J Comput Aided Mol Des (2006) 20:145–157
DOI 10.1007/s10822-006-9042-6

ORIGINAL PAPER

# Application of artificial neural networks and DFT-based parameters for prediction of reaction kinetics of ethylbenzene dehydrogenase

**Maciej Szaleniec · Małgorzata Witko · Ryszard Tadeusiewicz · Jakub Goclon**

Received: 15 December 2005 / Accepted: 6 March 2006 / Published online: 16 June 2006


**Abstract** Artificial neural networks (ANNs) are used for classification and prediction of enzymatic activity of ethylbenzene dehydrogenase from EbN1 *Azoarcus* sp. bacterium. Ethylbenzene dehydrogenase (EBDH) catalyzes stereo-specific oxidation of ethylbenzene and its derivates to alcohols, which find its application as building blocks in pharmaceutical industry. ANN systems are trained based on theoretical variables derived from Density Functional Theory (DFT) modeling, topological descriptors, and kinetic parameters measured with developed spectrophotometric assay. Obtained models exhibit high degree of accuracy (100% of correct classifications, correlation between predicted and experimental values of reaction rates on the 0.97 level). The applicability of ANNs is demonstrated as useful tool for the prediction of biochemical enzyme activity of new substrates basing only on quantum chemical calculations and simple structural characteristics. Multi Linear Regression and Molecular Field Analysis (MFA) are used in order to compare robustness of ANN and both classical and 3D-quantitative structure–activity relationship (QSAR) approaches.

**Keywords** Artificial neural network modeling · DFT · Enzyme activity · Ethylbenzene dehydrogenase · Multiple linear regression · QSAR

M. Szaleniec (✉) · M. Witko · J. Goclon
Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, ul. Niezapominajek 8, 30-239 Krakow, Poland
E-mail: ncszalen@cyf-kr.edu.pl
Tel.: +48-12-6395155
Fax: +48-12-4251923

R. Tadeusiewicz
AGH University of Science and Technology, al.Mickiewicza 30, 30-059 Krakow, Poland

**Abbreviations**

| | |
|---|---|
| ANN | Artificial Neural Network |
| DFT | Density Functional Theory |
| EBDH | Ethylbenzene Dehydrogenase |
| G/PLS | Genetic Partial Least Square |
| LFER | Linear Free Energy Relationship |
| MLP | Multi-Layer Perceptron |
| MLR | Multiple Linear Regression |
| MFA | Molecular Field Analysis |
| QSAR | Linear Quantitative Structure–Activity Relationship |
| SNN | Statistica Neural Networks |
| 3D-QSAR | Three-Dimensional Linear Quantitative Structure–Activity Relationship |

## Introduction

Ethylbenzene dehydrogenase (EBDH) is a key enzyme of the anaerobic metabolism in denitrifying bacterium *Azoarcus* sp. EbN1. It catalyzes an oxygen-independent, stereo-specific hydroxylation of ethylbenzene to (*S*)-1-phenylethanol (Fig. 1). It is the first known example of direct anaerobic oxidation of a non-activated hydrocarbon. EBDH promises potential applications in chemical and pharmaceutical industries, as the enzyme is enantioselective and seems to react with a relatively wide spectrum of substrates [1, 2]. Pure enantiomers of alcohols are of a high value as building blocks for physiologically active compounds. Therefore, it becomes very important to find an easy method, which is characterized by a low cost and low consumption of enzyme, for selection of potential substrates.

It is a common approach in chemistry to analyze the reaction kinetics in terms of Linear Free Energy Relationships

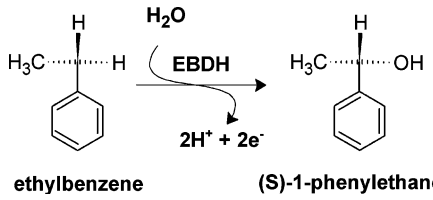


**Fig. 1** Oxidation of ethylbenzene by EBDH in *Azoarcus* sp. EbN1

(LFERs) in order to detect reaction similarities in the group of substrates and to find a tool for predicting a chemical behavior of new, not studied compounds. Such methods result in a wide range of equations that describe electronic and steric factors influencing reaction kinetics [3–5] and although are proved to be valuable for studying many reaction systems, their accuracy is frequently put to the test by enzymatic catalysis, where many parameters both from enzyme and substrate sides combine and interact with each other. Moreover, there are some factors that limit the application of Multiple Linear Regression (MLR), i.e. common reference structure must be shared in the whole studied population and usually only *meta* and *para* substituents parameters (Taft steric constant Es, Hammett's sigma) are available. Where linear methods fail to detect correlations the ANNs with their abilities to locate non-linear patterns frequently find applications as tools of analysis and prediction [6–8]. Combination of theoretical and topologic descriptors with the experimental kinetic rate constants forms a database for training ANN models. Trained ANNs can predict activities of substrates in a given reaction system and provide the theoretical model for studying factors that influence the enzyme–substrate interaction.

It is a common approach to use semi-empirical quantum chemical methods in a computation of QSAR theoretical parameters [9, 10]. Although these methods are faster—the ab initio density functional theory (DFT) approach was chosen due to the facts that orbital energies obtained on this level of theory have direct physicochemical interpretation with a strong background in chemical tradition. The concept introduced by Fukui in the Frontier Molecular Orbitals (FMO) theory [11] allows assignment of FMO energies such as Highest Occupied and Lowest Unoccupied Molecular Orbital (HOMO and LUMO, respectively) to ionization potential and electronic affinity, respectively. Moreover based on these energies absolute hardness, electro-negativity and chemical potential can be easily calculated [12, 13].

The main aim of our research was to build a neural model, which could be used for screening of EBDH activity with new compounds without need of expensive experimental test. We also hoped to get more insight into the mechanistic aspects of the reaction catalyzed by the enzyme. Due to a high structural diversity in EBDH substrates spectrum, an abundance of meta-substituted compounds as well as the presence of heteroaromats (Figs. 2 and 3) the LFER correlations with the classical parameters are of limited use. Therefore the theoretical DFT-based descriptors and ANN as an analysis tool were selected for screening of the commercially available compounds for their possible activities. After testing the activities of 24 compounds, the systems for classification and prediction (with regression approach) of their biological activity were developed. At first for prediction of EBDH activity we developed classifying ANN system. The aim of the classifying network is to estimate so called membership functions of the case to the particular category. The activity was divided into four categories, '0' for inhibitors, '1' for substrates exhibiting relative activity below 50% of activity with ethylbenzene, '2' for the activity of the order 51–150% and '3' for these with relative activity higher than 150%. The output value, which is produced by the neural network, assigns the compound (described by the input parameters) to one of the above-mentioned categories. The output values, calculated by the network, give us some measures of similarity between a new compound (under investigation) and all compounds classified to particular categories in training set. These measures are not (in precise sense) probabilities of belonging of a new compound to the particular class, but in practice it can be taken as an empirical approximation of such a probability.

The advantage of this approach is the simplification of prediction problem by translating continuously changing output value (in this case enzymatic activity) into discrete nominal categories. This method of interpretation results in increased robustness of the whole system and in decrease of number of cases, which are required to train ANN. In fact training of classification neural network is always faster and more successful than training of regression ANN, when we must expect exact values of estimated parameter (e.g. activities of the researched compounds).

