JOC The Journal of Organic Chemistry

Article

pubs.acs.org/joc

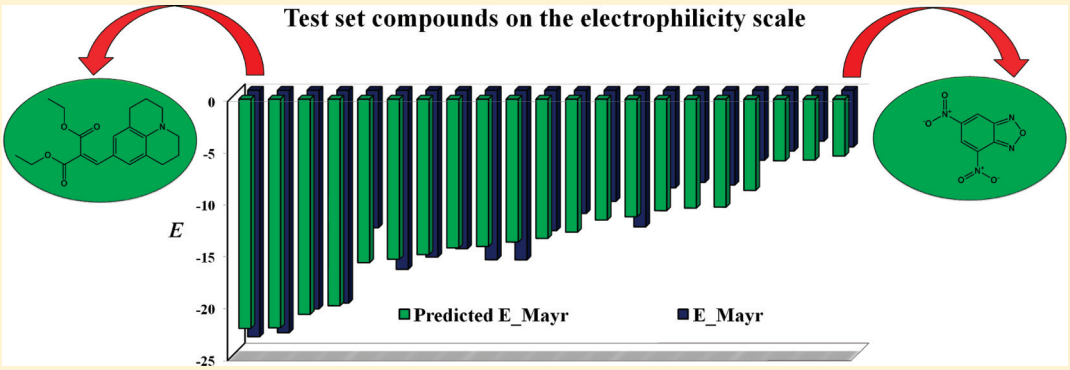
Estimation of Mayr Electrophilicity with a Quantitative Structure–Property Relationship Approach Using Empirical and DFT Descriptors

Florbela Pereira,[†] Diogo A. R. S. Latino,[†,‡] and Joao Aires-de-Sousa*[,†]

[†]CQFB and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

[‡]CCMM, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

Supporting Information



ABSTRACT: Quantitative structure–property relationships (QSPRs) were investigated for the estimation of the Mayr electrophilicity parameter using a data set of 64 compounds, all currently available uncharged electrophiles in Mayr's Database of Reactivity Parameters. Three collections of empirical descriptors were employed, from Dragon, Adriana.Code, and CDK. Models were built with multilinear regressions, *k* nearest neighbors, model trees, random forests, support vector machines (SVMs), associative neural networks, and counterpropagation neural networks. Quantum chemical descriptors were calculated with density functional theory (DFT) methods and incorporated in QSPR models. The best results were achieved with SVM using seven empirical and DFT descriptors; an R^2 of 0.92 was obtained for the test set (21 compounds). The final seven descriptors were the Parr electrophilicity index, ε_{LUMO}, hardness, and four CDK descriptors (FNSA-3, ATSc5, Kier2, and nAtomLAC). Screening of correlations between individual descriptors and Mayr electrophilicity revealed the highest absolute value of correlation for DFT ε_{LUMO} ($R = -0.82$) and comparable correlations for some empirical descriptors, e.g., Dragon's folding degree index ($R = -0.80$), Kier flexibility index ($R = -0.78$), and Kier S2K index ($R = -0.78$). High correlations were observed in the training set between reactivity descriptors calculated by the PM6 semiempirical and DFT methods ($R = 0.96$ for ε_{LUMO} and 0.94 for the electrophilicity index).

INTRODUCTION

The concepts of electrophilicity and nucleophilicity have an enormous impact on the understanding and prediction of a wide range of organic reactions. Nucleophilic and electrophilic substitution reactions, Michael additions, Schiff base formation, acylations, and Diels–Alder reactions are just a handful of examples. The interest and motivation in predicting these types of reactions go far beyond the specific boundaries of organic chemistry. In fact, the assessment of (eco)toxicological end points[1] 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. The whole manner in which chemicals are and will be regulated is changing dramatically, and that is culminating in many more substances needing to be evaluated and the realization that traditional approaches are likely to be too expensive. The EU chemicals legislation REACH[2] calls for the use of alternatives to minimize animal testing, and the 7th Amendment to the Cosmetics Directive[3] calls for a complete ban on animal testing for a number of end points. Alternative approaches[4] are thus needed, and in silico methods appear to be promising components of new testing strategies to be adopted by regulatory agencies.

Received: July 27, 2011
Published: October 4, 2011

Computer-assisted predictions of reactivity, particularly those based on fast algorithms, are also desirable for the development of new products by chemical and pharmaceutical industries to filter out compounds with a high risk of toxicity as early in the process as possible.

Even for the above-mentioned toxicity end points, reactivity is clearly not the only factor to be considered, but it is a valuable element to include in integrated approaches. Different levels of confidence are required depending on the end use of the information. For example, highly accurate and well-defined models would be required to replace, in some cases, experimental testing, while the prioritization of lists of compounds would tolerate lower levels of confidence.

Quantitative scales of electrophilicity and nucleophilicity are useful resources for both the rationalization of chemical reactivity and the prediction of new reactions. They are high-quality data for facilitating the benchmarking of computational approaches. Legon and Millen derived electrophilicity and nucleophilicity parameters from hydrogen-bond stretching force constants measured by rotational spectroscopy.[5] Other approaches based on experimental data have also employed, for example, ^{13}C NMR chemical shifts and ν_{CO} frequencies,[6] redox potentials,[7] or an HPLC assay of covalent bonding interactions.[8] Probably, the most established scale has been proposed by Mayr and co-workers to explain diverse types of reactions.[9,10] It was demonstrated for series of electrophile−nucleophile combinations that the kinetic rate constants of reactions can be fit to the following linear relationship

