FULL-LENGTH PAPER

Quantitative structure–activity relationship study of antitubercular fluoroquinolones

Nikola Minovski · Marjan Vračko · Tom Šolmajer

Abstract Quantitative structure–activity relationship study on three diverse sets of structurally similar fluoroquinolones was performed using a comprehensive set of molecular descriptors. Multiple linear regression technique was applied as a preprocessing tool to find the set of relevant descriptors (10) which are subsequently used in the artificial neural networks approach (non-linear procedure). The biological activity in the series (minimal inhibitory concentration (μg/mL) was treated as negative decade logarithm, pMIC). Using the non-linear technique counter propagation artificial neural networks, we obtained good predictive models. All models were validated using cross validation leave-one-out procedure. The results (the best models: *Assay1*, $R = 0.8108$; *Assay2*, $R = 0.8454$, and *Assay3*, $R = 0.9212$) obtained on external, previously excluded test datasets show the ability of these models in providing structure–activity relationship of fluoroquinolones. Thus, we demonstrated the advantage of non-linear approach in prediction of biological activity in these series. Furthermore, these validated models could be proficiently used for the design of novel structurally similar fluoroquinolone analogues with potentially higher activity.

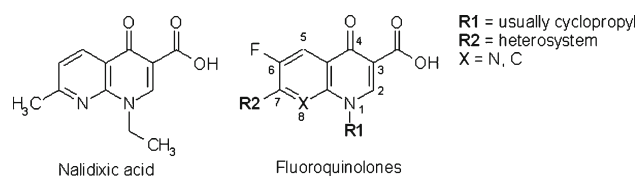**Keywords** Tuberculosis · Fluoroquinolones · DNA gyrase · QSAR · CP ANN

N. Minovski (✉) · M. Vračko · T. Šolmajer
National Institute of Chemistry, Hajdrihova 19, Ljubljana 1001, Slovenia
e-mail: nikola.minovski@ki.si

Abbreviations

ATP	Adenosine triphosphate
MIC	Minimal inhibitory concentration
SAR	Structure–activity relationship
QSAR	Quantitative structure–activity relationship
CODESSA	Comprehensive descriptors for structural and statistical analysis
MLR	Multiple linear regression
NN	Neural networks
ANN	Artificial neural networks
CP ANN	Counter propagation artificial neural networks
KANN	Kohonen artificial neural networks
SOM	Self organizing maps
CVLOO	Cross validation leave-one-out

Introduction

In the last decades, the increase in the incidence of tuberculosis is evident, and now it represents one of the major risk factors for lethality among the patients suffering from infectious disease. In order to control this situation, chemotherapy is used as a weapon of first choice. Enhanced drug resistance development in microorganisms, as well as the side effects and toxicity of already known antitubercular agents, direct the research efforts today toward the design of a new generation antitubercular drugs with better activity, and possible lower toxicity [1] (http://www.immunisation.nhs.uk/files/TB_factsheet.pdf). According to the World Health Organization, one-third of human population is infected by *Mycobacterium tuberculosis* and around two million people die from tuberculosis every year. The tuberculosis itself is a disease mainly caused by the microorganism *M. tuberculosis*, but in some cases it can be caused by other *Mycobaterium* species such as *M. fortuitum*,

Fig. 1 Nalidixic acid versus generic structure of fluoroquinolones

M. smegmatis, and *M. avium–intracellulare complex*. The treatment of tuberculosis [2] is long and faces many problems. Tuberculosis associated with AIDS (due to HIV infection) forms a fatal combination and is currently responsible for 13% of the number of deaths due to HIV infection [3]. In the past, several types of targets have been identified, as well as a wide spectrum of antitubercular agents [4]. The microbes continuously develop resistance toward drugs, many antituberculosis drugs show side effects in clinical use, and there is a lack of preparations active agains stable mycobacteria. These factors are stimulating the development of new generation anti-tuberculosis drugs.

In the search for new therapeutic targets and new anti-infective agents, fluoroquinolones are particularly interesting because of their broad spectrum of activity against various bacteria, mycobacteria, and parasites [5–11]. The quinolones are a family of broad-spectrum antibiotics originating from nalidixic acid. The majority of quinolones in clinical use belong to the subset of fluoroquinolones, which have a fluoro group attached to the central ring system, typically at the 6-position (Fig. 1).

According to structure–activity relationship (SAR) studies, the 1,4-dihydro-4-oxo-3-pyridinecarboxylic acid scaffold is essential for anti-mycobacterial activity [12]. The pyridone system must be annulated with an aromatic ring system. Substitutions at position 2 greatly reduce activity, but the substituents at positions 5, 6, 7 (especially), and position 8 of the annulated ring system may result in increased anti-mycobacterial activity. Fluorine atom substitution at position 6 is very important and also will result in significantly enhanced anti-mycobacterial activity. Substitutions of alkyl substituents at the position 1 are essential for activity (the lower alkyl substituents such as methyl, ethyl, and especially cyclopropyl, enhance the potency and efficiency). Ring condensations at the (1,8), (5,6), (6,7), and (7,8) positions are also important and can significantly increase the activity [12].

One of the well-established targets of antibacterial agents is DNA gyrase [13], which is unique bacterial type-II topoisomerase (topo II) that catalyzes the introduction of negative supercoils into DNA using the free energy of ATP hydrolysis [14]. The enzyme consists of two subunits GyrA and GyrB that form a functional heterodimer A_2B_2. The GyrA subunit is responsible for DNA breakage and reunion, where GyrB is also involved. This catalytic process involves large conformational movements of the enzyme, enabled by the binding

and hydrolysis of ATP on the GyrB subunit. Notably, similar conformational changes are also involved in the ATP-independent relaxation process. Another closely related bacterial enzyme from the topoisomerase-II family is topoisomerase-IV (topo IV), which also forms a heterodimer C_2E_2 consisting of two ParC subunits and two ParE subunits [15]. Despite being highly homologous enzymes, gyrase and topo IV possess different vital biological functions in the cell [16,17]. They are both involved in control of the topological state of DNA molecules, where gyrase is required for initiation of DNA replication and elongation of nascent DNA, while topo IV is able to relax DNA supercoils and decatenates daughter chromosomal DNA. Fluoroquinolones are the only inhibitors of gyrase/topo IV commonly used for treatment of bacterial infections in hospital and clinics. These synthetic compounds belong to the class of GyrA/ParC inhibitors [18]. However, reports of already known fluoroquinolone side effects and toxicity alerts have revived a growing interest in providing structurally similar analogues of these drugs with better antitubercular activity and lower toxicity.

