Chapter 8
Integrated *in Silico* Methods for the Design and Optimization of Novel Drug Candidates: A Case Study on Fluoroquinolones – *Mycobacterium tuberculosis* DNA Gyrase Inhibitors

Nikola Minovski
National Institute of Chemistry, Slovenia

Marjana Novič
National Institute of Chemistry, Slovenia

ABSTRACT

Although almost fully automated, the discovery of novel, effective, and safe drugs is still a long-term and highly expensive process. Consequently, the need for fleet, rational, and cost-efficient development of novel drugs is crucial, and nowadays the advanced in silico drug design methodologies seem to effectively meet these issues. The aim of this chapter is to provide a comprehensive overview of some of the current trends and advances in the in silico design of novel drug candidates with a special emphasis on 6-fluoroquinolone (6-FQ) antibacterials as potential novel Mycobacterium tuberculosis DNA gyrase inhibitors. In particular, the chapter covers some of the recent aspects of a wide range of in silico drug discovery approaches including multidimensional machine-learning methods, ligand-based and structure-based methodologies, as well as their proficient combination and integration into an intelligent virtual screening protocol for design and optimization of novel 6-FQ analogs.

DOI: 10.4018/978-1-4666-8136-1.ch008

INTRODUCTION

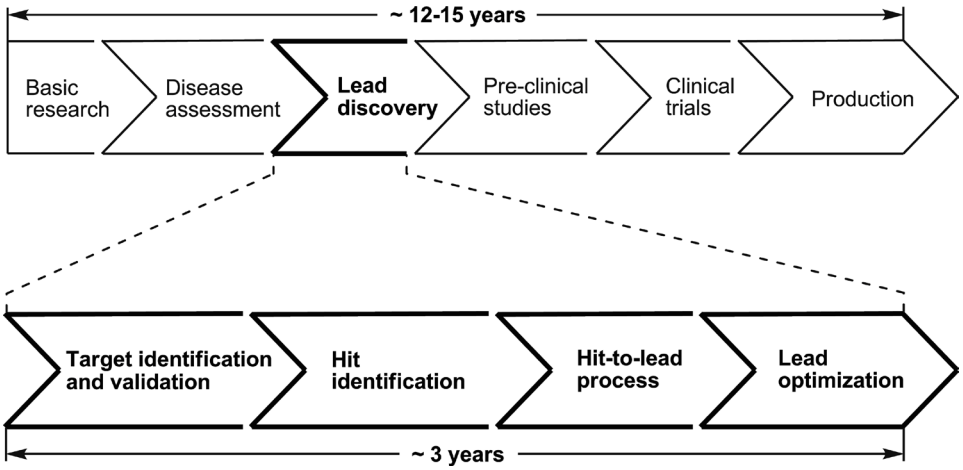
The discovery process of novel, effective and safe drugs, from day to day is becoming more advanced and sophisticated. Today, the profit- and innovation-based competition between the pharmaceutical companies is increasingly growing (Cavazzani, 2010). In addition to the profit, the process of discovery a new drug is not only long-term, but also highly expensive. It was estimated that a new drug discovery program is approximately 12-15 years long and takes around 200-300 millions to one billion dollars (Rawlins, 2004; Bartfai & Lees, 2006; Hughes et al., 2011). To alleviate this growing problem, nowadays the drug research efforts are mainly directed toward reducing the discovery costs as well as the time required. In that regard, the *in silico* drug discovery methods were found as particularly important to effectively meet these issues.

As depicted in Figure 1, the lead discovery part (which is constructed of several interconnected sub-phases) of the drug discovery pipeline can be considered as the essential one (Kenny et al., 1998; Langer & Hoffmann, 2001). Owing to the rapid development of various computational methods that today could readily and efficiently be applied in different lead discovery segments, one can observe a significant progress in reducing its durability to amazing 2-3 years (Figure 1). However, within the framework of the entire drug discovery process it is still a long period of time and therefore a further time reduction is certainly welcomed.

In the seventies of the previous century, the discovery of novel lead compounds was substantially based on a random screening of large chemical libraries comprised of chemicals of different origin. This approach has been initially used for discovery of new antibiotics. The drug discovery practice demonstrated that on average only one potential lead from a library containing around 20.000 molecules could be identified using the random screening approach (Young et al., 1996). Since the 1980s, with the growth and development of robotics and miniaturization of the *in vitro* testing methods, it became possible to screen hundreds of thousands of compounds on a large number of biological targets (Gribbon & Sewing, 2005), i.e., a methodology widely known as high-throughput screening (HTS). Nevertheless, such a philosophy postulated under the idea, the greater is the starting chemical library, the higher are the chances to identify a biologically-active molecule, was disappointing for many pharmaceutical companies (O'Driscoll, 2004). These failures as well as the daily progresses in molecular and structural biology were the major driving force to elevate the discovery of novel drugs to a significantly higher, knowledge-based level commonly known as the rational drug design (Mavromoustakos et al., 2011). Moreover, the advances in the computation and strategies such as 2D/3D computer-aided molecular design (CAMD) opened a new perspective into the drug discovery world. Nowadays, the *in silico* screening methods are indeed an efficient supplement to the experimentally grounded HTS methods, becoming an integral segment of the hit identification and lead generation processes (Klebe, 2006; Stahura & Bajorath, 2004; Bajorath, 2002; Shoichet, 2004; Bleicher et al., 2003).

The present text aim to report some of the recent trends and advances in the *in silico* design of novel drug candidates – from chemical sketches to predicted active conformations. Specifically organized in two, tutorial-like parts (theoretical and practical), this chapter is not strictly intended to expose the thorough theoretical and mathematical background behind the *in silico* methodologies employed, but rather to guide the reader through the different steps of the *in silico* design and prediction of novel biologically-active hits. Put differently, the theoretical part gives an overview of various attentively selected computational methodologies proficiently integrated into an efficient *in silico* screening framework including quantitative structure-activity relationship (QSAR) methods, virtual combinatorial library

Figure 1. Schematic representation of a classical industrial drug discovery pipeline



design, property-based screening and construction of focused "drug-like" libraries as well as ligand- and structure-based methodologies coupled with virtual screening (VS) for identification and selection of novel hit compounds. On the other hand, the second part of this chapter is primarily focused on the practical implementation of these methods for design of 6-fluoroquinolone (6-FQ) antibacterials as potential novel *Mycobacterium tuberculosis* DNA gyrase inhibitors. It covers all the aspects described in the theoretical part, but proficiently assembled and integrated into an intelligent VS protocol that could aid not only the development of new drugs in general, but also to shorten the time required.

Background

During the last five decades, a wide variety of *in silico* drug design approaches have been invented that undoubtedly altered the drug discovery paradigm – a rapid and cost-efficient development of potent and safe drugs. Nowadays, we have witnessed the effectiveness and benefits of these methods as some of them have not only facilitated the drug discovery in its entirety, but also became a golden standard to success (Ekins et al., 2007; Bharath et al., 2011; Cumming et al., 2013).

From the plethora of computational methods available today, just a few can be regarded as of exceptional importance for the purposes of modern drug discovery. From simple numerical predictions of biological activity values for compounds, which have not yet been synthesized, to possible intuitively based revelation of compound-protein interactions, these methods could be recognized as pivotal indivisible approaches within the framework of the entire drug discovery process. This list of *in silico* drug design approaches includes machine-learning (statistical) methods such as the QSAR, the methodology for virtual combinatorial library design, the property-based filtering approaches as well as the ligand- and structure-based methodologies (Ekins et al., 2007). Notwithstanding their confirmed high potential in various lead discovery segments, one cannot expect a successful identification of novel drug candidates when these methods are used separately. Consequently, their proficient assemblage into powerful *in silico* screening protocols for fast and accurate predictions of novel compounds can be regarded as a major challenge of many computational chemists (Salemme et al., 1997; Lewis, 2005; Nevin et al., 2012; Tian et al., 2013; Chen, 2013).