$$\log k(20\,^{\circ}\mathrm{C}) = s(N + E) \quad (1)$$

where E and N are the electrophilicity and nucleophilicity parameters, respectively, and s is a system specific parameter, which is dependent on the reference nucleophile. In the derivation of the original Mayr E and N scales, reference carbon electrophiles, e.g., alkenes, arenes, alkynes, enol ethers, enamines, diazo compounds, carbanions, hydride donors, phosphanes, amines, and alkoxides, were employed to compare the nucleophilicities of a large variety of compounds using eq 1.[11] This was also used to derive the electrophilicity parameter E for different types of electrophiles, such as carbocations, typical Michael acceptors, and electron-deficient arenes.[11] The obtained E, N, and s parameters can be used for semi-quantitative prediction of rates and selectivities of polar organic reactions. They are compiled in Mayr's Database of Reactivity Parameters,[12] currently comprising information for 621 nucleophiles and 182 electrophiles spanning a nucleophilicity (N) range from −4.47 to 28.95 and an electrophilicity (E) range from −23.80 to 6.16 for a wide variety of molecule classes.

From the theoretical side, many efforts to define and quantify the electrophilicity of molecules using quantum calculations have been reported. Electronegativity and hardness were rigorously defined using conceptual density functional theory (DFT) to arrive at an electrophilicity index.[13−15] Using Koopman's theorem,[16] these parameters can be calculated from the HOMO and LUMO energies. In the past several 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.[15] High correlations ($R > 0.94$) between the Mayr electrophilicity parameter and the electrophilicity index within series of compounds such as benzene diazonium ions[17] and benzhydryl cations[18] were found.

In this work, we explored the application of quantitative structure−property relationship (QSPR) techniques in predicting the Mayr electrophilicity parameter, E, from the molecular structural formula, using all the available uncharged electrophiles in the Mayr database. These were chosen to avoid the additional issue of charge in the calculations and are presumed to better represent large groups of compounds for which relationships between toxicity and reactivity have been discussed.[1] Fast empirical molecular descriptors were tested, as well as DFT reactivity descriptors, with state-of-the-art machine learning algorithms such as associative neural networks, support vector machines, and random forests. The calculation of thousands of commonly used empirical molecular descriptors provided the opportunity to identify those most correlated with Mayr electrophilicity in this reference data set.

METHODS

Data Set. The electrophilicity parameter E and the corresponding molecular structures were extracted from the Mayr's Database of Reactivity Parameters[12] for all 64 available uncharged compounds. The data set comprises the following classes of compounds: 25 acceptor-substituted alkenes, 14 quinone methides, 12 acceptor-substituted arenes, six 1,2-diaza-1,3-dienes, three chlorinating agents, and four azodicarboxylate compounds. The molecular structures of all compounds were drawn using MarvinSketch version 5.2 (ChemAxon Ltd., Budapest, Hungary) and saved as SMILES strings (available as Supporting Information).

Molecular Descriptors. Three-dimensional models of the molecular structures were generated with CORINA version 2.4 (Molecular Networks GmbH, Erlangen, Germany). Molecular descriptors were then calculated by ADRIANA.Code 2.2.2 (Molecular Networks GmbH), CDK Descriptor Calculator 1.1.1,[19,20] and Dragon Professional 5.5 (Talete srl, Milan, Italy). The numbers of descriptors calculated by each program were 1195 (ADRIANA.Code), 289 (CDK), and 1666 (Dragon). The three collections of descriptors overlap to some extent and include descriptors expected to be related to electronegativity such as atomic partial charges and their distribution or the codification of local structural patterns. CDK and Dragon are representative of the descriptors most commonly used in QSAR studies. ADRIANA.Code focuses on distribution functions, autocorrelation, and surface descriptors derived from calculated atomic properties developed in the Gasteiger group.

The calculation of the DFT descriptors was done in a semi-automatic way using the following steps: initial generation of the most stable conformer with the JChem CXCALC tool (ChemAxon Ltd.), optimization of the three-dimensional (3D) geometry with MOPAC2009[21] using the PM6 semiempirical method,[22] calculation of the harmonic vibrational frequencies to determine if the optimized geometries were minima on the potential energy surface (all real frequencies) at the same level of theory, and single-point energy calculations with the GAMESS package[23,24] using the hybrid B3LYP method[25,26] with the 6-31G* basis set.[27,28] The quantum chemical descriptors extracted directly from the GAMESS output were the energy of the highest occupied molecular orbital (ε_{HOMO}) and the energy of the lowest unoccupied molecular orbital (ε_{LUMO}). From these two orbital energies, the following descriptors were calculated: hardness, $\eta = \varepsilon_{LUMO} - \varepsilon_{HOMO}$; chemical potential, $\mu = -(\varepsilon_{HOMO} + \varepsilon_{LUMO})/2$; electrophilicity index, $\omega = \mu^2/(2\eta)$ as defined by Parr et al.[14]

Selection of Training and Test Sets. The whole data set was divided into a training set of 43 compounds and a test set of 21 compounds, which were used for the development and external validation of the QSPR models. The approximate 2:1 partition was assisted by a Kohonen Self-Organizing Map (SOM)[29] in such a way that both sets span the chemical diversity of the data set. The 64 structures were mapped on a SOM on the basis of Dragon

dx.doi.org/10.1021/jo201562f | J. Org. Chem. 2011, 76, 9312−9319

Table 1. Exploration of Three Collections of Empirical Descriptors for the QSPR of E_{Mayr}[a]

descriptor source	no. of descriptors	no. of selected descriptors	selected descriptors	R^2	rmse
Adriana[51]	914	20	Diameter; 2DACorr_SigChg_5; RDF_Ident_16; RDF_SigChg_31; RDF_SigChg_50; RDF_SigChg_100; RDF_SigChg_102; RDF_SigChg_110; RDF_SigChg_111; RDF_SigChg_115; RDF_PiChg_21; RDF_PiChg_27; RDF_PiChg_28; RDF_PiChg_33; RDF_PiChg_70; RDF_PiChg_73; RDF_TotChg_42; RDF_PiEN_56; RDF_Polariz_34; SurfACorr_HBP_5	0.74	2.06
CDK[52]	117	9	BCUTc-1l; BCUTc-1h; FPSA-2; FNSA-2; FNSA-3; ATSc5; MDEN-33; geomShape; nRotB	0.72	2.15
Dragon[53]	1061	12	Lop; X3Av; X5Av; FDI; PJI3; G2u; H7e; RTu+; R4v+; C-006; GVWAI-50; B10[C-N]	0.65	2.39