As the experiments with classification networks gave promising results, the more demanding regression approach was also applied. In regression ANN, the objective is to estimate the value of a continuous output variable, which is calculated for the known values of the input parameters. In this approach a higher accuracy is required than in classification ANN, which results in much longer and difficult training process. Also number of learning examples (input data assembles with known proper output values) must be richer for regression task than for classification one. Moreover, regression neural model is expected to show extrapolation capabilities, which can be the next source of problems (e.g. we must use ANN of the MLP type instead of RBF networks, because RBF networks are not suitable for extrapolation activities). As in our case we expected that ANN should be able to predict correctly enzyme activity exciding the range of activities for training compounds this issue was also of major importance. The

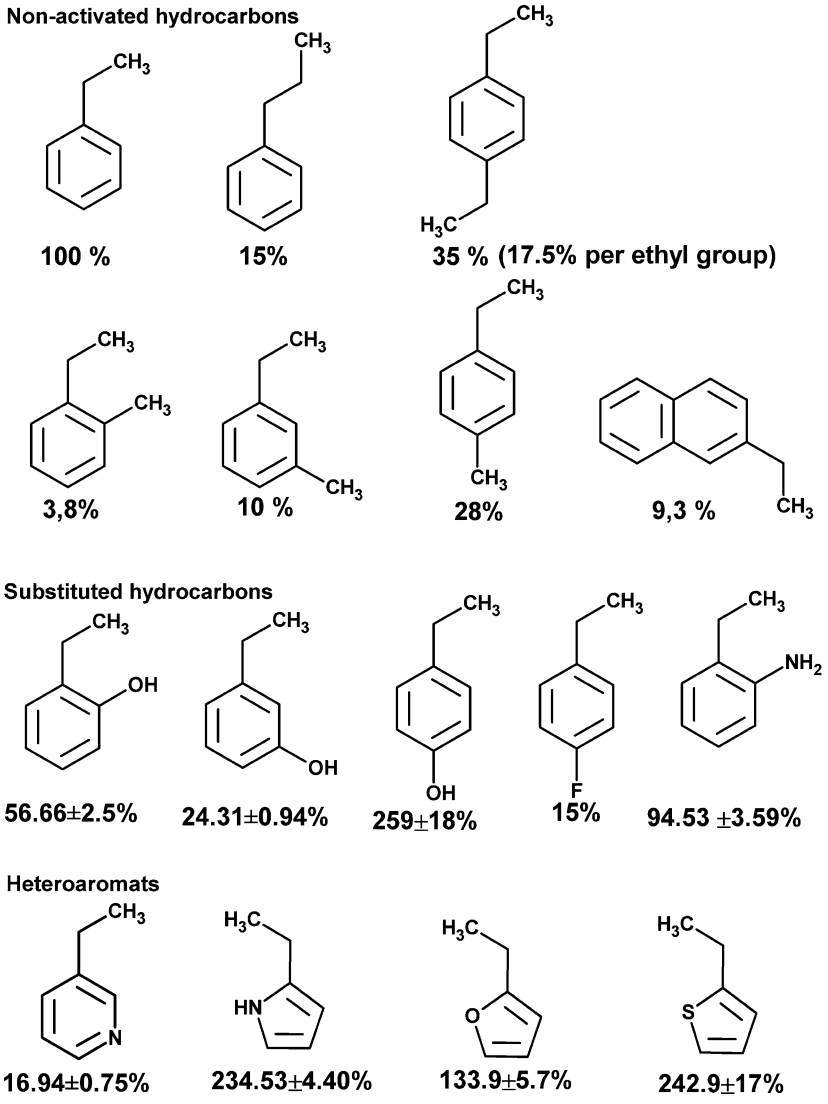


**Fig. 2** EBDH substrates. Relative activity is given as a percentage value below structures

application of regression network yielded with developing of model system, which can directly predict normalized relative reaction rates ($rk_{cat}$).

Basing on the developed systems, the activity of five new compounds (Fig. 4) was assessed. For promising cases the kinetic measurements were performed, thus verifying theoretical forecasting along with an immediate application of ANN systems in our biochemical research. The experimentally measured activities for 4-propylphenol and 4-ethylaniline were in very good agreement with the values provided by regression neural system. Moreover, the post-processing sensitivity analysis of ANN input parameters showed, that the charges differences, the dipole moments as well as the occupation of *ortho* and *para* positions are the most important for the catalytic behavior of the substrates. The neural network approach was complemented by multiple linear regression QSAR analysis for a group of 10 substrates and genetic partial least square (G/PLS) 3D-QSAR for all superimposed substrates. The standard QSAR model showed that hardness and dipole moments of EBDH substrates, as well as electron donating capabilities and steric hindrance of substituents are crucial for describing variation in reactivity. 3D-QSAR model identified regions around substrates that decrease or increase reaction rate through steric or electrostatic interactions with the enzyme's active site.

## Experimental

### Enzyme activity test

The reaction system comprised of three components: organic ethylbenzene derivate, electron acceptor (ferricenum tetrafluoroborate) and enzyme. Two molecules of ferricenum re-oxidize EBDH that was reduced in the reaction

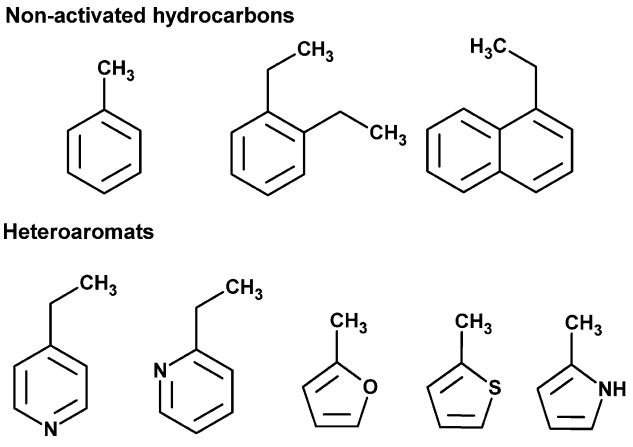


**Fig. 3** EBDH inhibitors

with organic substrate. The reaction kinetics was followed at 290 nm ($\Delta\epsilon$=6,200 $M^{-1}$ $cm^{-1}$). The kinetic parameters were determined by a non-linear fitting to the Michaelis–Menten equation. Each analysis was based on more than 10 concentrations and in every case at least two measurements were performed. As different batches of the enzyme exhibited variations in the specific activity, the $k_{cat}$ of ethylbenzene dehydrogenase towards given substrate was related ($rk_{cat}$) to the $k_{cat}$ of EBDH towards ethylbenzene (typically 0.258 $s^{-1}$) that was measured on the same occasion at saturating concentration of ~60 μM. It was assumed that reaction stoichiometry is analogical (2 electron reaction) in case of all substrates. For the 1,4-diethylbenzene (two ethyl substituents) the activity was calculated per ethyl group.

## Substrate spectrum

For an identification of EBDH substrate spectrum the ferricenum tetrafluoroborate activity assay was applied. Approximately 30 substances were tested for potential activity with EBDH and then 24 substances, which interacted with EBDH, were carefully analyzed in function of their concentration. Among them 16 proved to be substrates whereas 8 turned out to be inhibitors. Introduction of the relative activity ($rk_{cat}$) provided the scale of substrates' activity, which ranges from 3.8% (the worst substrate 2-ethyltoluene) to 259% (the best substrate 4-ethylphenol). As the inhibition mechanism varied between inhibitors we did not provide any discrepancy in the inhibitory effect (assuming 0 activities for all of them). Studied compounds are presented in Figs. 2 and 3.

