WPSA-1, the total surface area weighted partial positive surface area (WPSA-1 = (SA/1000) · PPSA-1),[24] appears as the third variable in CM2 (AM1) and CM4 (B3LYP/6-31G**). By construction, WPSA-1 encodes the molecule's readiness to undergo electrostatic interactions with excess negative charge, and DPSA-2 (third variable of CM3) as the difference between PPSA-2 (partial positive surface area weighted by the sum of positive atomic charges)[24] and PNSA-2 (partial negative surface area weighted by the sum of negative atomic charges)[24] represent the molecule's capacity to interact with positive or negative excess charge. As such, both parameters appear useful for modeling the molecule's movement in response to an external electric field.

Thus, all molecular parameters used for CM1 to CM4 are mechanistically reasonable in the context of the Sumoto effect. Nonetheless, the total number of descriptors tested (83) and the significant difference between the calibration and prediction performance raises the question whether and to what degree chance correlations might have played a role in deriving the classification models. In order to address this question, 10 new calibration runs with 50 % scrambled activity categories were performed for the descriptor combinations of CM1 to CM4 (Table VII). Note that for all 10 artificial sets (Set 1 to Set 10 in Table VII), the chance of a randomly selected compound showing a high flow rate is 50 %. The resultant LDA and BLR concordance values of around 60 % (as opposed to the perfect situation of 50 %) show that for the presently derived classification models, chance correlation is not a crucial factor.

TABLE VIII. Contingency table statistics of five LDA classification models (CMs) for the flow direction (left *vs.* right, 26 compounds)[a],[b],[c]

| Association parameter | CM1 | CM2 | CM3 | CM4 | CM5 |
|---|---|---|---|---|---|
| All chemicals | | | | | |
| Concordance | 0.654 | 0.808 | 0.731 | 0.769 | 0.846 |
| $\kappa$ | 0.291 | 0.607 | 0.449 | 0.530 | 0.690 |
| $\lambda_B$ | 0.250 | 0.583 | 0.417 | 0.500 | 0.667 |
| Group1 calibration | | | | | |
| Concordance | 0.846 | 0.846 | 0.692 | 0.846 | 0.923 |
| $\kappa$ | 0.682 | 0.682 | 0.365 | 0.690 | 0.843 |
| $\lambda_B$ | 0.667 | 0.667 | 0.333 | 0.667 | 0.833 |
| Group1 prediction | | | | | |
| Concordance | 0.538 | 0.615 | 0.615 | 0.615 | 0.846 |
| $\kappa$ | 0.070 | 0.216 | 0.216 | 0.216 | 0.690 |
| $\lambda_B$ | 0 | 0.167 | 0.167 | 0.167 | 0.667 |
| Group2 calibration | | | | | |
| Concordance | 0.615 | 0.615 | 0.615 | 0.692 | 0.923 |
| $\kappa$ | 0.235 | 0.235 | 0.235 | 0.380 | 0.843 |
| $\lambda_B$ | 0.167 | 0.167 | 0.167 | 0.333 | 0.833 |
| Group2 prediction | | | | | |
| Concordance | 0.615 | 0.615 | 0.692 | 0.615 | 0.615 |
| $\kappa$ | 0.216 | 0.216 | 0.206 | 0.216 | 0.235 |
| $\lambda_B$ | 0.167 | 0.167 | 0.333 | 0.167 | 0.167 |

[a] The subset group1 and group2 each contain 13 compounds as specified in Table III.

[b] CM1 = –1.19 HD – 168.77 SP/V + 18.72

CM2 = 1.22 + 167.99 SP/V – 0.0105 DPSA-2 – 17.99

CM3 = –114.23 SP/V + 0.110 WPSA-3 + 7.05

CM4 = 111.40 SP/V – 0.116 WPSA-3 + 0.0311 PNSA-3 – 6.29

CM5 = 1.12 HD + 146.93 SP/V – 0.201 $WPSA\text{-}3^{B3LYP}$ – 26.74 $QPSX^{B3LYP}$ – 13.60

[c] Descriptor abbreviations are explained in the legend to Table III.

