Exploration of Quantitative Structure–Reactivity Relationships for the Estimation of *Mayr* Nucleophilicity

by **Diogo A. R. S. Latino***[a])[b])[1]) and **Florbela Pereira**[a])

[a]) CQFB and LAQV-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, PT-2829-516 Caparica
(phone: +351-21-2948300; fax: +351-21-2948550; e-mail: diogolatino@gmail.com)

[b]) CCMM, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa,
Campo Grande, PT-1749-016 Lisboa

Quantitative structure–reactivity relationships (QSRRs) were investigated for the estimation of the *Mayr* nucleophilicity parameter N using data sets with 218 nucleophiles (solvent: CH_2Cl_2) and 88 compounds (solvent: MeCN) extracted from the *Mayr*'s Database of Reactivity Parameters. The best predictions were observed for consensus models of random forests and associative neural networks, trained with empirical 2D and 3D CDK molecular descriptors, which yielded RMSE of 1.54 and 1.97 for independent test sets of the two solvent data sets, respectively. Compounds with silicon atoms were more difficult to predict, as well as classes of compounds with a reduced number of examples in the training set. The models' predictions were consistently more accurate than estimations simply based on the average of the N parameter within the class of the query compound. The possibility of calculating rate constants using the obtained models was also explored.

Introduction. – Nucleophilicity and electrophilicity are useful concepts for rationalizing the electronic aspects of reactivity, selectivity, and substitution effects in organic reactions. Ultimately, they can be applied for reaction prediction, not only within the specific boundaries of organic chemistry, but also in the assessment of biological mechanisms involving chemical reactivity. In fact, the prediction of toxicological endpoints [1][2] such as skin sensitization, cytotoxicity, genotoxicity, chromosomal aberration, hepatotoxicity, or acute aquatic toxicity, which include mechanisms triggered by covalent bonding to biological molecules, requires methodologies for reaction prediction, which have often incorporated electrophilicity parameters [3]. Biological molecules, such as peptides, proteins or enzymes, lipids, or DNA are typically targets for covalent binding of small electrophiles, xenobiotic molecules [1]. Biochemical mechanisms involving nucleophilic xenobiotics are also known. The enzyme acetylcholinesterase (AChE, EC 3.1.1.7) has a peripheral anionic site (PAS) comprising a set of aromatic residues which provides a binding site for nucleophilic allosteric modulators and nucleophilic inhibitors [4]. Organophosphorus (OP) compounds (*e.g.*, tabun, soman, diisopropyl fluorophosphates, sarin, cyclosarin, pesticides) are known as nerve agents and considered potential warfare threats due to their high intrinsic toxicity [4][5]. The acute toxicity of OP compounds in mammals is due to inhibition of the enzyme AChE [5]. In the last years, several efforts have been made to

[1]) Current address: Environmental Chemistry, Eawag, Überlandstrasse 133, CH-8600 Dübendorf.

develop small nucleophilic inhibitors of AChE for medical management of nerve agent poisoning. Nucleophilic inhibitors of AChE have also been envisaged for the treatment of *Alzheimer*'s disease and glaucoma [4].

Quantitative scales of electrophilicity/nucleophilicity have been proposed both for the rationalization of chemical reactivity and for the prediction of new reactions. A most established scale has been proposed by *Mayr* and co-workers to explain diverse types of reactions [6][7]. It was demonstrated for a series of electrophile–nucleophile combinations that the kinetic rate constants of reactions can be fit to the following linear relationship

$$\log k\ (20^\circ) = s_{\mathrm{N}}(N + E) \qquad (1)$$

where E and N are the electrophilicity and nucleophilicity parameters, respectively, and s_{N} is a system-specific parameter, the sensitivity parameter, which is dependent on the reference nucleophile. The original *Mayr* E and N scales were derived from reference carbon electrophiles, *e.g.*, alkenes, arenes, alkynes, enol ethers, enamines, diazo compounds, carbanions, hydride donors, phosphanes, amines, and alkoxides, that were employed to compare the nucleophilicities of a large variety of compounds using *Eqn. 1* [8]. This was also used to quantify the electrophilicity parameter E for different types of electrophiles, such as carbocations, typical *Michael* acceptors, and electron-deficient arenes [8]. The obtained E, N, and s_{N} parameters can be used for semiquantitative prediction of rates and selectivities of polar organic reactions. The *Mayr*'s Database of Reactivity Parameters [9] contains a compilation of published reactivity parameters that in July 2012 comprised information on 706 nucleophiles and 218 electrophiles spanning a nucleophilicity range of $-4.47 \leq N \leq 28.95$ and an electrophilicity range of $-23.80 \leq E \leq 6.16$ for a wide variety of molecule classes.

On the theoretical point of view, many efforts have been reported to define and quantify the electrophilicity of molecules using quantum calculations. Electronegativity and hardness were rigorously defined using conceptual density functional theory (DFT) to arrive at an electrophilicity index [3][10][11]. On the basis of the assumption that electrophilicity and nucleophilicity are inversely related to each other, *Chattaraj* and co-workers [12] proposed that the nucleophilicity index can be considered as inverse of the electrophilicity index. Over the last years, many reports have appeared, in which the electrophilicity index and derivatives could be successfully correlated with experimental chemical reactivity, spectroscopic data, toxicological end points, and biological activities [3]. High correlations ($R > 0.94$) were presented between the *Mayr* electrophilicity parameter and the electrophilicity index within series of compounds such as benzene diazonium ions [13] and benzhydryl cations [14], as well as between the *Mayr* nucleophilicity N parameter and the nucleophilicity index for a specific pyridine series [15]. In spite of the reasonable agreement ($R > 0.70$) between the *Mayr* nucleophilicity parameter and the theoretical nucleophilicity index reported in some works [16][17], lower correlations were usually observed as compared with those in electrophilicity studies. Recently, *Chamorro et al.* [18] proposed relative electrophilicity and nucleophilicity electronic theoretical indices for electrophile–nucleophile pairs of combining species.

Following our previous work [19] on a QSPR approach for the estimation of *Mayr* electrophilicity, we wondered if *Mayr* nucleophilicity parameters could be quickly and accurately estimated by data-driven QSPR approaches exclusively from empirical molecular descriptors. Here, we report the results of that investigation using state-of-the art machine learning (ML) techniques and well-established empirical molecular descriptors. It is to emphasize that the main purpose of this work is to present QSRR models for the estimation of *Mayr* nucleophilicity rather than the theoretical interpretation of the involved physico-chemical phenomena. Additionally, the application of these models to the estimation of rate constants is illustrated.

Data Sets and Computational Methods. – Mayr *Nucleophilicity Data Set.* The *Mayr* nucleophilicity parameter *N* and the sensitivity parameter s_N were extracted from the *Mayr*'s Database of Reactivity Parameters (July 2012) [6] for all available uncharged compounds, except P-nucleophiles and ylides. The corresponding molecular structures were drawn using MarvinSketch 5.2. (*ChemAxon Ltd.*, Budapest, Hungary) and saved as SMILES strings (available as Supplementary Information from the author).