The most common side effects caused by the quinolones involve gastrointestinal tract (GIT, nausea, and diarrhea) and central nervous system (CNS, headache, and dizziness) disorders. These side effects are generally benign and therefore the discontinuation of the therapy is usually not needed [19]. Some idiosyncratic adverse reactions are more specific to individual agents that may share some structural similarity, such as the 1-(2,4)-difluorophenyl group located in trovafloxacin (associated with hepatitis), temafloxacin (associated with hemolytic-uremic syndrome), and tosufloxacin (associated with eosinophilic pneumonitis). On the other hand, the quinolones with higher discontinuation rates, such as trovafloxacin (7.0%) and grepafloxacin (6.4%), are no longer available for general use (trovafloxacin was withdrawn from the market due to the risk of hepatotoxicity).

The structure–activity (SAR) and structure–toxicity (STR) relationship analyses of the fluoroquinolone antibiotics show that one of the most significant structural changes for activity was the increasing the steric bulk at R7 position of the main 6-fluoroquinolone scaffold. This steric bulk, usually obtained by alkylation, results in reducing the CNS effects, drug–drug interactions, and genotoxicity. Simultaneously, the improved potency and in vivo efficacy against Gram-positive species is observed. However, the activity against the Gram-negative bacteria is lowered. The halogen substitutions at the position 8, provided large increases in the antibacterial activity, but these substitutions have also been associated with side effects such as increased phototoxicity and decreased bacterial selectivity.

In some of the new quinolone antibiotics reported in [20] the halogen substitution at position 8 has been eliminated. Two other substituents at R8 (N, $COCH_3$) have also been used in combination with the new alkylated side chains at R7

position of the main quinolone scaffold in order to maintain efficacy. At position R1, cyclopropyl substituent is still the optimal one and it appears to be important for increasing the metabolic stability. On the other hand, this substituent represents the risk factor for increasing theophyllin interactions (drug–drug interactions) and clastogenicity. According to the structure–activity (SAR) and structure–toxicity (STR) relationship studies, the 2,4-difluorophenyl group would be an ideal bioisosteric replacement for the cyclopropyl group [20].

In this study, we have used three diverse datasets consisting of 65–145 fluoroquinolones and a large set of molecular descriptors (more then 600 constitutional, topological, geometrical, and electrostatic parameters) to construct structure–activity relationship models which could subsequently be used for prediction of the biological activity of novel fluoroquinolones.

Materials and methods

The biological assay data used in our study are from in vitro tests for inhibitory activity against *M. tuberculosis*. Initially, we collected three diverse datasets of fluoroquinolones and their related activity values [Minimal inhibitory concentration, MIC (μg/mL)] measured in the same laboratory for each dataset separately. The datasets were collected from online database source [21] using the advanced search criteria by chemical/therapeutic class. The datasets are named *Assay1*, *Assay2*, and *Assay3*. *Assay1* consists of 66 fluoroquinolone structures and their activity values. These fluoroquinolones were collected using the search criteria "fluoroquinolones" in the NIAID therapeutics database [21]. According to SAR analysis, each structure in the dataset has 1,4-dihydro-4-oxo-3-pyridinecarboxylic acid moiety and fluorine atom substitution at the position 6 in the main ring system which are believed to be essential for anti-mycobacterial activity (Fig. 1). The MIC values of the collected fluoroquinolones are in the range from 0.003 to 256.0 μg/mL. Some of the structures do not correspond to the class of classical fluoroquinolones because of the third ring system condensed to the bicyclic ring system. These fluoroquinolone structures have MIC values in the range (0.9–64.0 μg/mL). From the same online database source, we collected *Assay2* consisting of 145 fluoroquinolone structures and their biological activity values. These fluoroquinolones are collected using the search criteria "antituberculous agents" in the NIAID therapeutics database. The MIC values of the collected fluoroquinolones are in range from 0.00121 to 7.12 μg/mL. Because several fluoroquinolone structures in the *Assay2* dataset had identical values for activity (MIC, μg/mL), we eliminated these structures and their corresponding activity values and constructed a reduced dataset named *Assay3* consists of 65 diverse fluoroquinolone structures. There are similarities in the chem-ical structures between the *Assay1*, *Assay2*, and *Assay3*. The chemical structures of the compounds and the related biological activity values are available in the Supporting information (Tables S1–S3). The method of collecting the compounds and their properties is based on using Microsoft Excel and the specifically integrated toolbox for ChemBio-Office Ultra 2008 (v.11.0) (http://www.cambridgesoft.com/software/ChemBioOffice). This unusual method of collecting data is suitable for generating the structure database input file format (*.sdf) for each of the datasets, which is subsequently used for simultaneous calculation of the molecular descriptors in DRAGON software package [22]. In Microsoft Excel, MIC (μg/mL) values are converted into pMIC ($-\log_{10}$(MIC)) values, and used in CODESSA software package for multiple linear regression (MLR) analysis [23,24].

Construction of a quantitative structure–activity relationships model

To obtain a QSAR model, compounds are presented by the molecular descriptors. For calculation of the 2D molecular descriptors for each compound of the datasets obtained, we used two software packages, DRAGON [22] and CODESSA [23,24]. Using an in-house developed software application [25], we made a conversion of the DRAGON's list of calculated molecular descriptors (*.txt output format) into a CODESSA descriptor input file (*.txt input format). The rest of the descriptors used in the MLR analysis were calculated directly in CODESSA before running the MLR analysis. The pool of the calculated 2D molecular descriptors (about 600), which were considered for further calculations, can be separated into four classes: constitutional, topological, geometrical, and electrostatic. All descriptors were analyzed using the CODESSA software with heuristic option, which is a suitable option for selection of descriptors [24].

