

Investigation of 6-fluoroquinolones activity against *Mycobacterium tuberculosis* using theoretical molecular descriptors: a case study

Research Article

Nikola Minovski^{1*}, Aneta Jezierska-Mazzarello²,
Marjan Vračko¹, Tom Šolmajer¹

¹National Institute of Chemistry, POB 660, 1001 Ljubljana, Slovenia

²Faculty of Chemistry, University of Wrocław, 50-383 Wrocław, Poland

Received 3 September 2010; Accepted 23 May 2011

Abstract: A quantitative structure-activity relationship (QSAR) study on a set of 66 structurally-similar 6-fluoroquinolones was performed using a large pool of theoretical molecular descriptors. *Ab initio* geometry optimizations were carried out to reproduce the geometrical and electronic structure parameters. The resulting molecular structures were confirmed to be minima via harmonic frequency calculations. Obtained atomic charges, HOMO and LUMO energies, orbital electron densities, dipole moment, energy and many other properties served as quantum-chemical descriptors. A multiple linear regression (MLR) technique was applied to generate a linear model for predicting the biological activity, Minimal Inhibitory Concentration (MIC), treated as negative decade logarithm, ($pMIC$). The heuristic method was used to optimize the model parameters and select the most significant descriptors. The model was tested internally using the CV LOO procedure on the training set and validated against the external validation set. The result ($Q^2_{ext} = 0.7393$), which was obtained on an external, previously excluded validation data set, shows the predictive performances of this model ($R^2_{tr} = 0.7416$, $Q^2_{tr} = 0.6613$) in establishing (Q)SAR of 6-fluoroquinolones. This validated model could be proficiently used to design new 6-fluoroquinolones with possible higher activity.

Keywords: Tuberculosis • Fluoroquinolones • DNA gyrase • Molecular descriptors • QSAR

© Versita Sp. z o.o.

1. Introduction

In the last few years, the incidence of tuberculosis has dramatically increased. The global statistics from the World Health Organization (WHO) indicate that today approximately one third of the human population is infected by *Mycobacterium tuberculosis* and around 8 million people die from tuberculosis every year [1]. *M. tuberculosis*, the causative agent of tuberculosis, is a persistent pathogen microorganism. Although tuberculosis itself is a disease mainly caused by the microorganism *M. tuberculosis*, in some cases it can be caused by other *Mycobacterium* species such as *M. fortuitum*, *M. smegmatis* and *M. avium-intracellulare* complex [2-4]. Tuberculosis can be treated with chemotherapy. The most commonly used antitubercular agents in tuberculosis therapy belong to three common

classes: first-line antitubercular drugs (isoniazid, ethambutol, pyrazinamide, rifampicin, streptomycin), second-line antitubercular drugs (aminoglycosides: amikacin, kanamycin; polypeptides: capreomycin; quinolones: ciprofloxacin, levofloxacin, moxifloxacin; thioamides: ethionamide; *p*-aminosalicylic acid), and third-line antitubercular drugs (rifabutin, clarithromycin, linezolid) [5]. Nevertheless, the whole treatment is quite long taking approximately 6-9 months. The durability of the treatment as well as the toxicity and the poor patient compliance, are risk factors which frequently lead to selection of drug resistant and very often deadly multi-drug resistant strains. This increasing problem of multi-drug resistant strains is the major challenge for the investigation and design of novel drug candidates which are not only active against stable drug resistant mycobacteria, but also shorten the length of therapy [6].

* E-mail: nikola.minovski@ki.si

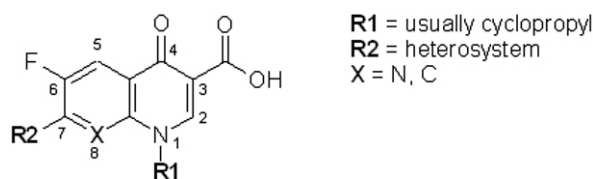
In the search for new therapeutic targets and new anti-tubercular agents, fluoroquinolones which are a family of broad-spectrum antibiotics are particularly interesting. They originate from the nalidixic acid which is the parent of the group. The majority of quinolones in clinical use belong to the subset of 6-fluoroquinolones, which have an F atom attached to the main ring scaffold, typically at the 6-position (Fig. 1).

According to structure-activity relationship (SAR) studies, the main scaffold, *i.e.*, the 1,4-dihydro-4-oxo-3-pyridinecarboxylic acid moiety, is essential for anti-mycobacterial activity [7-9]. Substitutions at position 2 of the annulated ring system greatly reduce activity (a measure of the physiological response which a drug produces) and potency (a measure of a range of effective doses in selected and not totally specified situations) [10], but positions 5, 6, 7 (especially), and position 8 of the fused ring system may be substituted with good effect. These substitutions will result in increasing the anti-mycobacterial activity and potency. The substitution with the F atom at position 6 is very important and also will result in significantly enhanced anti-mycobacterial activity. Position 1 of the main ring system, can also be substituted (small alkyl substituents such as methyl, ethyl, and especially cyclopropyl are known to enhance the potency and efficiency of 6-fluoroquinolones). These substitutions will result in increased activity and metabolic stability of the drug due to steric bulk. Ring fusions at the positions (1,8), (5,6), (6,7) and (7,8) are also very important and can significantly increase the activity [11].

One of the well established molecular targets of antitubercular agents in mycobacteria is DNA gyrase. It is a unique bacterial type II topoisomerase enzyme responsible for the catalysis of the process of introduction of negative supercoils into the double-stranded DNA molecule using the free energy that comes from the hydrolysis of ATP [12]. This bacterial enzyme consists of two major subunits, GyrA and GyrB which form the functional heterodimer A_2B_2 .

The GyrA subunit is responsible for the process of DNA breakage and reunion, where GyrB is also involved. Another closely related bacterial enzyme is type IV topoisomerase that also forms a heterodimer [13]. Both enzymes are involved in the process of controlling the topological state of DNA molecules. The gyrase is required for initiation of the process of DNA replication and elongation, while topoisomerase IV is responsible for relaxation of DNA [14,15].

Fluoroquinolones are the only inhibitors of gyrase/topoisomerase IV. They instantly inhibit the process of DNA synthesis in mycobacteria through a cleavage of the nascent mycobacterial DNA molecule in the



Fluoroquinolones

Figure 1. Generic structure of 6-fluoroquinolones.

complex formed between the DNA gyrase and type IV topoisomerase, resulting in topological perturbation and bacterial cell death [16]. These synthetic compounds belong to the class of GyrA/ParC inhibitors [17].