[a] R^2 and rmse values are for 10-fold cross-validation experiments with the training set.

constitutional descriptors (atom and group counts). No information about the electrophilicity parameter was used. A trend for clustering according to structural classes of compounds was observed. Compounds belonging to the various clusters were selected for the test set from singly occupied neurons.

Selection of Descriptors and Optimization of QSPR Methods. After the removal of constant descriptors, multilinear regressions (MLR)[30] were built with Weka 3.6.3[31,32] to select descriptors by the M5 method, using the training set. With this method, a first regression model is built with all descriptors, and then descriptors with the smallest standardized regression coefficients are removed in a stepwise manner until no improvement is observed in the estimate of the average prediction error given by the Akaike information criterion (AIC).[30] The procedure was separately applied to the three collections of descriptors for a preliminary selection of descriptors: 914 descriptors were selected from ADRIANA.Code, 117 from CDK, and 1061 from Dragon.

In the quest for QSPR models with the minimum possible number of descriptors, feature selection was further performed with the CFS (Correlation-based Feature Subset Selection) algorithm[33] implemented in Weka 3.6.3. This heuristic takes into account the usefulness of individual descriptors for predicting the property (electrophilicity) together with the level of intercorrelation among them. The experiments for comparing the three collections of descriptors, as well as those for comparing different machine learning techniques, were conducted with the AttributeSelectedClassifier routine of Weka with the CfsSubsetEval option for evaluator and BestFirst option for search. This procedure selects descriptors with the CFS algorithm within a 10-fold cross-validation procedure.

Machine Learning Techniques. The k nearest neighbor (KNN) algorithm[34] predicts the property for a compound by the average of the values for the k most similar compounds in the training set. It was applied here with the Weka 3.6.3 software using a k of 10, Euclidean distances, and contributions of neighbors weighted by the inverse of distance.

Model trees (i.e., M5 decision trees) were grown with the Quinlan M5 algorithm[35,36] implemented in Weka 3.6.3 using the default parameters and unpruned trees. Model trees are constructed by first using a decision tree induction algorithm to build the initial tree, and then a multilinear regression model is constructed for each node of the tree. Each linear model is improved by eliminating descriptors that are removed in a stepwise manner until no improvement is observed in the estimate of the average prediction error. A tree is sequentially constructed by partitioning compounds from a parent node into two child nodes. Each node is produced by a logical rule, defined for a single descriptor, where compounds below a certain descriptor's value fall into one of the two child nodes and compounds above fall into the other child node.

Random forests (RF)[37] are ensembles of unpruned regression trees created by using bootstrap samples of the training data. The best split at each node is defined among a randomly selected subset of descriptors. Prediction is conducted with an average of the individual regression trees in the forest. In this study, RF of 100 trees were grown with Weka 3.6.3.

Support vector machines (SVMs)[38] map the data into a hyperspace through a nonlinear mapping (a boundary or hyperplane) and then run a linear regression in this space.[39] The boundary is positioned using examples in the training set that are known as the support vectors. With nonlinear data, kernel functions can be used to transform it into a hyperspace where the linear regression can be conducted. In this study, SVMs were established with Weka 3.6.3, using the LIBSVM software.[40−42] 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. The parameter C of the ε-SVM-regression was optimized in the range of 10−500. Data were normalized.

A counterpropagation neural network (CPGNN)[43] consists of a Kohonen Self-Organizing Map (Kohonen SOM)[29] linked to an output layer of neurons aligned with the Kohonen layer. A Kohonen SOM distributes objects over a two-dimensional (2D) surface (a grid of neurons) in such a way that objects bearing similar descriptors are mapped onto the same or adjacent neurons. The input data are stored in the 2D grid of neurons, each containing as many elements (weights) as there are input variables (molecular descriptors). The output data (E_{Mayr}) are stored in the output layer that acts as a look-up table. CPNNs of toroidal topology with a size of 9 × 9 (number of neurons approximately twice the number of training cases) were trained with default parameters: a linear decreasing triangular scaling function, an initial learning rate of 0.1, and an initial learning span of 4. The training was performed over 150 cycles, with the learning span and the learning rate linearly decreasing until they reached zero. CPNNs were implemented with a Java application, developed in house, derived from the JATOON Java applets.[44,45]

Associative neural networks (ASNNs)[46,47] integrate an ensemble of feed-forward neural networks (FFNNs) with a memory of experimental data. The ensemble consists of independently trained FFNNs, which contribute to a single prediction. The ASNN scheme is employed for composing a prediction from (a) the outputs produced by the ensemble of NNs and (b) the most similar cases in the memory (here, the training set). The ASNN program from Tetko[48] was used and employed the Levenberg−Marquardt algorithm for training fully connected FFNNs with an input layer (including a bias equal to 1), one hidden layer (also including a bias equal to 1), and one output neuron. The number of hidden neurons was optimized and in the final experiments 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. Prior to the training of each network, the program randomly divided the training set into a validation set and a reduced training set that were approximately the same size. Full cross-validation of the training set was performed using the leave-one-out (LOO) method. The training was stopped when there was no further improvement in the root-mean-square error (rmse) for the validation set. After the training, the results were calculated for the reduced training set, for the validation set, for the LOO method, and for an external test set.