First, the following selection algorithm calculating the one-parameter correlation equations between descriptors and activity and eliminating all descriptors that do not fulfill the criteria below was used [23].

The *F* test's value for the one-parameter correlation with the descriptor is below 1.00.

The squared correlation coefficient of the one-parameter equation is less than $R^2_{\min}$ (in our case $R^2_{\min} = 0.1$).

The parameter's *t* value is less than *t*1 (where $R^2_{\min}$ and *t*1 = 1.5).

The descriptor is highly intercorrelated (above r_{full}, where $r_{\text{full}} = 0.99$), with another descriptor and this other descriptor has a higher squared correlation coefficient in the one-parameter equations based on these descriptors.

With the remaining descriptors all possible two- and more-parameter linear models were calculated. We examined the models with 2, 3, 4, 5, and 10 parameters, presented in

Table 1 Linear models obtained by CODESSA

ID	Dataset	Str.	D.No./D.Re.	Params.	R	R_{cv}	F	s^2	Descriptors involved
1	Assay1	66	594/38	2	0.7081	0.6661	31.68	0.8636	EEig10r, Tei
				3	0.7968	0.7601	35.96	0.6423	EEig10r, Tei, MATS4m
				4	0.8341	0.8008	34.87	0.5443	EEig09r, Tei, nR09, MATS4m
				5	0.8489	0.8141	30.96	0.5080	EEig10r, Tei, MATS4m, SEigZ, MATS8e
				10	0.9049	0.8566	24.86	0.3594	EEig10r, Tei, nR09, MATS4e, MATS3v, GGI2, GATS3e, EEig01r, EEig04d, EEig07d
2	Assay2	145	605/26	2	0.5929	0.5686	38.50	0.7086	X1A, nR10
				3	0.6336	0.6068	31.53	0.6586	X1A, nR10, JGI2
				4	0.6637	0.6356	27.57	0.6200	X1A, nR10, JGI2, EEig03d
				5	0.6801	0.6484	23.93	0.5999	X1A, nR10, JGI2, EEig03d, ZXS/ZXR
				10	0.7661	0.7205	19.04	0.4783	X1A, nR10, PW3, MATS8e, JGI2, GATS8v, EEig04d, nR09, nR03, YZS/YZR, JGI3
3	Assay3	65	603/33	2	0.7270	0.6951	34.76	0.8134	X1A, nR10
				3	0.7840	0.7500	32.46	0.6755	X1A, MSD, D/Dr03, nR09
				4	0.8430	0.8136	36.87	0.5156	X1A, nR10, EEig03d, GGI2
				5	0.8551	0.8255	32.10	0.4874	X1A, nR10, EEig03d, GGI2, MPCO, EEig01r
				10	0.9172	0.8796	28.65	0.3142	X1A, nR10, EEig03d, GGI2, MPCO, piPC08, GATS5v, MATS8m, ESpm13d

Str. number of structures per dataset, *D.No.* starting pool of molecular descriptors, *D.Re.* number of descriptors retained from the starting pool, *Params.* parameters used in heuristic option. Statistically significant molecular descriptors are italicized

Table 1. The definition of best descriptors used for the correlations are listed in Table 2. The selection procedure resulted in ten best descriptors, which were subsequently used in neural network modeling.

Neural networks approach

As further modeling techniques we applied Kohonen Artificial Neural Networks (KANN) alias Self Organizing Maps (SOM) and the Counter Propagation Artificial Neural Networks (CP ANN). The details of both the techniques were described previously and therefore we give here only a brief summary [26]. Due to their relatively simple structure the SOM and CP ANN are suitable tools for QSAR/QSPR modeling [27,28]. The models were developed and tested with two strategies. SOM represents a basic type of ANN and is a mapping from multi-dimensional descriptor space onto two-dimensional network of neurons. The mapping runs over a non-linear algorithm known as training. The result of the training is two-dimensional array of neurons, which are occupied with objects in a way that similar objects are located close to each other (or on the same neuron). With the visual inspection of the map one can recognize similarity relationships among objects with the respect on the entire dataset. A trained network is not only a look up table of objects but also a model. During the mapping from multi-dimensional descriptor space onto a limited number of neurons (training) different parts of space overlap and build the complex model. It was shown that the SOM is a suitable tool to divide the set of compounds into the training set and the test set. In this procedure, the two-dimensional map was divided in sub-parcels and from each sub-parcel some objects were put into the training set and others into the test set [29–32]. It is expected that after such division both the sets cover entire information space equivocally what is a condition for a reliable test. If one wish to validate a QSAR model the points in the test set must be close to the points in the training set in the multidimensional descriptor space [30]. To adjust the optimal technical parameters the training of the SOMs was carried out with different dimensions of the network architecture and trained with different number of epochs for each of our datasets. We selected the network with such dimensions that all neurons were occupied. In this way about 80% of objects were put into the training set and 20% of objects into the test set. It is important to emphasize that the division was performed considering only descriptors, i.e., independent variables and before the modeling.