TABLE IX. Contingency table statistics of five BLR classification models (CMs) for the flow direction (left *vs.* right, 26 compounds)[a],[b],[c]

| Association parameter | CM1 | CM2 | CM3 | CM4 | CM5 |
|---|---|---|---|---|---|
| All chemicals | | | | | |
| Concordance | 0.692 | 0.808 | 0.769 | 0.808 | 0.808 |
| $\kappa$ | 0.373 | 0.607 | 0.530 | 0.611 | 0.611 |
| $\lambda_B$ | 0.333 | 0.583 | 0.500 | 0.583 | 0.583 |
| Group1 calibration | | | | | |
| Concordance | 1 | 1 | 0.692 | 0.846 | 1 |
| $\kappa$ | 1 | 1 | 0.365 | 0.690 | 1 |
| $\lambda_B$ | 1 | 1 | 0.333 | 0.667 | 1 |
| Group1 prediction | | | | | |
| Concordance | 0.538 | 0.615 | 0.615 | 0.615 | 0.846 |
| $\kappa$ | 0.070 | 0.216 | 0.216 | 0.216 | 0.690 |
| $\lambda_B$ | 0 | 0.167 | 0.167 | 0.167 | 0.667 |
| Group2 calibration | | | | | |
| Concordance | 0.615 | 0.615 | 0.615 | 0.692 | 0.769 |
| $\kappa$ | 0.235 | 0.235 | 0.235 | 0.380 | 0.530 |
| $\lambda_B$ | 0.167 | 0.167 | 0.167 | 0.333 | 0.500 |
| Group2 prediction | | | | | |
| Concordance | 0.615 | 0.538 | 0.769 | 0.615 | 0.615 |
| $\kappa$ | 0.235 | 0.092 | 0.530 | 0.216 | 0.252 |
| $\lambda_B$ | 0.167 | 0 | 0.500 | 0.167 | 0.167 |

[a] The subset group1 and group2 each contain 13 compounds as specified in Table III.

[b] CM1 = 1.44(± 0.86) + 243.27(± 113.06) SP/V – 25.49(± 12.11)

CM2 = 1.64(± 0.95) + 250.38(± 117.66) SP/V – 0.010(± 0.001) DPSA-2 – 26.10(± 12.69)

CM3 = 124.79(± 67.69) SP/V – 0.092(± 0.080) WPSA-3 – 7.90(4.98)

CM4 = 163.11(± 92.61) SP/V – 0.137(± 0.110) WPSA-3 + 0.065(± 0.060) PNSA-3 – 9.25(± 6.08)

CM5 = 4.56(± 2.43) HD + 682.23(± 324.68)SP/V – 0.729(± 0.424)$WPSA\text{-}3^{B3LYP}$ – 96.60(± 49.86) $QPSX^{B3LYP}$ – 62.44(± 30.75)

[c] Descriptor abbreviations are explained in the legend to Table III.

### *Flow Direction Modeling*

For four of the 30 compounds, the observed flow rate was zero (s.a.). As a consequence, no flow direction was available in these cases, reducing the training set to 26 compounds.

In Tables VIII and IX, the LDA and BLR statistics are summarized for both calibration (total set, group1, group2) and prediction (group1, group2). With LDA, CM5 yields the overall best performance, and is the only model with an acceptable discrimination power for predicting group1, probably because it is significantly superior to all other models in calibrating group2 (note that group1 prediction is achieved through application of the group2 model). However, all LDA models including CM5 are very poor for predicting group2 (concordance $\geq 0.692$, $\kappa \geq 0.235$, $\lambda_B \leq 0.333$). It indicates that the subset-specific property profiles of group1 (13 compounds) and group2 (13 compounds) differ significantly with respect to some characteristics affecting the flow direction.