In the quest for more sophisticated and efficient ways toward the computational design of novel drug candidates, various integrated *in silico* drug design methodologies are developed and a bundle of useful examples can be found in the literature. A straightforward and frequently utilized *in silico* integrations is the methodology of extrapolation of a pre-devised QSAR model for selection and/or prediction of the biological activity values for compounds comprising a virtual combinatorial library (Liu et al., 1998; Burden & Winkler, 1999; Agrafiotis, 2000; Andersson et al., 2001; Grzybowski et al., 2002; Cruz-Monteagudo et al., 2008). As demonstrated, principal component analysis (PCA), artificial neural networks (ANNs) or 3D-QSAR models such as the Comparative Molecular Field Analysis (CoMFA) are the most widely exploited as robust *in silico* screening devices for selection of hit candidates, while genetic algorithm (GA) or desirability-based methods (e.g., implementation of multi-objective optimization for simultaneous mapping of desired drug properties within QSAR models) are commonly applied for ranking purposes (Cruz-Monteagudo et al., 2008; Nicolaou & Brown, 2013). Notwithstanding their favorable outcome in many cases (Grzybowski et al., 2002), it must be taken into account that not every QSAR model is suitable. The lack of validation of the established QSAR model, the wrongly assessed or not assessed model's applicability domain, or even the use of mis- and/or non-interpretable molecular descriptors are just a few instances which could lead the drug discovery process into a wrong direction (Scior et al., 2009; Bajot, 2010; Sacan et al., 2012). Even in cases where the QSAR model is properly established, its implementation just as a rough *in silico* filtering tool for selection of drug candidates from a massive pool of possibilities without taking into consideration their druggability properties (Oprea, 2000) could also outcome in disappointingly low hit rates. Moreover, the absence of any structural data for the biological target or even an *a priori* knowledge for the structure-activity relationship (SAR) between the investigated entities, could additionally contribute to a poor epilogue of the screening strategy.

Conceptually similar, but improved strategies that cover all of these missing aspects or just partially, are the methods of employing three-dimensional (3D) structural concepts such as the ligand-based, structure-based, or their combination as *in silico* filters for virtual combinatorial libraries (Hecker et al., 2002; Salemme et al., 1997; Shaikh et al., 2007; Krovat et al., 2005; Schlegel et al., 2007; Vilar et al., 2009; Zhang et al., 2013; Drwal & Griffith, 2013). As implied by their name, the availability of 3D structural information for the ligand(s) and/or biological target(s) usually obtained by X-ray crystallography or NMR is of critical significance for their practical utilization as *in silico* ligand filters. More importantly, these methods could also give a structural insight into the possible ligand-protein interactions – an invaluable information that indicate to, what might be expected or what will be the next step toward the design of better hits. Over the years, various straightforward, but also advanced approaches have been invented and some of them became a cornerstone of the modern drug discovery. Among them, the ligand-based and its more advanced counterpart the structure-based pharmacophore modeling concept as well as the methodology of molecular docking of ligands into a defined protein binding site are the most commonly utilized for virtual ligand screening (Wolber & Langer, 2005; Wolber & Kosara, 2006; Schlegel et al., 2007).

However, the majority of these methods are principally not designed for *in silico* screening of extremely massive compound libraries (small-scale screening methods), but rather reduced target-focused libraries (Harris et al., 2011), which can be obtained by integration of various rational knowledge-based

Table 1. A summary of various successful in silico drug design integrations (LBD, ligand-based design; SBD, structure-based design)

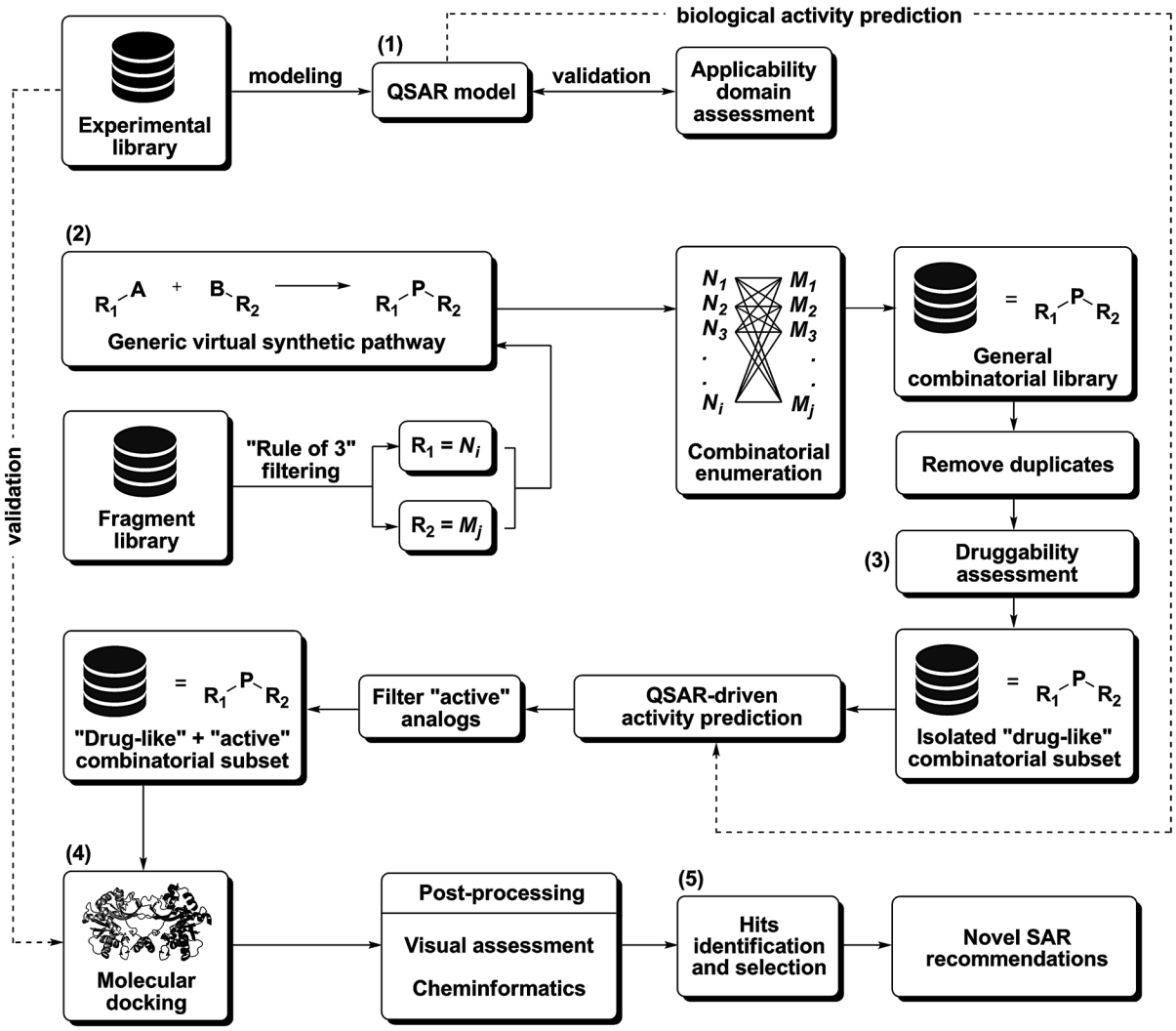
***In Silico* Drug Design Methodologies**				**References**
QSAR	**Drug-Likeness Assessment**	**LBD**	**SBD**	
+	+	+	-	Khalaf et al., 2010; & Abuhamdah et al., 2013
+	-	+	-	Yang et al., 2013; & Kamaria et al., 2014
+	+/-	-	+	Hu et al., 2013; Tan et al., 2013; Ul-Haq et al., 2013; & Zheng et al., 2014
+	+/-	+	+	Musmuca et al., 2010; Coi et al., 2013; Gozalbes et al., 2013
+	+	+	+	Shaikh et al., 2007; Tian et al., 2013

pre-processing strategies (e.g., QSAR-based activity predictions and screening, "drug-likeness" filtering, etc.). Numerous examples demonstrating a high successful rate when using such combined approaches are available elsewhere in the literature and some of the recent are summarized in Table 1.