CP ANN represents an extension of SOM [26,33]. In variation to SOM the property or target values (in our case the MIC activities) are included in the modeling. The network architecture shows an additional layer, which is associated to the activity values. The learning algorithm works in two steps. In the first step, which is the same as described above,

Table 2 Specification of descriptors used in our two to ten-parameter QSAR models

Descriptor	Source	Definition	Class
EEig01r	DRAGON	Eigenvalue 01 from edge adj. matrix weighted by resonance integrals	T
EEig09r	DRAGON	Eigenvalue 09 from edge adj. matrix weighted by resonance integrals	T
EEig10r	DRAGON	Eigenvalue 10 from edge adj. matrix weighted by resonance integrals	T
EEig03d	DRAGON	Eigenvalue 03 from edge adj. matrix weighted by dipole moments	T/E
EEig04d	DRAGON	Eigenvalue 04 from edge adj. matrix weighted by dipole moments	T/E
EEig07d	DRAGON	Eigenvalue 07 from edge adj. matrix weighted by dipole moments	T/E
MATS4m	DRAGON	Moran autocorrelation-lag4/weighted by atomic masses	T
MATS8m	DRAGON	Moran autocorrelation-lag8/weighted by atomic masses	T
MATS4e	DRAGON	Moran autocorrelation-lag4/weighted by atomic Sanderson electronegativities	T/E
MATS8e	DRAGON	Moran autocorrelation-lag4/weighted by atomic Sanderson electronegativities	T/E
MATS3v	DRAGON	Moran autocorrelation-lag4/weighted by atomic van der Waals volumes	T
GATS3e	DRAGON	Geary autocorrelation-lag3/weightred by atomic Sanderson electronegativities	T/E
GATS5v	DRAGON	Geary autocorrelation-lag5/weightred by atomic van der Waals volumes	T
GATS8v	DRAGON	Geary autocorrelation-lag8/weightred by atomic van der Waals volumes	T
nR03	DRAGON	Number of 3-membered rings	C
nR09	DRAGON	Number of 9-membered rings	C
nR10	DRAGON	Number of 10-membered rings	C
JGI2	DRAGON	Mean topological charge index of order 2	E
JGI3	DRAGON	Mean topological charge index of order 3	E
GGI2	DRAGON	Topological charge index of order 2	T/E
YZS/YZR	CODESSA	YZ Shadow/YZ Rectangle	T
ZXS/ZXR	CODESSA	ZX Shadow/ZX Rectangle	T
X1A	DRAGON	Average connectivity index chi-1	T
PW3	DRAGON	Path/Walk 3-Randic shape index	T
MSD	DRAGON	Mean square distance index (Balaban)	T
MPCO	CODESSA	Max partial charge for a O atom [Zefirov's PC]	E
SEigZ	DRAGON	Eigenvalue sum from Z weighted distance matrix (Barysz matrix)	T
D/Dr03	DRAGON	Distance/detour ring count of order 3	T
piPC08	DRAGON	Molecular multiple path count of order 08	T
ESpm13d	DRAGON	Spectral moment 13 from edge adj. matrix weighted by dipole moments	T/E
Tei	CODESSA	Topographic electronic index (all pairs) [Zefirov's PC]	T/E

C constitutional, *T* topological, *G* geometrical, *E* electrostatic

the descriptors determine the distribution of objects in the two-dimensional network. In the second step, the positions are projected to the additional layer where the response surface is constructed. In the process of training all neurons are affected. The output neurons neighboring the central neuron may remain unoccupied at the end of training. The output layer enables the predictive activity of the CP ANN. The prediction of property for an unknown input object is performed by positioning the input object into the input layer of the trained neural network, locating it by finding the most similar or central neuron and disclosing the weights in output layer at the position of the central neuron [34]. The training process encompasses all neurons, also those that remain empty after the input of objects in the training set to the trained network. These empty neurons may better accommodate a new, unknown object (in the test set), not equal or similar to any of the objects from the training set.

Different models which were developed with training sets were used to assess the ability of the obtained ANN models to predict biological activity (MIC) of the compounds from each fluoroquinolone dataset. We have performed ample experiments and test calculations to obtain the best models with respect to dimensions of the network, number of epochs and learning rates. The criterion for the model selection was the maximal value of correlation coefficient R for cross validation leave-one-out (CV LOO, R_{cv}) and test set results (see Supporting Information in file, Table S4). When these models were optimized they were again tested on their predictive ability with the test sets. The models we have selected in Table 4 (marked in red in Supporting Information in the file, Table S4) produce the best test set predictions as evaluated by R_{cv} criterium. The apparently better models which give a better value for training set (marked in blue) all result in inferior prediction for the test set.

Results and discussion

Quantitative structure–activity relationships (QSAR) which have been examined to rationalize the various biological activities of fluoroquinolones as well as to design new power-

Table 3 Details of dividing the datasets using SOM

ID	Dataset	No	NNarch	Num. Ep.	Training set	Test set
1	*Assay1*	66	10×10	1200	48	18
2	*Assay2*	145	18×18	1200	115	30
3	*Assay3*	65	10×10	1200	46	19

No total number of objects, *NNarch* the dimensions of the network, *Num. Ep.* number of Epochs

ful compounds are well documented [35–39]. The aim of our study is to approach the mechanism of anti-bacterial activity in these series on the basis of a comprehensive set of 2D molecular descriptors (about 600) and to construct a robust model for use in the design of new potential candidates. In our QSAR analysis, we initially examined the models with 2, 3, 4, 5, and 10 parameters performed on three diverse sets of fluoroquinolones active against *M. tuberculosis* for which a good agreement was observed between the predicted and experimental *p*MIC values. The two to ten-parameter models for each of the datasets (*Assay1*, *Assay2*, and *Assay3*) are described in Table 1.

Since we start our procedure with a set of 600 descriptors a possibility exists to encounter a chance correlation in case that the number of screened variables would exceed the number of observations. As shown in Table 1, the number of observations (column 3, number of observations is 66, 145, and 65 in *Assay 1*, *Assay 2*, and *Assay 3*, respectively) is sufficient in order to screen the number of retained descriptors or variables (column 4: number of descriptors 38, 26, 33) to keep the probability of encountering a chance correlation with $R^2 > 0.8$ at the 1% level or less [40].