When comparing LDA (Table VIII) and BLR (Table IX), the group1 calibration of BLR is clearly better than the one of LDA (*e.g.* $\lambda_B$ 1.000 *vs.* 0.667 for CM1 and CM2), whilst the associated group2 prediction (the models which were trained on group1) is about equally poor for both methods. Here, LDA is even slightly superior to BLR in the case of CM2, keeping in mind that CM2 is apparently unable to (externally) predict the group2 flow direction properly (concordance, 0.615 *vs.* 0.538; $\kappa$, 0.216 *vs.* 0.092; $\lambda_B$, 0.167 *vs.* 0.000).

With CM3, only BLR achieves still acceptable prediction statistics for group2, which is at the same time superior to group1 prediction, and interestingly also to group1 calibration (which is used for group2 prediction). Note further that all other LDA and BLR classification models are inferior to the BLR-based CM3 in (externally) predicting the group2 flow directions. This is remarkable considering the fact that CM3 contains only two variables (like CM1 and CM2), whilst CM4 and CM5 are based on three and four molecular descriptors, respectively. The latter finding makes it difficult to qualify a particular CM as best CM.

CM4 differs from CM3 through additional inclusion of PNSA-3 (atomic charge weighted partial negative surface area)[24] as the third descriptor. As a consequence, CM4 performs significantly better for group1 calibration with both LDA and BLR than CM3 (*e.g.* $\lambda_B$ 0.667 *vs.* 0.333, Table IX), and there is also some improvement in the total set training (*e.g.* BLR: $\lambda_B$, 0.583 *vs.* 0.500), however, for group1 prediction as well as for the calibration and prediction of group2, CM3 and CM4 yield mostly similar (and partly identical) results.

Like with the models for classifying the flow rate, the question of chance correlation was addressed through calibrating the descriptor combinations on data sets where 50 % of the flow direction results were allocated wrongly. The resultant analysis as summarized in Table X shows several cases where the prediction deviates from the perfect situation (50 %) by more than 20 % (*e.g.* BLR CM4 descriptor set, scrambled Sets 1, 5 and 6). It suggests that for the models derived to predict the flow direction, the impact of chance correlations on the calibration success is probably somewhat greater than for the flow rate classification models.

As regards molecular descriptors, the variables selected for predicting the flow direction differ (except for DPSA-2) from the ones used for the flow rate categories. However, the lower part of Table IV shows that an acceptable discrimination between the left and right flow direction cannot be achieved with any of the parameters when used as single variables.

All final LDA and BLR models contain the self polarizability (SP)[26] normalized with respect to molecular volume, SP/V. SP indicates how easily the molecule can accommodate local changes in the electron distribution,[26] and the division through V aims at eliminating the associated molecular size component (as a general trend, polarizability increases with increasing molecular size). Moreover, three of the five LDA and BLR models (CM1, CM2, CM5) include molecular hardness, HD, which is also a measure of how easily the electron density of the compound can be changed.[25] The greater HD, the lower is the readiness of the molecule to donate or accept electron charge. For the purpose of modeling the flow direction induced by an external electric field, however, the mechanistic meaning of these two parameters is unclear, except that both of them represent different aspects of the electronic flexibility of the molecules.

The other descriptors are all CPSA parameters,[24] for which a relationship to the flow direction is at least more reasonable. WPSA-3 (CM3, CM4, CM5) is the total surface area weighted PPSA-3, which is defined as the sum of all positively charged surface areas weighted by their associated atomic charges. Similarly, PNSA-3 (CM4) is the sum of all negatively charged surface areas multiplied by their associated charges, and the newly introduced QPSX is the maximum value of all products of a positively charged atom and its fractional surface area (atomic