The present study involves structure-activity relationships and development of predictive models using the MLR method and a comprehensive set of calculated theoretical molecular descriptors which could subsequently be used for prediction of the biological activity of novel unknown 6-fluoroquinolone analogs.

The main goal of the presented study was to develop a robust QSAR model based on a set of structurally similar 6-fluoroquinolones, which can be further used to estimate the activities of novel compounds.

2. Experimental procedure

2.1. Data set

The biological assay data used in our study were obtained from classical *in vitro* tests for inhibitory activity against *M. tuberculosis*. The data we used was extracted from an online structural database [18]. The constructed dataset is named *TBCData* and consists of 66 fluoroquinolone analogs and their corresponding activity values MIC ($\mu\text{g mL}^{-1}$). There are several searching criteria for extracting the needed data. These fluoroquinolones were collected using the search criteria 'fluoroquinolones' in the NIAID (National Institute of Allergy and Infection Diseases) therapeutics database [18]. The search procedure employed, resulted in listing of all available 6-fluoroquinolone analogs (total 856 records). Detailed visual inspection of the listed compounds shows that not every structure in the fluoroquinolone's database has a measured MIC value. Although the majority of the fluoroquinolones provided are presented as electroneutral forms, there are also some compounds that exist as charged forms, salt forms, and double/triple forms. These limitations (missing MIC values and descriptors computing limitation of non-individualistic forms (charged forms, salts, double/triple forms)) resulted in construction of a final dataset of a total of 66 single molecule compounds.

The detailed investigation of *in vitro* tests used for MIC determination, showed a uniformity in performing the test (all details are available in Supporting Information in Supplemental Table 1). In other words, it is a standard procedure (agar/broth-dilution turbidimetric method) where a temperature of 37°C for incubation (growth) of the microorganisms, the 7H(9,10,11) agar for preparing the suspension/broth, as well as the chemicals used as supplements (glycerol and albumin-dextrose-catalase) are used in all tests. This *in vitro* procedure is in accordance with the CLSI international standards for susceptibility testing, *i.e.*, a standard named M24-A that describes the method in detail (media, inoculum, incubation, *etc.*) [19].

The MIC values of the collected fluoroquinolones are in the range from 0.003 to 256.0 µg mL⁻¹. According to SAR analysis, each structure in the dataset has 1,4-dihydro-4-oxo-3-pyridinecarboxylic acid moiety and F atom substitution at the position 6 in the main scaffold which are believed to be essential for anti-mycobacterial activity (Fig. 1). Some of the compounds in the *TBCData* do not correspond to the group of classical 6-fluoroquinolones because of the fusion of a third ring system to the bicyclic 1,4-dihydro-4-oxo-3-pyridinecarboxylic acid moiety. These fluoroquinolone structures have MIC values in the range (0.9-64.0 µg mL⁻¹). The chemical structures of the compounds and the related biological activity values are available in the Supporting information (Supplemental Table 2). The collection of the compounds and the activity values was performed using Microsoft Excel and the specifically integrated toolbox for ChemBioOffice Ultra 2008 (v.11.0) [20]. This method of collecting data is suitable for generating the structure database input file format (*.sdf), which is subsequently used for simultaneous calculation of the molecular descriptors in the DRAGON software package [21]. The activity values MIC (µg mL⁻¹) were converted into pMIC (-log₁₀(MIC)) values, and used in the CODESSA software package for multiple linear regression (MLR) analysis [22,23].

2.2. Geometry optimization and calculation of quantum-chemical descriptors

The molecular structures previously collected were in 2D format (ChemBioOffice 2D sketch (*.mol)). Using the MOLDEN software package [24] each structure was subsequently re-sketched in a 3D environment, checked by visual inspection in order to ensure that the 3D geometry is correct, and saved as GAUSSIAN03 input file format (*.com) for geometry optimization. The *ab initio* geometry optimization calculations for the dataset of investigated compounds were carried out using the Gaussian03 suite of programs [25]. The

Hartree-Fock-Roothaan method [26] and split-valence double-zeta (6-31G(d,p)) basis set with polarization functions on heavy atoms and hydrogens [27] were applied to reproduce the geometrical and electronic structure parameters. Resulting molecular structures were confirmed to be minima via harmonic frequency calculations. Population analysis was performed by application of Mulliken [28] and Merz-Kollman [29] schemes. The correct 3D geometry of the compounds is very important to find a possible relationship between the electronic properties of the compounds (rendered from quantum-chemical calculations) and their macroscopic properties (pMIC) [30] as well as for subsequent calculation of 2D/3D molecular descriptors using the DRAGON software package [21].

Obtained atomic charges, highest occupied and lowest unoccupied molecular orbital energies (ϵ_{HOMO} and ϵ_{LUMO}), orbital electron densities, total dipole moment, polarizability, energy ($\Delta E = \epsilon_{\text{HOMO}} - \epsilon_{\text{LUMO}}$) and many other properties served as quantum-chemical descriptors for further MLR analysis.

2.3. Calculation of 2D/3D descriptors

For calculation of the 2D/3D molecular descriptors for each compound of the training/validation sets obtained, we used the DRAGON [21] software package.

Using the MinoSuite's [31] integrated part DragCOD v2.0, an *in-house* developed software application, 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 (orbital energies (ϵ_{HOMO} and ϵ_{LUMO}), total dipole moment, polarizability, and Mulliken atomic charges) used in the multiple linear regression analysis, were extracted from each optimized structure (training/validation set) obtained by Gaussian03 (*.log files) using the MinoSuite's [31] integrated tool GaussExtractor v3.0.

The final pool of the calculated 2D/3D/QC theoretical molecular descriptors (1718), which were considered for further calculations, can be separated into nine classes: topological descriptors, electrostatic descriptors, Randic molecular profiles, geometrical descriptors, RDF descriptors (Radial Distribution Function descriptors, *i.e.*, molecular descriptors obtained by radial basis functions centered on different interatomic distances (from 0.5 to 15.5Å)), 3D-MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction derived from infrared spectra simulation using a generalized scattering function), WHIM descriptors (Weighted Holistic Invariant Molecular descriptors, based on the statistical indices calculated on the projections of atoms along principal axes. These

descriptors are obtained utilizing a Principal Components Analysis on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the atoms), GETAWAY descriptors (GEometry, TOpology, and Atom-Weights Assembly; molecular descriptors derived from the Molecular Influence Matrix (MIM)) and quantum-chemical descriptors. After the calculation, all descriptors were analyzed using the CODESSA software with the pre-integrated Heuristic algorithm, which is a suitable option for selection of most important molecular descriptors [22,23].