RESULTS AND DISCUSSION

Exploration of Empirical Molecular Descriptors for QSPRs. Three wide sets of molecular descriptors, calculated by Adriana, CDK, and Dragon, were calculated for all the molecules in the data set, starting from the 3D model generated with CORINA. These descriptors are fast to calculate and take into account different structural features, including

physicochemical and geometrical properties. The performances of the three sets of descriptors in QSPR experiments in predicting Mayr electrophilicity parameter E were compared. These exploratory QSPR experiments employed selection of descriptors with the CFS filter[31−33] followed by the simple k nearest neighbor (KNN) prediction of E_{Mayr}, within a 10-fold cross-validation procedure (Table 1). The CFS filter (correlation-based feature selection) maximizes the correlation with the variable to predict and minimizes intercorrelation between descriptors. CDK descriptors were chosen for further experiments as they allowed good predictions with the fewest descriptors. The nine selected CDK descriptors include two BCUT descriptors[49] (eigenvalue-based molecular descriptors), three CPSA[50] (charged partial surface area) descriptors, three topological descriptors, and a constitutional descriptor. A rationalization of this selection from the physical meaning of the descriptors can be attempted.

BCUT descriptors have been useful in molecular diversity-related tasks.[49] The BCUT descriptors calculated by CDK incorporate both connectivity information and atomic properties of the molecule (atomic weight, atomic charge, and polarizability). The highest and lowest eigenvalues of Burden matrices have been shown to be discriminating descriptors because they contain contributions from all atoms and thus reflect the topology of the whole molecule. BCUTc-1l and BCUTc-1h are the lowest and highest eigenvalues of the Burden matrix, respectively (weighted by partial charge).

FPSA-2, FNSA-2, and FNSA-3 are charge partial surface area (CPSA) descriptors that combine molecular surface area and partial atomic charge information. The molecular representation used for deriving CPSA descriptors views molecule atoms as hard spheres defined by the van der Waals radius and their solvent-accessible surface area.[50] An electron distribution is considered, consequently accounting for the charged contact surface where polar intermolecular interactions can take place. FPSA-2 is the partial positive surface area multiplied by the total positive charge on the molecule and divided by the total molecular solvent-accessible surface area. FNSA-2 is similar to FPSA-2 but uses negative charges. FNSA-3 is the ratio between the charge-weighted partial negative surface area and the total molecular solvent-accessible surface area. The three descriptors can reflect how charge is distributed in the molecule and are also related to the hydrogen bonding capability. Interestingly, both positively and negatively charged surface area descriptors were chosen.

ATSc5 is a 2D Moreau−Broto autocorrelation descriptor, defined for the path of five bonds and weighted by partial charges, that is an indicator of spatial partial charge association. The MDEN-33 descriptor is defined as the molecular distance edge between all tertiary nitrogen atoms. Inspection of the compounds belonging to the training set reveals that this descriptor provides an indication of the presence of electron-withdrawing groups such as nitro groups, hetereocyclic N-oxides, or N-heterocyclic rings, which activate aromatic rings or double bonds in aromatic nucleophilic substitutions or nucleophilic additions, respectively. The geomShape descriptor is the geometric shape index of Petitjean, which reflects the anisotropy of a molecule, and nRotB is calculated from the analysis of nonrotatable bonds.

From the three sets of descriptors, only CDK descriptors were used in further investigations. Different representative machine learning techniques available in Weka for building a QSPR model with these descriptors were compared (Table 2).

Table 2. Comparison of Different Statistical and Machine Learning Techniques for Building QSPR Models with CDK Descriptors[a]

	MT	MLR	KNN	RF	SVM
			Training Set[b]		
R^2	0.69	0.80	0.72	0.69	0.89
rmse	2.26	1.84	2.15	2.26	1.35
			Test Set		
R^2	0.70	0.82	0.77	0.72	0.74
rmse	3.17	2.41	2.75	3.08	2.98

[a]Legend: MT, model tree; MLR, multiple linear regression; KNN, k nearest neighbor; RF, random forest; SVM, support vector machine. [b]Tenfold cross-validation.

As before, selection of descriptors was performed with the CFS algorithm within a 10-fold cross-validation on the training set. The same nine descriptors were used for all the models that predicted the test set.

Inclusion of DFT-Based Reactivity Descriptors. After the exploration of models derived with empirical descriptors, we investigated the inclusion of DFT reactivity descriptors. These have a well-established relationship with reactivity and are expected to increase both the accuracy and robustness of the predictive models. SVM was used because of its performance in previous experiments. Five quantum chemical descriptors were calculated [ε_{LUMO} (energy of the lowest unoccupied molecular orbital), ε_{HOMO} (energy of highest occupied molecular orbital), hardness, chemical potential, and Parr electrophilicity index], and their usefulness in predicting Mayr electrophilicity was assessed with SVM and 10-fold cross-validation of the training set. With the five DFT descriptors alone, predictions yielded an R^2 of 0.73 and an rmse of 2.11 (the test set was predicted with an R^2 of 0.90 and an rmse of 2.23). When DFT and CDK descriptors were taken together, the performance was improved to an R^2 of 0.90 and an rmse of 1.27 for the training set and an R^2 of 0.92 and an rmse of 1.64 for the test set. From 117 initial CDK descriptors and the five DFT descriptors, the CFS filter selected eight descriptors: ε_{LUMO}, hardness, Parr electrophilicity index, FNSA-3, ATSc5, khs.sBr, Kier2, and nAtomLAC (five CDK and three DFT descriptors) (Figure 1).