TABLE X. LDA and BLR calibration for the flow direction using scrambled activity categories[(a)]

| Descriptor set | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 | Set 7 | Set 8 | Set 9 | Set 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LDA | | | | | | | | | | |
| CM1 (HD, SP/V) | 65.4 | 57.7 | 34.6 | 50.0 | 50.0 | 65.4 | 73.1 | 61.5 | 69.2 | 53.8 |
| CM2 (HD, SP/V, DPSA-2) | 69.2 | 57.7 | 50.0 | 50.0 | 57.7 | 65.4 | 73.1 | 57.7 | 57.7 | 61.5 |
| CM3 (SP/V, WPSA-3) | 65.4 | 61.5 | 53.8 | 61.5 | 73.1 | 69.2 | 61.5 | 69.2 | 53.8 | 65.4 |
| CM4 (SP/V, WPSA-3, PNSA-3) | 73.1 | 61.5 | 50.0 | 61.5 | 73.1 | 69.2 | 46.2 | 65.4 | 57.7 | 69.2 |
| CM5 (HD, SP/V, WPSA-3$^{B3LYP}$, QPSX$^{B3LYP}$) | 61.5 | 73.1 | 65.4 | 61.5 | 69.2 | 76.9 | 65.4 | 50.0 | 65.4 | 61.5 |
| BLR | | | | | | | | | | |
| CM1 (HD, SP/V) | 65.4 | 57.7 | 34.6 | 50.0 | 50.0 | 73.1 | 73.1 | 61.5 | 69.2 | 57.7 |
| CM2 (HD, SP/V, DPSA-2) | 69.2 | 53.8 | 50.0 | 50.0 | 61.5 | 73.1 | 73.1 | 57.7 | 57.7 | 57.7 |
| CM3 (SP/V, WPSA-3) | 65.4 | 61.5 | 53.8 | 65.4 | 73.1 | 69.2 | 57.7 | 69.2 | 53.8 | 65.4 |
| CM4 (SP/V, WPSA-3, PNSA-3) | 73.1 | 61.5 | 50.0 | 61.5 | 73.1 | 76.9 | 50.0 | 69.2 | 57.7 | 65.4 |
| CM5 (HD, SP/V, WPSA-3$^{B3LYP}$, QPSX$^{B3LYP}$) | 61.5 | 76.9 | 69.2 | 61.5 | 73.1 | 76.9 | 73.1 | 50.0 | 65.4 | 61.5 |

[(a)] For each of the 10 sets, a randomly selected half of the compounds showing (truly) left flow are allocated wrongly to the right flow direction, and *vice versa* for the compounds with a (true) right flow. Descriptor abbreviations are explained in the legend to Table III.