2.4. Development of a Quantitative Structure-Activity Relationships Model

In order to construct a QSAR model, the compounds must be represented by molecular descriptors. Initially we divided our dataset *TBCData* into a training and an external validation set using the random dividing approach. This division procedure resulted in 51 compounds in the training set, and the rest previously excluded 15 compounds in the external validation set. The descriptor selection procedure runs in several steps. First, the following selection algorithm was used for calculating the one-parameter correlation equations between descriptors and activity and eliminating all descriptors that do not fulfill the criteria below [23]:

- (1) The *F*-test's value for the one-parameter correlation with the descriptor is below 1.00.
- (2) The squared correlation coefficient of the one-parameter equation is less than $R^2_{\min}$ (in our case $R^2_{\min} = 0.1$).
- (3) The parameter's *t*-value is less than *t*₁ (where $R^2_{\min}$ and *t*₁=1.5).
- (4) The descriptor is highly inter-correlated (above r_{full} , where r_{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.

The number of retained descriptors (Supporting Information, Supplemental Table 3) is 13 for models 2D and 3D, 14 for 3D and 3D+QC, 15 for 2D+QC, 21 for 2D+3D and 2D+3D+QC. With the remaining descriptors after each inter-correlation and one-parameter calculation, all possible two- and more-parameter linear models were calculated (R^2_{tr} , Q^2_{tr}). The squared correlation coefficient of the models (R^2_{tr} , correlation between the observed and predicted activity values) was calculated employing the pre-integrated CODESSA's equation (Eq. 1) [32]:

$$R^2_{tr} = \left(\frac{N_r \sum_{i=1}^{N_r} x_{tr} y_{tr} - \left(\sum_{i=1}^{N_r} x_{tr} \right) \left(\sum_{i=1}^{N_r} y_{tr} \right)}{\sqrt{N_r \sum_{i=1}^{N_r} x_{tr}^2 - \left(\sum_{i=1}^{N_r} x_{tr} \right)^2} \sqrt{N_r \sum_{i=1}^{N_r} y_{tr}^2 - \left(\sum_{i=1}^{N_r} y_{tr} \right)^2}} \right)^2 \quad (1)$$

where N_{tr} is the total number of training set objects, x_{tr} are experimental ($pMIC_{exp}$) values, and y_{tr} are predicted ($pMIC_{pred}$) values. All the models were internally validated using the cross-validation leave-one-out procedure (CV LOO, Q^2_{tr}) using the equation (Eq. 2) [33]:

$$Q^2_{tr} = 1 - \frac{\sum_{i=1}^{N_{tr}} (y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{N_{tr}} (y_{i,exp} - \overline{y_{i,exp}})^2} \quad (2)$$

where N_{tr} is the total number of training set objects; $y_{i,exp}$ and $y_{i,pred}$ are the experimental and predicted values, respectively; $\overline{y_{i,exp}}$ is the average response value of the training set.

The assessment of the model expansion and selection of the smallest optimal number of molecular descriptors was performed employing the "breaking point" rule, *i.e.*, a simple Cartesian plot which shows how the squared correlation coefficient of the model (R^2_{tr}) is changing as a function of the number of molecular descriptors involved in the modeling procedure. Namely, if the enhancement between the models with *n* and (*n* + 1) descriptors is insignificant, then the optimal model will be the one with *n* descriptors [34-37]. Initially we constructed several linear models with up to nine descriptors ("breaking point" rule, Fig. 2), of which only the models with 2, 3, 4, and 5 descriptors were presented (improvement in model development until the one point, *i.e.*, the breaking point (5 descriptors, the optimal model)). The Heuristic selection algorithm resulted in five best descriptors, which were subsequently used for testing the predictive performances using the previously excluded validation set objects (Eq. 3) [38]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{N_{ext}} (y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{N_{ext}} (y_{i,exp} - \overline{y_{i,exp}})^2} \quad (3)$$

where N_{ext} is the total number of external validation set objects; $y_{i,exp}$ and $y_{i,pred}$ are the experimental and predicted values, respectively; $\overline{y_{i,exp}}$ is the average response value of the external validation set.

2.5. Applicability domain

According to OECD QSAR Validation Principles, a QSAR model is usable in the boundaries of its applicability domain [39]. The applicability domain of a (Q)SAR model is defined as physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds [40,41]. It should be described in terms of molecular descriptors of the model which are the most relevant parameters. The activity predictions can be made only within the domain's boundaries. Therefore, the applicability domain can be defined as a theoretical region in the space represented by the model's descriptors and the response (predicted activity values) in which a (Q)SAR model gives reliable outcome.

The applicability domain of the investigated 6-fluoroquinolones (five-descriptor MLR model, Table 1c, Supporting Information Supplemental Table 4) was calculated using the leverage approach [42]. The scatter plot defined by standardized residuals as a function of the leverage (Williams plot) was employed to assess and visualize the quality of the predictions. The leverage is defined as a compound's distance from the centroid of X . Mathematically, the leverage (h_i) of a given compound in the multidimensional descriptor space, can be calculated as (Eq. 4):

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (4)$$

where x_i is the descriptor vector of the compound under investigation, whereas the X is the descriptor matrix rendered from the descriptor values of the training set [43]. According to Eriksson *et al.* [44], the cut-off leverage value (h^*) is defined as (Eq. 5):

$$h^* = \frac{3(p+1)}{n} \quad (5)$$

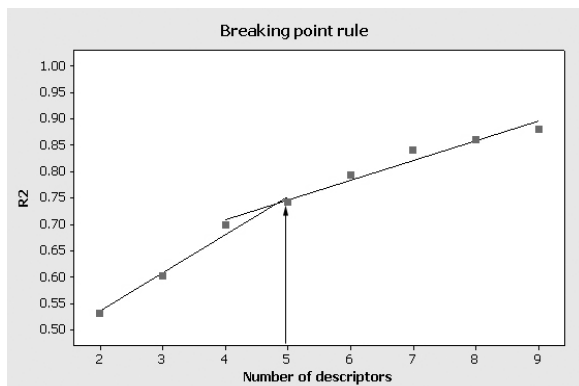


Figure 2. Breaking point rule for determination of optimal number of descriptors (R^2_{tr} of the models vs. number of descriptors).

where n is the number of compounds in the training set (51), while p is the number of descriptors used for modeling (5) [45]. In our model the cut-off value is $h^* = 0.353$ (Eq. 5). Eriksson *et al.*, proposed that the prediction for compounds with ($h_i > h^*$) can be considered as unreliable, and vice-versa. A value of 3 for standardized residuals in the Williams plot is frequently used as a limit (cut-off value) for accepting predictions (3.0 standard deviation units, $\pm 3.0\sigma$). The compounds that lie in this region cover 99% of the normally distributed data [40].