Finally, backward removal of descriptors was performed to find the best subset on the basis of the rmse for 10-fold cross-validation of the training set with SVM. This procedure removed the khs.sBr descriptor yielding an rmse of 1.07 and an R^2 of 0.93. The test set was then submitted, yielding an rmse of 1.55 and an R^2 of 0.92. Two of the four selected CDK descriptors were also selected for the models without DFT descriptors, and two are new. This is not unexpected, as the new information introduced by the DFT descriptors required the recalculation of intercorrelations between descriptors, as well as their relative merit.

Some interpretation of the physical meaning of the descriptors is possible. Kier2 is a Kier and Hall κ molecular shape index.[54] These descriptors are intended to capture different aspects of molecular shape. The Kier2 is the second κ shape index, which encodes the spatial density of atoms in the molecule, and nAtomLAC is related to the number of carbon atoms in the longest nonaromatic chain.

The selection of the three quantum chemical descriptors (ε_{LUMO}, hardness, and Parr electrophilicity index) was not surprising, and their relation to electrophilicty is soundly

dx.doi.org/10.1021/jo201562f|J. Org. Chem. 2011, 76, 9312−9319

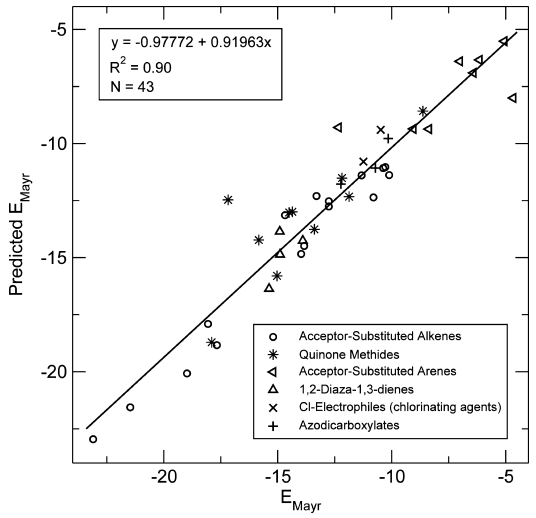


Figure 1. Database values vs SVM predictions for the Mayr electrophilicity parameter E obtained for the training set in a 10-fold cross-validation experiment with selection of descriptors among 117 CDK and five DFT descriptors.

founded. The energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of molecules to attack by nucleophiles.[1,15] Many authors have used this descriptor for modeling reactive toxicity.[1] The hardness descriptor corresponds to the HOMO−LUMO gap. It is related to stability: a large hardness implies high stability for the molecule and thus low reactivity.

The electronic CPSA (charged partial surface area) descriptors appear to be highly relevant in the prediction of electrophilicity, being selected with and without DFT descriptors. The importance of CPSA descriptors in modeling electrophilicity is in accordance with the results of Stanton et al.,[55] who interpreted CPSAs as descriptors of electrophilicity, and alternative to LUMO energy.

Consensus Model. Individual models sometimes exhibit weaknesses in different regions of the input space, as well as fluctuations over different random training conditions. Consensus models can often overcome these differences, producing better predictions than each of the individual models alone. Therefore, two other machine learning algorithms not implemented in the Weka package were also tried with the nine CDK-selected descriptors: counter-propagation neural networks (CPGNNs) and associative neural networks (ASNNs). As they exhibited performance similar to that of SVM, a consensus model was produced with these three machine learning techniques and the final seven CDK and DFT descriptors (Table 3). Even though some classes include just a few compounds, the results suggest that some methods predict better one class of compounds while the opposite occurs for other classes. Overall, we can say the consensus model generally performs better than individual models (Table 3 and Figure 2).

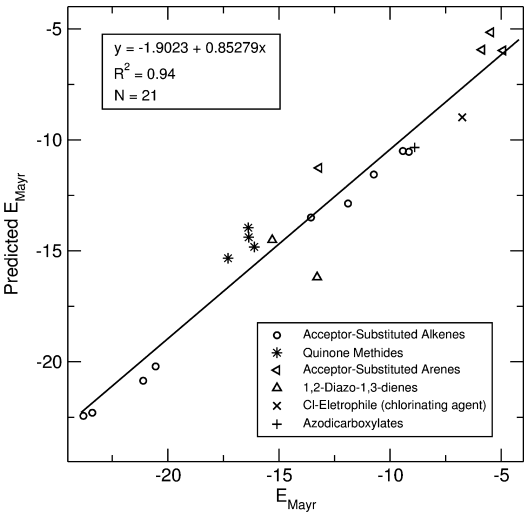


Figure 2. Database values vs consensus predictions for Mayr electrophilicity parameter E (test set).

A comparison of the distribution of predictions with E_{Mayr} values over the electrophilicity scale is shown in Figure 3.

The three individual models, as well as the consensus model, were further validated using the y-randomization technique. The models were rebuilt with a modified training set; the y-column data (E_{Mayr}) were scrambled, keeping the descriptor matrix unchanged. The random models were found to have a considerably lower R^2 and at the same time a considerably higher rmse compared to those of the original models (R^2 range of 0.03−0.06 and rmse range of 6.99−8.24 on the test set), corroborating the statistical reliability of the original models.