In order to improve the predictive performance of our models, we applied the CP ANN modeling technique. As input data we used previously obtained best ten-parameter models for each dataset separately. Initially, SOM was applied to divide our datasets (*Assay1*, *Assay2*, and *Assay3*) into the training set and the test set as described above (Table 3). The models were built with the training set using the Cross Validation Leave-One-Out (CV LOO) testing method for optimization of technical parameters (the details are given in Table 4). For these purposes, we used different neural network architectures and different number of epochs. The neural network architecture is determined by the dimension of the two-dimensional network (number of neurons in X and Y direction). The optimal dimensionality of the network ($X \times Y$) depends on the number of structures in the training set in order to cover all investigated objects (structures). The results of the test set predictions (R_{best}) which are obtained on the basis of the best training set models are presented in Table 5a. The mean values (R_{mean}) of the test set predictions show that for all training set models (different NN architectures and number of epochs), the corresponding R values for the test sets are statistically significant (*Assay1*:

$0.90396 < R < 0.93289$; *Assay2*: $0.82088 < R < 0.91770$; *Assay3*: $0.84116 < R < 0.93969$). In this way any possible bias in test set predictions is eliminated. For more information, see Supporting Information (Table S4). The MLR models developed in the first linear part of the manuscript are of minor importance for testing the predictive performances in comparison with non-linear models developed in the second part.

These results suggest that *p*MIC can be correlated to a sum of interactions of constitutional, topological, and electrostatic nature. The frequency analysis of the descriptors involved in our models, suggest the list of parameters (Table 2) which emerged as most important for establishing a QSARs. Namely, the ten most frequent descriptors in our models are: EEig10r, EEig03d, Tei, MATS4m, nR09, nR10, X1A, JGI2, GGI2, and MPCO that belong to topological, constitutional, and electrostatic class of molecular descriptors. The parameters EEig10r, MATS4m, X1A, Tei, EEig03d, and GGI2 belong to the class of topological and electro-topological parameters, and describe the importance of the shape, i.e., the scaffold 6-fluoroquinolone moiety (1,4-dihydro-4-oxo-3-pyridinecarboxylic acid) for activity.

On the other hand, Tei parameter defined as Topographic electronic index of order 2 (statistically significant parameter) is of great importance for activity of 6-fluoroquinolones on the binding site and points out the electrostatic interaction between fluoroquinolones and the GyrA subunit which may result in increased stability of the fluoroquinolone binding to the complex [41]. The parameters, JGI2 and MPCO, denoting Mean topological charge index of order 2 and Max partial charge for a O atom [Zefirov's PC], respectively, belong to the class of pure molecular electrostatic parameters. These molecular descriptors are of significant importance for activity and indicate that in vitro antibacterial activity against *M. tuberculosis* is strongly dependent on the charge indices for the F atom at position 6 and the O (sp^2) atom of the carbonyl and carboxyl groups in the scaffold 6-fluoroquinolone moiety (Fig. 1) and suggest the possibility of establishing an electrostatic interaction between the F atom and the receptor, as well as the formation of hydrogen bonds between the carbonyl groups of the fluoroquinolones and the receptor.

The three datasets (*Assay1*, *Assay2*, and *Assay3*) in the article were observed separately as three different cases. This non-uniformity is described also with the occurrence

Table 4 Details of training set learning strategy using CV LOO and selection of the best models

Assay1	NN-architecture							
	7 × 7		8 × 8		9 × 9		10 × 10	
	R	R_{cv}	R	R_{cv}	R	R_{cv}	R	R_{cv}
48-structures								
Epochs								
400	0.87	0.48	*0.94*	*0.51*	0.98	0.45	0.96	0.55
600	0.87	0.40	0.92	0.53	0.98	0.43	0.98	0.50
800	0.87	0.55	0.93	0.51	0.98	0.40	0.96	0.51
1000	0.87	0.58	0.92	0.43	0.98	0.33	0.98	0.41
Assay2	11 × 11		12 × 12		13 × 13		15 × 15	
115-structures								
Epochs								
400	0.91	0.71	0.93	0.72	0.93	0.69	0.94	0.73
600	0.91	0.69	0.91	0.70	0.94	0.69	0.97	0.71
800	0.92	0.68	0.93	0.68	0.95	0.71	0.95	0.69
1000	0.91	0.66	0.94	0.70	*0.96*	*0.70*	0.95	0.67
Assay3	7 × 7		8 × 8		9 × 9		10 × 10	
46-structures								
Epochs								
400	0.94	0.67	0.95	0.70	0.96	0.73	0.99	0.71
600	*0.96*	*0.73*	0.96	0.77	0.95	0.69	0.98	0.73
800	0.94	0.70	0.95	0.79	0.95	0.74	0.97	0.70
1000	0.96	0.73	0.96	0.72	0.96	0.72	0.96	0.68

Italicized values correspond to the best training set models

Table 5 Technical parameters and test set predictions obtained by CV LOO

ID	Dataset	NoTes	Param	Learning rates		Trs (models)		NN definition		Tes (predictions)	
				η_{min}	η_{max}	R	R_{cv}	NNarch	Number of epochs		
(a) *Ten-parameter predicted pMIC*										R_{best}	R_{mean}
1	Assay1	18	10	0.01	0.5	0.94	0.51	8 × 8	400	0.9329	0.9148
2	Assay2	30	10	0.01	0.5	0.96	0.70	13 × 13	1000	0.9177	0.8748
3	Assay3	19	10	0.01	0.5	0.96	0.73	7 × 7	600	0.9397	0.9059
(b)*Reduced-parameter predicted pMIC*										R	
1	Assay1	18	5	0.01	0.5	0.96	0.50	8 × 8	400	0.8108	
2	Assay2	30	7	0.01	0.5	0.96	0.62	13 × 13	1000	0.8454	
3	Assay3	19	6	0.01	0.5	0.95	0.62	7 × 7	600	0.9212	

NoTes number of objects in the test set, *Param* number of parameters, *Trs* training set, *R* coefficient of correlation, R_{cv} cross validated coefficient of correlation, *NNarch* neural network architecture, *Tes* test set, R_{best} best test set predictions, R_{mean} mean values of all test set predictions

of different molecular descriptors through the datasets. The analysis of the molecular descriptors selected by Heuristic algorithm, showed these differences in the main scaffold, described for example with the nR09 and nR10 parameters (number of nine, i.e., ten-membered rings). We believe that such differences enhance the robustness of our models for possible prediction of activity for classical and non-classical 6-fluoroquinolones. In order to ensure that the selected descriptors of the ten-parameter models are intercorrelated between the datasets, we calculated the intercorrelation matrix. Table S5 presented in the Supporting Information, shows the intercorrelation table for ten descriptors of *Assay1*

and *Assay2* used in the models. Two of them are the same, further seven from eight descriptors from *Assay2* are strongly correlated with at least one descriptor from *Assay1* ($R^2 > 0.40$) [42].