3. Results and discussion

3.1. Assessment of the calculated theoretical molecular descriptors

Quantitative structure-activity relationships (QSAR) which have been examined to rationalize the biological activity values of structurally-similar 6-fluoroquinolone analogs as well as to design novel compounds with possible enhanced activity are well documented [46-51]. In order to construct a reliable QSAR model in a statistical manner, first the compounds must be represented by molecular descriptors [52], and second, a good selection algorithm is needed for selection of the most important molecular descriptors from the large multi-descriptor space.

The main goal of our study is to approximate the mechanism of anti-mycobacterial activity in this series of known fluoroquinolones using a comprehensive set of theoretical molecular descriptors (2D/3D/QC) as well as to build a robust and statistically significant model which could subsequently be used to design new potential drug candidates, with possible better activity.

One of the most important steps in QSAR modeling is to define the number of independent variables in the model equation obtained. In this way the over-parameterization of the mathematical model as well as the chance correlation between the molecular descriptors is avoided [53].

Since we start our modeling procedure with a pool of 1718 theoretical molecular descriptors, a possibility exists to encounter a chance correlation in a case where the number of examined variables is higher than the number of observations.

Initially we investigated seven different cases employing the classical MLR method and all possible combinations of the calculated theoretical molecular descriptors (2D, 3D, QC, 2D+3D, 2D+QC, 3D+QC, and 2D+3D+QC) and selected the model which not only has the highest coefficient of correlation (R^2_{tr}), but also good and acceptable predictive performances.

Table 1. Linear models obtained by CODESSA (ID, Case number; Tr. Str., Number of structures in the training set; D. No., Starting pool of molecular descriptors; D.Re. number of descriptors retained from the starting pool; Params., Parameters used in heuristic option; R^2_{tr} , squared coefficient of correlation for the training set; Q^2_{tr} , the predictive squared correlation coefficient). Statistically significant molecular descriptors are highlighted.

a) MLR models obtained by employing only 3D molecular descriptors (Supporting Information Supplemental Table 3, case 2).

ID	Dataset	Tr.Str.	D.No./D.Re.	Params.	R^2_{tr}	F	s^2	Q^2_{tr}	Descriptors involved				
2	TBCData	51	682/14	2	0.5317	27.24	0.8042	0.4807	Mor21p	Mor09u			
				3	0.6014	23.63	0.6991	0.5424	Mor21m	Mor09u	HOMT		
				4	0.6982	26.62	0.5406	0.6290	Mor21m	Mor09u	HOMT	Mor16e	
				5	0.7416	25.83	0.4733	0.6613	Mor21p	Mor09u	HOMT	Mor16e	R2p+

b) MLR models obtained by employing only QC molecular descriptors (Supporting Information Supplemental Table 3, case 3).

ID	Dataset	Tr.Str.	D.No./D.Re.	Params.	R^2_{tr}	F	s^2	Q^2_{tr}	Descriptors involved				
3	TBCData	51	494/14	2	0.2394	7.56	1.3060	0.1133	HOMO13	HOMO5			
				3	0.2574	5.43	1.3021	0.1057	HOMO13	HOMO5	TDM		
				4	0.2574	5.43	1.3021	0.1057	HOMO13	HOMO5	TDM	/	
				5	0.2574	5.43	1.3021	0.1057	HOMO13	HOMO5	TDM	/	/

c) MLR models obtained by employing a combination of 3D+QC molecular descriptors (Supporting Information Supplemental Table 3, case 6).

ID	Dataset	Tr.Str.	D.No./D.Re.	Params.	R^2_{tr}	F	s^2	Q^2_{tr}	Descriptors involved				
6	TBCData	51	1176/14	2	0.5317	27.24	0.8042	0.4807	Mor21p	Mor09u			
				3	0.6014	23.63	0.6991	0.5424	Mor21m	Mor09u	HOMT		
				4	0.6982	26.62	0.5406	0.6290	Mor21m	Mor09u	HOMT	Mor16e	
				5	0.7416	25.83	0.4733	0.6613	Mor21p	Mor09u	HOMT	Mor16e	R2p+

These results are presented in Supporting Information in Supplemental Table 3 (models) and Supplemental Table 4 (predictions). Supplemental Table 3 shows seven different cases (all possible combinations of molecular descriptors) where we examined four models (with 2, 3, 4, and 5 descriptors) for each case separately (total 28 linear models).

The analysis of 5-descriptor models for each case, shows that six out of seven models are good (case 1, case 2, case 4, case 5, case 6, and case 7) according to the R^2_{tr} values ($R^2_{tr} > 0.7$). According to the R^2_{tr} values, one would say that 5-descriptor models built with 2D (case 1, $R^2_{tr} = 0.7894$), 2D+3D (case 4, $R^2_{tr} = 0.7623$), 2D+QC (case 5, $R^2_{tr} = 0.7942$), and 2D+3D+QC (case 7, $R^2_{tr} = 0.7623$) descriptors are better than the model built only with 3D (case 2, $R^2_{tr} = 0.7417$) or 3D+QC (case 6, $R^2_{tr} = 0.7417$) descriptors (highlighted in green). The linear plots ($pMIC_{experimental}$ vs. $pMIC_{predicted}$ for training/validation set) represented in the Supporting Information in (Supplemental Fig. 1) show better fitting for case 2 (only 3D descriptors) and case 6 (3D+QC descriptors) in comparison with the other cases. The R^2_{tr} values of these two models as well as the molecular descriptors selected by the Heuristic algorithm are the same.

This result apparently shows the domination, *i.e.*, significance, of the 3D descriptors over QC descriptors (Tables 1abc) and therefore selection of the best 5-descriptor linear model (Table 1c). The MLR analysis between QC descriptors alone (494) and the inhibitory activity (Supporting Table 3, case 3) shows that the descriptor combination of ϵ_{HOMO} and Total Dipole Moment (TDM) was the most important for elucidating the activity (50.74% of the variation). This observation clearly indicates that quantum-chemical descriptors alone were not sufficient for explanation of $pMIC$ variation. Therefore the additional pool of 682 3D DRAGON descriptors was used in order to enhance the linear model as well as to explain the structural diversity of the fluoroquinolones used.