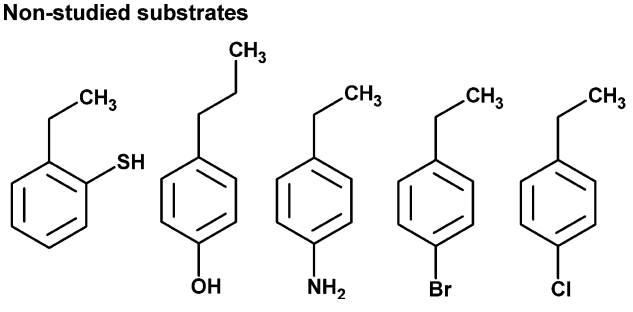


**Fig. 4** Non-studied substrates, which $rk_{cat}$ were predicted with ANN

## DFT modeling

The theoretical parameters for the substrates were calculated by ab initio DFT method using Gaussian 2003 package [14]. Electron correlation and exchange were described by the exchange–correlation functional of the restricted B3LYP [15] type whereas Kohn-Sham orbitals were represented by linear combinations of atomic orbitals using the 6–31G** basis sets. The geometry of the substrates was optimized and the vibration analysis was employed in order to check if the minimum was found. The conformational analyses for chosen substrates (ethylbenzene, 2-ethylnaphthalene, 2-ethylpyrrole, 2-ethyltoluene, 2-ethylphenol, 1,2-diethylbenzene and *n*-propylobenzene) were performed and the lowest energies were found for ethyl (or propyl) group perpendicular (or near-perpendicular) to the aromatic ring for tested compounds. Superimposed structures of all substrates are presented in Fig. 11. Therefore in all cases optimized perpendicular conformers were used in calculation of electronic parameters. Two independent population analyses were performed: Mulliken [16] and Natural Bond Orbital analysis [17]. NMR shielding tensors were computed with the Gauge-Independent Atomic Orbital (GIAO) method [18] and related to tetramethylsilane (TMS) shielding in order to obtain chemical shift.

For each compound the following quantum-chemical parameters were computed and considered in ANN analysis: the partial alpha-carbon charge, the highest and the lowest atomic charges, the difference between the highest (positive) and the lowest (negative) charge in both Mulliken and NBO analyses, the $\nu_{sC-H}$ symmetrical stretching frequency, the HNMR shift of substituted hydrogen, the CNMR shift of alpha-carbon, the dipole moment $\mu$, frontier orbital energies i.e. $E_{LUMO}$ (as an approximation of electron affinity) and $E_{HOMO}$ (as an approximate measure of ionization potential) and GAP—the energy difference between $E_{LUMO}$ and $E_{HOMO}$ (as a measure of absolute hardness). Moreover, as a measure of bulkiness, the molecular weight (MW), total energy of the molecule (SCF) and zero point energy (ZPE) were provided [9].

## Topologic parameters

As EBDH activity seems to be limited to aromatic compounds, in our study we investigated ethylbenzene-core substances. In order to describe structural variations in the particular substituent simple numerical topologic descriptors

such as the number of substituents in the ring, the number of heavy atoms (not hydrogen) in the longest substituent and the number of heavy atoms in all substituents were provided (Fig. 5a). The localization of substituents (*para*, *meta*, *ortho*) in the aromatic ring in relation to the active center (i.e. ethyl group) was described with a variation of one-of-N encoding (1 for presence and 0 for absence of substituent in *para*, *meta* and *ortho* location). The localization of heteroatoms in heterocyclic compounds was not considered. Such an approach allowed straightforward encoding of structural properties of almost all compounds that were taken into consideration. The only artificial encoding was applied for ethylnaphthalenes, where fused aromatic ring was encoded as two substituents in *ortho* and *meta* (or *meta* and *para*) positions with the length of the longest substituent set to 3 atoms.

## ANN analysis

For developing the ANN architecture as well as for training and data validation the commercially available software package namely Statistica Neural Networks 6.0 (www.statsoft.com) was applied. The molecular weight, the relative specific activity, and all DFT-based descriptors were subjected to min-max normalization before feeding into ANN. It is well known that the reduction of the dimensionality of input is the most powerful method for decreasing the number of internal storage elements (synaptic weights) in the network, which leads to better results of the learning process. As only limited set of data was available for the network training it was very important to determine an optimal input dataset, guarantying the fast and effective learning process. For reducing dimensionality of the input vector the forward and backward stepwise feature selection algorithms as well as the genetic algorithm were applied. The performed analysis gave incoherent results suggesting that differences of the information value of the consecutive inputs and differences between corresponding parameters importance were relatively small. The detailed analysis, which was performed, has indicated that only HOMO, LUMO and GAP values as well as charge minimum, maximum and difference should be correlated. As differences were calculated from singular values one of three dependent values was skipped (for example minimal charge, when difference and maximum charge was left). In conclusion, we decided to use an automatic tool: SNN Intelligent Problem Solver (IPS)—the experimental algorithm build-in into Statistica Neural Networks 6.0 software package, which tests thousands of ANN and excludes these input parameters that occur of limited usefulness in the previous models.

At the starting point of each training all cases were randomly divided into 3 subsets: learning, validation and testing. For classification problem the above subsets were composed of 16, 4 and 4 cases whereas for regression model of 18, 3 and 3 cases. Supervised training was used in both cases with standard back-propagation (100 epochs) and conjunct gradients (1–100 epochs) learning algorithm. Approximately 3,000 models were tested by IPS for classification and almost 4,000 models for regression. The IPS-based experiments showed that Multi-Layer Perceptions architecture with one hidden layer were the most appropriate for solving our problem. The best models obtained in

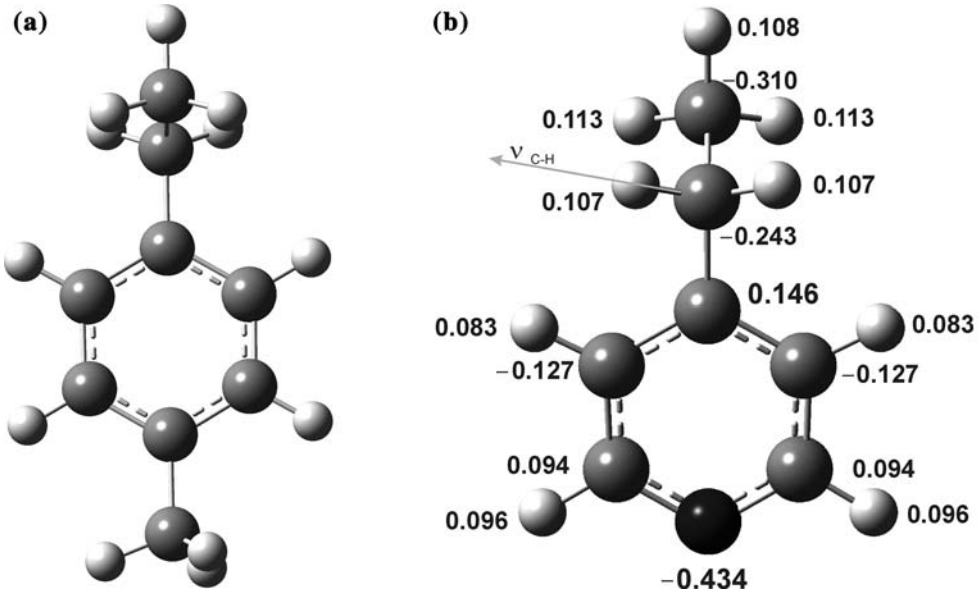


**Fig. 5** (**a**) Example of topologic encoding: No. of substituents: 2, No. of heavy atoms in the longest substituent: 2, number of heavy atoms in all substituents: 3, localization of substituents (*para*, *meta*, *ortho*) in the aromatic ring in relation to ethyl group: (1,0,0), (**b**) Mulliken charge analysis of 4-ethylpyridine (the highest charge: 0.146 and the lowest charge: −0.434) and vibration mode of C–H bond stretching

these experiments were additionally retrained and modified manually. Manual modification of network architecture was based upon the tuning the number of neurons in the hidden layer, in such a way that when the given ANN's training pattern suggested frequent over-fitting, the number of neurons was decreased by one. The optimization of input layer was also conducted manually taking into account the results of the post-processing sensitivity analysis. If, after retraining, the ratio of particular input value was significantly below 1, the parameter was deleted and the network retrained. However, this approach was applied only when the robustness of final ANN increased afterwards. The manual optimization of the retraining process comprised of the modification of training algorithm's type and the number of epochs applied. Quick propagation, conjunct gradients, Levenberg–Marquardt algorithms were used along with standard back propagation procedure.