As the *Mayr* nucleophilicity parameter *N* was defined as solvent-dependent, two data sets were extracted for CH_2Cl_2 and MeCN solvents separately. The CH_2Cl_2 data set consists in 168 C-nucleophiles (13 conjugated 1,3-dienes, 39 mono-enes, 1 alkyne, 19 allyl compounds, 5 carbocyclic arenes, 4 indoles, 5 isonitriles, 8 other heterocyclic arenes, 3 pyrroles, 7 diazo compounds, 22 enamines and enamides, 3 enol ethers of type C=C-OR, 29 enol ethers of type C=C–OSi, 9 enol ethers of type C=C(OR)(OSi), 1 enol ether of type C=C–$(OR)_2$), 17 H-nucleophiles (H–C hydride donors), 33 N-nucleophiles (4 aliphatic amines, 1 amidine, 8 guanidines, 11 isothioreas, and 9 pyridines, quinolones, *etc.*). The data set was randomly partitioned yielding 151 compounds in the training set and 67 compounds in the test set. The data set for solvent MeCN consists in 41 C-nucleopiles (3 mono-enes, 1 carbocyclic arene, 13 indoles, 4 pyrroles, 17 enamines and enamides, 3 enol ethers of type C=C(OR)(OSi)), 3 H-nucleophiles (H–C hydride donors) and 44 N-nucleophiles (18 aliphatic amines, 2 amidines, 4 aromatic amines, 10 azoles, 1 compound from the class of hydrazines, hydroxylamines *etc.*, and 9 pyridines, quinolines, *etc.*). The data set was randomly partitioned into a training set with 61 compounds and a test set with 27 compounds.

Calculation and Selection of Molecular Descriptors. CORINA version 2.4. (*Molecular Networks GmbH*, Erlangen, Germany) was used for the generation of three-dimensional models of the molecular structures from SMILES strings. Then, the CDK Descriptor Calculator 1.3.2 [20] was used in the calculation of empirical molecular descriptors. All descriptors available in the CDK software were calculated with exception of ionization potential. The CDK Descriptor calculator includes electronic, topological, geometrical, constitutional, and hybrid (BCUT and WHIM) descriptors that implicitly encode properties expected to be nucleophilicity related.

After the removal of constant or *quasi*-constant descriptors (223 and 208 descriptors were obtained from the CH_2Cl_2 and MeCN solvents data sets, resp.), an independent selection of the most relevant descriptors to establish QSRRs for the *N* parameter was performed with the CFS (Correlation-based Feature Subset Selection) algorithm [21–23] available in Weka 3.6.5. This filter simultaneously maximizes the correlation with the dependent variable to predict and minimizes intercorrelation between descriptors. The selection of descriptors was performed with the CFS algorithm within a ten-fold cross-validation procedure on the training set and *k nearest neighbor* (KNN) algorithm as the ML technique. All experiments were performed with the same partitions of the data sets. For the CH_2Cl_2 and MeCN data sets 12, and 8 descriptors were respectively selected and were used to develop SVM and AsNN models. The selected descriptors are available as Supporting Information.

Support Vector Machines (SVM). Support vector machines (SVM) [24] map the data into a hyperspace through a non-linear mapping (a boundary or hyperplane) and then run a linear regression in this space [25]. The boundary is positioned using examples in the training set, which are known as the support vectors. With non-linear data, kernel functions can be used to transform it into a hyperspace where the linear regression can be done. In this study, SVMs were established with the Weka 3.6.5, using the LIBSVM software [26]. The type of SVM was set to ε-SVM-regression, the kernel function was the radial basis function. The default γ parameter in the kernel function was used and the parameter C of the ε-SVM-regression was set in the range of 10–500. Data were normalized (descriptors selected by the CFS procedure).

Random Forests. A Random Forest [27], RF, is an ensemble of unpruned trees which was created using bootstrap samples of the training set. In this process, for each individual tree the best split at each node is defined using a randomly selected subset of descriptors. Each individual tree is created using a different training and validation set and also a different set of descriptors. The final prediction for an object from a Random Forest is obtained as an average of the predictions of the individual regression trees in the forest. RFs were grown with R program [28], version 2.13.1 and using the Random Forest library [29]. The number of trees in the forest was set to 1000, the number of variables tested at each split was set to the square root of the total number of variables or higher. RF models were built for the CH_2Cl_2 and MeCN solvent data sets using 223 and 208 CDK descriptors, respectively.

Associative Neural Networks. Associative Neural Networks (AsNNs) [30] integrate an ensemble of Feed-Forward Neural Networks (FFNNs) with a memory of data. The ensemble consists of independently trained FFNNs, which contribute to a single prediction. The final prediction for an object from an AsNN is obtained from *a*) the outputs produced by the ensemble of individual FFNNs and *b*) the most similar cases in the memory (here, the training set). The *Levenberg–Marquardt* learning algorithm [31] was used for training FFNNs with an input layer, one hidden layer, and one output neuron. The number of hidden neurons was optimized for each data set using the internal validation data sets and was set to 5. The logistic activation function was used, and each input and output variable was linearly normalized between 0.1 and 0.9 on the basis of the training set. Before the training of each NN, the training set was randomly divided into a learning set and validation set; each one with 50% of the objects. Full cross-validation of the training set was performed using the leave-one-out (LOO) method. The maximum number of iterations in the training was set to 1000. The training was stopped when there was no further improvement in the root-mean-square error (RMSE) for the validation set. The experiments were performed using the AsNN program of *Igor Tetko* [32]. The AsNN models were built with the sets of 12 and 8 descriptors selected by the CFS procedure for the CH_2Cl_2 and MeCN solvent data sets, respectively.

Results and Discussion. – *Establishment of QSRR Models for* Mayr *Nucleophilicity* N *Parameter.* The descriptors used to build each model were selected using the CFS filter [21–23] (except for the Random Forest model, where the 223 and 208 descriptors were used with the CH_2Cl_2 and MeCN data sets, respectively). The performance of different machine learning techniques using empirical CDK descriptors to model the *N* parameter were compared – *Table 1* and *Table 2*. All experiments were performed with the same partitions of the data sets and the models were optimized using only the training set (results from the internal validation procedures of each method).

For the data set of CH_2Cl_2 solvent, the SVM and AsNN models were built using twelve CDK descriptors. The AsNN model performed best among individual models, with a RMSE of 1.96 and a R^2 of 0.88 in internal validation. When the developed model was applied to the independent test set, a RMSE of 1.70 and a R^2 of 0.90 were observed. In this article, R^2 refers to the squared *Pearson* correlation coefficient between the predicted and experimental values of *N*.

For the MeCN solvent data set, eight CDK descriptors were selected by the CFS filter of Weka and used to train the SVM and AsNN models. Similarly to the previous experiments, AsNN provided the best models achieving RMSE of 2.08 and R^2 of 0.79 for the internal validation set, and RMSE of 2.19 and R^2 of 0.80 for the external test set.

Table 1. *Comparison of Different Machine Learning Techniques for QSRR of* Mayr *Nucleophilicity* N *Parameter* (CH_2Cl_2 solvent).

Method[a])	R^2/RMSE	
	Training Set[b])	Test Set
RF	0.84/2.32	0.90/1.72
SVM	0.82/2.38	0.87/1.99
AsNN	0.88/1.96	0.90/1.70
CM Model (RF and AsNN)	0.88/2.02	0.92/1.54

[a]) RF – Random Forest; SVM – Support Vector Machine; AsNN – Associative Neural Networks; CM Model – Consensus model. [b]) Results from out-of-bag (OOB) estimation in RF, ten-fold Cross-Validation in SVM, and internal validation in AsNN.

Table 2. *Comparison of Different Machine Learning Techniques for QSRR of* Mayr *Nucleophilicity* N *Parameter* (MeCN solvent).