The current state-of-the-art in the *in silico* design of novel drug candidates includes more powerful, large-scale, integrative methods, which take into consideration various drug design aspects specifically constructed to operate in a simultaneous fashion. The workflow and automation technologies available today are capable for concurrent, complex processing of billions of compounds in an astonishingly fast manner (KNIME[1], Accelrys Pipeline Pilot[2]; Warr, 2012). Based on an elegant and simple idea to visually connect a plenitude of available pre-constructed open-source and commercial component collections (e.g., CDK[3], RDKit[4], Indigo[5], Enalos[6], HCS-Tools[7], SeqAn[8], Molecular Networks[9], BioSolveIT[10], Cresset[11]) into powerful, effective, and fully customizable data pipelines, these scientific data workflow systems were initially developed for solving various chem- and/or bioinformatic problems (e.g., Taverna[12]; Kuhn et al., 2010; Truszkowski et al., 2011). However, over the years much broader scientific fields are covered and this trend is growing constantly. Nowadays, the breakthroughs in the high-performance computing also enable virtualization of the traditional HTS methods through implementation of high-throughput virtual docking (HTD) experiments on multiple various biological targets (Ellingson et al., 2013; Ellingson, Dakshanamurthy et al., 2013). Moreover, the recent explosion in the internet-based cloud technologies allow not only cloud-deposition, integration, and global accessibility of these cutting-edge *in silico* drug design solutions (Ellingson & Baudry, 2012; Hsu et al., 2013), but also facilitate the complete drug discovery process *via* simplified, shareable, collaborative research (InhibOx[13], ScienceCloud[14]).

This chapter illustrates a useful, small-scale integrated *in silico* screening approach for fast and efficient design of novel drug candidates. For this purpose, various mindfully selected 2D/3D *in silico* methodologies are covered and their proficient integration into a powerful knowledge-based virtual screening protocol is discussed. Special emphasis is given to the data flow and their gradual knowledge-based evolvement from a simple 2D to 3D environment followed by a rational concomitant data reduction for final selection of hit candidates.

Figure 2. Graphical overview of our proposed integrated in silico screening platform for design and identification of novel hit candidates



QSAR-AIDED *IN SILICO* INTEGRATIONS FOR SMALL-SCALE LIGAND SCREENING

As reviewed so far, the integrated *in silico* approaches for small-scale ligand screening are probably the most frequently exploited. To demonstrate their practical utilization for selection of promising hit candidates from a large number of possibilities (e.g., a combinatorially-generated compound library), we propose here an effective virtual screening platform (Minovski et al., 2012; Minovski et al., 2013) by integrating various *in silico* drug design methodologies (Figure 2).

As represented in Figure 2, one can perceive that the overall *in silico* screening strategy is fundamentally constructed of five crucial interconnected levels:

1. Predictive QSAR modeling.
2. Virtual combinatorial library design by SAR-based fragments examination.
3. Property-based druggability assessment and QSAR-driven construction of a focused "drug-like" combinatorial library.
4. Structure-based virtual screening and post-processing of QSAR prioritized "drug-like" combinatorial compounds.
5. Hits identification, selection, and novel SAR recommendations.

The main idea behind such an integrated screening concept lies in the generation of a small molecular subspace, which can be regarded as a part of the available chemical space as well as its subsequent knowledge-based gradual reduction for identification of novel hit molecules. Nevertheless, it was estimated that the entire chemical space, which can be regarded as an ensemble of all possible molecular entities contain around 10^{60} biologically-relevant molecules that form the so-called "drug-like" chemical space (Raymond et al., 2010; Deng et al., 2013). Consequently, the objective is not to explore such a large molecular pool, but rather to generate, identify, isolate, and select a small, discrete molecular subspace that fulfill our needs – interaction with the biological system under study (Hopkins, 2008). Therefore, a crucial question arises: Are we capable to create such a small "drug-like" molecular subspace that will serve as a source for hit(s) mining? The answer to this question resides in the distinction between two substantial aspects:

1. Generation of an isolated combinatorial chemical space comprised of all possible structural analogs (e.g., novel compounds under investigation) through implementation of the virtual combinatorial library design approach, and
2. Encirclement and extraction of its embedded "drug-like" combinatorial subspace by sequential *in silico* knowledge-based evaluation supported by property-based filtering, QSAR-driven activity prediction, and structure-based virtual screening methodologies.

If carefully done, these two aspects could not only enable the generation of a high-quality "drug-like" chemical library, but could also significantly contribute to a favorable outcome at the end of the screening strategy.

Predictive QSAR Modeling

One of the earliest, but still popular and useful computational strategies in the modern drug discovery is assuredly the methodology of establishing a quantitative relationship between the chemical structures for a given compound's class and their experimentally-determined property values (Katritzky et al., 1997), i.e., an *in silico* modeling (statistical) technique commonly known as a quantitative structure-property relationship (QSPR). In the scope of the medicinal chemistry, the term "property" usually refers to an experimentally measured biological activity (e.g., IC_{50}) or even toxicity (e.g., TD_{50}) value and consequently the acronym QSPR could be referred as QSAR (quantitative structure-activity relationship) or QSTR (quantitative structure-toxicity relationship). It was found that the concept of modeling the structure-activity relationship (SAR) of chemicals in a quantitative manner is of vital importance in various drug discovery segments (Gussio et al., 1996; Ekins et al., 2002; Lee et al., 2004; Kuz'min et

al., 2007). Regardless of its wide application for plain numerical predictions of biological activities or physicochemical properties for novel compounds, mechanistic interpretation of their chemical and/or biological nature encoded in a form of molecular descriptors, or even understanding the physicochemical features influencing the compound's biological response, nowadays QSAR could be recognized as a viable approach within the framework of the modern multidimensional lead optimization (Lewis, 2005; Gedeck & Lewis, 2008).

Conceptually, QSAR is based on the essential principles of medicinal chemistry, by which the biological activity of a compound is in relation to its molecular structure. As a consequence, one could expect that structurally similar compounds may have similar biological activities (Esposito et al., 2004). In QSAR, the structural features of the compounds under study are usually encoded in a numerical form commonly known as molecular descriptors. Therefore, QSAR could also be regarded as a mathematical relationship between the calculated molecular descriptors (independent variables) and *in vitro* measured biological activity values (dependent variables) for a set of known molecules. The obtained result is commonly expressed in a form of predictive mathematical model, which could readily be utilized for estimation of the biological activity values for novel, not yet synthesized compounds (Figure 2).

When QSAR as a paradigm was introduced for the first time in its entirety in the middle of the previous century (Free & Wilson, 1964; Hansch & Fujita, 1964), a plethora of QSAR approaches have been devised. Except the traditional 2D-QSAR methods (e.g., the popular Free-Wilson and Hansch-Fujita models), which are still very useful, a further breakthrough in the QSAR progress was the introduction of 3D-QSAR methods (Cramer et al., 1988) such as the comparative molecular field analysis (CoMFA) and its relative, the comparative molecular similarity indices analysis (CoMSIA). Since then, the development of QSAR approaches drastically evolved and several multidimensional QSAR congeners (e.g., 4D-, 5D, and 6D-QSAR approaches) were introduced recently (Lill, 2007; Vedani et al., 2005).

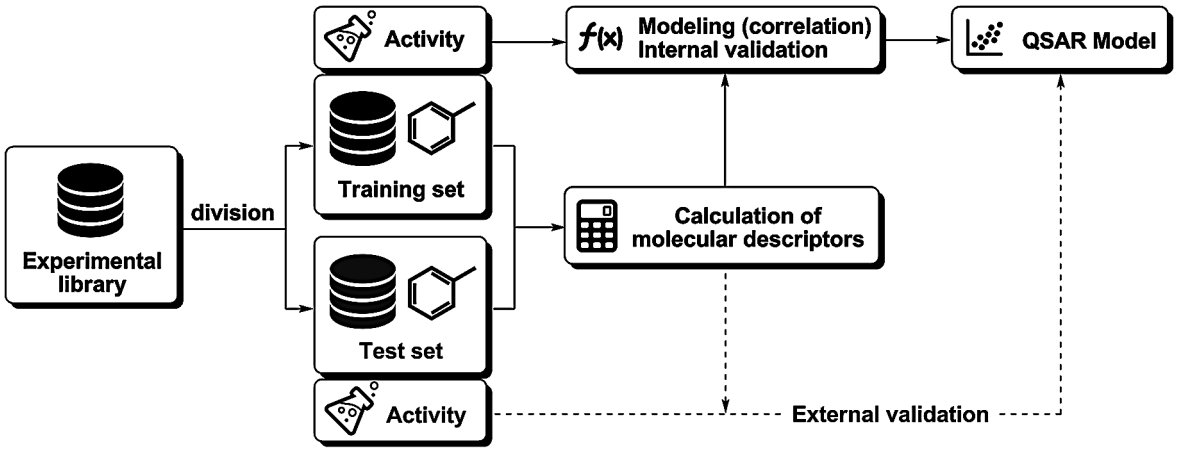
No matter which modeling method is employed, the general procedure for construction of a statistically-reliable predictive QSAR model is assembled of several consecutive steps: dataset preparation, dataset division on the so-called training and test set, calculation of molecular descriptors for both sets, modeling on the training set compounds with a simultaneous internal validation and selection of statistically-significant molecular descriptors, and finally assessment of the predictive power of the constructed QSAR model for estimation of the biological activity values for previously excluded test set compounds (Figure 3).