### MLR analysis

In constructing several linear models the commercially available Statistic 6.0 (www.statsoft.com) Multi Linear Regression Package was applied. The relative kinetic constants (scaled in % from 0 to 250 where 100 is for ethylbenzene) were provided in logarithmic scale. The correlation analysis of log $rk_{cat}$ with various parameters allowed us the selection of parameters characterized by the highest correlation coefficients, namely Taft steric constants Es, the Hammett constants $\sigma$, the hydrophobic parameter $\pi$ [19], the HOMO energy, the dipole moment $\mu$, the charge on alpha carbon (from Mulliken analysis), the frontier orbital energy difference GAP, and the Mulliken charge difference $\Delta q$. As Mulliken and NBO charges exhibit a high degree of co-linearity only the former analysis was used. The limitation posed by Es and $\sigma$ allowed only 10 cases to be used in the analysis. Statistically significant model was obtained by means of step-wise regression. From obtained regression model the $rk_{cat}$ values were calculated and correlated with the experimental ones. As a final test, the activity of 4-ethylaniline was calculated and compared with prediction performed by ANN.

### MFA analysis

Molecular Field Analysis (MFA) was performed with Cerius$^2$ molecular modeling package [20]. All 16 substrates' conformers, optimized by Gaussian 2003, were included in the analysis. Rigid-body, least-square fitting of methin carbon and aromatic ring heavy atoms of each molecule to the corresponding atoms in ethylbenzene was performed. In case of five-member ring compounds only carbon atoms were used in the superposition. Partial atomic charges were computed by the Gasteiger algorithm [21]. The energies of steric and electrostatic interactions were calculated in the universal force field (UFF) [22] respectively with $CH_3$ and H+ probe molecules, in a rectangular grid of 384 points with 2 Å step size. Logarithm of relative activity (scaled in % from 0 to 250 where 100 is ethylbenzene) was provided as $y$-value. Only these data points, which exhibited square correlation with log $k_{cat}$ above 0.2 level, were included in the subsequent analysis. The QSAR equations were constructed by the genetic partial least squares (G/PLS) [23, 24] method with 5,000 crossovers. In order to check the internal predictivity of the derived models, the five best equations were cross-validated with leave-one-out G/PLS algorithm. Finally, as the model's external cross-validation, prediction of the enzyme activity for 4-ethylaniline and 4-propylphenol was conducted for the first five best models and the last one.

## Results

The simplest model, which performs flawless classification, consisted of 15 neurons in input layer, 10 neurons in hidden layer and 4 neurons in output layer (e.g. one output neuron for each classification value 0, 1, 2 or 3). The most excellent ANN was trained with quick propagation algorithm with momentum for 5 epochs and was characterized by 100% correctness in the classification. The learning, validation and testing errors were: 0.137217, 0.190977 and $1\times10^{-6}$, respectively. The architecture and learning curves are presented in Fig. 6.

The prediction of the activity of not studied substrates yields in classification the 4-ethylaniline to the class 3 ($rk_{cat}$ more then 150%), the 4-propylphenol and 3-chlorobenzene to the group 1 ($rk_{cat}$ less then 50%) and the 2-ethylbenzenthiol and 4-bromoethylbenzene to inhibitors (class 0).

### Regression ANN

In case of regression ANN the best architecture, which was found after broad searching, turn to be the MLP characterized by 19 input neurons, 10 neurons in hidden layer and 1 in output layer (Fig. 7a). Several clones of that architecture were retrained with different algorithms yielding very good correlation between the predicted relative specific activities and the experimental data. The choice of the most useful network was based mainly on its quality of prediction for the testing group, as we intended to use them to forecast the activity of non-studied substrates in our experiments. The best ANN, which does not show the over fitting characteristics, was trained with the quick propagation algorithm (37 epochs) with momentum 0.3 and