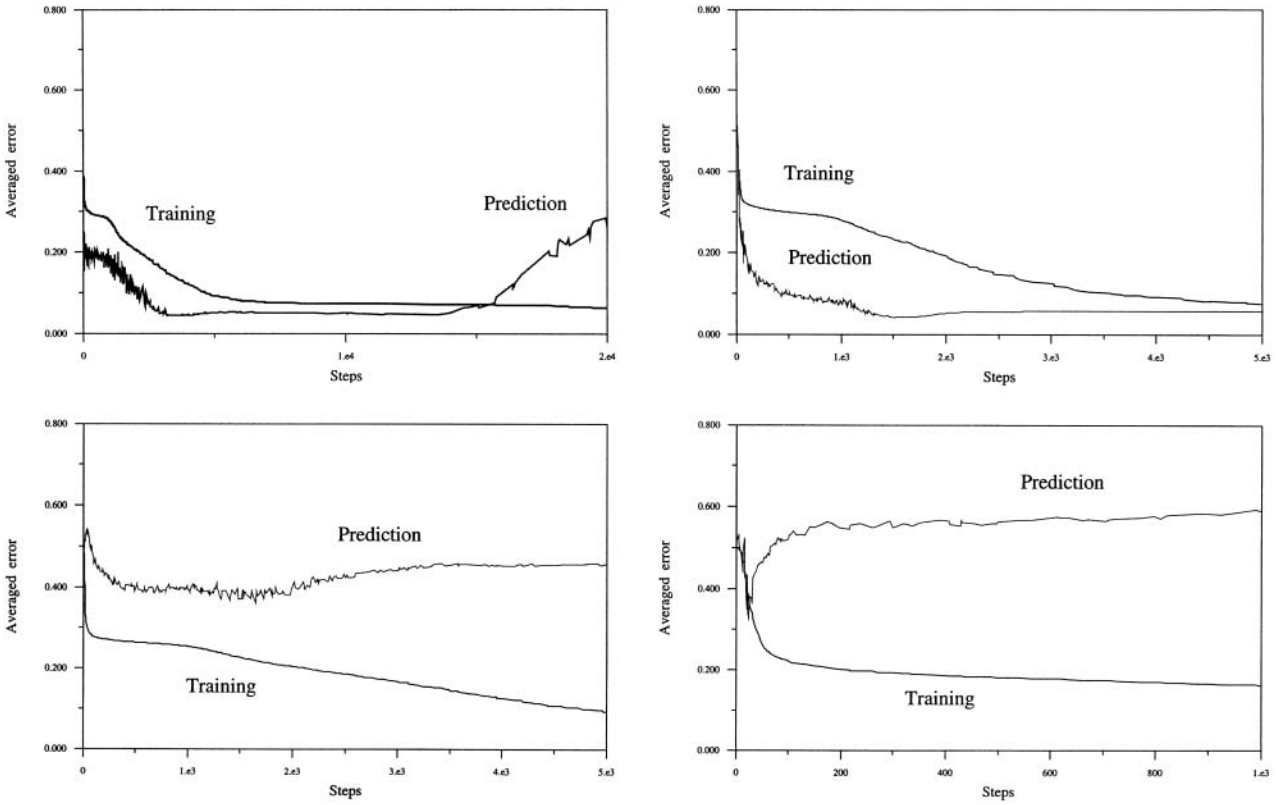


Figure 2. Dependence of the 3-layer neural network training and prediction performance on the number of iteration cycles for four different randomly selected separations of the flow rate data set into 25 training compounds and 5 prediction compounds.

surface area divided by the total surface area of the molecule). The parameter DPSA-2 (= PPSA-2 – PNSA-2) used in CM2 was already discussed in the flow rate section. Interestingly, both positively and negatively charged surface area descriptors contribute to discriminating between the two flow directions.

*Neural Network Model Analysis of the Flow Rate*

The limited size of the data set (30 and 26 compounds, respectively) poses severe limits on the possibility to elucidate the potential impact of nonlinear relationships between molecular descriptors and the flow pattern under an external electric field. Under the given circumstances, a 3-layer neural network (NN) model with three input nodes (for three molecular descriptors) plus a separate bias, two hidden-layer nodes plus a separate bias, one output node and a full connection between all non-bias nodes is considered as the maximum model size for the flow rate data set. The short-cut notation for its architecture is (3+1):(2+1):1, and it contains 10 weights as model parameters to be calibrated. It follows that when training the full flow rate data set, only three compounds are available per model parameter (note that a (4+1):(2+1):1 NN model for four descriptors would already contain 12 model parameters).

Since the iterative model training depends to some degree on the initial values selected for the model weights, all runs were performed three times with different randomly selected starting weights. More importantly, for NN backpropagation models there is a trade-off between training success and prediction capability with an increasing number of iteration cycles.[30] Thus it is mandatory to check the prediction power during the model training. Since the number of 10 model parameters does not allow applying the group1-group2 approach, we generated 10 sets with random selections of 25 training and 5 prediction compounds (keeping in mind that now the ratio of compounds to model parameters is already below 3). For the same reason, NN analyses of the flow direction were not undertaken simply due to the too small ratio of compounds to model parameters.