As shown in Figure 3, only training set compounds are considered in the construction of the QSAR model. In a mathematical sense, the aim is to establish a correlation between the chemical structures of the investigated compounds (training set objects) expressed in a form of calculated molecular descriptors and their corresponding biological activity values (Equation 1).

$$pA = c_0 + c_1 d_1 + c_2 d_2 + c_3 d_3 + ... + c_i d_i \quad (1)$$

where pA denotes the biological activity A in the series expressed as a negative decade logarithm, c_i are coefficients (the fitted parameters), while d_i are the calculated molecular descriptors for each compound i comprising the training set. For the purpose of modeling, different correlation methods (modeling algorithms) are available (Golbraikh & Tropsha, 2002), which can be generally divided in two major categories: linear modeling methods (e.g., multiple linear regression (MLR), partial least squares (PLS), etc.) and non-linear modeling methods (e.g., *k*-Nearest Neighbors (*k*NN), artificial neural networks

Figure 3. A general QSAR methodology flowchart



(ANN), support vector machines (SVM), etc.). Moreover, the squared correlation coefficient is frequently used as a measure for the quality of the established predictive QSAR model, which can be calculated as follows (Equation 2):

$$R^2 = 1 - \frac{\left[\sum_{i=1}^{n_{tr}} \left(\hat{y}_i - y_i\right)^2\right] \Big/ n_{tr}}{\left[\sum_{i=1}^{n_{tr}} \left(y_i - \overline{y}_{tr}\right)^2\right] \Big/ n_{tr}} \tag{2}$$

where n_{tr} indicates the total number of objects (compounds) in the training set, $\hat{y}_i$ and y_i are the predicted and experimentally-determined (observed) biological activities for the *i*-th compound in the training set, respectively, while $\overline{y}$ designates the average of the observed values.

Another important aspect, which directly determines the predictive quality of the established QSAR model, refers to its properly performed validation (internal and external). It has been demonstrated that during the modeling, if the number of independent variables (molecular descriptors) is comparable to or higher than the number of objects (training set compounds), the probability to encounter a chance correlation between the experimental and predicted biological activity for the series significantly increases (Topliss & Edwards, 1979). In order to avoid such undesirable events, a validation of the established QSAR model is required and nowadays various validation criteria are proposed (Consonni et al., 2009; Consonni et al., 2010; Chirico & Gramatica, 2011; Roy et al., 2013). During the modeling, some commonly accepted internal validation techniques such as cross-validation leave-one-out (CV LOO) or cross-validation leave-many-out (CV LMO) are frequently utilized for assessing the statistical stability of the QSAR model (Wold, 1991). The estimation parameter carrying the statistical stability of the QSAR model is expressed as a cross-validated squared correlation coefficient (usually designated as R^2_{cv} or Q^2), which can be calculated by implementation of the same formula above (Equation 2).

In addition to the internal model validation, the robustness of the established predictive QSAR model should also be assessed. In that context, a broadly accepted methodology known as *Y*-randomization is commonly applied (Wold & Eriksson, 1995). The method is grounded on re-building of the QSAR model several times (usually around ten or even more) by iterative shuffling (randomizing) the values comprising the vector of dependent variables *Y* (e.g., the biological activity values) within the original data matrix, while the independent variables (e.g., molecular descriptors used in the construction of the original QSAR model) remain intact. Therefore, one should expect that the newly established randomized QSAR models have significantly lower statistical parameters (R^2 and Q^2) comparing to those of the original non-randomized one – a result that clearly demonstrates the significance and robustness of the constructed predictive QSAR model. In case this requirement is not satisfied, it generally indicates that the established QSAR model is obtained by chance and consequently could be considered as a model of a questionable quality for its further application.

The predictive QSAR model thus established and properly validated is now ready for testing of its external predictive performances on previously excluded test set compounds, which were not used during the model development (external validation of the model). For this purpose, the external validation quantifier ($Q^2_{(te/ext)}$) is determined that reflects the external predictability of the established QSAR model and can be calculated by using the following equation (Equation 3):

$$Q^2_{(te/ext)} = 1 - \frac{\left[\sum_{i=1}^{n_{(te/ext)}} \left(\hat{y}_i - y_i\right)^2\right] \Big/ n_{(te/ext)}}{\left[\sum_{i=1}^{n_{tr}} \left(y_i - \overline{y}_{tr}\right)^2\right] \Big/ n_{tr}} \tag{3}$$

where n_{tr} is the total number of training set objects, while $n_{(te/ext)}$ designates the total number of test, i.e., external validation set objects, respectively.

Alongside the validation, the assessment of applicability domain (AD) for the established QSAR model is another important aspect, which must be taken under consideration as well (Figure 2). It directly reflects the model's reliability for its further practical application (Eriksson et al., 2003). As stated by the OECD principles for validation of the QSAR models for regulatory purposes, the QSAR model should be used within the boundaries of its clearly defined AD (OECD, 2004). To date, several approaches for estimation of the AD for QSAR models have been introduced (Zhang et al., 2006; Sahigara et al., 2012; Minovski, Župerl et al., 2013), which implementation generally vary on the modeling routine employed (linear or non-linear). Once the predictive QSAR model is rigorously validated and its robustness, predictability, and reliability are properly established, it could be readily extrapolated for estimation of the biological activity values for novel compounds (Figure 2).

Virtual Combinatorial Library Design by SAR-Based Fragments Examination

The methodology of compound library design usually refers to the generation of a list of molecules that could be synthesized utilizing either solid- or solution-phase combinatorial chemistry approach (Chabala, 1995; Coe & Storer, 1999). Traditionally, the synthetic combinatorial chemistry alias high-throughput synthesis was the major production "device" for boosting the chemical space not only in a quantitative