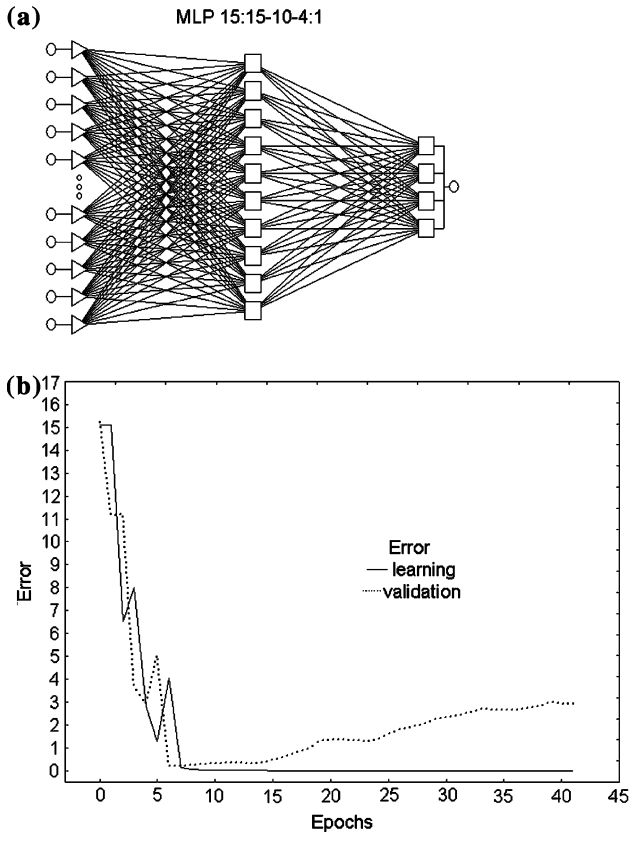


**Fig. 6** (**a**) Architecture of the best classification network and (**b**) Learning curve of MLP 15:15-10-4:1

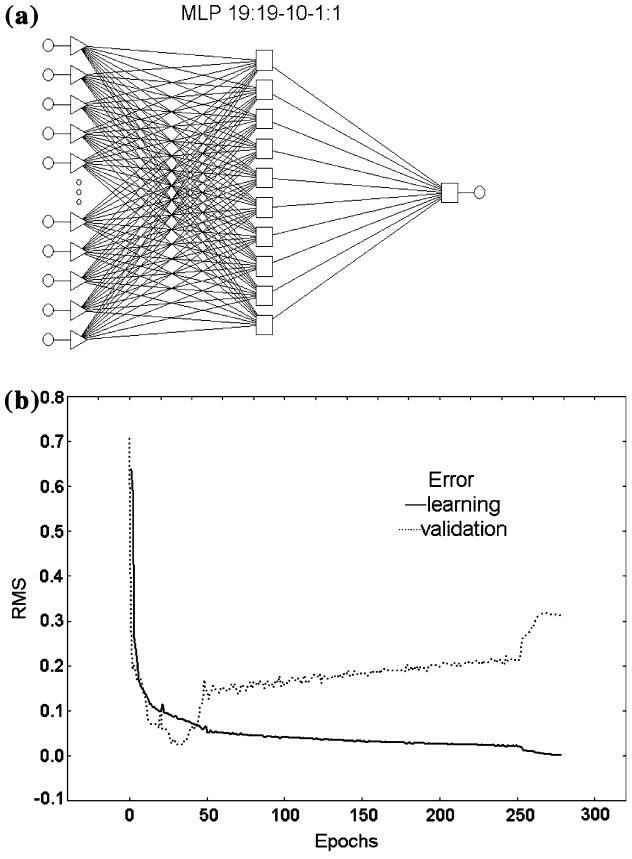


**Fig. 7** (**a**) Architecture of the best regression network and (**b**) Learning curve of MLP 19:19-10-1:1

learning coefficient 0.1. The obtained errors are as follows: learning error=$8.14\times10^{-2}$, validation error=$2.4\times10^{-2}$, testing error=$6.29\times10^{-2}$.

As it is seen from the Fig. 8 the predicted $rk_{cat}$ correlates quite well with the experimental data ($R$=0.97) even the network has a relatively wide variation for inhibitors (with 0 activity). This, however, was to be expected, as we had not provided any variation in inhibitors potency. One should stress that for the non-studied substrates the predicted activities seem reasonable in terms of current understanding of the reaction system [Szaleniec et al. unpublished]. The ANN predicts both the 4-ethylaniline and 4-propylphenol to be highly active ($rk_{cat}$ 0.7), whereas the 2-ethylobenzenthiol with having traceable or zero activity (0.02) and the 4-chloro and 4-bromoethylbenzenes to be inhibitors (with negative values).

Such results suggest that electron-withdrawing substituents stabilize a transition state. This observation is consisted with a common experience from organic chemistry for electrophilic substitution of the aromatic and heterocyclic compounds where a stabilization of carbocation transition state decides about the reactivity. This type of the transition state for ethylbenzene oxidation by EBDH was previously suggested in works of Spormann et al. [2]. If stabilization of carbocation by a vicinal aryl or heterocyclic group is supposed, substituents that stabilize positive charge in *ortho* and *para* position to ethyl groups should increase the reaction rate and decrease the activation energy. The same should take place in case of the five member heterocycles, where ethyl group is in *ortho* position. On the other hand, the reactivity should be decreased in the case of compounds with the electron withdrawing substituents localized in *para* or *ortho* position (such as 4-fluoroethylbenzene) and six member heterocycles such as 4-ethylpyridine or 2-ethylpyridine. What we observed in our kinetic study (data not published) was a decrease of the activation energy for 4-ethylphenol by 24.8 kJ/mol and 2-ethylpyrrole by 7.25 kJ/mol. The decrease of the activation energy correlates with an increase of the reaction rate. For weak electron donor substituents, such as a methyl group in 4-ethyltoluene, we observed only small decrease of activation energy (1.83 kJ/mol) and consequently, possibly due to steric effects, overall decrease in reaction rate. Moreover, it should be pointed out that 4-ethylpyridine and 2-ethylpyridine similarly to ethylbenzene, do not have any steric hindrances, and are not oxidized by EBDH. Instead they are very good enzyme inhibitors. Therefore it can be

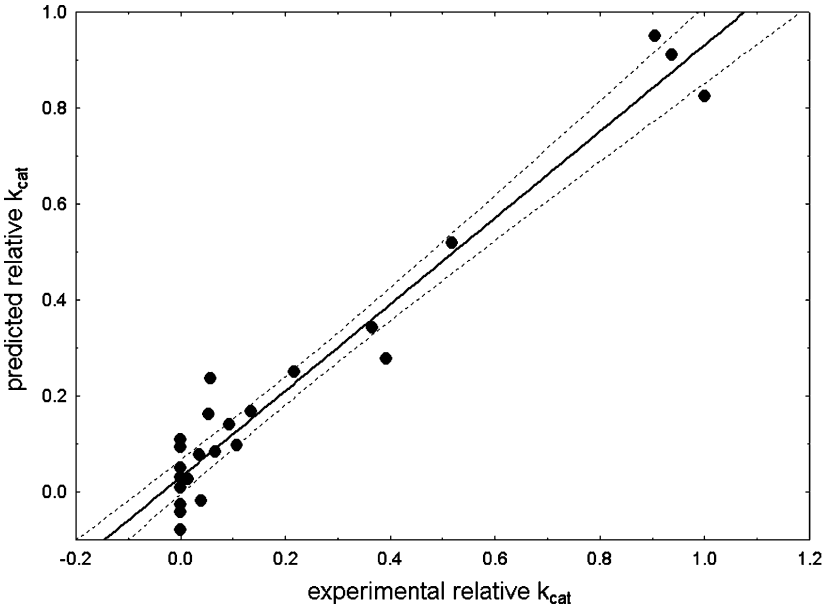


**Fig. 8** Correlation of predicted relative $k_{cat}$ with experimental values ($R^2$=0.9465; $R$=0.9729; $p$=0.0000); Solid line: regression line; dotted line: confidence interval $p$=0.95

assumed, that both compounds are bound to the active center, and cannot be oxidized due to increased activation energy.

## Sensitivity analysis

Additionally to the statistics obtained by means of the ANN, the sensitivity analyses for input variables, conducted by Statistica Neural Network module, were carried out (see Tables 1 and 2). The sensitivity analyses identify key variables in the particular model created by the network. Key variables are these input parameters that have the highest influence on output values obtained from the model, which solves the problem. Therefore, the key variables must be retained in following ANN experiments, while other variables can be excluded form the input vector without significant decrease in the networks performance. The procedure of sensitivity analysis, provided by Hunter et al. [25], rates variables according to the deterioration in modeling performance that occurs if that the variable is no longer available to the model. In order to define the sensitivity of particular input variable the SNN runs the network on a set of test cases, and accumulates the network error for particular input vector. Two types of input vector are used: with (original vector) and without (incomplete vector) assayed input variable. The basic measure of sensitivity is the ratio of the error accumulated for the

**Table 1** Results of sensitivity analysis for classifying neural network (MLP 15-10-4)

| Variable name | Ratio | Rank |
|---|---|---|
| Alpha carbon charge in NBO analysis | 145.01 | 1 |
| $\mu$ | 39.06 | 2 |
| Occupation of *ortho* position | 35.48 | 3 |
| Highest charge in Mulliken analysis | 32.23 | 4 |
| Number of substituents | 17.49 | 5 |
| Δq NBO | 13.00 | 6 |
| Highest charge in NBO analysis | 4.28 | 7 |
| SCF | 4.06 | 8 |
| Lowest charge in NBO analysis | 3.28 | 9 |
| Occupation of *para* position | 2.72 | 10 |
| Lowest charge in Mulliken analysis | 2.49 | 11 |
| ZPE | 1.71 | 12 |
| HOMO | 1.57 | 13 |
| Number of atoms in the longest substituent | 1.05 | 14 |
| Number of heavy atoms in substitutes | 0.41 | 15 |