Method[a])	R^2/RMSE	
	Training Set[b])	Test Set
RF	0.73/2.39	0.79/2.28
SVM	0.72/2.36	0.77/2.40
AsNN	0.79/2.08	0.80/2.19
CM Model (RF and AsNN)	0.83/1.98	0.86/1.97

[a]) RF – Random Forest; SVM – Support Vector Machine; AsNN – Associative Neural Networks; CM Model – Consensus model. [b]) Results from out-of-bag (OOB) estimation in RF, ten-fold Cross-Validation in SVM, and internal validation in AsNN.

In both cases, consensus models combining predictions from the three models were investigated, considering that an individual model may exhibit weaknesses in some regions of the chemical space, which may be overcome by a consensus model. For both data sets, the use of a consensus model with the two best performing individual models (AsNN and RF) yielded more accurate predictions for the test sets than the best individual models. Inclusion of the SVM predictions was found to deteriorate the accuracy of the predictions. *Fig. 1* shows plots of predicted *vs.* experimental *N* parameter using the AsNN-RF consensus models.

In the experiments with the CH_2Cl_2 solvent data set, predictions for the training set appear consistently worse than those for the test set. This could be a consequence of the structure of our data set, and the existence of outliers. If the three worst predictions in the training set are removed, the performance improves from R^2/RMSE of 0.88/2.02 to 0.90/1.79. Another explanation for the better performance with the test set relatively to

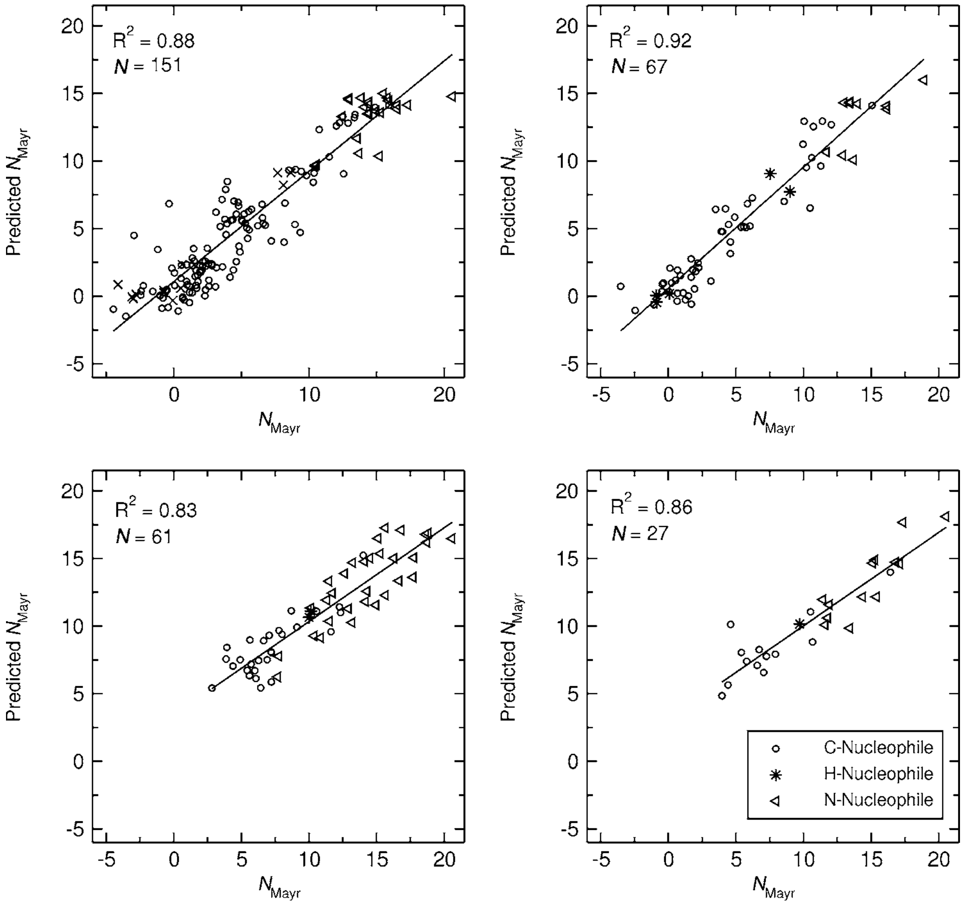


Fig. 1. *Predicted* vs. *experimental values for parameter* N *obtained from: training set in internal validation procedures* (left) *and test set* (right). The upper row corresponds to predictions for the CH_2Cl_2 solvent and the bottom row to predictions for the MeCN solvent.

the training set could be the distribution of the data by classes of compounds. There are classes with only a small number of nucleophiles in the training set that were not represented in the test set, or were represented by only one nucleophile, *e.g.* alkynes, C=C–OR, C=C(OR)$_2$, pyrroles, aliphatic amines, and amidines for the CH_2Cl_2 solvent, and carbocyclic arenes, mono-enes, H–C hydrides, amidines, and hydrazines for the MeCN solvent data set. In classes of compounds with two or three nucleophiles exhibiting a similar parameter *N*, and one with a very different parameter *N*, the possibility of the models to learn is significantly reduced.

Fig. 2 represents the distribution of the errors for the predictions obtained by the consensus models for CH_2Cl_2 and MeCN solvents data sets. The error varies between -5.82 and 7.45 in parameter *N* units for the CH_2Cl_2 solvent data set, and between -4.11 and 4.49 for the MeCN solvent data set. For the CH_2Cl_2 solvent data set, 43% of

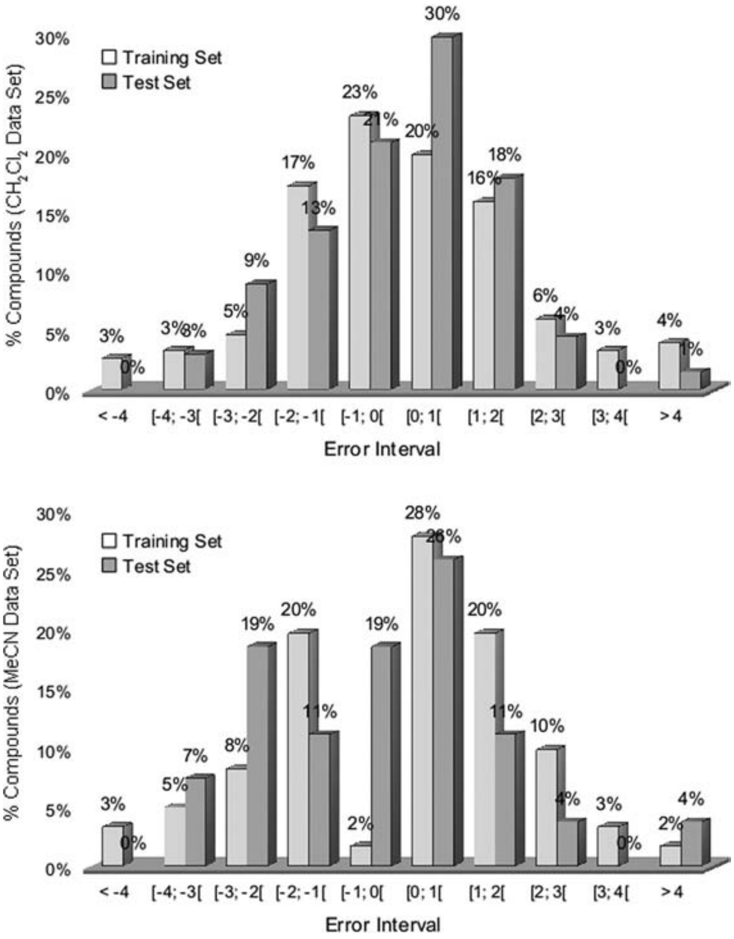


Fig. 2. *Distribution of the errors for the predictions of* N *obtained by the consensus models for CH_2Cl_2* (top) *and MeCN* (bottom) *solvents data sets*

the nucleophiles in the training set and 51% of the nucleophiles in the test set were predicted with an absolute error lower than 1. On the opposite side, only 13% of the nucleophiles in the training set and 4% of the nucleophiles in the test set were predicted with absolute errors larger than 3.