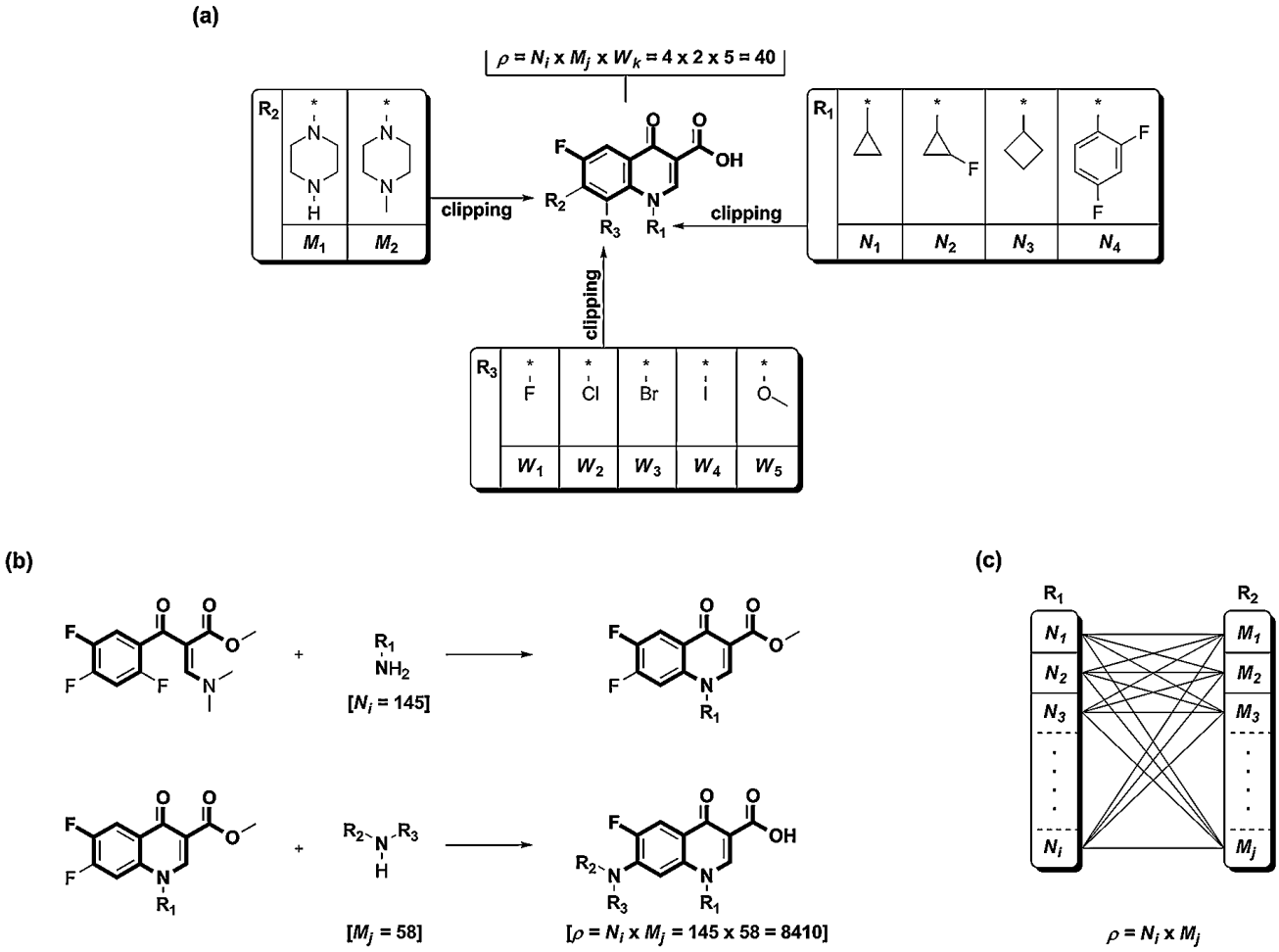
manner, but also in the terms of its structural diversity. As mentioned previously, this concept was generally based on the hypothesis: the bigger and more diverse is the generated chemical space, the higher is the probability to identify a potential lead by high-throughput screening (HTS). Unfortunately, such a strategy was disappointing for the pharmaceutical industry, since very frequently resulted in an unsatisfactory cost-benefit ratio (Lahana, 1999; Ramesha, 2000; O'Driscoll, 2004). Over the years, the drug discovery practice has evolved in a direction of swapping such unreasonable strategies with novel, cost-effective, and smart knowledge-based approaches, and the methodology of virtual combinatorial library design (known as combinatorial enumeration) was found as a suitable one (Aronov, 2002). Nowadays, by using this *in silico* approach, one can rapidly and efficiently generate thousands of novel compounds in a virtual environment that could effortlessly be manipulated.

Two generally accepted strategies are commonly used for virtual enumeration of combinatorial libraries such as the Markush enumeration approach (Leland et al., 1997) as well as the reaction-based enumeration approach (Leach et al., 1999; Lobanov & Agrafiotis, 2002). In the Markush enumeration, the combinatorial library is characterized by using a Markush structural representation (widely known as "*Core plus R-groups*" molecular representation) that is the main molecular scaffold (the structural core of the combinatorial library) with one or more defined variable attachment points expressed as R-groups (e.g., R_1-, R_2-, R_3-, etc.) and a set of reagents that should be "*clipped*" (transformed) into an appropriate set of substituents (Figure 4a). On the other hand, in the reaction-based enumeration, the combinatorial library is defined by using one- or multi-step generic reaction pathway where the variable entities are represented as reactants (e.g., R_1-OH, R_2-NH_2, etc.) as illustrated in Figure 4b. Which one of these two approaches will be used, mainly depends on the complexity of the problem as well as the user's preferences (Agrafiotis et al., 2002). For instance, if the synthetic pathway is long and complex, and the scaffold representing the final product is clearly defined (no significant inter-reaction scaffold changes exist), then the Markush enumeration approach would be appropriate. Conversely, if the scaffold is variable and must be progressively constructed during the virtual synthesis, then the reaction-based enumeration approach may be the method of choice.

A key part of both enumeration approaches that could directly influence the quality as well as diversity of the generated combinatorial library is the selection of synthetically feasible reactants (e.g., fragments, building-blocks) to be attached at the pre-defined variable scaffold positions. For this purpose, a variety of *in-house* developed or commercial reactants databases are available (e.g., Accelrys ACD[15]). Nevertheless, for the sake of medicinal chemistry, a viable approach is to pre-screen the reactants database by using cheminformatically-based fragment-likeness filters such as the widely known "*Rule of 3*" ("Ro3") filtering set (MW $\leq$ 300 and *n*HBD, *n*HBA, *n*RB $\leq$ 3), where MW is the molecular weight, while *n*HBD, *n*HBA, and *n*RB, designate the number of hydrogen bond donors, hydrogen bond acceptors, and rotatable bonds, respectively (Congreve et al., 2003). This procedure assures the location of the obtained fragments within the "lead-like" chemical space (Hann & Oprea, 2004). Thus pre-treated, the obtained fragments are routinely assembled into fragment libraries, which are now ready to use in the next step – fragments selection for combinatorial enumeration.

The fragments selection is frequently done by substructure search (SSS) procedure enforced by the fragment(s) designation. For instance, the R_1-OH designation determines the SSS-supported retrieval of all organic fragments from a fragment library that contain –OH group(s) including alcohols, phenols, carboxylic acids, etc. However, if one takes into account all the fragments retrieved by SSS without their subsequent examination, then the generated combinatorial library would be so diverse so it can be compared to the chemical space obtained by the traditional high-throughput synthesis that is not fully

Figure 4. Combinatorial enumeration approaches (examples of 6-fluoroquinolones combinatorial libraries): (a) Markush enumeration, (b) reaction-based enumeration, and (c) schematic representation of a combinatorial enumeration process for a simple reaction system configured of two subsets of reactants R_1 (N_1…N_i) and R_2 (M_1…M_j)



in agreement with our expectations. In order to avoid this, the retrieved fragment subsets should be carefully examined (automatically and visually, if possible). In practice, the fragments examination is usually SAR-based, and therefore the availability of any SAR knowledge for the molecular entity (e.g., a drug with well-established potency) used as a template for analogs design would be of high significance. Moreover, the SSS procedure can also be additionally employed for efficient elimination of those fragments with undesirable reactive functionalities and thereby the risk for late determination of *in vitro* false positives could be minimized (Rishton, 1997; Rishton, 2003). The fragment subsets prepared by using this way can finally enter the combinatorial enumeration process.

The combinatorial enumeration (virtual combinatorial generation) can be regarded as a straightforward procedure of statistical non-repetitive fragmental permutation where each fragment of a given reactants subset "interact" with each fragment of the other reactants subsets within the previously defined molecular scaffold (Figure 4c). Let's consider a simple reaction system configured of two subsets of reactants R_1 ($N_1 \ldots N_i$) and R_2 ($M_1 \ldots M_j$). Following the combinatorial enumeration process (Figure 2 and Figure 4c), one could easily estimate the total number of combinatorial analogs obtained by the enumeration process as a multiplication product between each pair of reactants ($N_1 \ldots N_i$ and $M_1 \ldots M_j$). Mathematically, this process can be expressed by using the following equation (Equation 4):

$$\rho = N_i \times M_j \tag{4}$$

where N_i are the total number of fragments within the R_1 subset, M_j are the total number of fragments within the R_2 subset, while ρ is the total number of products (combinatorial analogs) obtained by the combinatorial enumeration (Wieland, 1997). Furthermore, an additional check of the constructed virtual combinatorial library for duplicate molecular entities and their subsequent removal is frequently a good practice (Figure 2).