**Table 2** Results of sensitivity analysis for regression neural network (MLP 19-10-4)

| Variable name | Ratio | Rank |
|---|---|---|
| Highest charge in NBO analysis | 3.89 | 1 |
| Alpha carbon charge in NBO analysis | 3.18 | 2 |
| Highest charge in Mulliken analysis | 2.79 | 3 |
| Δq NBO | 2.44 | 4 |
| Occupation of *ortho* position | 2.44 | 5 |
| Number of atoms in the longest substituent | 2.31 | 6 |
| SCF | 2.19 | 7 |
| Occupation of *meta* position | 2.11 | 8 |
| Number of substituents | 2.08 | 9 |
| C NMR | 1.93 | 10 |
| GAP | 1.89 | 11 |
| ZPE | 1.68 | 12 |
| $\mu$ | 1.64 | 13 |
| Occupation of *para* position | 1.56 | 14 |
| Number of heavy atoms in substitutes | 1.24 | 15 |
| LUMO | 1.10 | 16 |
| Lowest charge in NBO analysis | 1.10 | 17 |
| $\nu_s$C–H | 1.08 | 18 |
| Molecular weight | 1.03 | 19 |

incomplete input vector to the original error. Assuming that analysis removes the input variable that provided valuable information to the model, some deterioration in error might be expected. The more sensitive the network is to a particular input parameter, the greater the deterioration can be expected, and therefore the greater is the evaluated ratio. If the ratio is high above one, the analysis suggests huge loss of model performance after switching off the considered input variable. If it is one or lower, it indicates that the variable has no effect on the network's performance or even cutting it out from the input vector enhances ANN's robustness. Once sensitivities have been calculated for all variables, they are ranked in order.

According to the sensitivity analysis performed for our ANNs the most important parameter in classification problem is the atomic charge (derived from NBO analysis) that is located on reacting carbon (alpha). The next three variables namely the dipole moment, the occupation of the *ortho* position by substituent in benzene ring and the highest atomic charge following from the Mulliken analysis turn out to be almost of the same importance. The least important variables are the number of atoms in the longest substituent and the number of heavy atoms in substitutes.

In the regression network the most important are: the highest charge and the charge on alpha carbon (both following from the NBO charge analysis) whereas the highest charge from the Mulliken analysis is the next one. The difference $\Delta q$ in NBO charges as well as the occupation of *ortho* position turn out to be of the same major. The least important variables are: the symmetric stretching frequency of the C–H bond and the molecular weight.

Standard correlation analysis shows that only the highest charge and the charge difference $\Delta q$ (both arising from the NBO analysis) correlate significantly with the relative $k_{cat}$ ($R$=0.4707, $p$=0.020, $R$=0.4059, $p$=0.049). The rest of parameters do not linearly couple with relative kinetic constant.

## Multiple linear regression

In the multiple linear regression analysis the substrates such as 1,4-diethylbenzene, 2-ethylaniline, 2-ethylphenol, 2-ethyltoluene, 3-ethylphenol, 3-ethyltoluene, 4-ethylphenol, 4-ethyltoluene, 4-fluorethylbenzene, ethylbenzene were considered. The best-obtained model has very high (0.99) correlation coefficient and corrected $R^2$ (0.987) value. The regression equation takes into account two classical and two quantum chemical parameters. Its final form with $\beta$ coefficients is as follow:

$$\begin{aligned}\log k_{\mathrm{cat}} = &-0.18\ (\pm 0.05)\sigma + 0.97\ (\pm 0.05)\\ &\times \mathrm{Es} - 1.04\ (\pm 0.1)\mu - 1.09\ (\pm 0.09)\ \ \mathrm{GAP}\\ &+ 14.84\ (\pm 1.1)\\ R^2 = &0.98;\ \ \mathrm{No.\ \ cases} = 10\end{aligned}$$

Figure 9 shows the graphical correlation of experimental and predicted values.

The accuracy of the model was tested in a similar manner as for the ANN approach i.e. the relative activity of a substrate with a structural core analogical to ethylbenzene (4-ethylaniline) was calculated. The obtained value for $k_{cat}$ is 398% (1.53 in normalized ANN scale).

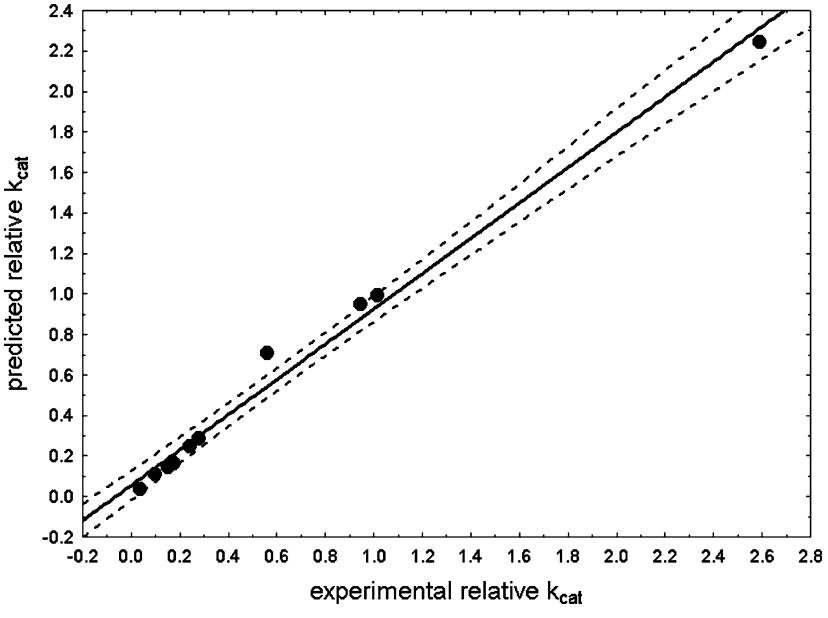


**Fig. 9** Correlation of predicted relative $k_{cat}$ with experimental values in MLR model ($R^2$=0.98; $R$=0.99; $p$=4×10$^{-9}$). Solid line: regression line; dotted line: confidence interval $p$=0.95

Molecular field analysis

In MFA all substrates were considered. The used G/PLS procedure generated a set of 99 equations. For the five top models $R^2$ reached a value of 0.994. The cross-validation procedure, which was performed for these equations, allowed selection of the best model (PRESS=0.0482, cross-validated $R^2$=0.990). The activities calculated from the model exhibit very high correlation with experimental data ($R^2$=0.9928) with only one, significant outliner present, namely 1,4-diethylbenznene (Fig. 10). However, when obtained models were confronted with external validation problem, their performance was no longer that excellent. The studied models estimated the activity of 4-propylphyenol in the range of 6–18% (the best model 18.3%, 0.07 of normalized ANN output value) and 4-ethylaniline 12–13% (0.05 of normalized value) in case of the best top models, and 6% in case of the last model.

The obtained model comprises of the set of points in space, surrounding substrate structure, which defines a positive or a negative steric and electrostatic interaction with the enzyme active site (Fig. 11). Therefore, in our case, the ability of the model to predict enzyme activity of not-studied compounds was less important than the knowledge provided by point's positions and their importance in the model. However, in order to correctly interpret results of MFA, one should look at loading values instead of variables co-efficiencies from the QSAR equation. The analysis of the loading values, provided in square brackets in Fig. 11, shows that the positive influences on kinetics come from electrostatic interactions in position *para* (loading value 0.497) and *meta* (loading value 0.232), and finally, the least important *ortho* (loading value 0.108) one. These influences are balanced by steric negative interactions in the vicinity of *para* (CH3/131 loading value 0.497) and *ortho* (CH3/321 loading value 0.282) positions as well as negative electrostatic interaction (H/254 loading value 0.255). To some extent this analysis explains, for example, a variation of activity of 4-ethylphenol, where steric hindrance, introduced by phenolic substituent, is outweighed by strong positive electrostatic interactions. It is also true in case of *ortho*-substituted ethylphenol although not to the same extent (as overall steric influences are stronger). Finally, in case of the least active 3-ethylphenol, negative steric and electrostatic interactions in the substituent vicinity seem to decrease the activity while electrostatic interaction (H+/226), localized above the aromatic ring, is of a smaller importance due to the distance to the substituent.