In an additional modeling experiment we reduced the number of significant parameters in our models using the intercorrelation matrix for each of the training sets (*Assay1*, *Assay2*, and *Assay3*) and eliminated the molecular descriptors for which intercorrelation coefficients were more then or equal then 0.40, i.e., $R^2(P_i, P_m) \leq 0.40$. An upper limit of 0.40 was proposed for R^2 to eliminate the chance correlation [42]. According to this, we removed the insignificant param-
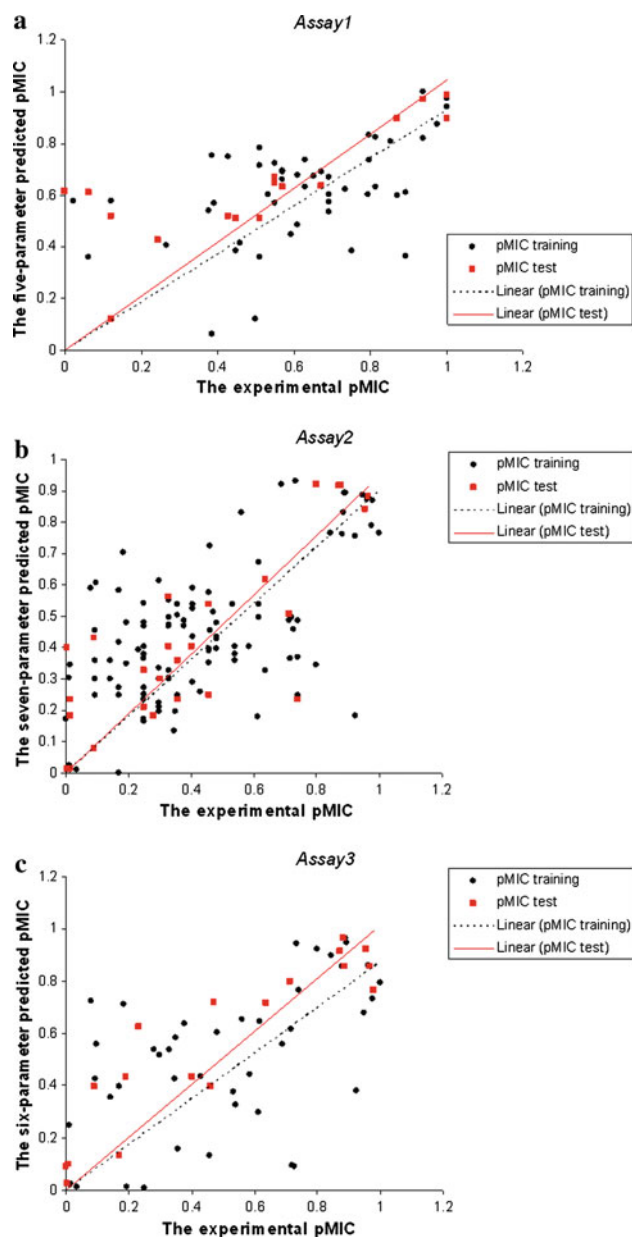
Fig. 2 The predicted *p*MIC versus experimental values using CV LOO method. **a** *Assay1*, **b** *Assay2*, **c** *Assay3*

eters and obtained the following configuration (number of parameters + property) for the training/test set (*Assay1*, 5 + 1; *Assay2*, 7 + 1; *Assay3*, 6 + 1). Then we used the same NN settings (NN architecture and number of epochs) for the training set developed before parameter reduction, for development of the new training set models. The obtained models were used for test set predictions employing CV LOO method (Table 5b). The results on the test sets (the best models: *Assay1*, $R = 0.8108$; *Assay2*, $R = 0.8454$, *Assay3*, $R = 0.9212$) demonstrate the good predictive ability of these models in providing novel structurally similar fluoroquinolone analogues with possible higher activity (Fig. 2).

In summary, for the modeling the biological activity in the fluoroquinolone datasets we applied linear and non-linear techniques. The linear method was used, first as heuristic algorithm for descriptor selection, and second to build the models in form of linear equations with up to ten descriptors. As non-linear method the CP ANN method was selected mostly because of its transparent structure when compared to other NN techniques [33,43]. The CV LOO results indicate that both the models show comparable performance. The output layer of the CP ANN model is positioned directly below the input layer, with a one-to-one correspondence of the neurons. This means that each neuron from the input layer has an ascribed property stored in the output layer at the same position. Such 2D map is used for the representation of the distribution of compounds (Fig. 3).

When the model is used for prediction of the compound from external dataset a new compound is first situated into the map, and second, its position is projected into the output layer, which gives the activity value. The coordinates of the position of each object in the map (Kohonen layer) are correlated to the target property to create the response surface in the output layer. The analysis of the closest neighbors in the map allows us to justify the prediction. Figure 3 shows the output layers and the positions of test set compounds in the Kohonen networks. More details are given in Table S6 of Supporting Information which shows for each compound of test sets its predicted value, the coordinates of the neuron in the network, and the neighbors in the network. Under "neighbors" we report the compounds of training set, which are situated on the same neuron, or when it is empty (not occupied with the compound of the training set) on the neurons in the closest vicinity. These compounds determine the predictions. In most cases, the neighborhood consists of compounds with similar activities. This means that a compound which is situated into a neighborhood of similar compounds is expressing also similar activities. Different situation occurs when a compound is situated in the neighborhood occupied with compounds of different activities. Such examples are the compounds 4 and 8 of *Assay3* with activities 0.4603 and 0.0902, respectively. For both the compounds, the model predicts the activity 0.3005, which was determined with training compounds 1, 2, and 5 with activities 0.0970, 0.7250, and 0.0800, respectively. We know a priori that the predictions for compounds situated in such conflict neighborhood are not reliable [44].