Property-Based Druggability Assessment and QSAR-Driven Construction of a Focused "Drug-Like" Combinatorial Library

As exemplified above, the methodology of virtual combinatorial library design resulted in obtaining a general virtual combinatorial library comprised of all possible combinations of structural analogs for the compound(s) under study (Figure 2); therefore, it can be regarded as an isolated part of the available chemical space (Bohacek et al., 1996). However, the essential question here is, whether this isolated molecular pool is useful (at all) for further selection of novel drug molecules?

Previously, we demonstrated that the so-called "drug-like" chemical space was estimated to contain approximately 10^{60} biologically relevant molecules with MW ~ 500 (Raymond et al., 2010; Deng et al., 2013). Thus, one should primarily assess whether our isolated virtual combinatorial library represents a part of it. Moreover, it was found that the combinatorial algorithm in first instance increases the molecular complexity (Hann et al., 2001); as a consequence, one should also bear in mind the low probability that each generated compound within such a virtual combinatorial library possess "drug-like" properties (Walters et al., 1999; Muegge, 2003). To eliminate these dilemmas as well as to identify and distill those compounds representing the "drug-like" chemical space, the constructed general virtual combinatorial library should be further subjected to a thorough property-based druggability assessment (Barril, 2012).

It was found that the "drug-like" properties confer favorable ADMET (absorption, distribution, metabolism, excretion, and toxicity) attributes to a compound. However, it is important to clarify which are those properties that describe a single molecule as a potential drug? In a broader sense, the "drug-like" properties are defined as intrinsic molecular features that can strongly influence the optimization of their pharmacological characteristics (Borchardt, 2004). During the last two decades, various attempts were made for identification of these molecular properties and clarification of their

role. Owing to the seminal work of Christopher A. Lipinski who pioneered the famous "Lipinski's rule of 5" ("Ro5"), the concept of "drug-likeness" acquired a totally new dimension in the drug discovery world (Lipinski et al., 1997). It states that the probability for a compound to be poorly absorbed after oral administration increases if any two of the following rules are violated:

- Molecular weight (MW) is less than 500 Da.
- Number of hydrogen bond donors (*n*HBD: OH/NH groups) is equal or less than 5.
- Number of hydrogen bond acceptors (*n*HBA: O/N) is less than 10.
- Calculated logP is less than 5.0 (by using ClogP) or 4.15 (by using MlogP).

Nevertheless, an important aspect to be taken with precaution when using the "Ro5" as "drug-like" filter for druggability assessment is the lipophilicity parameter (logP). Namely, in the "Ro5" druggability filter the logP parameter refers to the lipophilicity of a compound in its neutral state. On the other hand, the majority of the drug molecules (~ 95%) prevail in an ionizable state. Since the orally administered drugs are mainly absorbed in the small intestine where the environment is slightly acidic, the pH parameter should also be considered. Therefore, a modified "Ro5" concept was introduced recently, favoring the implementation of the pH dependent version of logP at intestinal pH ~ 5.5 ($logD_{5.5}$), instead of the classical logP parameter for estimation of the compounds lipophilicity (Bhal et al., 2007).

Another critical molecular parameter that strongly correlates with the compounds membrane permeability and consequently their oral bioavailability is the polar surface area (PSA) designated as a sum of the van der Waals surface areas of the polar atoms (O and N) within the molecule. Accordingly, Daniel F. Veber assembled a useful two-parameter rule set, i.e., "*Veber rule*" by which a molecule is likely to have a good bioavailability after oral administration if the following criteria passed (Veber et al., 2002):

- PSA is equal or less than 140 $Å^2$ (or ≤ 12 *n*HBD + *n*HBA).
- Number of rotatable bonds (*n*RB) is equal or less than 10.

In addition to these "drug-like" filters, there are also some other similar rule sets carrying various drug-discriminating physicochemical properties including the Pardridge blood-brain barrier (BBB) permeability filter (Pardridge, 1995), Clark-Lobell's BBB permeability filter (Clark, 2003; Lobell et al., 2003), the Pharmacophore Point Filter (Muegge et al., 2001), and many others. It is important to stress out that these *in silico* filters should not be used indiscriminately, i.e., there must be a solid ground why to use one filter, and not another. For instance, if one needs to screen a set of intravenously administered compounds for their BBB penetrating abilities, then one of the available *in silico* BBB permeability filters should be applied (e.g., Clark-Lobell's BBB filter), and not the filters for prediction of intestinal absorption (e.g., Lipinski's Ro5). Once the appropriate "drug-like" filter is selected, it can be proficiently employed for *in-depth* screening of our previously constructed general virtual combinatorial library. Those combinatorially-generated compounds (if any) which successfully pass all the pre-defined filter's criteria represent our so-called isolated "drug-like" combinatorial subspace.

The selected compounds are further subjected to the previously constructed predictive QSAR model (Figure 2) to obtain *in silico* prediction of the biological activity values. After sorting the compounds by decreasing the predicted biological activities, the priority list could be further reduced if some activity range data are available (usually obtained by *in vitro* functional and/or biochemical studies of known

drugs). This could be considered as an additional filter for extraction of those "drug-like" combinatorial analogs with predicted biological activity values within the desired pre-defined activity range. For example, the *in vitro* inhibition assays of 6-fluoroquinolone (6-FQ) antibacterials against *M. tuberculosis* DNA gyrase enzyme revealed that these inhibitors are active in the range of $0.0 \leq \text{MIC}\ [\mu g/mL] \leq 1.0\text{-}2.0$, where MIC refers to the minimal inhibitory concentration. Therefore, one can apply this activity range as *in silico* filter for selection of those combinatorially-generated 6-FQs which QSAR-predicted MIC values fall within the pre-defined MIC boundaries (predicted "active" 6-FQ analogs). The obtained combinatorial subset comprised of all "drug-like" plus "active" (as predicted by the QSAR model) combinatorial analogs can be regarded as a focused "drug-like" combinatorial library which is now prepared to enter the structure-based virtual screening stage of our *in silico* integrated protocol (Figure 2).

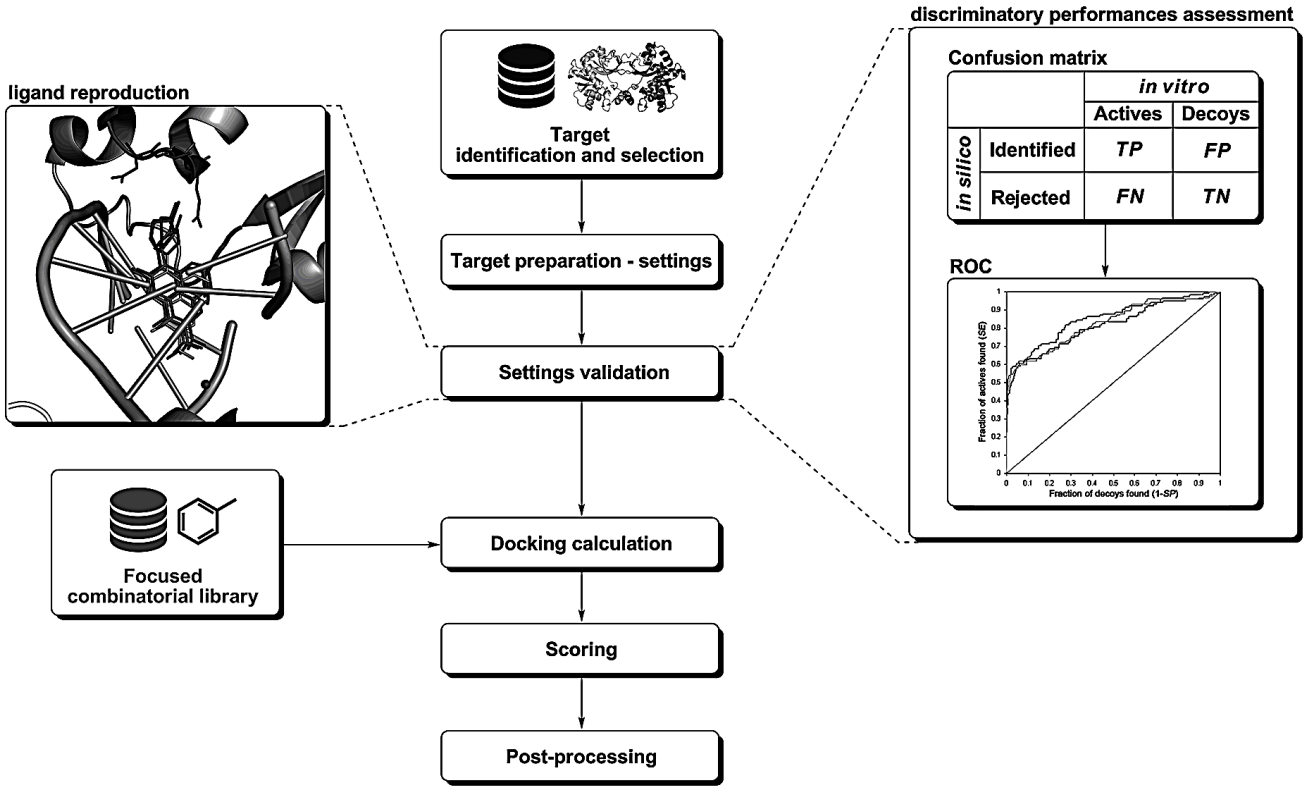
Structure-Based Virtual Screening and Post-Processing of QSAR Prioritized "Drug-Like" Combinatorial Compounds