Screening of Correlations between Individual Molecular Descriptors and Mayr Electrophilicity. The large number of calculated widely used molecular descriptors presents an opportunity to screen for correlations between individual descriptors and electrophilicity. Although nor directly used in the models presented here, such correlations

Table 3. Prediction of Electrophilicity by Three Machine Learning Techniques[a] for the 21 Compounds of the Test Set[b]

	SVM	CPGNN	ASNN	consensus model
acceptor-substituted alkenes	0.79	1.46	1.01	0.94
quinone methides	2.35	1.8	2.06	1.95
acceptor-substituted arenes	0.89	1.18	1.37	1.11
1,2-diaza-1,3-dienes	2.23	2.18	2.33	2.13
chlorinating agent	3.13	1.2	2.37	2.23
azodicarboxylate	0.071	3.32	0.96	1.45
test set (rmse/R^2)	1.55/0.92	1.68/0.94	1.56/0.93	1.45/0.94

[a]Legend: SVM, support vector machine; CPGNN, counterpropagation neural network; ASNN, associative neural network. [b]For the individual classes, only rmse values are given.

dx.doi.org/10.1021/jo201562f | J. Org. Chem. 2011, 76, 9312–9319

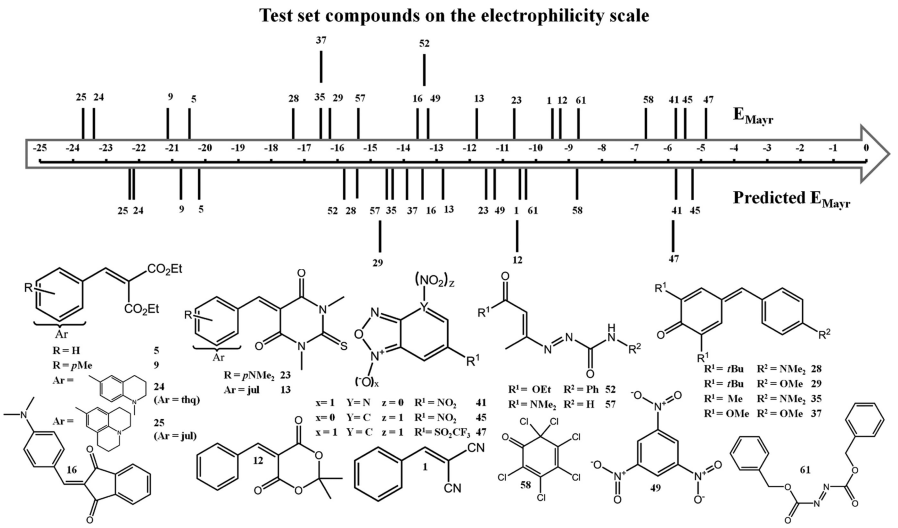


Figure 3. Comparison of database and predicted values of E_{Mayr} over the electrophilicity scale.

could be useful in other QSPR−QSAR studies, in linking specific descriptors to electrophilicity, and suggest a possible physical meaning for their use. With this purpose, all molecular descriptors from the four sets (i.e., from Adriana, CDK, Dragon, and quantum chemical descriptors) were analyzed with respect to their Pearson and Spearman correlation coefficient against Mayr electrophilicity, E, for the training set.

The five descriptors yielding the highest absolute value of the Pearson coefficient were (1) ε_{LUMO} (quantum chemical descriptor; R = −0.82), (2) FDI (geometrical descriptor calculated by Dragon, which encodes the folding degree index; R = −0.80), (3) Parr electrophilicity index (quantum chemical descriptor; R = 0.79), (4) PHI (topological descriptor calculated by Dragon, which encodes the Kier flexibility index; R = −0.78), and (5) S2K (topological descriptor calculated by Dragon, which encodes the second-path Kier α-modified shape index; R = −0.78).

In terms of Spearman's rank correlation coefficient, r_S, the highest correlations were observed for (1) the R7e GETAWAY descriptor (GEometry, Topology and Atom-Weights AssemblY, a 3D autocorrelation descriptor of lag 7 weighted by atomic Sanderson electronegativities, calculated by Dragon; r_S = −0.83), (2) R7u GETAWAY descriptor (calculated by Dragon, a 3D autocorrelation descriptor of lag 7 unweighted; r_S = −0.82), (3) H7e GETAWAY descriptor (calculated by Dragon, a 3D autocorrelation descriptor of lag 7 weighted by atomic Sanderson electronegativities; r_S = −0.78), (4) H7u GETAWAY descriptor (calculated by Dragon, a 3D autocorrelation descriptor of lag 7 unweighted; r_S = −0.78), and (5) Mor17u (3D-MoRSE descriptor calculated by Dragon, which encodes signal 17 unweighted; r_S = −0.77).

Inspection of plots between some of the descriptors and E_{Mayr} reveals that in some cases, e.g., the PHI descriptor, the correlation is partially explained by the simultaneous ability of the descriptor to discriminate between different classes of compounds and to rank the classes in an order that grossly correlates with their median E_{Mayr} values. In other cases, e.g., Mor17u and ε_{LUMO}, high correlations exist within some classes of compounds (e.g., acceptor-substituted alkenes and quinone methides) (Figure 4).