In Figure 2, the dependence of the model training and prediction error on the number of iteration cycles is shown for four such settings, employing $\mu^{exp}$, log $K_{ow}$ and WPSA-1$^{B3LYP}$ as molecular descriptors. Taking the bottom right plot as an example, the training error (referring to 25 randomly selected compounds) decreases over the whole range up to 1000 steps (with only small improvements for the range above 100 steps), whilst the prediction error (referring to 5 compounds left out for the training) shows an early sharp minimum at about 20 steps, with a subsequent error increase over the starting error.

The error pattern in the top right plot (referring to a different separation into 25 training and 5 prediction com-

pounds) represents a rather unusual situation, in which the prediction error is below the training error for the whole range monitored (up to 5000 steps). The situation in the bottom left plot is again less common, since the increase in prediction error with the increasing number of iteration cycles fluctuates and is on the whole rather moderate. Finally, the top left plot also shows an unusual pattern, where both training and prediction error decrease similarly over a greater number of steps, and the prediction error remains below the training error for quite a long training period up to *ca*. 50000 steps.

These four results differ significantly from each other, and additional, rather unique, patterns are achieved with the other six settings (results not shown). This suggests that the number of compounds is simply too small to derive a stable NN model for predicting the flow rate.

In Table XI, contingency table statistics are shown as value ranges for 10 runs using thresholds for the NN output node of < 0.333 (low flow rate) and > 0.666 (high flow rate; top) and < 0.5 (low flow rate) and > 0.5 (high flow rate; bottom), respectively. For the statistical evaluation, the optimum number of iteration cycles was specified for each run separately, keeping in mind the large differences as illustrated in Figure 2. As can be seen from the table, good calibrations contrast with predictions ranging from perfect (*e.g.* $\lambda_B = 1$) to useless ($\lambda_B = 0$), where the latter depends strongly on the particular subset of 5 compounds selected for the prediction mode. It confirms that for the present data set, the number of compounds is indeed too small to allow derivation of predictive NN models.

TABLE XI. Contingency table statistics of a 3-layer neural network model employing $\mu^{exp}$, log $K_{ow}$ and WPSA-1$^{B3LYP}$ for discriminating between the flow rate categories high and low[(a)]

| Association parameter | Calibration | Prediction |
|---|---|---|
| Thresholds < 0.333 and > 0.666 | | |
| Concordance | 0.840...1.0 | 0.600...1.0 |
| $\kappa$ | 0.841...1.0 | 0.286...1.0 |
| $\lambda_B$ | 0.667...1.0 | 0...1.0 |
| Threshold 0.5 | | |
| Concordance | 0.960...1.0 | 0.600...1.0 |
| $\kappa$ | 0.920...1.0 | 0.286...1.0 |
| $\lambda_B$ | 0.900...1.0 | 0...1.0 |

[(a)] The 3-layer neural network (NN) contains bias nodes in the input and hidden layer, resulting in an architecture (3+1):(2+1):1. Thresholds < 0.333 and > 0.666 indicate that respective NN output values are interpreted as allocating the compound to the categories of low and high flow rate, respectively, thus implying that NN outputs in the medium value range 0.333 – 0.666 are interpreted by definition as wrong allocation (upper part of the table). By contrast, the threshold 0.5 indicates that NN output values below and above 0.5 allocate the compound to the category of low and high flow rate, respectively. The descriptor abbreviations are explained in the legend of Table III.

## CONCLUSIONS

In organic fluids, an external electric field can induce spatial movement of dielectric compounds that depends on the field strength and polarity. Our analysis shows that this flow pattern is related to physicochemical and electronic properties of the compounds available or can be calculated using routine quantum chemical methods. The flow rate appears to be governed by the molecular polarity as well as by electrostatic interactions, which is also reasonable from the mechanistic viewpoint. Whilst chance correlations were shown to play no significant role, the currently achieved prediction capability is only moderate, which is probably caused by the still quite limited data set available for model derivation. It follows that with a more extended set of experimental data, with compounds covering a wider range of chemical functionalities, the identified types of significant molecular descriptors are likely to yield more predictive models. As regards CPSA descriptors and other parameters related to net atomic charges, however, the level of computation (semiempirical *vs*. *ab initio*) should be selected judiciously depending on the compound classes under investigation.[25]