The *in silico* strategies elaborated so far generally belong to the class of ligand-based methodologies which usefulness and versatility in various drug discovery segments are broadly reviewed (Vidal et al., 2011). Notwithstanding their mounting success, a major drawback that follows all these approaches resides in their incapability to account for the possible intermolecular interactions between a given ligand and its biomolecular target. In that context, the structure-based virtual screening approaches were found as particularly important to proficiently complement the missing gap (Ghosh et al., 2006; Kroemer, 2007; Villoutreix, et al., 2009) and here the methodology of automated molecular docking was recognized as a pivotal one in the framework of the modern *in silico* drug discovery (Lybrand, 1995; Morris & Lim-Wilby, 2008).

As mentioned previously, the availability of 3D structural information for the biological target at atomic resolution (preferably ≤ 2.5 Å, obtained by experimental techniques such as X-ray crystallography or NMR) is critical for its implementation as virtual ligand filter. Herein, the generic term "biological target" usually refers to a biomacromolecular structure (e.g., a protein, nucleic acid, or their complex) into which the ligand (e.g., a small drug molecule) is being docked. This concept ordinarily relates to a method commonly known as protein-ligand docking, however, during the last few years an expanding interest was observed for protein-protein docking methods as well (Gray et al., 2003; Wang et al., 2007; Moal et al., 2013). Nevertheless, for the purpose of *in silico* design of small molecular entities as possible future drug candidates, the protein-ligand docking methods were found as the most frequently exploited.

As illustrated in Figure 5, a key part of the protein-ligand docking methods is the target identification and its subsequent selection. This imposes a profound understanding of the nature of the biological system under study including its function and mechanism in physiological or pathophysiological purposes, the location of the catalytic and/or binding site(s), and preferably the ligand-binding mechanism if available (Seifert & Lang, 2007). Moreover, one should be aware of some possible scenarios that could frequently follow the target selection and later can strongly affect the final epilog of the docking study. Ideally, the availability of a protein-ligand complex structure originating from the designated organism is of invaluable importance in the protein-ligand docking studies. This often enables a successful run of the entire docking process in a straightforward way, and consequently obtaining a productive outcome at the end of the screening strategy. Unfortunately, this is not always a case. Some protein targets are not or cannot be co-crystallized with the ligand, even if they originate from the desired organism. There are also some cases where the structure of the protein-ligand complex is available, but originates from other

Figure 5. A typical methodology workflow covering the essential steps of a protein-ligand docking



species, or the desired protein structure is not available as a whole, but as separated subunits. In such cases, some alternative approaches in the scope of the target preparation stage could be applied including automated homology modeling and/or ligand's binding site recognition (Xiang, 2006; Minovski et al., 2013; Yang et al., 2013).

Once the designated biological target is identified and selected, it should be properly prepared for virtual screening. Except some preparation routines related to the entire target structure such as adding hydrogen atoms and elimination of the water molecules, the target preparation stage usually refers to the settings of the binding site for docking calculation. This often includes assignment of the protonation, stereoisomeric, and tautomeric states of the amino acid residues covering the binding pocket (Anderson, 2003). Additionally, small molecular entities (e.g., water molecules, metal ions, and other co-factors) present in the binding pocket should also be removed, unless there is a strong evidence for their specific role in the protein-ligand binding mechanism (Cheng et al., 2012). For instance, the water molecules present in the binding pocket could serve as hydrogen bonding bridges between the ligand and surrounding amino acid residues. Furthermore, if a co-crystallized ligand conformation is available, it can be easily employed for assignment of the screening area, i.e., the ligand search space that is usually a sphere-shaped constrain with a centroid defined by the Cartesian coordinates (x, y, z) of a selected ligand's atom. Otherwise, the screening area can algorithmically be assigned as an integrated part of the binding site recognition routine.

It should be stressed, however, that in cases where a protein-ligand complex is available (experimentally determined or assembled by homology modeling), an initial docking validation run is required. This involves reproduction of the spatial orientation and conformation for the co-crystallized ligand that directly reflects the quality of all the settings performed previously. The ligand reproduction assessment usually consists of re-docking of the co-crystallized ligand conformation and its subsequent comparison to the calculated docking solutions. As judging criteria for the performed ligand reproduction, the all-atom or heavy-atoms root-mean-square deviation (RMSD) value between each calculated docking pose and the co-crystallized ligand conformation is usually calculated. The calculated RMSD values below 2.0 Å are generally accepted and often indicate a successfully performed ligand reproduction (Verdonk et al., 2003). Moreover, if a library comprised of compounds with experimentally-determined *in vitro* biological activities is on hand (e.g., similar to the one used for QSAR modeling), then it could be proficiently employed in an additional docking validation experiment for evaluation of the target's discriminatory performances – a valuable guideline that indicates the capability of our biological target to effectively discriminate between known active and inactive ligands (Figure 2). It was demonstrated that such a validation is of exceptional importance where one needs to rank two or more analogous protein homology models for their ligands discriminating capabilities (Minovski et al., 2013). In that sense, a widely accepted strategy is the so-called *enrichment* of the library of known actives with an incomparably higher number of decoy molecules (Huang et al., 2006). Here, a decoy is defined as a molecular entity that physicochemically matches the known active compounds, i.e., shares the same physicochemical features, but at the same time is topologically distinct. The main idea behind this concept is to evaluate how good a constructed protein homology model can identify the known actives in the wave of decoy molecules. For that purpose, the "receiver operating characteristic" (ROC) curve methodology proved to be the most suitable one for solving binary classification problems (Triballeau et al., 2005; Hevener et al., 2009). The ROC curve covers the sensitivity (*SE*) and specificity (*SP*) of the model, where *SE* is graphically expressed as a function of (1-*SP*). These two features one could easily determine from the confusion matrix representing a binary classification (Manallack et al., 2002). Let's consider a case where a docking-based virtual screening protocol was applied for evaluation of the discriminatory power of two analogous protein homology models to correctly identify active ligands among a large pool of decoy molecules. In the context of binary classification, four classes could be distinguished: true positives (*TP* = *in vitro* actives and *in silico* identified), true negatives (*TN* = *in vitro* inactives and *in silico* rejected), false positives (*FP* = *in vitro* inactives and *in silico* identified), and false negatives (*FN* = *in vitro* actives and *in silico* rejected). Consequently, *SE* and *SP* parameters for both homology models can be calculated as follows (Equation 5 and Equation 6):

$$SE = \frac{TP}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{TN + FP} \tag{6}$$

where *SE* is defined as the ratio of *in silico* correctly identified actives (positives) over all compounds which are known to be truly active *in vitro*, while *SP* is defined as the ratio of *in silico* correctly identified inactive molecules (negatives) over all compounds which are known to be truly inactive *in vitro*. Finally, the area under the ROC curve (ROC-AUC) is computed directly from the ROC plot – a convenient metric that reflects the discriminatory performances of a protein homology model as a virtual ligand filter (Equation 7).

$$\text{ROC-AUC} = 1 - \frac{1}{N_A}\sum_{i=1}^{N_A}\frac{N^*_{i,D}}{N_D} \tag{7}$$

where N_A represents the total number of actives, N_D is the total number of decoy molecules, while $N^*_{i,D}$ depicts the number of decoy molecules that are higher ranked than the *i*-th active compound.