3.2. Model development

As presented in Table 1c, the number of examinations (column 3, number of examinations is 51 (training set)) is enough for screening the number of retained descriptors (column 4: number of descriptors 14) in order to keep the probability of encountering a chance correlation with $R^2 > 0.8$ at the 1% level or less [53]. The modeling procedure is based on the breaking point rule

(the slope change). This rule shows the enhancement threshold of correlation coefficient (R^2_{tr}) over the number of descriptors in each model.

Initially we developed preliminary linear models with up to 9 descriptors and found that the threshold, i.e., the maximal enhancement of R^2_{tr} was when the number of descriptors is equal to five (optimal number of descriptors, the breaking point, Fig. 2) [34–37].

The heuristic algorithm implemented in modeling, results in a five-descriptor model which can be described with the following QSAR equation (Eq. 6):

$$\begin{aligned} \text{pMIC} = & -0.94 - 3.575\text{Mor21p} - 1.118\text{Mor09u} - \\ & -1.3883\text{Mor16e} + 0.263\text{HOMT} - 52.91\text{R2p}^+ \\ N_{tr} = & 51; R^2_{tr} = 0.7416; Q^2_{tr} = 0.6613; \\ s = & 0.4733; F = 25.83, \end{aligned} \quad (6)$$

where N_{tr} is the number of examinations (compounds included into the training set used for modeling), R^2_{tr} is the squared coefficient of correlation, Q^2_{tr} is the predictive squared correlation coefficient, s is the standard deviation, and F is the Fisher value. Following the frequency analysis of the descriptors involved in our models (Table 1c), the most important molecular descriptors for establishing a quantitative relationship between the molecular structure and activity can be visually determined.

According to this postulate as well as the five-descriptor model equation described above, the five most frequent and relevant molecular descriptors for activity are: Mor21p, Mor09u, Mor16e, HOMT and R2p⁺. The parameters Mor21p, Mor09u, Mor16e, and R2p⁺ belong to the class of electrostatic and electro-topological descriptors (3D-MoRSE-signal 21/weighted by atomic polarizabilities, 3D-MoRSE-signal 09/unweighted, 3D-MoRSE-signal 16/weighted by atomic Sanderson electronegativities, R maximal autocorrelation of lag 2/weighted by atomic polarizabilities (3D-GETAWAY descriptor), respectively) whereas the HOMT parameter (HOMA-total; Harmonic Oscillator Model of Aromaticity Index) belongs to the class of geometrical descriptors.

3.3. Mechanistic interpretation

The electrostatic descriptors listed above are of significant importance for activity and indicate that *in vitro/in vivo* anti-mycobacterial activity against *M. tuberculosis* is strongly dependent on the electrostatic and aromaticity properties of the properly substituted main 6-fluoroquinolone scaffold.

These parameters also corroborate with findings in the literature that position 6 of the F atom in the main quinolone core which is present in all inhibitors of this study is of significant importance for accommodation of

the inhibitor into the active binding site and suggests that a possible establishment of an electrostatic interaction between the F atom and the target (possible intermolecular electrostatic interactions with the amino acid residues of the GyrA subunit active site) [54–57] may result in increased stability of the fluoroquinolone binding to the complex [58]. The importance of the Mor09u and R2p⁺ parameters (pure electrostatic descriptors) for anti-mycobacterial activity also suggests that possible intermolecular electrostatic interactions between the carbonyl and carboxyl groups of the main quinolone core and the corresponding amino acid residues of the active site within the GyrA subunit are responsible for tight binding. The recently published crystal structure of the complex levofloxacin with GyrA fully substantiates this notion [56]. It has to be clearly pointed out that the molecular descriptors selected by the Heuristic procedure are an interpolation of various structural features introduced in the 6-fluoroquinolone scaffold by substituents with implicit electrostatic properties. For example, *in vitro/in vivo* anti-mycobacterial activity against *M. tuberculosis* appears to be dependent also on the electronegativity of the O atom (sp^2) of the carboxyl and carbonyl group of the main core substituents (Fig. 1).

On the other hand, HOMT is also an important molecular descriptor indicating the importance of aromaticity of the main 6-fluoroquinolone scaffold (the annulated pyridone system) for activity as well as the optimal basicity for better intestinal permeability after possible oral administration. At the molecular level, the correct 3D geometry (molecular shape) of the ligand is one of the important factors for good accommodation into the binding pocket as well as for establishing good interactions with the surrounding amino acid residues and forming a more stable complex. The aromaticity of the main scaffold is also of significant importance for establishing π - π stacking interactions between the main 6-fluoroquinolone scaffold and the planar aromatic systems of the bacterial DNA [56]. This relationship is also suggested in our QSAR equation through the importance of the HOMT parameter that belongs to the group of 3D geometrical descriptors.

According to the values obtained for the R^2_{tr} and Q^2_{tr} , one can observe the good predictive performances of our five-descriptor linear model. As mentioned previously, the predictive power of our five-descriptor QSAR model [59] was observed using an external, previously excluded validation data set (Fig. 3). The results of the validation set predictions ($Q^2_{ext} = 0.7393$) that was obtained on the basis of the best (five-descriptor) training set model as well as the numerical values for each of the selected descriptors are shown in Table 2.

Table 2. The experimental vs. predicted activity values (pMIC) and numerical values of the descriptors (3D+QC) involved in the modeling procedure.