## Conclusions and application

The ANNs turn out to be a very useful tool for predicting the variation of the activity in the studied group of ethylbenzene derivatives. The approach surpasses the multiple linear regression method by a number of factors. First, as it predicts directly the kinetic constant, not its logarithm, it lowers the prediction error. Second, thanks to the theoretical parameters, it lacks the limitation to the family of compounds with common structural core, which is characteristic for the standard QSAR. Third, both substrates and inhibitors could be analyzed with ANN. Finally, and most importantly, the ANNs exhibit higher prediction

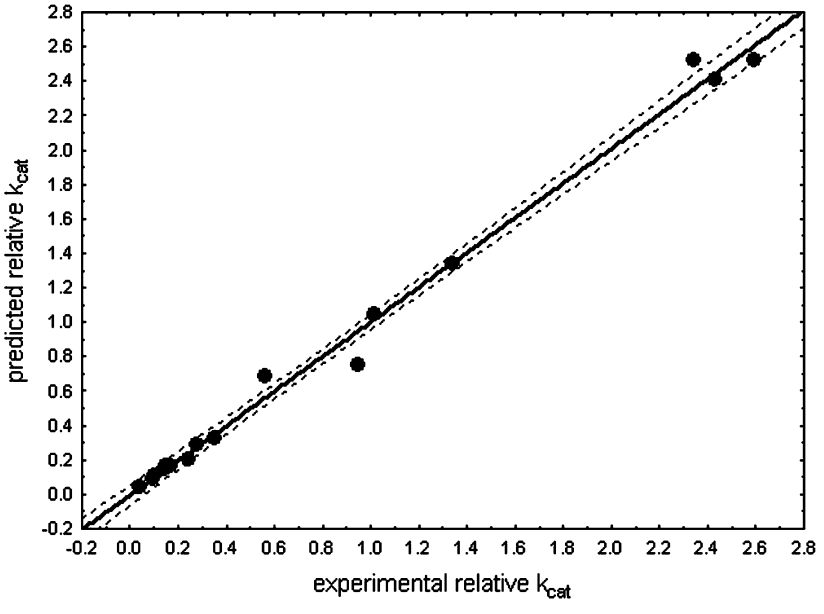


**Fig. 10** Correlation of predicted relative $k_{cat}$ with experimental values in MFA model ($R^2$=0.9928; $R$=0.9964; $p < 1\times10^{-5}$). Solid line: regression line; dotted line: confidence interval $p$=0.95

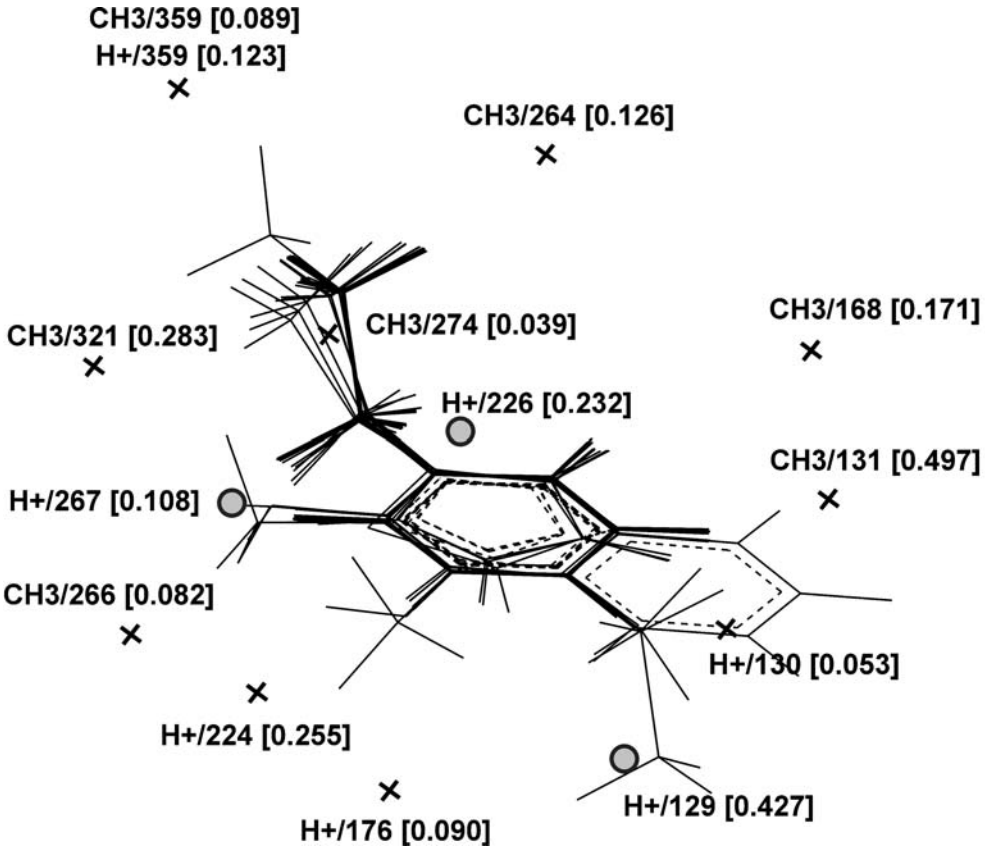


**Fig. 11** Localization of steric (CH3) and electrostatic (H+) points around the superimposed substrates. Crosses/circles denote interactions that decrease/increase enzyme activity, respectively. Loading values (importance in the model) are provided in the square brackets

capabilities for cases, which were not included into the learning (or regression) process.

As the conclusion it can be clearly stated that ANN approach is capable of supporting chemist intuition in the quantitative prediction of the enzyme activity and is far much superior to the traditional 2D-QSAR. To some extent the MFA in 3D-QSAR [26] approach can complement ANN methods. This approach bases on the calculation of steric (van der Waals) and electrostatic (Coulombic) interaction between the compound of interest, and a ''probe atom'' placed at the various intersections of a regular 3D lattice, large enough to surround all of the compounds in the series. The construction of 3D-QSAR equation allows not only the quantitative prediction of chemical activity but also localization of points that can show favorable and unfavorable steric regions around the molecules as well as favorable and unfavorable regions for electropositive or electronegative substituents in certain positions.

This approach certainly lacks second, mentioned above, disadvantage of traditional QSAR, that is the restriction to the same core molecular architecture (as superimposing of all pharmacophore atoms is the compulsory condition in MFA) and the limited availability of experimental descriptors (such as Hammett sigma) for some isomers. Still ANN surpasses that approach by the fact that it can incorporate into the model both substrates and inhibitors. As it is sometime difficult to define one scale for both groups (especially when the type of inhibition varies) the construction of regression model describing all compounds might be very tedious, if ever possible. On the other hand, the flexibility of ANN systems provides relatively easy method to build a model, which describes well even such non-evenly distributed dataset as was in our case. Moreover, there is no need to superimpose molecules for ANN input, and the only prerequisite is to optimize the whole dataset into the same type of minimum.