Table 3 discriminates the results according to classes of nucleophiles. Several classes that present large errors in the test set are represented only by one compound in the test set and by few compounds in the training set (*e.g.* carbocyclic arenes, pyrroles, isonitriles, aliphatic amines, diazo compounds). Pyridines and quinolones are an exception. The classes of nucleophiles with the highest RMSE were aliphatic amines and the diazo compounds in both solvents data sets. The diazo class has five compounds in the training set with the N parameter varying from -0.35 to 9.35. This is the widest range of N in classes represented by a low number of compounds, making this one of the most difficult classes to establish a QSRR. *Fig. 3* shows the structures, *Mayr* nucleophilicity, and predictions by the RF-AsNN consensus model for this class.

Table 3. *RMSE of AsNN-RF Consensus Model Predictions by Class of Compound.*

Nucleophile Type	Class	RMSE			
		Training Set[a])		Test Set	
		CH_2Cl_2	MeCN	CH_2Cl_2	MeCN
C-Nucleophiles	Conjugated 1,3-dienes	1.49	–	1.16	–
	Mono-enes	1.33	3.45*[b])	1.07	1.55*
	Alkynes	1.43*	–	–	–
	Allyls	1.41	–	1.18	–
	Carbocyclic arenes	3.13	2.26*	1.94*	–
	Indoles	0.59	1.88	1.05*	1.54
	Other heterocyclic arenes	0.88	–	1.09*	–
	Pyrroles	2.42*	2	0.99*	1.84*
	Isonitriles	2.37	–	2.31*	–
	Diazo compounds	4.63	–	2.88*	–
	Enamines and enamides	1.5	1.71	1.77	2.56
	C=C–OSi	2.24	–	1.54	–
	C=C(OR)(OSi)	2.47	1.11*	1.05	0.53*
	Enol ethers C=C–OR	1.34*	–	0.88*	–
	Enol ethers C=C–$(OR)_2$	0.88*	–	–	–
H-Nucleophiles	H–C hydride donors	2.2	0.81	1.01	0.44
N-Nucleophiles	Aliphatic amines	4.06	1.88	2.91*	1.71
	Amidines	0.51*	0.89*	–	–
	Guadinines	1.08	–	1.81	–
	Isothioureas	1.48	–	1.05	–
	Aromatic amines	–	2.07	–	3.57*
	Azoles	–	1.13	–	0.77
	Hydrazines, Hydroxylamines, *etc.*	–	2.69*	–	–
	Pyridines, quinolones, *etc.*	2.51	3.14	2.6	2.38
Total (RMSE/R^2)		2.02/0.88	1.98/0.83	1.54/0.92	1.97/0.86

[a]) Results from out-of-bag (OOB) estimation in RF and internal validation in AsNN. [b]) * Classes of compound represented only by one or two compounds in training or test set.

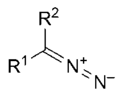
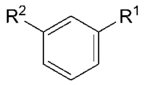
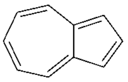
	Structures	ID (Tr/Te)	N_{Mayr}/Predicted N_{Mayr}
Diazo Compounds		**19** (Tr): R^1 = H, R^2 = Me_3Si	8.97/5.31
		129 (Tr): R^1 = H, R^2 = Me(O)C	3.96/8.48
		130 (Te): R^1 = R^2 = H	10.48/6.52
		131 (Tr): R^1 = R^2 = EtOOC	–0.35/6.84
		132 (Tr): R^1 = R^2 = Ph	5.29/5.41
		136 (Te): R^1 = H, R^2 = EtOOC	4.91/5.85
		165 (Tr): R^1 = H, R^2 = Ph	9.35/4.72
Carbocyclic Arenes		**29** (Tr): R^1 = R^2 = MeO	2.48/3.55
		100 (Te): R^1 = Me, R^2 = MeO	0.13/2.07
		120 (Tr): R^1 = H, R^2 = MeO	–1.18/3.45
		141 (Tr): R^1 = Me, R^2 = Me	–3.54/–1.48
		172 (Tr): R^1 = H, R^2 = Me	–4.47/–0.96
		**70** (Tr) for MeCN data set	6.66/8.91

Fig. 3. *Molecular structures of diazo compounds and carbocyclic arene classes, respective parameter* N, *and predictions from the RF-AsNN consensus model.* Tr, training set; Te, test set.

The carbocyclic arenes class for the CH_2Cl_2 data set is also represented by four compounds in the training set (Tr) encompassing a wide range of *N* values, from – 4.47 to 2.48 (only one compound has a positive value of *N*). Carbocyclic arenes were predicted by the RF-AsNN consensus model with a RMSE of 3.13 and 1.94 for the training and test sets, respectively.

Non-accurate predictions for other problematic cases could be explained by their structure or value of *N* – compounds ID95, ID19 (**19**), ID131 (**131**), ID70 (**70**), ID10, ID101, and ID12. Compound ID95 yielded the highest error in the training set – 7.45 in *N* units. This compound belongs to the class C=C–OSi and is, at the same time, the only compound in this class with the fragment CF_3 in the structure, and the only compound with a negative value (–2.94) of *N*. *Fig. 4* shows the structure and *N* values of some silicon compounds that were predicted with higher errors than the average in their class.

Compound ID19 (**19**) illustrates a similar situation – it is a diazo compound and the only with a Si atom in its class. Compound ID131 (**131**) is the only diazo compound with a negative *N*, while the most similar compound, ID129 (**129**), (by visual inspection within the class) has *N* = 3.96 (see *Fig. 3*). Compound ID165 (**165**) has the highest value of *N* in its class (9.35) and the most similar compound (ID132; **132**) has a value of *N* = 5.29 (see *Fig. 3*). Compound ID130 (**130**), in the test set (Te), is the simplest and most distinct structure of the diazo compounds, and the compound with the largest value of *N* (10.48) – as expected, it obtained one of the worst predictions in the test set. The class of enamines and enamides compounds were relatively well predicted. Nonetheless, compounds ID70 (**70**), ID10, and ID101 were predicted with higher error than the average in their class (see *Fig. 4*). These three compounds have Si atoms in their

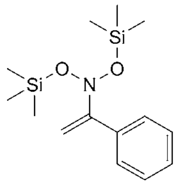
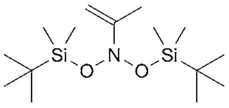
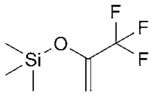
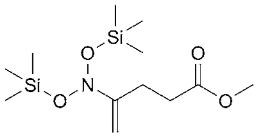
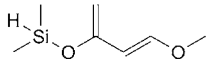
Structures	ID (Tr/Te)	Class	N_{Mayr}/Predicted N_{Mayr}
	**10** (Tr)	Enamines/Enamides	4.80/6.68
	**70** (Te)	Enamines/Enamides	4.23/6.46
	**95** (Tr)	C=C–OSi	–2.94/4.51
	**101** (Tr)	Enamines/Enamides	3.84/7.89
	**128** (Te)	1,3-Dienes	8.57/7.00

Fig. 4. *Molecular structures of silicon compounds with their experimental and predicted* (by the RF-AsNN consensus model) *values of* N

structure. A similar situation occurs for compound ID128 (**128**) of the 1,3-dienes class, which is the only compound of this class with Si in their composition.