For the flow direction, both calibration and prediction statistics are inferior to the flow rate models, which may be partly caused by the smaller number of compounds available for this property. Besides electrostatic interaction, electronic softness appears to affect the flow direction, which is somewhat surprising and requires further investigation. For future studies aimed to understand and predict the flow pattern of organic compounds under an external electric field, the present findings provide guidance as regards mechanistically relevant molecular descriptors.

*Acknowledgement.* – This work was supported in part by the European Union IMAGETOX Research Training Network, HPRN-CT-1999-00015.

## REFERENCES

1. I. Sumoto, *Oyo Butsuri* **25** (1956) 264–265.
2. Y. Otsubo and K. Edamura, *Appl. Phys. Lett.* **71** (1997) 318–320.
3. A. Yabe and H. Maki, *Int. J. Heat Mass Transfer* **31** (1988) 407–417.
4. Y. Iwai, N. Ito, K.Yoshida, Y. Arai, and H. Itahara, *Ind. Eng. Chem. Res*. **37** (1988) 3782–3785.
5. S. Yokota, A. Sadamoto, Y. Kondoh, Y. Otsubo, and K. Edamura, *Trans. Jpn. Soc. Mech. Eng.* **66** (2000) 273–279.
6. Y. Iwai, K.Yoshida, Y. Arai, G. Schüürmann, B. Loeprecht, W. M. F. Fabian, and T. Suzuki, *J. Chem. Inf. Comput. Sci*. **40** (2000) 988–993.
7. G. Schüürmann, A. O. Aptula, R. Kühne, and R. U. Ebert, *Chem. Res. Toxicol.* **16** (2003) 974–987.
8. G. Klopman and A. N. Kalos, *J. Comput. Chem*. **6** (1985) 492–506.
9. A. O. Aptula, R. Kühne, R.-U. Ebert, M. T. D. Cronin, T. I. Netzeva, and G. Schüürmann, *QSAR Comb. Sci.* **22** (2003) 113–128.