Even our two neural systems were not in prefect agreement towards the activity of non-studied substrates, which were modeled in silico (Fig. 4), we decided to investigate activities of two promising compounds, which exhibited activities in both analysis, namely the 4-ethylaniline and 4-propylphenol. The preliminary kinetic measurements show that 4-ethylaniline exhibits 136% of ethylbenzene $k_{cat}$ (0.51 of normalized value), which expose slight ANN overestimation (0.7 predicted). However, one should have in mind that this value might be much nearer the truth due to a fact that in the experiment the effect of substrate inhibition was detected. This means that evaluation of maximum activity needs more careful investigation, and what follows the actual value might be lowered by that effect. It should be

underlined that prediction made by the MLR model yields the normalized activity equal to 1.53, which is three times more than the experimental value.

The normalized relative activity that was obtained for the second substrate equals to 0.69, and is in a very good agreement with the predicted value (0.7). Therefore, we can conclude that the ANN screening has proved to be of a high practical usefulness as it enabled selection of new and very active substrates.

Our genetically optimized MFA models exhibited much better correlation of predicted to experimental log $k_{cat}$ than it was seen from ANN models. However, although the G/PLS algorithm was used for equation development, which should prevent the over-fitting, and in addition models had exhibited the high internal cross-validation quality, the activities of not studied compounds, 4-ethylaniline and 4-propylphenol were far much lower than these seen in the experiment (ANN normalized values: 0.07 instead of 0.51 and 0.05 instead 0.69, respectively).

Some conclusion supporting the understanding of the reaction mechanism can also be drawn from the sensitivity analyses. Although exact ranking of variables changes strongly from one to other ANN systems, one can notice that charges, charge differences and dipole moments have very frequently high position in the global ranking. Therefore, one can assume that these features are important for the catalytic behavior of the substrates. From the topological parameters the occupation of *ortho* and *para* positions seems to be of the highest importance, possibly due to the steric effects influencing binding to the active center and positioning of the substituent with their directional electronic effect on benzene ring (and therefore stabilization of charge in reaction intermediate). This assumption found its confirmation in 3D-QSAR model, which showed that there is a strong diversification of steric and electrostatic influences around the substrate's aromatic ring. For example, localization of MFA points show, that *meta* position is unfavorable because of the in-plane steric and electrostatic interactions, while the steric hindrance, introduced in the *para* position, can be counterweighted by the electrostatic interaction (such as in 4-ethylphenol).

From MLR models we can clearly assume that both electronic and steric effects are of the great significance for the reaction system. The big advantage of the MLRs models over the ANNs is the relative simplicity of the results interpretation. Even the selection of the parameters sometimes depends on the method of regression that is used and on the assembly of starting parameters (in a very analogical manner as in ANN), the straightforward information, which is provided by beta coefficients, has much higher value than parameters ranking in case of ANN specially in terms of understanding of the reaction mechanism from the chemical point of view.

The results of performed studies allow us to conclude that the bigger (more negative Es) and the more electron-withdrawing (higher $\sigma$) constituent is the slower reaction proceeds (lower log $k_{cat}$). This result quantitatively supports our experimental kinetic observation and activation energy measurements and is backing up the hypothesis of the carbocationic transition state as the reaction rate-limiting step. Moreover, even as the exact binding mode of substrate at the active center is unknown, it is believed that there is some type of direct interaction of molybdenum metal center with substrate carbon atom, which activates chemically inert hydrocarbon. The negative coefficients for GAP (the difference of energy of frontier orbitals), which can be understood as absolute hardness, indicates that softer the substrate is the faster it reacts. Supposing that some electronic interaction of substrates and the metal center indeed takes place, and soft character of the active center of the molybdenum enzyme is taken into account, this result is in the perfect agreement with Hard and Soft Acid and Bases theory. These cannot be deduced from complicated ANN models. Therefore, it can be stated, that where the prediction is crucial, the ANN surpasses QSAR models, and where the simplistic induction from data is needed, the regression still finds useful application if only it is applicable.

Future work is going to be directed into the bearing of a new type of self-optimizing neural networks [27]. These ANNs can be used both with supervised and unsupervised learning methods. Some results presented recently on the Neural Networks Methodology [28] show that in case of problems similar to the discussed above, such a type of neural tool can be found as more flexible and more effective.

**Acknowledgments** State Committee for Scientific Research (KBN) has supported this research under grant KBN/SGI2800/PAN/037/2003 and 3T09A06228. Maciej Szaleniec acknowledges a PhD grant of Polish Academy of Sciences.

## References

1. Kniemeyer O, Heider J (2001) J Biol Chem 276:21381
2. Johnson HA, Pelletier DA, Spormann AM (2001) J Bacteriol 183:4536
3. Hammett LP (1936) J Chem Phys 4:613
4. Taft RW, Lewis IC (1958) J Am Chem Soc 80:2436
5. Shorter J, Chapman NB (ed) (1978) Correlation analysis in chemistry recent advances. Plenum Press, New York and London
6. Funar-Timofei S, Suzuki T, Paier JA, Steinreiber A, Faber K, Fabian WMF (2003) J Chem Inf Comput Sci 43:934
7. Bravo S, Diez MC, Shene C (2004) Braz J Chem Eng 21:509
8. Mager PP, Weber A (2003) Drug Des Discov 18:127
9. Bucinski A, Nasal A, Kaliszan R (2000) Comb Chem High Throughput Screen 3:525
10. Nasal A, Bucinski A, Baczek T, Wojdelko A (2004) Comb Chem High Throughput Screen 7:313

11. Fukui K (1982) Science (Washington DC) 218:747
12. Singh PP, Srivastava HK, Pash FA (2004) Bioorgan Med Chem 12:171
13. Nalewajski R (2001) Podstawy i metody chemii kwantowej. Wydawnictwo Naukowe PWN, Warszawa
14. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision D.01. Gaussian, Inc., Wallingford, CT
15. Becke ADJ (1993) Chem Phys 98:5648
16. Mulliken RS (1955) J Chem Phys 23:1833
17. Reed AE, Curtiss LA, Weinhold F (1988) Chem Rev 88:899
18. Wolinski K, Hilton JF, Pulay P (1990) J Am Chem Soc 112:8251
19. Hansch C, Leo A (1995) Exploring QSAR. Fundamentals and application in chemistry and biology. ACS Processional Reference Book, American Chemical Society, Washington DC
20. Accelrys Inc. (2005) Cerius$^2$ modeling environment, release 4.8. Accelrys Software Inc., San Diego
21. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
22. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM (1992) J Am Chem Soc 114:10024
23. Rogers D, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854
24. Glen WG, Dunn WJ III, Scott DR (1989) Tetrahedron Comput Methodol 2:349
25. Hunter A, Kennedy L, Henry J, Ferguson RI (2000) Comput Methods Prog Biomed 62:11
26. Cramer RD III, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
27. Horzyk A, Tadeusiewicz R (2005) In: Mira J, Alvarez JT (eds) Mechanism, symbols, and models underlying cognition. Lecture Notes in Computer Science, vol 3561, Part I. Springer-Verlag, Berlin Heidelberg New York, pp 156–165
28. Braspenning PJ, Thuijsman F, Weijters AJMM (ed) (1995) Artificial neural networks an introduction to ANN theory and practice. Lecture Notes in Computer Science, vol 931. Springer Verlag