ID	pMIC _{exp}	pMIC _{pred}	Mor21p	Mor09u	Mor16e	HOMT	R2p ⁺
Training set							
2	0.5229	0.8875	-0.477	0.038	-0.461	4.021	0.029
3	-0.5441	-0.1400	-0.564	0.069	-0.408	-0.250	0.031
4	1.0000	0.0902	-0.517	-0.215	-0.339	-0.182	0.028
5	-0.1461	-0.0990	-0.505	0.084	-0.407	-0.229	0.026
6	-0.5052	-0.1297	-0.624	0.145	-0.284	-0.250	0.030
7	1.3010	1.3320	-0.600	-0.570	-0.819	-0.229	0.030
8	0.3010	0.4930	-0.716	-0.053	-0.624	-0.361	0.037
9	-0.2041	-0.2781	-0.428	-0.005	-0.284	0.206	0.025
11	-0.3010	-1.8846	-0.187	-0.120	0.559	3.343	0.035
12	0.3010	0.6400	-0.581	-0.586	-0.641	-0.322	0.037
13	0.6021	-0.2269	-0.330	0.591	0.029	5.711	0.024
14	2.0000	1.7479	-0.854	-0.707	-0.045	-0.206	0.022
15	1.5086	0.3169	-0.650	-0.263	-0.449	-0.504	0.035
16	2.0000	1.5034	-0.444	-0.825	-0.040	3.958	0.022
18	2.3979	2.4936	-0.971	-0.856	-0.199	-0.206	0.023
19	-0.3010	-0.4755	-0.323	-0.612	0.256	3.564	0.037
20	-1.0969	-0.5191	-0.506	1.543	-0.398	4.604	0.027
21	0.1079	0.2678	-0.573	0.915	-0.260	4.544	0.026
23	0.9031	1.0072	-0.450	-0.621	0.260	4.038	0.020
24	2.5229	1.8796	-0.767	-1.032	-0.137	-0.389	0.022
25	0.6021	0.3867	-0.846	-0.374	0.019	-0.100	0.039
26	1.7959	2.0591	-0.664	-0.991	0.242	3.261	0.019
27	0.2218	-0.3768	-0.708	0.442	0.147	-0.202	0.023
28	0.4089	-0.8549	-0.302	0.039	-0.365	-0.112	0.027
30	1.5229	1.7064	-0.731	-0.532	-0.465	0.036	0.023
31	-0.5052	0.1537	-0.314	-1.024	-0.170	4.090	0.047
32	-2.3010	-0.9789	-0.367	-0.395	0.356	-0.310	0.023
33	-2.1004	-2.7446	-0.616	2.257	0.104	0.336	0.027
34	0.1079	0.4046	-0.439	0.043	-0.664	-0.155	0.020
36	0.9031	0.6620	-0.925	0.556	-0.051	0.036	0.022
37	0.8069	0.4884	-0.539	-0.794	-0.507	-0.504	0.037
38	0.4089	0.1632	-0.268	-0.982	0.442	4.740	0.030
39	-1.2041	0.1212	-0.366	-0.299	-0.131	3.530	0.032
41	2.2218	1.4623	-0.641	-0.787	-0.342	-0.304	0.022
43	2.2218	1.5336	-0.627	-0.780	-0.329	-0.147	0.020
44	2.5229	2.1496	-0.974	-0.722	-0.011	-0.194	0.022

Continued Table 2. The experimental vs. predicted activity values (pMIC) and numerical values of the descriptors (3D+QC) involved in the modeling procedure.

ID	pMIC _{exp}	pMIC _{pred}	Mor21p	Mor09u	Mor16e	HOMT	R2p ⁺
Training set							
46	1.8861	1.4446	-0.758	-0.428	-0.258	-0.194	0.021
47	2.2218	1.7312	-0.726	-0.596	-0.419	-0.233	0.021
50	1.2219	0.6336	-0.853	0.690	-0.368	-0.197	0.022
51	0.6990	1.4933	-0.759	0.204	-0.838	-0.197	0.022
53	0.1079	0.7245	-0.823	0.191	-0.109	-0.197	0.022
54	-0.1931	0.8958	-0.767	0.143	-0.376	-0.197	0.023
55	0.4089	0.8823	-0.842	0.169	-0.156	-0.197	0.022
56	1.0000	1.0215	-0.838	-0.527	-0.005	0.036	0.031
58	1.0000	1.3341	-0.801	-0.446	-0.483	-0.250	0.032
60	2.5229	2.4969	-0.971	-0.859	-0.199	-0.206	0.023
61	1.6021	1.7003	-0.731	-0.529	-0.463	0.036	0.023
62	0.4089	0.4705	-0.689	0.895	-0.876	-0.197	0.023
63	-1.8062	-0.8973	-0.413	-0.220	0.344	0.055	0.023
64	-1.8062	-0.5435	-0.519	-0.318	0.441	0.055	0.023
66	-2.1072	-1.4587	-0.384	-0.502	0.458	-0.471	0.032
External validation set							
1	-0.4771	-0.4820	-0.604	0.086	0.042	-0.250	0.028
10	0.3010	0.1498	-0.719	0.653	-0.283	0.076	0.022
17	0.0458	0.6802	-0.626	-0.210	-0.420	-0.229	0.026
22	0.1079	0.1644	-0.377	0.366	-0.135	4.538	0.023
29	0.3010	-0.1131	-0.531	-0.310	-0.423	-0.584	0.035
35	-1.8062	-0.3000	-0.432	-0.124	-0.213	0.139	0.026
40	2.2218	2.7892	-0.949	-0.973	-0.382	-0.246	0.023
42	1.8861	2.0438	-0.802	-0.619	-0.385	-0.197	0.020
45	2.5229	2.4392	-0.914	-0.485	-0.545	-0.290	0.021
48	1.6021	0.9918	-0.852	0.497	-0.397	-0.197	0.020
49	1.5229	0.4146	-0.877	1.287	-0.591	-0.197	0.021
52	0.4089	1.3753	-0.741	0.016	-0.648	-0.197	0.022
57	1.0000	1.1112	-0.768	-0.687	-0.137	-0.048	0.031
59	0.6990	0.0733	-0.514	0.776	-0.953	-0.040	0.024
65	-2.4082	-1.5143	-0.302	-0.341	0.115	-0.493	0.033

The applicability domain (AD) of the five-descriptor linear model previously selected (Williams plot) was assessed utilizing the well known leverage approach (Fig. 4). Training set objects (51 compounds with experimental activity values) used in the model development are presented as solid dots, whereas the

external validation set objects (15 compounds) as solid rectangles labeled with the corresponding number (ID signature).

The analysis of AD for the training set objects shows that only one compound labeled with (ID33) signature can be identified as a typical X-outlier ($h > h^* = 0.353$).

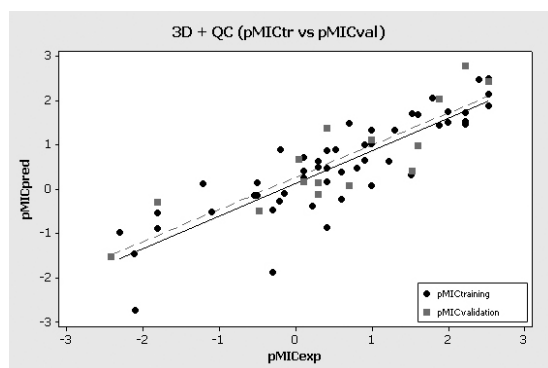


Figure 3. The experimental vs. predicted pMIC values using MLR method.