Conclusions

QSAR models were derived for three diverse datasets of 6-fluoroquinolones. Initially, linear QSAR models with different number of parameters were developed, and subsequently non-linear CP ANN approach was used. The cor-
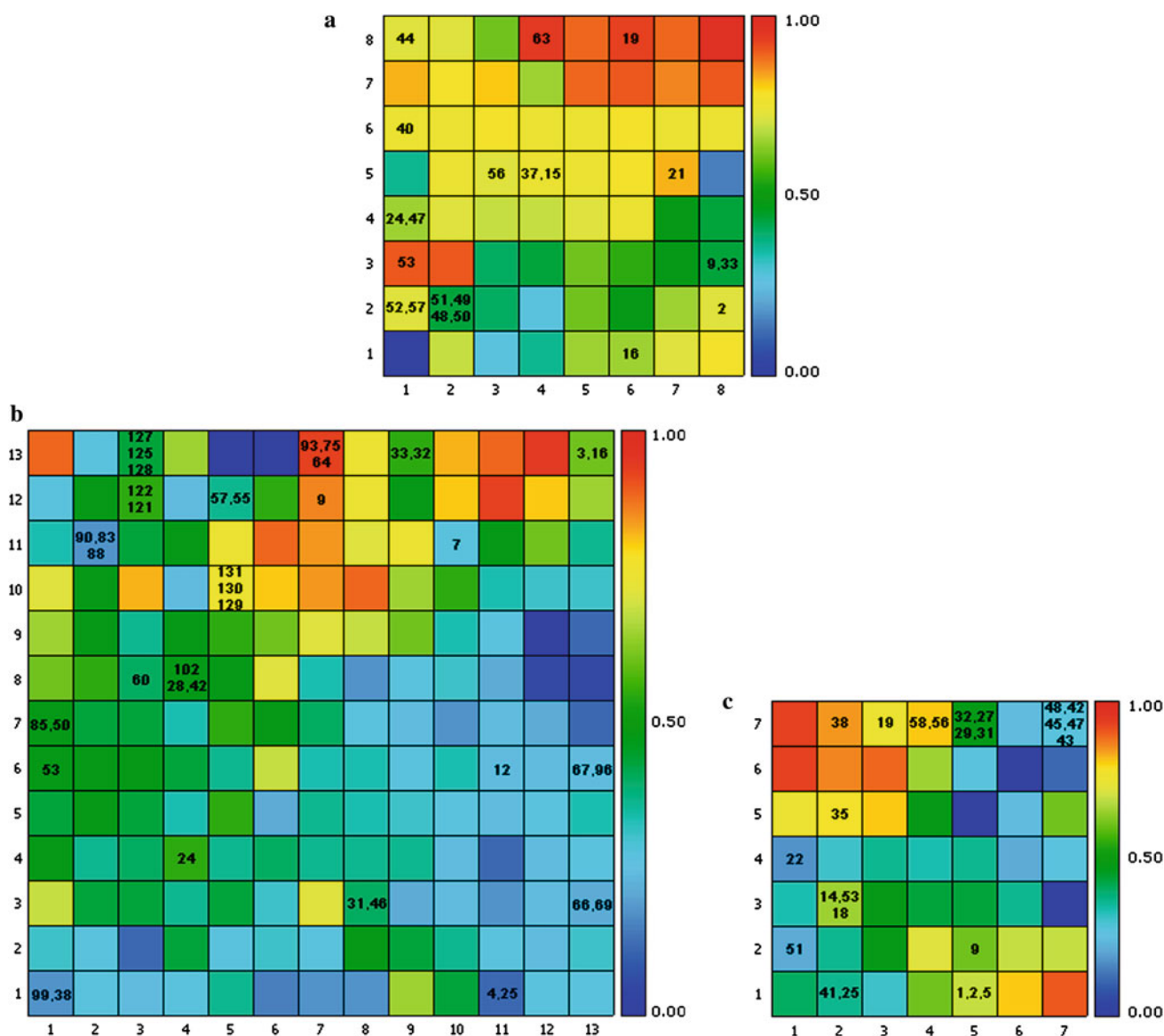
Fig. 3 The output layers of models. **a** *Assay1*, **b** *Assay2*, **c** *Assay3*. The *p*MIC values were normalized to the interval from 0 to 1. The value 0.00 corresponds to minimum activity, whereas value 1.00 cor-

responds to maximum activity. The dimensions of the networks were **a** 8×8, **b** 13×13, **c** 7×7. The *number* labels denote the test set objects in the output layer

relation coefficients between measured and predicted values for test sets are large $R > 0.9$. This indicates a good predictive ability of presented models. Among all the descriptors, topological and electrostatic descriptors were found to have high coding capabilities for the *p*MIC values and were selected to represent the chemical structures.

The present work provides an effective model for the prediction of the *p*MIC values for 6-fluoroquinolones. This study also shows that the utility of the QSAR treatment involving descriptors derived solely from chemical structure and the correlation equation and descriptors can be used for the prediction of the *p*MIC values for unknown structures.

We would like to stress that our linear and non-linear results are not directly compared. The linear procedure served as a preprocessing tool into find the set of relevant descriptors (10) which are subsequently used in the ANN approach (non-linear procedure). The models show comparable statistical results, however, CP ANN models enable visualization of data and prediction.