The RF model for the CH_2Cl_2 data set was used to perform a further validation with the *y*-randomization procedure. The training set was modified by scrambling the *y*-column (the parameter *N*) and keeping the descriptor matrix unchanged. The RF model was then retrained with the modified data. The performance of the obtained model was significantly worse than the real RF model – a R^2 of 0.09 against 0.84 and a RMSE of 5.99 against 2.32 in OOB estimation. When the scrambled RF model was applied to the unchanged test set, the results were again significantly worse – R^2 of 0.02 (against 0.90) and RMSE of 6.05 (against 1.72).

The RF-AsNN consensus model (CH_2Cl_2 solvent data set) was further validated by comparing the predictions both in the internal validation and for test set with the average and median values of the parameter *N* for each class of compounds. The statistical parameters were calculated using the average value of *N* for the compound in the same class as the query nucleophile. The predictions for the test set were performed using the average values of the training set classes.

For the training set, a RMSE of 2.45 and R^2 of 0.81 were obtained, which compare to 2.02 and 0.88 with the consensus model. For the test set, a RMSE of 2.43 and R^2 of 0.80 were obtained, which compare to 1.54 and 0.92 with the consensus model.

The results using the median instead of the average were even worse – RMSE of 2.57 and R^2 of 0.80 for the training set, and RMSE of 2.63 and R^2 of 0.77 for the test set.

The applicability domain of the obtained models was simply defined by the physicochemical space of the descriptors, delimited by the maximum and minimum of each descriptor, and the property space (*Mayr* nucleophilicity parameter *N*) for the objects of the training set. As the initial data sets were randomly partitioned in training and test sets, some of the test set nucleophiles fell outside this applicability domain – four nucleophiles of the CH_2Cl_2 solvent data set and two of the MeCN solvent dataset.

One of the nucleophiles of the latter data set is the nucleophile ID51, an aromatic amine that shows one of the highest errors in the test set. In this specific case, there are only three aromatic amines in the training set and the *Mayr* nucleophilicity parameter *N* of nucleophile ID51 is outside the domain of the training set for this specific class of compounds. However, the other five nucleophiles were predicted with no worse accuracy than the average. One of the reasons may be that the nucleophiles outside the descriptors' space have descriptors' values only slightly higher or lower than the maximum and minimum, respectively, of the training set for the corresponding descriptor, and only for one or two descriptors.

The results and performance of the models of the present study are at the same level as the ones presented in our previous publication about modelling of *Mayr* electrophilicity [19]. The best model for prediction of *Mayr* electrophilicity using CDK descriptors was obtained with a multiple linear regression model which achieved a RMSE of 2.41 and a R^2 of 0.82 for an independent test set of 21 compounds. These results were improved to a RMSE of 1.45 and a R^2 of 0.94 using a combination of CDK and DFT-based descriptors and a consensus of models. The present study for *Mayr* nucleophilicity show better results for the two data sets (*i.e.* the CH_2Cl_2 and the MeCN data sets) using only CDK descriptors and a single AsNN model. For the CH_2Cl_2 test set, with 67 compounds, a RMSE of 1.70 and a R^2 of 0.90 were achieved and for the MeCN test set, with 27 compounds, a RMSE of 2.19 and a R^2 of 0.82. Using a consensus of the RF and AsNN models these results were improved to a RMSE of 1.54 and 1.97 and a R^2 of 0.92 and 0.86, respectively, for the CH_2Cl_2 and MeCN test sets. It is also to emphasize that, in the present study, the structural diversity of the data sets, class of compounds, is much more diverse than in the previous study about *Mayr* electrophilicity.

Analysis of Molecular Descriptors Identified as Relevant for the Prediction of Mayr *Parameter* N. *Table 4* lists the descriptors selected by the CFS algorithm for both data sets, and the top most important descriptors for the RF models. The twelve selected descriptors by the CFS filter with the CH_2Cl_2 data set include three electronic descriptors (two charged partial surface area descriptors – CPSA [33] – and a descriptor that calculates the number of H-bond acceptors), four topological descriptors, and five constitutional descriptors related to fragments. In the RF model, the ten descriptors with the highest importance by the %IncMSE measure also include two electronic descriptors, six topological descriptors and one constitutional descriptor. A large number of descriptors is common to both %IncMSE and IncNodePurity measures (seven of ten descriptors) and the two most important descriptors are the same.

For the MeCN models, eight descriptors were selected by the CFS filter including two topological descriptors, four molecular descriptors, and two constitutional descriptors. For the CH_2Cl_2 models, there are four descriptors present in both selection

Table 4. *Descriptors Selected to Build QSRR Models for the Prediction of* Mayr *Nucleophilicity Parameter* N.

Data Set	Selection Procedure	CDK Descriptors
N_{Mayr} CH_2Cl_2 Solvent	CFS[a])	FPSA-1; RNCG; ATSc2; SCH-6; VCH-6; nHBAcc; khs.dCH2; khs.dsCH; khs.dsN; khs.aaN; khs.sssN; WTPT-5
	% IncMSE[b])	khs.sssN; nHBAcc; BCUTp-1l; WTPT-5; ATSc3; MDEO-22; ATSc2; ATSc1; WTPT-4; DPSA-1
	IncNodePurity[b])	khs.sssN; nHBAcc; WTPT-5; BCUTp-1l; ATSc2; ATSc1; ATSc3; nBase; ATSc4; MDEO-22
N_{Mayr} MeCN Solvent	CFS[a])	BCUTw-1h; BCUTc-1h; WD.unity; ATSc4; HybRatio; khs.dsCH; khs.aaNH; PetitjeanNumber
	% IncMSE[b])	MDEC-33; TopoPSA; HybRatio; WD.unity; BCUTc-1l; PPSA-1; RNCG; Weta2.unity; LOBMIN; FNSA-1
	IncNodePurity[b])	MDEC-33; WD.unity; HybRatio; khs.ssCH2; FNSA-1; FPSA-1; Wnu1.unity; nAtomP; DPSA-1; PPSA-1

[a]) Selection of descriptors with the CFS filter from Weka. [b]) The Mean Decrease in Accuracy (% IncMSE) and Mean Decrease in Gini (IncNodePurity) are two measures of importance for the descriptors using the RF algorithm.

approaches: khs.sssN (the most important descriptor in the RF model), NHBAcc, ATSc2, and WTPT-5.