10. M. Boháč, B. Loeprecht, J. Damborský, and G. Schüürmann, *Quant. Struct. Act. Relat.* **21** (2002) 3–11.
11. A. O. Aptula, T. I. Netzeva, I. V. Valkova, M. T. D. Cronin, T. W. Schultz, R. Kühne, and G. Schüürmann, *Quant. Struct. Act. Relat.* **21** (2002)12–22.
12. J. Sangster, LOGKOW – A databank of evaluated octanol-water partition coefficients. Sangster Research Laboratories, Montreal, Canada, 1993.
13. J. A. Dean (Ed.), *Lange's Handbook of Chemistry*, 13th ed., McGraw Hill, New York, 1985.
14. SYBYL Molecular Modeling Software 6.4. Tripos Asociates Inc., St. Louis, Missouri, 1998.
15. MOPAC 93, Rev. 2, 1994, Fujitsu Ltd., 9-3, Nakase 1-Chome, Mihama-ku, Chiba-city, Chiba 261, Japan, and Stewart Computational Chemistry, 15210 Paddington Circle, Colorado Springs, Colorado 80921, USA.
16. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* **107** (1985) 3902–3909.
17. Gaussian 98 (Revision A.7) 1999. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, J. A. Pople, Gaussian Inc., Pittsburgh, PA, USA.
18. A. D. Becke, *J. Chem. Phys.* **98** (1993) 5648–5652.
19. C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B* **37** (1988) 785–789.
20. P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta* **28** (1973) 213–222.
21. W. J. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople, *Ab initio Molecular Orbital Theory*, John Wiley, New York, 1986, p. 548.
22. G. M. Smith, MOLSV. QCPE program No. 509, 1985.
23. G. Schüürmann, *Assessment of semiempirical quantum chemical continuum-solvation models to estimate $pK_a$ of organic compounds*, in: F. Chen and G. Schüürmann (Eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences* – VII, SETAC Press, Pensacola (FL), USA, 1997, pp. 225–242.
24. D. T. Stanton and P. C. Jurs, *Anal. Chem*. **62** (1990) 2323–2329.
25. G. Schüürmann, *Quantum chemical descriptors in structure-activity relationships – Calculation, interpretation and comparison of methods*, in: M. T. D. Cronin and D. J. Livingstone (Eds.), *Predicting chemical toxicity and fate*, CRC Press, Boca Raton (FL), USA, 2004, Chapter 6, in press.
26. G. Schüürmann, *Ecotoxic modes of action of chemical substances*, in: G. Schüürmann and B. Markert (Eds.), *Ecotoxicology*, John Wiley and Spektrum Akademischer Verlag, New York, USA, 1998, pp. 665–749.
27. STATISTICA for Windows '99 Edition, Statsoft Inc., Tulsa, OK, USA, 1999.
28. SPSS for Windows, Rel. 10.0. SPSS Inc., Chicago, IL, USA, 1999.
29. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design. An Introduction*, 2nd edn., Wiley-VCH, Weinheim, Germany, 1999, p. 194.
30. G. Schüürmann and E. Müller, *Environ. Toxicol. Chem.* **13** (1994) 743–747.
31. J. Hartung, B. Elpelt, and K.-H. Klösener, *Statistik*, 13th edn., Oldenbourg Verlag, München, 2002, p. 975.
32. J. Cohen, *Educ. Psychol. Measure* **20** (1960) 37–46.
33. T. Suzuki, *J. Comput. Aided Mol. Des.* **5** (1991) 149–166.

## SAŽETAK

### Kategorijsko modeliranje protoka tekućih organskih spojeva između elektroda s krilcima pomoću semiempirijskih i *ab intio* kvantnih kemijskih deskriptora

**Takahiro Suzuki, Kohei Yoshida, Hiroya Onizuka, Yoshio Iwai, Yasuhiko Arai, Aynur Aptula, Ralph Kühne, Ralf-Uwe Ebert i Gerrit Schüürmann**

Kategorijsko modeliranje primijenjeno je na skup podataka od 30 organskih tekućina kako bi se prognozirao njihov protok u uvjetima vanjskog električnog polja. U tu svrhu prethodno sastavljen skup podataka proširen je s 10 spojeva s novim eksperimentalnim rezultatima te su primijenjene kvantne kemijske metode za karakterizaciju elektronske strukture molekula na semiempirijskoj i *ab initio* razini teorije. Linearna diskriminacijska analiza (LDA) i binarna logistička regresija (BLR) upotrijebljene su za modeliranje brzine protoka (velika / mala) i smjera protoka (ulijevo / udesno). Za brzinu protoka, dobra kalibracijska statistika za LDA i BLR uporabom dipolnog momenta, hidrofobnosti i deskriptora dijela površine pod nabojem (CPSA) popraćena je umjerenom prognostičkom statistikom, što je utvrđeno simuliranom vanjskom provjerom, a poremećivanje aktivnosti pokazuje da slučajna korelacija nije važna. Dodatne analize neuronske mreže nisu dale stabilne modele radi ograničenja zbog veličine skupa podataka. Za smjer protoka, kalibracijska i prognostička statistika za LDA i BLR pokazuju veće razlike između izrađenih modela, a opći im je učinak lošiji od onog za brzinu protoka. Uz CPSA deskriptore tu su uključena još dva parametra koji karakteriziraju mekoću elektronske strukture. Općenito, BLR je nešto bolja od LDA za oba svojstva. O rezultatima se raspravlja u kontekstu statistike tablice kontingencije i u odnosu na mehanističko značenje molekularnih deskriptora.