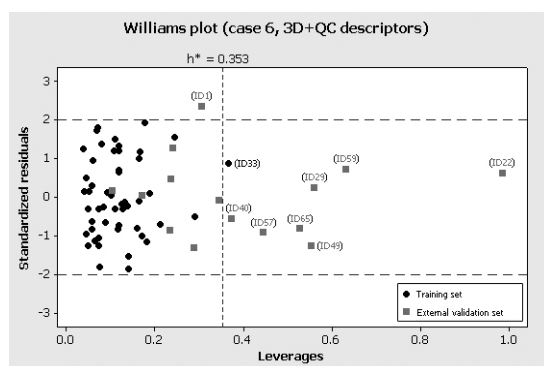


Figure 4. Graphical representation (Williams plot) of the five-descriptor MLR model's applicability domain (AD) together with external validation set objects.

Surprisingly, the pMIC prediction for this compound was quite good as presented in Table 2. On the other hand, seven compounds from the external validation set (ID22, ID29, ID40, ID49, ID57, ID59, and ID65) are typical X-outliers with $h > h^*$. Only one compound from the external validation set with signature (ID1) can be identified as a typical Y-outlier if the cut-off value for standard deviation is $\pm 2.0\sigma$. The predicted activity values for the compounds signed as (ID1, ID22, ID40, and ID57) are good, whereas the rest of the compounds identified as outliers (ID29, ID49, and ID65) are poorly predicted. The structural investigation of these compounds shows that compounds (ID29 and ID65) have an *o*, *p*-difluoroaromatic moiety attached into position 1 of the main 6-fluoroquinolone scaffold, whereas the compound (ID49) has a chirally-modified 2-fluorocyclopropyl moiety. According to the accepted SAR rules for 6-fluoroquinolone antibiotics, the substitutions in position 1 of the main quinolone core with smaller alkyl substituents (methyl, ethyl, and especially cyclopropyl) will greatly enhance the anti-mycobacterial activity [11]. Since the compounds (ID29) and (ID65) have attached a significantly larger group than the substituents proposed, we suggest that these structural changes induce lower activity. The MIC value

of the compound (ID65) is $256 \mu\text{g mL}^{-1}$. In addition, these two compounds are structural analogs of tosufloxacin (photocytotoxic and photohaemolytic agent) in which the presence of the *o*, *p*-dihalogenated aromatic ring system at position 1 increases the probability of emerging possible photocytotoxic/photohaemolytic adverse effects [60]. Compound (ID49), a chiral analog of ciprofloxacin, has a predicted activity value that is too low. Apparently, the five-descriptor (3D+QC) model cannot adjust the prediction of chiral/achiral substituted fluoroquinolones at position 1.

4. Conclusion

A quantitative structure-activity relationships (QSAR) study on a set of 66 structurally-similar 6-fluoroquinolone analogs was performed using a comprehensive set of theoretical molecular descriptors. The MLR method was employed for the construction of a robust model for prediction of the inhibitory activity (pMIC) against *M. tuberculosis*. The robustness and the predictive ability of the model were verified using a method for internal validation (cross-validation leave-one-out). The predictive power of the model was tested through the extrapolation of the model over the external previously excluded validation data set [59]. The result obtained in this study ($Q^2_{ext} = 0.7393$) suggests that QSAR models utilizing calculated theoretical molecular descriptors can be successfully used for the design of novel 6-fluoroquinolone analogs with possible higher anti-mycobacterial activity. Furthermore, the selected most frequent theoretical molecular descriptors describe some of the crucial inter-molecular interactions between the 6-fluoroquinolones and the GyrA subunit and DNA [56]: π - π stacking interaction of the main 6-fluoroquinolone scaffold and the planar aromatic systems of the bacterial DNA, as well as possible electrostatic interactions with the amino acid residues of the GyrA subunit active site.

Acknowledgments

Authors thank Agency of Research of R. Slovenia (ARRS) for the financial support through the Grants P1-0017 and 1000-07-310016. We are sincerely grateful to Dr. Marjana Novic for valuable insights, discussion and her continuing support of this research.

Anetka Jezierska-Mazzarello gratefully acknowledges the Wrocław Center for Networking and Supercomputing (WCSS) and Poznań Supercomputing and Networking Center for providing computer time and facilities.

Supporting Information Available

In vitro test details of the 6-fluoroquinolones collected (Supplemental Table 1).

Table of chemical structures of the compounds and the related biological activity values (Supplemental Table 2).

MLR models obtained by employing all possible combinations of theoretical molecular descriptors (Supplemental Table 3).

Prediction results for training/external validation sets for all models and numerical values of the molecular descriptors involved in the modeling procedure (Supplemental Table 4).

Graphical representation of each case 1-7, *i.e.*, five-descriptor models (the experimental vs. predicted *p*MIC values) using MLR method (Supplemental Fig. 1).