The khs.sssN descriptor belongs to the KierHallSmartsDescriptors, which is basically a fragment count descriptor that uses e-state fragments – it codifies the presence of a tertiary nitrogen group in which it has three single bonds. This simple descriptor can provide an indication of the presence of electron-donating groups, such as tertiary amine groups, as well as the presence guanidines, amidines, and isothioureas. For all compounds in the training set, the *Pearson* correlation coefficient between the khs.sssN and the *Mayr*'s nucleophilicity is very high ($R = +0.6342$), but within the class of N-nucleophile compounds, a decrease in the absolute value of the correlation ($R = -0.1345$) is observed.

nHBAcc counts the number of hydrogen bond acceptors using a slightly simplified version of the PHACIR atom types. The ATSc1, ATSc2, ATSc3, and ATSc4 descriptors are 2D Autocorrelation of Topological Structure (ATS) descriptors. The *Moreau–Broto* autocorrelation, ATSc2, defined for the path of length two and weighted by partial charges, is an indicator of space-charge association. Decreasing values of ATSc2 (*i.e.*, increasing absolute values) correspond to increasing nucleophilicity. Within N-nucleophiles and H-nucleophiles, an increase in the correlations between the absolute value of ATSc2 and N_{Mayr} ($R = 0.6553$ and $R = 0.6553$, resp.) was observed, as compared with the correlation of this descriptor for all the training set compounds ($R = 0.4963$).

WTPT-4 and WTPT-5 are based on identifying all paths between pairs of atoms [34], and characterize molecular branching. WTPT-5 is the sum of path lengths starting from N-atoms. The WTPT-5 and khs.sssN correlate similarly with *N*, although the correlation between WTPT-5 and *N* is slightly higher ($R = 0.6392$). But, again, within

N-nucleophiles there is an inverse correlation ($R = -0.4691$). Inspection of the training set reveals that the WTPT-5 descriptor can discriminate well the various classes of compounds, as well as their *Mayr*'s nucleophilicity. All the five compounds with WTPT-5 >8 are guanidine derivatives with a low average N value (13.85). On the other hand, there are fourteen compounds with $4 \leq$ WTPT-5 < 8, and all are isothiourea, amidine, pyridine or quinolidine derivatives with an intermediate average N value (14.81). Finally all the four compounds with WTPT-5 < 4 are aliphatic amines or pyridine derivatives with a high average N value (17.02).

The ten most important descriptors in the RF model of the MeCN data set include two descriptors that are present in both approaches (*i.e.*, RF selection by importance and CFS filter) – WD.unity (a global WHIM descriptor) and HybRatio. The WD.unity descriptor is a global WHIM density descriptor unweighted [35], defined as the total density of the atoms within a molecule. The WHIM descriptors are built in such a way as to capture relevant molecular 3D information with respect to molecular size, shape, symmetry, and atom distribution [35]. The HybRatio descriptor calculates the fraction of sp^3 C-atoms to sp^2 C-atoms using the value of $\mathrm{Nsp}^3/(\mathrm{Nsp}^3 + \mathrm{Nsp}^2)$. This value relates to molecular complexity, especially for natural compounds, which usually have a high value of the sp^3 to sp^2 ratio. For all compounds in the training set, the *Pearson* correlation coefficient between the WD.unity, HybRatio descriptors, and the *Mayr*'s nucleophilicity are among the highest ($R = 0.6629$ and $R = 0.6674$, resp.).

The BCUT descriptors appear as highly relevant in both data sets. BCUT descriptors have been useful in molecular diversity-related tasks [36]. The BCUT descriptors calculated by CDK incorporate both connectivity information and atomic properties of the molecule (atomic weight, atomic charge, polarizability). The importance of BCUT descriptors in modeling chemical reactivity parameters is in accordance with our recent work [19], in which the BCUTs were selected without density functional theory (DFT) descriptors, to build QSRR models for the prediction of the *Mayr* electrophilicity.

Steric effects can obviously affect the nucleophilicity parameter. The models were developed with descriptors that encode 3D features of a compound and by this way, in principle, the model is enabled to learn steric effects on nucleophilicity. However, the data available do not allow to further elaborate. Series of compounds with related structures and varying steric hindrance of the nucleophilic site would be required in the training set.

Calculation of Rate Constants. The application of the developed nucleophilicity QSRR models to predict rate constants of reactions was explored with the aliphatic amines of the MeCN test set. This class of compounds presented a RMSE of 1.71, which is close to the RMSE of 1.97 for the total test set (*Table 3*). For these aliphatic amines, experimental rate constants and *Mayr* electrophilicity for several reference electrophiles were collected from the literature [37–39]. The final set consists in 37 rate constants for 19 reference electrophiles. The rate constants were calculated using the database *Mayr* electrophilicity of the corresponding electrophiles, the database s_N parameters, and the ASNN-RF-predicted *Mayr* nucleophilicity, using *Eqn. 1.*

Table 5 shows the experimental rate constants and those obtained using both the ASNN-RF-predicted *Mayr* nucleophilicity and the experimental *Mayr* nucleophilicity. The 37 experimental rate constants varied between 0.125 and 5.44×10^9. The absolute deviation between the experimental and predicted log k is < 1.7 for 95% of the reactions, and < 1 for 54% of the reactions. Three out of the four largest deviations were for the reactions with k values at the endpoints of the interval.

Fig. 5 shows the experimental $\log(k_2)$ against the predicted $\log(k_2)$ from N and from predicted N.

From the results on *Table* 5 and *Fig. 5*, it is to point out *the steep* rise of the plot toward $\log(k_{\text{exp}}) \sim 10$. Probably for $\log(k_{\text{exp}}) \sim 10$, the reaction rate of the pair nucleophile/electrophile is controlled by the diffusion of the nucleophile and electrophile (diffusion limited reactions), and not by the rate of formation of products from the nucleophile and electrophile, independently of this rate.

However, the calculation of the rate constants, with *Eqn. 1*, using the *Mayr* electrophilicity and nucleophilicity do not take into account the possibility of the diffusion control. For the mentioned reactions, with high values of nucleophilicity and electrophilicity, the predicted rate constants are $> 10^{10}$ $\text{M}^{-1}\text{s}^{-1}$, but the experimental rate constants are close to 10^{10} $\text{M}^{-1}\text{s}^{-1}$.

This example shows the possibility to calculate rate constants using a nucleophilicity QSRR model and experimental electrophilicity values. With the growing number of nucleophiles and electrophiles in the *Mayr* Database of Reactivity Parameters, work is planned to develop new models for prediction of both parameters, using the extended data sets of compounds. The obtained models would then be used for the calculation of

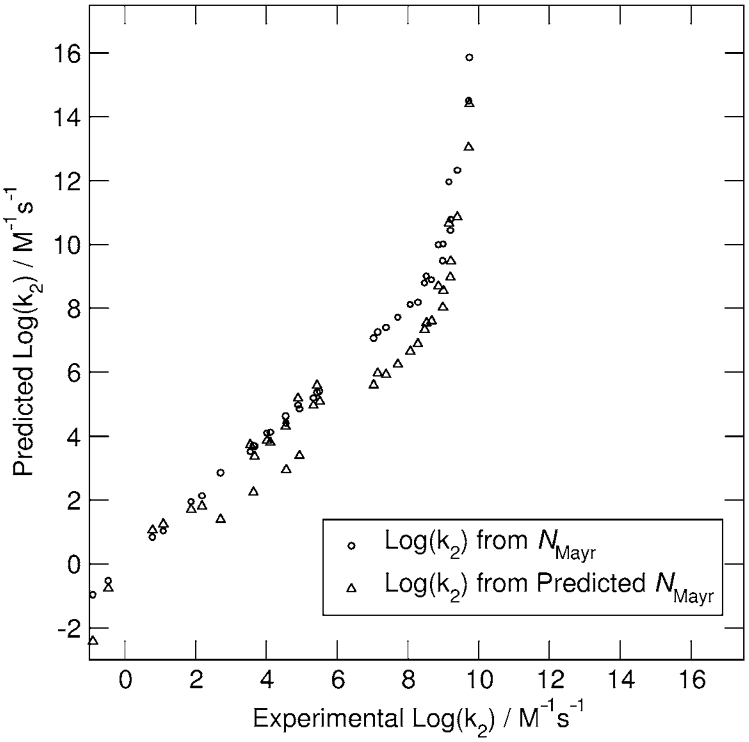


Fig. 5. *Predicted log*(k_2) *from parameter* N *and predicted parameter* N vs. *experimental values of log*(k_2)