On the other hand, descriptors that globally correlated little with electrophilicity may be useful in predicting electrophilicity

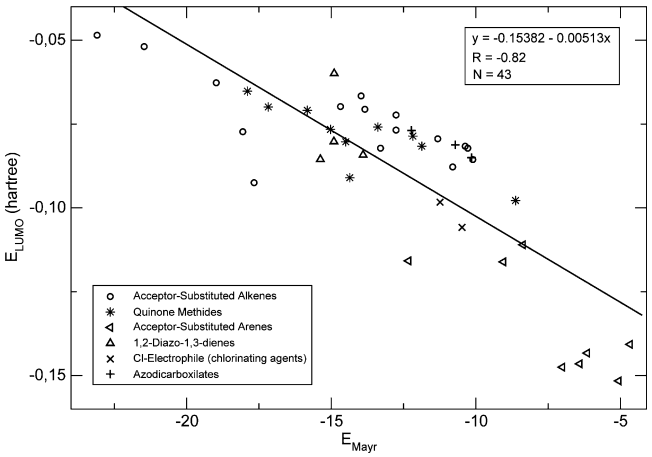


Figure 4. Representation of E_{Mayr} values (database) vs DFT ε_{LUMO}.

within certain classes of compounds. This was observed for the DFT hardness descriptor, which was used in the final seven-descriptor model, although its global correlation with E_{Mayr} is weak.

Interpretation of these individual correlations must be done with caution, as they can be influenced by artifacts of the relatively small data set. However, we cannot avoid noticing the strong presence of geometrical descriptors among those most correlated with electrophilicity. This may be related to steric effects, as well as to the influence of resonance effects that require coplanarity of conjugated bonds.[56] The folding degree index (FDI) encodes the conformational variability of the molecule.[45] The Kier flexibility index (PHI) is a measure of bond rigidity because it represents the increase in flexibility of fragments with respect to the parent molecule.[57] The two-path Kier α-modified shape index (S2K) is related to the molecular shape.[57] GETAWAY[57] and 3D-MoRSE[57] descriptors are also geometrical.

The 3D structures optimized by the semiempirical PM6 method provide data for calculating, at that level, the electrophilicity index and the LUMO energy and for comparing their values with those obtained by DFT. Within the training set, we observed that the PM6 and DFT LUMO energies correlate with an R of 0.96 while the electrophilicity index gave

dx.doi.org/10.1021/jo201562f|J. Org. Chem. 2011, 76, 9312–9319

an R of 0.94. Although the PM6 values are less correlated with E_{Mayr} than the DFT counterparts, the PM6 LUMO–E_{Mayr} correlation coefficient is −0.75 (vs −0.82), while the PM6 electrophilicity index–E_{Mayr} correlation coefficient is 0.72 (vs 0.79).

QSPR Prediction of DFT Descriptors. Having empirical and DFT descriptors in hand, we also explored the possibility of building QSPR models for predicting relevant DFT descriptors from empirical descriptors. Such a model would allow very fast access to DFT descriptors, their application with extremely large data sets, and their incorporation in systems requiring fast predictions. The prediction of the LUMO energy and Parr electrophilicity index was tried using the methodology employed for the prediction of E_{Mayr} using the CDK descriptors. The best models achieved predictions of the LUMO energy and Parr electrophilicity index up to R^2 values of 0.87 and 0.79, respectively (rmse values of 0.013 and 0.019 hartree, respectively) (test set).

The results suggest that it may be possible to calibrate QSPR models to predict DFT descriptors, such as orbital energies and electrophilicity indices, from fast empirical descriptors. High levels of accuracy will be expected if large data sets of molecules are calculated to train the models. We are currently investigating this strategy.

CONCLUSIONS

The combination of empirical and DFT reactivity descriptors allowed us to build QSPR models to estimate Mayr electrophilicty with accuracies up to an R^2 of 0.94 and an rmse of 1.5 for an independent test set belonging to the same chemical classes of the training set. In a manner different from that of previous studies, several classes of compounds were processed together. Considering the relatively small size of the data set, it must be underlined that a test set with one-third of the whole data was a highly challenging procedure.

The predictive power of DFT descriptors, such as Parr's electrophilicity index and the energy of the LUMO, was confirmed. However, a number of fast empirical molecular descriptors were identified with almost the same relevance within this data set, and their combination with DFT descriptors proved to be advantageous. Some empirical descriptors can encode important molecular features (like steric effects or scaffold rigidity) that are not identified by the DFT descriptors. They may also allow machine learning algorithms to explore the specific predictive capacities of other descriptors within certain subsets of compounds.

Methods for estimating the Mayr electrophilicity parameter have application in the design of reactions with electrophiles not yet evaluated or synthesized. They can also assist in the curation and data mining of reactivity databases, pinpointing compounds with suspicious or unexpected experimental values. Better models for predicting electrophilicity would probably require more data for calibration. While the number of experimental E_{Mayr} values available is not expected to dramatically increase soon, DFT electrophilicity indices can be calculated for large databases of compounds; their usefulness for the development of fast empirical procedures is to be investigated.

ASSOCIATED CONTENT

Supporting Information

SMILES strings, E_{Mayr}, final subset of descriptors, and predictions for all molecular structures. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Telephone: (+351) 212948300. Fax: (+351) 212948550. E-mail: joao@airesdesousa.com.

ACKNOWLEDGMENTS

We acknowledge Fundação para a Ciência e a Tecnologia (Lisboa, Portugal) for financial support, including individual research grants for DARSL (SFRH/BPD/63192/2009). We thank ChemAxon Ltd. for access to JChem and Marvin, Molecular Networks GmbH for access to CORINA and Adriana.Code, and Igor Tetko (Helmholtz Zentrum München, Neuherberg, Germany) for providing the ASNN program.

DEDICATION

Dedicated to Professor Sundaresan Prabhakar.