References

- [1] C. Dye, S. Scheele, P. Dolin, V. Pathania, M.C. Raviglione, J. Am. Med. Assoc. 282, 677 (1999)
- [2] N.V. Bhanu, D. van Soolingen, J.D.A. van Embden, P. Seth, Diagn. Microbiol. Infect. Dis. 48, 107 (2004)
- [3] O. Dussurget, M. Rodriguez, I. Smith, Tubercle. Lung. Dis. 79(2), 99 (1998)
- [4] S.K. Field, D. Fisher, R.L. Cowie, Chest 126, 566 (2004)
- [5] J.M. Beale, In: J.H. Block, J.M. Beale (Eds.), Whilson and Gisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry, 11th edition (Lippincott Williams & Wilkins, Baltimore, MD, 2004) 266
- [6] Y. Zhang, K.P. -Martens, S. Denkin, Drug Discov. Today 11, 21 (2006)
- [7] G. Anquetin, J. Greiner, P. Vierling, Curr. Drug Targets: Infect. Disord. 5, 227 (2005)
- [8] M.J. Suto, J.M. Domagala, G.E. Roland, G.B. Mailloux, M.A. Cohen, J. Med. Chem. 35, 4745 (1992)
- [9] L.R. Peterson, Clin. Infect. Dis. 33 (Suppl. 3), S180 (2001)
- [10] A.A. Miles, Bull. World Hlth. Org. 6, 131 (1952)
- [11] J.M. Beale, In: J.H. Block, J.M. Beale (Eds.), Whilson and Gisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry, 11th edition (Lippincott Williams & Wilkins, Baltimore, MD, 2004) 247
- [12] J.J. Champoux, Annu. Rev. Biochem. 70, 369 (2001)
- [13] H. Peng, K.J. Mariani, J. Biol. Chem. 268, 24481 (1993)
- [14] R.J. Reece, A. Maxwell, Crit. Rev. Biochem. Mol. 26, 335 (1991)
- [15] C. Levine, H. Hiasa, K.J. Mariani, Biochim. Biophys. Acta 1400, 29 (1998)
- [16] D.C. Hooper, Drugs 58 (suppl. 2), 6 (1999)
- [17] M. Oblak, M. Kotnik, T. Šolmajer, Curr. Med. Chem. 14, 2033 (2007)
- [18] N. Minovski, M. Vračko, T. Šolmajer, Mol. Divers. 15(2), 417 (2011)
- [19] Clinical Laboratory Standards Institute/NCCLS. Susceptibility Testing of Mycobacteria, Nocardiae, and Other Aerobic Actinomycetes; Approved Standard. CLSI/NCCLS document M24-A (Clinical Laboratory Standards Institute, Wayne PA, 2003)
- [20] ChemBioOffice Ultra v11.0, CambridgeSoft (Cambridge, England, 2008) <http://www.cambridgesoft.com/software/ChemBioOffice>
- [21] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON Version 5.4 (Talete srl, Milan, Italy) <http://www.talete.mi.it/dragon.htm>
- [22] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA, Training Manual (University of Florida, Gainesville, 1994)
- [23] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA, Reference Manual (University of Florida, Gainesville, 1994)
- [24] G. Schaftenaar, J.H. Noordik, J. Comput.-Aided Mol. Des. 14, 123 (2000)
- [25] M.J. Frisch et al., Gaussian 03, Revision A.02 (Gaussian, Inc., Wallingford CT, 2004)
- [26] C.C. J. Roothaan, Rev. Mod. Phys. 23, 69 (1951)
- [27] R. Ditchfield, W.J. Hehre, J.A. Pople, J. Chem. Phys. 54, 724 (1971)
- [28] R.S. Mulliken, J. Chem. Phys. 23, 1833 (1955)
- [29] B.H. Besler BH, K.M. Jr. Merz, P.A. Kollman, J.

Abbreviations

ATP, Adenosine triphosphate; MIC, Minimal Inhibitory Concentration; SAR, Structure-Activity Relationship; QSAR, Quantitative Structure-Activity Relationship; QC, Quantum-Chemical; CODESSA, COmprehensive DEscriptors for Structural and Statistical Analysis; MLR, Multiple Linear Regression; CV LOO, Cross-Validation Leave-One-Out; AD, Applicability Domain.

- Comp. Chem. 11, 431 (1990)
- [30] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, Chem. Phys. Lett. 23, 59 (2000)
- [31] N. Minovski, MinoSuite v2.0, Laboratory for Chemometrics (National Institute of Chemistry, Ljubljana, Slovenia, 2009)
- [32] J. Rebehmed, F. Barbault, C. Teixeira, F. Maurel, J. Comput. Aided. Mol. Des. 22, 831 (2008)
- [33] V. Consonni, D. Ballabio, R. Todeschini, J. Chem. Inf. Model. 49, 1669 (2009)
- [34] A.R. Katritzky, O.V. Kulshyn, I. Stoyanova-Slavova, D.A. Dobchev, M. Kuanar, D.C. Fara, M. Karelson, Bioorg. Med. Chem. 14, 2333 (2006)
- [35] A.R. Katritzky, L.M. Pacureanu, D.A. Dobchev, D.C. Fara, P.R. Duchowicz, M. Karelson, Bioorg. Med. Chem. 14, 4987 (2006)
- [36] A.R. Katritzky, D.A. Dobchev, I. Tulp, M. Karelson, D.A. Carlson, Bioorg. Med. Chem. Lett. 16, 2306 (2006)
- [37] A.R. Katritzky, D.A. Dobchev, E. Hur, D. Fara, M. Karelson, Bioorg. Med. Chem. 13, 1623 (2005)
- [38] G. Schüürmann, E. Ralf-Uwe, J. Chen, B. Wang, R. Kühne, J. Chem. Inf. Model. 48 (11), 2140 (2008)
- [39] Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models (Head of Publication Service, OECD, Paris, France, 2007) <http://www.oecd.org/dataoecd/55/35/38130292.pdf>
- [40] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, Altern. Lab. Anim. 33, 445 (2005)
- [41] J. Jaworska, M. Comber, C.V. Leewen, C. Auer, Environ. Health Perspect. 11, 1358 (2003)
- [42] H. Liu, E. Papa, P. Gramatica, Chem. Res. Toxicol. 19, 1540 (2006)
- [43] T.I. Netzeva et al., Altern. Lab. Anim. 33, 155 (2005)
- [44] L. Eriksson et al., Environ. Health Perspect. 11, 1361 (2003)
- [45] E. Papa, S. Kovarich, P. Gramatica, QSAR Comb. Sci. 28(8), 790 (2009)
- [46] L.A. Mitscher, Chem. Rev. 105, 559 (2005)
- [47] L.A. Mitscher, Z. K. Ma, In: A.R. Ronald, D.E. Low (Eds.), Fluoroquinolone Antibiotics (Birkhaeuser, Basel, 2003) 11
- [48] M.R. Jacobs, Curr. Pharm. Des. 10, 3213 (2004)
- [49] L.R. Peterson, Clin. Infect. Dis. 33, S180 (2001)
- [50] B. Ledoussal et. al, Curr. Med. Chem.: Agents 2, 13 (2003)
- [51] M.C. Bagchi, D. Mills, S.C. Basak, J. Mol. Model. 13, 111 (2007)
- [52] M. Randić, J. Chem. Inf. Comput. Sci. 41, 639 (2001)
- [53] J.G. Topliss, R.P. Edwards, J. Med. Chem. 22, 1238 (1979)
- [54] J.G. Heddle et al., Nucleosides, Nucleotides Nucleic Acids 19, 1249 (2000)
- [55] I. Laponogov et al., Nat. Struct. Mol. Biol. 16, 6, 667 (2009)
- [56] I. Laponogov et al., Plos one 5, 6, e11338 (2010)
- [57] J. Piton et al., Plos one 5, 8, e12245 (2010)
- [58] J.T. Smith, Eur. J. Clin. Microbiol. 3, 347 (1984)
- [59] P. Gramatica, QSAR Comb. Sci. 26, 694 (2007)
- [60] T. Yamamoto et al., Toxicol. In Vitro 15, 721 (2001)