References

1. Kayukova LA, Praliev KD (2000) Main directions in the search for new antituberculous drugs. Pharm Chem J 34: 11–18. doi:10.1007/BF02524551

2. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC (1999) Consensus statement. Global burden of tuberculosis estimated incidence, prevalence, and mortality by country. JAMA 282:677–686

3. Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C (2003) Tuberculosis. Lancet 362:887–889. doi:10.1016/S0140-6736(03)14333-4

4. Janin YL (2007) Antituberculosis drugs: ten years of research. Bioorg Med Chem 15:2479–2513. doi:10.1016/j.bmc.2007.01.030

5. Derouin F (2001) Anti-toxoplasmosis drugs. Curr Opin Invest Drugs 2:1368–1374

6. Drlica K, Zhao X (1997) DNA gyrase, topoisomerase IV, and the 4-quinolones. Microbiol Mol Biol Rev 61:377–392

7. Owens RC Jr, Ambrose PG (2000) Clinical use of the fluoroquinolones. Med Clin North Am 84:1447–1469

8. Hooper DC (2001) Mechanisms of action of antimicrobials: focus on fluoroquinolones. Clin Infect Dis 32:S9–S15

9. Hooper DC (2001) Emerging mechanisms of fluoroquinolone resistance. Emerg Infect Dis 7:337–341

10. Ruiz J (2003) Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection. J Antimicrob Chemother 51:1109–1117. doi:10.1093/jac/dkg222

11. Anquetin G, Greiner J, Vierling P (2005) Quinolone-based drugs against Toxoplasma gondii and Plasmodium spp. Curr Drug Targ: Infect Disord 5:227–245

12. Block JH, Beale JM (eds) (2004) Whilson and Gisvold's textbook of organic medicinal and pharmaceutical chemistry. Lippincott Williams & Wilkins, Philadelphia, pp 247–252

13. Maxwell A (1997) DNA gyrase as a drug target. Trends Microbiol 5:102–109

14. Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annu Rev Biochem 70:369–413

15. Peng H, Marians KJ (1993) Escherichia-coli topoisomerase-iv-purification, characterization, subunit structure, and subunit interactions. J Biol Chem 268:24481–24490

16. Reece RJ, Maxwell A (1991) DNA gyrase—structure and function. Crit Rev Biochem Mol 26:335–375

17. Levine C, Hiasa H, Marians KJ (1998) DNA gyrase and topoisomerase IV: biochemical activities, physiological roles during chromosome replication, and drug sensitivities. Biochim Biophys Acta 1400:29–43

18. Oblak M, Kotnik M, Solmajer T (2007) Discovery and development of ATPase inhibitors of DNA gyrase as antibacterial agents. Curr Med Chem 14:2033–2047

19. Owners RC, Ambrose PG (2005) Antimicrobial safety: focus on fluoroquinolones. Clin Infect Dis 41:S144–S157

20. Domagala JM (1994) Structure-activity and structure-side-effect relationships for the quinolone antibacterials. J Antimic Chemother 33:685–706

21. NIAID-Division of AIDS, HIV/OI/TB Therapeutics Database. http://chemdb2.niaid.nih.gov

22. Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations), version 5.4, 2006. http://www.talette.mi.it

23. Katritzky AR, Lobanov VS, Karelson M (1994) CODESSA, Reference Manual, Gainsville

24. Katritzky AR, Lobanov VS, Karelson M (1995) CODESSA, Training Manual, Gainsville

25. Minovski N (2009) DragCOD v1.0, laboratory for chemometrics. National Institute of Chemistry, Ljubljana

26. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley-VCH, Weinheim, 380pp

27. Vracko M (2005) Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. Curr Comp-Aid Drug Des 1:73–78

28. Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Mazzatorta P, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A (2006) Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR QSAR Environ Res 17: 265–284. doi:10.1080/10659360600787650

29. Simon V, Gasteiger J, Zupan J (1993) A combined application of two different neural network types for prediction of chemical reactivity. J Am Chem Soc 115:9148–9159

30. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25: 235–251. doi:10.1002/qsar.200510161

31. Valkova I, Vračko M, Basak SC (2004) Modeling of structure-mutagenicity relationships: counter propagation neural network approach using calculated structural descriptors. Anal Chim Acta 509:179–186. doi:10.1016/j.aca.2003.12.035

32. Kotnik M, Oblak M, Humljan J, Gobec S, Urleb U, Solmajer T (2004) Quantitative structure–activity relationships of Streptococcus pneumoniae MurD transition state analogue inhibitors. QSAR Comb Sci 23:399–405. doi:10.1002/qsar.200430875

33. Zupan J, Novic M, Gasteiger J (1995) Neural networks with counter propagation learning strategy used for modeling. Chemometr Intell Lab Systems 27:175–187

34. Novic M (2008) Artificial neural networks. In: Livingstone DS (ed) Methods and protocols. Humana Press, Totowa, pp 45–60

35. Mitscher LA (2005) Bacterial topolsomerase inhibitors: quinolone and pyridone antibacterial agents. Chem Rev 105: 559–592. doi:10.1021/cr030101q

36. Jacobs MR (2004) Fluoroquinolones as chemotherapeutics against mycobacterial infections. Curr Pharm Des 10:3213–3220

37. Mitscher LA, Ma ZK (2003) Structure-activity relationships of quinolones In: Ronald AR, Low DE (eds) Fluoroquinolone antibiotics. Birkhaeuser, Basel, pp 11–48

38. Ledoussal B, Almstead J-IK, Gray JL, Hu EX, Roychoudhury S (2003) Discovery, structure-activity relationships and unique properties of non-fluorinated quinolones (NFQs). Curr Med Chem: Agents 2:13–25. doi:10.2174/1568012033354531

39. Peterson LR (2001) Quinolone molecular structure-activity relationships: What we have learned about improving antimicrobial activity. Clin Infect Dis 33:S180–S186

40. Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity-relationships. J Med Chem 22:1238–1244

41. Smith JT (1984) Mutational resistance to 4-quinolone antibacterial agents. Eur J Clin Microbiol 3:347–350

42. Topliss JG (1983) In: Topliss JG (ed) Quantitative structure-activity relationships of drugs. Academic Press, New York, pp 497–504

43. Zupan J (2003) Neural networks. In: Gasteiger J (ed) Handbook of cheminformatics. Wiley-VCH, Weinheim, pp 1167–1215

44. Vračko M, Mills D, Basak SC (2004) Structure-mutagenicity modelling using counter propagation neural networks. Environ Toxicol Pharmacol 16:25–36. doi:10.1016/j.etap.2003.09.004