REFERENCES

(1) Schwöbel, J. A. H.; Koleva, Y. K.; Enoch, S. J.; Bajot, F.; Hewitt, M.; Madden, J. C.; Roberts, D. W.; Schultz, T. W.; Cronin, M. T. D. *Chem. Rev.* **2011**, *111*, 2562.
(2) http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed September 2011).
(3) Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC. OJ L066, 26–35, 11 March 2003.
(4) Bassan, A.; Worth, A. P. *QSAR Comb. Sci.* **2008**, *27*, 6.
(5) Legon, A. C.; Millen, D. J. *J. Am. Chem. Soc.* **1987**, *109*, 356.
(6) Neuvonen, H.; Neuvonen, K.; Koch, A.; Kleinpeter, E.; Pasanen, P. *J. Org. Chem.* **2002**, *67*, 6995.
(7) Topol, I. A.; McGrath, C.; Chertova, E.; Dasenbrock, C.; Lacourse, W. R.; Eissenstat, M. A.; Burt, S. K.; Henderson, L. E.; Casas-Finet, J. R. *Protein Sci.* **2001**, *10*, 1434.
(8) Morris, S. J.; Thurston, D. E.; Nevell, T. G. *J. Antibiot.* **1990**, *43*, 1286.
(9) Mayr's Database website. http://www.cup.lmu.de/oc/mayr/reaktionsdatenbank/ (accessed September 2010).
(10) Mayr, H.; Patz, M. *Angew. Chem., Int. Ed.* **1994**, *33*, 938.
(11) Mayr, H.; Bug, T.; Gotta, M. F.; Hering, N.; Irrgang, B.; Janker, B.; Loos, R.; Ofial, A. R.; Remennikov, G.; Schimmel, H. *J. Am. Chem. Soc.* **2001**, *123*, 9500.
(12) Mayr, H.; Ofial, A. R. *Pure Appl. Chem.* **2005**, *77*, 1807.
(13) Maynard, A. T.; Huang, M.; Rice, W. G.; Covel, D. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11578.
(14) Parr, R. G.; Szentpaly, L. V.; Liu, S. *J. Am. Chem. Soc.* **1999**, *121*, 1922.
(15) Chattaraj, P. K.; Giri, S.; Duley, S. *Chem. Rev.* **2011**, *111*, PR43.
(16) Koopmans, T. A. *Physica* **1933**, *1*, 104.
(17) Pérez, P. *J. Org. Chem.* **2003**, *68*, 5886.
(18) Pérez, P.; Toro-Labbe, A.; Aizman, A.; Contreras, R. *J. Org. Chem.* **2002**, *67*, 4747.
(19) *CDK Descriptor Calculator*, version 1.1.1. http://cdk.sourceforge.net/ (accessed September 2011).
(20) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Curr. Pharm. Des.* **2006**, *12*, 2111.
(21) Stewart, J. J. P. *MOPAC2009*; Stewart Computational Chemistry: Colorado Springs, CO, 2008.
(22) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(23) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.;

Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
(24) Gordon, M. S.; Schmidt, M. W. In *Theory and Applications of Computational Chemistry, the first forty years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 1167−1189.
(25) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
(27) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
(28) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
(29) Kohonen, T. *Self-Organization and Associative Memory*; Springer: Berlin, 1988.
(30) Akaike, H. *IEEE Trans. Autom. Control* **1974**, *19*, 716.
(31) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *SIGKDD Explorations* **2009**, *11*, 10.
(32) *Weka*. http://www.cs.waikato.ac.nz/ml/weka/ (accessed September 2011).
(33) Hall, M. A.; Smith, A. *Proceedings of the Twelfth International FLAIRS Conference*; AAAI Press: Menlo Park, CA, 1999; p 235.
(34) Aha, D. W.; Kibler, D.; Albert, M. K. *Mach. Learn.* **1991**, *6*, 37.
(35) Quinlan, R. J. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Singapore*; 1992, p 343.
(36) Wang, Y.; Witten, I. H. *Proceedings of the 9th European Conference on Machine Learning* **1997**, 128.
(37) Breiman, L. *Mach. Learn.* **2001**, *45*, 5.
(38) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 237.
(39) Wang, W. J.; Xu, Z. B.; Lu, W. Z.; Zhang, X. Y. *Neurocomputing* **2003**, *55*, 643.
(40) Chang, C.-C.; Lin, C.-J. *ACM Trans. Intelligent Syst. Technol.* **2011**, *2*, 27.
(41) *LIBSVM*. http://www.csie.ntu.edu.tw/∼cjlin/libsvm (accessed July 2011).
(42) El-Manzalawy, Y.; Honavar, V. WLSVM: Integrating LibSVM into Weka Environment, 2005. http://www.cs.iastate.edu/∼yasser/wlsvm (accessed September 2011).
(43) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.
(44) Aires-de-Sousa, J. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 167.
(45) JATOON applets. http://joao.airesdesousa.com/jatoon/ (accessed September 2011).
(46) Tetko, I. V. *Neural Process. Lett.* **2002**, *16*, 187.
(47) Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717.
(48) VCCLAB, Virtual Computational Chemistry Laboratory. http://www.vcclab.org, 2005.
(49) Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28.
(50) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(51) Details on ADRIANA descriptors can be found at http://www.molecular-networks.com/files/docs/adrianacode/adrianacode_manual.pdf (accessed September 2011).
(52) Details on CDK descriptors can be found at http://pele.farmbio.uu.se/nightly/dnames.html (accessed September 2011).
(53) Details on Dragon descriptors can be found at http://michem.disat.unimib.it/chm/Help/edragon/index.html (accessed September 2011).
(54) Kier, L. B. *Med. Res. Rev.* **1987**, *7*, 417.
(55) Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. *SAR QSAR Environ. Res.* **2002**, *13*, 341.
(56) Schwöbel, J. A. H.; Wondrousch, D.; Koleva, Y. K.; Madden, J. C.; Cronin, M. T. D.; Schüurmann, G. *Chem. Res. Toxicol.* **2010**, *23*, 1576.
(57) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009; Vol. 1 and 2.