Table 5. *Experimental and Calculated Second Order Rate Constants of Aliphatic Amines with Reference Eletrophiles in MeCN at 20°.*

Nuc. ID[a])	N[b])	Pred. N[c])	s_N[d])	Elect. ID[e])	E[f])	k_2/M^{-1}s^{-1} [g])		
						Exp.[h])	From N[i])	From Pred. N[j])
68	14.37	12.16	0.66	tol(^{t}Bu)$_2$QM	– 15.83	1.25×10^{-1}	1.09×10^{-1}	3.78×10^{-3}
				(lil)$_2$CH$^+$	– 10.04	5.09×10^{2}	7.21×10^{2}	2.51×10^{1}
				(ind)$_2$CH$^+$	– 8.76	4.32×10^{3}	5.04×10^{3}	1.75×10^{2}
				(pyr)$_2$CH$^+$	– 7.69	3.64×10^{4}	2.56×10^{4}	8.92×10^{2}
				(dma)$_2$CH$^+$	– 7.02	8.61×10^{4}	7.10×10^{4}	2.47×10^{3}
77	15.1	14.66	0.73	(ani)(Ph)$_2$QM	– 12.18	1.51×10^{2}	1.35×10^{2}	6.46×10^{1}
				(lil)$_2$CH$^+$	– 10.04	4.62×10^{3}	4.94×10^{3}	2.36×10^{3}
				(jul)$_2$CH$^+$	– 9.45	1.29×10^{4}	1.33×10^{4}	6.35×10^{3}
				(ind)$_2$CH$^+$	– 8.76	3.49×10^{4}	4.25×10^{4}	2.03×10^{4}
				(pyr)$_2$CH$^+$	– 7.69	3.24×10^{5}	2.57×10^{5}	1.22×10^{5}
83	15.27	14.89	0.63	(ani)(^{t}Bu)$_2$QM	– 16.11	3.41×10^{-1}	2.96×10^{-1}	1.70×10^{-1}
				(ani)(Ph)$_2$QM	– 12.18	7.49×10^{1}	8.85×10^{1}	5.10×10^{1}
				(ind)$_2$CH$^+$	– 8.76	1.03×10^{4}	1.26×10^{4}	7.28×10^{3}
				(dma)$_2$CH$^+$	– 7.02	2.12×10^{5}	1.58×10^{5}	9.08×10^{4}
88	17.35	17.67	0.68	(ani)(^{t}Bu)$_2$QM	– 16.11	6.04	6.97	1.15×10^{1}
				tol(^{t}Bu)$_2$QM	– 15.83	1.23×10^{1}	1.08×10^{1}	1.78×10^{1}
				(ani)(Ph)$_2$QM	– 12.18	3.52×10^{3}	3.28×10^{3}	5.41×10^{3}
				(lil)$_2$CH$^+$	– 10.04	7.85×10^{4}	9.35×10^{4}	1.54×10^{5}
				(jul)$_2$CH$^+$	– 9.45	2.69×10^{5}	2.36×10^{5}	3.89×10^{5}
90	20.54	18.10	0.60	(ind)$_2$CH$^+$	– 8.76	1.08×10^{7}	1.17×10^{7}	4.02×10^{5}
				(pyr)$_2$CH$^+$	– 7.69	5.22×10^{7}	5.13×10^{7}	1.76×10^{6}
				(dma)$_2$CH$^+$	– 7.02	1.18×10^{8}	1.29×10^{8}	4.45×10^{6}
				(mpa)$_2$CH$^+$	– 5.89	2.97×10^{8}	6.17×10^{8}	2.12×10^{7}
				(mor)$_2$CH$^+$	– 5.53	3.34×10^{8}	1.01×10^{9}	3.48×10^{7}
				(dpa)$_2$CH$^+$	– 4.72	9.70×10^{8}	3.10×10^{9}	1.07×10^{8}
				(mfa)$_2$CH$^+$	– 3.85	9.97×10^{8}	1.03×10^{10}	3.55×10^{8}
				(pfa)$_2$CH$^+$	– 3.14	1.59×10^{9}	2.75×10^{10}	9.46×10^{8}
				(Ph)$_2$CH$^+$	5.9	5.44×10^{9}	7.31×10^{15}	2.51×10^{14}
92	17.1	14.61	0.52	(Ph)$_2$CH$^+$	5.9	1.45×10^{9}	9.12×10^{11}	4.63×10^{10}
				(tol)$_2$CH$^+$	3.63	1.64×10^{9}	6.02×10^{10}	3.05×10^{9}
				(Ph)(ani)CH$^+$	2.11	7.31×10^{8}	9.75×10^{9}	4.95×10^{8}
				(ani)$_2$CH$^+$	0	4.66×10^{8}	7.80×10^{8}	3.96×10^{7}
				(fur)$_2$CH$^+$	– 1.36	1.91×10^{8}	1.53×10^{8}	7.76×10^{6}
				(pfa)$_2$CH$^+$	– 3.14	1.40×10^{7}	1.82×10^{7}	9.21×10^{5}

[a]) Nuc. ID – Nucleophile ID; [b]) N – *Mayr* nucleophilicity; [c]) Pred. N – *Mayr* nucleophilicity predicted by ASNN-RF consensus model; [d]) s_N – nucleophile-specific sensitivity parameter; [e]) Elect. ID – Reference electrophile ID; [f]) E – *Mayr* electrophilicity; [g]) k_2 – Second order rate constant; [h]) Exp. – Experimental second order rate constant; [i]) From N – Second order rate constant calculated using *Mayr* nucleophilicity; [j]) From Pred. N – second order rate constant calculated using *Mayr* nucleophilicity predicted by ASNN-RF consensus model.

Rate constants for compounds ID68, ID77, ID83 and ID88 were extracted from [37], for compound ID90 from [38], and for ID92 from [39]. The following abbreviations are used (by alphabetical order): ani: *p*-anisyl (= 4-methoxyphenyl); dpa: 4-(diphenylamino)phenyl; fur: 2,3-dihydrobenzofuran-5-yl; ind: *N*-methyl-2,3-dihydro-1*H*-indol-5-yl; jul: julolidin-9-yl (= 2,3,6,7-tetrahydro-1*H*,5*H*-pyrido[3,2,1-*ij*]quinolin-9-yl); lil: lilolidin-8-yl (= 1,2,5,6-tetrahydro-4*H*-pyrrolo[3,2,1-*ij*]quinolin-8-yl); mfa: 4-(methyl(trifluoroethyl)amino)phenyl; mor: 4-(*N*-morpholino)phenyl; mpa: 4-(methylphenylamino)phenyl; pfa: 4-(phenyl(trifluoroethyl)amino)phenyl; ph: phenyl; pyr: 4-(*N*-pyrrolidino)phenyl; ^{t}Bu: *tert*-Butyl; thq: *N*-methyl-1,2,3,4-tetrahydroquinolin-6-yl; tol: *p*-tolyl (= 4-methylphenyl).

new electrophiles and nucleophiles with desired values to be used in a wide range of applications related with chemical reactivity. The QSRR models could ultimately also be used to derive other properties – *e.g.*, to calculate rate constants of reactions purely based on the predicted values of the electrophilicity and nucleophilicity.

Unsuccessful Experiments. We also investigated QSRR models for the sensitivity parameter s_N, using the same methods. We could not, however, obtain models performing significantly better than simple estimations by average values within classes of compounds.

In the beginning of the work, experiments were also performed to model *Mayr* nucleophilicity using DFT-based descriptors alone or in combination with CDK descriptors. However, the preliminary experiments did not show significant improvement on the accuracy of the models contrary to what was observed in our previous work on *Mayr* electrophilicity.

Conclusions. – QSRR models could be established for the *Mayr N* nucleophilicity parameter, separately for CH_2Cl_2 and MeCN solvents, with CDK empirical descriptors and were most successful using random forests and associative neural networks – RMSE of 1.54 and 1.97 were obtained for independent test sets of the two solvents data sets. A consensus model of the RF and AsNN models was more accurate than individual models.

Several classes of compounds with the largest errors had a reduced number of examples in the training set. Compounds with silicon atoms were also more difficult to predict.

The models were consistently more accurate than estimations simply based on the average (or median) of the *N* parameter within the class of the query compound – and avoids the previous classification of compounds by classes. The observed difference in accuracy was *ca.* 0.9 in the RMSE for the test set.

SMILES strings, Experimental N_{Mayr}, final subset of descriptors, and predictions for all molecular structures can be obtained from the author as supporting information.

We acknowledge *Fundação para a Ciência e a Tecnologia* (Lisboa, Portugal) for financial support, including individual research grants for *DARSL* (SFRH/BPD/63192/2009). This work has been supported by *Fundação para a Ciência e a Tecnologia* (Ministério da Educação e Ciência, Portugal) through grant n° PEst-C/EQB/LA0006/2013. We thank *ChemAxon Ltd*. for access to JChem and Marvin, *Molecular Networks GmbH* for access to CORINA, and *Igor Tetko* for providing the ASNN program. Prof. *João Aires-de-Sousa* is acknowledged by his comments and suggestions on the dataset, models, and results. Prof. *Filomena Martins* is acknowledged for her helpful discussion on calculation of rate constants and diffusion controlled reactions.

REFERENCES

[1] J. A. H. Schwöbel, Y. K. Koleva, S. J. Enoch, F. Bajot, M. Hewitt, J. C. Madden, D. W. Roberts, T. W. Schultz, M. T. D. Cronin, *Chem. Rev.* **2011**, *111*, 2562.
[2] http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed November 2014).
[3] P. K. Chattaraj, S. Giri, S. Duley, *Chem. Rev.* **2011**, *111*, PR43.
[4] A. K. Bhattacharjee, E. Marek, H. T. Le, R. K. Gordon, *Eur. J. Med. Chem.* **2012**, *49*, 229.
[5] A. K. Bhattacharjee, K. Kuča, K. Musilek, R. K. Gordon, *Chem. Res. Toxicol.* **2010**, *23*, 26.

[6] M*ayr*'s Database web site: http://www.cup.lmu.de/oc/mayr/reaktionsdatenbank/ (accessed November 2014).

[7] H. Mayr, M. Patz, *Angew. Chem., Int. Ed.* **1994**, *33*, 938.

[8] H. Mayr, T. Bug, M. F. Gotta, N. Hering, B. Irrgang, B. Janker, B. Kempf, R. Loos, A. R. Ofial, G. Remennikov, H. Schimmel, *J. Am. Chem. Soc.* **2001**, *123*, 9500.

[9] H. Mayr, A. R. Ofial, *Pure Appl. Chem.* **2005**, *77*, 1807.

[10] A. T. Maynard, M. Huang, W. G. Rice, D. G. Covell, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11578.

[11] R. G. Parr, L. V. Szentpály, S. Liu, *J. Am. Chem. Soc.* **1999**, *121*, 1922.

[12] P. K. Chattaraj, B. Maiti, *J. Phys. Chem. A* **2001**, *105*, 169.

[13] P. Pérez, *J. Org. Chem.* **2003**, *68*, 5886.

[14] P. Pérez, A. Toro-Labbé, A. Aizman, R. Contreras, *J. Org. Chem.* **2002**, *67*, 4747.

[15] S. Deuri, P. Phukan, *Comp. Theor. Chem.* **2012**, *980*, 49.

[16] E. Chamorro, M. Duque-Noreña, P. Pérez, *J. Mol. Struct. (THEOCHEM)* **2009**, *896*, 73.

[17] E. Chamorro, M. Duque-Noreña, P. Pérez, *J. Mol. Struct. (THEOCHEM)* **2009**, *901*, 145.

[18] E. Chamorro, M. Duque-Noreña, R. Notario, P. Pérez, *J. Phys. Chem. A.* **2013**, *117*, 2636.

[19] F. Pereira, D. A. R. S. Latino, J. Aires-de-Sousa, *J. Org. Chem.* **2011**, *76*, 9312.

[20] CDK Software: http://sourceforge.net/projects/cdk/ (accessed November 2014).

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *SIGKDD Explorations* **2009**, *11*, 10.

[22] Weka software: http://www.cs.waikato.ac.nz/ml/weka/ (accessed November 2014).

[23] M. A. Hall, A. Smith, *Proceedings of the Twelfth International FLAIRS Conference*, Orlando, Florida. AAAI Press, Menlo Park, California, 1999, 235.

[24] C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.

[25] W. Wang, Z. Xu, W. Lu, X. Zhang, *Neurocomputing* **2003**, *55*, 643.

[26] C.-C. Chang, C.-J. Lin, *ACM Trans. Intel. Systems Techn.* **2011**, *2*, 1; LIBSVM software: http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed November 2014); Y. El-Manzalawy, V. Honavar, WLSVM: Integrating LibSVM into Weka Environment, 2005. Software available at http://www.cs.iastate.edu/yasser/wlsvm.

[27] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.

[28] R Development Core Team, 'R: A Language and Environment for Statistical Computing'; R Foundation for Statistical Computing, Vienna, 2004, ISBN 3-900051-07-0; http://www.R-project.org/ (accessed November 2014).

[29] Package 'randomForest', Fortran original by L. Breiman, A. Cutler, R port by A. Liaw and M. Wiener, 2004.

[30] I. V. Tetko, *Neural Process. Lett.* **2002**, *16*, 187.

[31] I. V. Tetko, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717.

[32] VCCLAB, Virtual Computational Chemistry Laboratory; http://www.vcclab.org, 2005 (accessed November 2014).

[33] D. T. Stanton, P. C. Jurs, *Anal. Chem.* **1990**, *62*, 2323.

[34] M. Randić, *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.

[35] R. Todeschini, V. Consonni, 'Molecular Descriptors for Chemoinformatics', Wiley-VCH, Weinheim, Vol. 1 and 2, 2009.

[36] R. S. Pearlman, K. M. Smith, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28.

[37] T. Kanzian, T. A. Nigst, A. Maier, S. Pichl, H. Mayr, *Eur. J. Org. Chem.* **2009**, 36, 6379.

[38] M. Baidya, S. Kobayashi, F. Brotzel, U. Schmidhammer, E. Riedle, H. Mayr, *Angew. Chem., Int. Ed.* **2007**, *46*, 6176.

[39] J. Ammer, M. Baidya, S. Kobayashi, H. Mayr, *J. Phys. Org. Chem.* **2010**, *23*, 1029.

Received November 20, 2014