The ROC-AUC values can vary between 0 (poor model's discriminatory performances) and 1 (excellent model's discriminatory performances), whereas ROC-AUC = 0.5 corresponds to a random selection (Triballeau et al., 2005). Therefore, the higher values for ROC-AUC (preferably above 0.5), pinpoint to better discriminatory performances of a given protein homology model, and *vice versa*. Nevertheless, while ROC-AUC metric evaluates the overall performance of a given protein homology model considering the entire data (actives and decoys), it is not capable to effectively account for the *early enrichment* of active molecules in small portions of the entire compound library (e.g., 0.5%, 1.0%, or 2.0%). Therefore, to fulfill the missing gap an additional metric commonly known as *enrichment factor* ($EF_{x\%}$) should be applied, which is derived from the ROC curve as well (Jahn et al., 2011). The $EF_{x\%}$ for a given library portion ($x\%$) can be calculated as follows (Equation 8)

$$\text{EF}_{x\%} = \frac{\dfrac{N^*_A}{N_{x\%}}}{\dfrac{N_A}{N_A + N_D}} \tag{8}$$

where N_A is the total number of active molecules in the entire library, N_D is the total number of decoy molecules in the entire library, $N_{x\%}$ is the total number of compounds in the observed portion of the library ($x\%$), and N^*_A represents the total number of actives found within the observed portion (Jahn et al., 2011). Thus validated, the target can now be used for performing molecular docking calculations on novel compounds such as those that comprise our previously constructed focused "drug-like" combinatorial library derived by QSAR-driven prioritization (Figure 2).

The molecular docking calculation refers to a virtual examination of the pre-assigned ligand search space (screening area) followed by ranking of the calculated docking solutions (i.e., the generated ligand binding poses) for subsequent determination of the correct binding modes with the target (Morris & Lim-Wilby, 2008). In the attempt to find the energetically most favorable binding solutions, the search algorithm runs several times for each ligand entering the docking calculation. The number of trials per ligand is often assigned by the user and usually corresponds to the total number of calculated poses at the end of the docking calculation (frequently 3-10). Whether the calculated ligand's binding pose is energetically favorable depends on the computed protein-ligand interaction energy

(known as a score) for that ligand. For this purpose, many docking tools are equipped with various scoring functions (Halperin et al., 2002; Jain, 2006; Huang et al., 2010). The scoring function initially evaluates the spatial orientation and conformation of the calculated ligand pose by calculating a simple energy function (e.g., a force field constructed of electrostatic and attractive/repulsive van der Waals energy terms), while later more complex scoring schemes are used for estimation of the ligand binding affinity (Gohlke & Klebe, 2001).

The docking calculation is considered finished once a sufficient number of solutions (docking poses) have been generated for each compound entering the docking process. The obtained results are usually stored in a compact form (e.g., a virtual compound library) where the estimated docking solutions for each docked ligand are organized as small clusters. The docking poses within each cluster are frequently ranked by the calculated scoring function in a descending order, i.e., the highest scored docking pose is located at the top of each cluster. Naturally, one would expect that these top-scored solutions are indeed the best binding poses. Unfortunately, it was found that the hit(s) selection based solely on the highest calculated dock score is not always sufficient, mainly as a consequence of the imperfection of some scoring functions to correctly account for the ligand's binding mode (Kitchen et al., 2004; Cheng et al., 2012). Therefore, irrespectively of the calculated dock scores, a post-processing of the obtained results, i.e., a thorough post-docking analysis should be accomplished (Minovski et al., 2012; Minovski et al., 2013). Thus, a good practice is to perform an *in-pocket* visual assessment of all the calculated docking solutions, which could raise the entire post-docking analysis to a significantly higher, chemically intuitive, knowledge-based level (Doman et al., 2002). From a drug design perspective, it may involve an evaluation of various useful attributes such as dock pose spatial orientation, conformation, fitness of the calculated dock pose with the co-crystallized ligand, as well as an *in-pocket* SAR-based pharmacophore assessment based on the estimation of how many common pharmacophoric features are shared between an investigated dock pose and the co-crystallized ligand (Minovski et al., 2013). The latter could be recognized as particularly interesting, since it could effectively account for the potential protein-ligand interactions. Nevertheless, while the visual inspection might be feasible in cases when a reasonable number of docking solutions should be checked (e.g., several hundreds of docking poses), it might be totally impractical when one needs to assess thousands of docking poses. In such circumstances, the automated solutions to this problem can be regarded as one of the most needed. Nowadays, owing to the advancements in cheminformatics, various useful tools for automated post-docking analysis based on mapping of the desired protein-ligand interactions are developed (Marcou & Rognan, 2007; Rastelli et al., 2009; Bouvier et al., 2010). However, within the framework of an integrated protocol for small-scale ligand screening (such the one illustrated in this chapter), where the starting compound library is significantly reduced throughout all the levels (Figure 2), only a reasonably small number of compounds are subjected to docking calculation. Therefore, taking into account all the attributes mentioned above, the post-docking analysis based on visual examination of the calculated docking solutions could be equally good for final identification and selection of potential novel hits.

HITS IDENTIFICATION, SELECTION, AND NOVEL SAR RECOMMENDATIONS

As mentioned above, the post-docking analysis based on visual examination provides some useful information about calculated dock poses as well as their putative binding mechanism with the key amino acid residues wrapping the target's binding site. Nonetheless, the question that arises here is how this information might be helpful for identification and selection of the most promising hit candidates? For

the purpose of hits identification, various *in silico* methods and tools are devised (Marcou & Rognan, 2007; Rastelli et al., 2009; Bouvier et al., 2010; Moldover et al., 2012; Liu et al., 2013), however, in the context of our proposed integrated *in silico* screening protocol, we introduce here a novel and simple semi-automated Boolean-based [T/F (true/false)] clustering method for hit(s) identification, covering all the attributes determined in the scope of the visually derived post-docking analysis (Minovski et al., 2013).

The Boolean-based clustering method is substantially organized in three consecutive levels:

- **Level 1, Geometric Properties Assessment:** Spatial examination of each calculated dock pose relative to the experimental co-crystallized ligand conformation and building a cluster of (T)-signed dock poses. Within this level, a total of three geometric properties should be visually assessed: (1) spatial orientation (how a calculated dock pose is spatially oriented relative to the position of the co-crystallized ligand), (2) pose fitness (how well a calculated dock pose fits the co-crystallized ligand), and (3) number of matching pharmacophoric features (how many common pharmacophoric features are shared between a calculated dock pose and the co-crystallized ligand). The latter implies to an *in-pocket* generation of a structure-based pharmacophore model for both the co-crystallized ligand conformation and the investigated dock pose. For this purpose, various powerful automatic pharmacophore generation tools are available (e.g., Catalyst[16], LigandScout[17]). Contrary to the properties (1) and (2), which are evaluated qualitatively by using the Boolean-based (T/F) signing scheme, the third property (number of matching pharmacophoric features) is evaluated quantitatively. Whether an investigated dock pose will pass to the next level or will be rejected, depends on the positive outcome for all three geometric properties. For instance, if the calculated dock pose is properly oriented (T), fits well (T), and shares the desired number of pharmacophoric features with the co-crystallized ligand conformation (e.g., 5), then as a consensus, the dock pose could be estimated as geometrically good (T) and consequently goes further. Otherwise, the dock pose will be rejected (F). This assessment should be repeated for all the calculated dock solutions. At the end, only those poses signed as (T) should be distilled as a separate cluster and used in the second level of the Boolean-based clustering method.
- **Level 2, Score-Based Clustering:** (T)-signing of the *Level 1* dock poses with calculated dock score within a pre-defined, desired range (e.g., highly scored dock solutions) and building a new cluster of highly scored (T)-signed hits. Those calculated dock solutions, which dock score is outside the boundaries of the pre-defined range, would be rejected (F).
- **Level 3, Activity-Based Clustering:** (T)-signing of the *Level 2* hits with predicted biological activity values within a pre-defined, desired range (e.g., highly "active" combinatorial analogs as estimated by the previously used predictive QSAR model) and building a new cluster of (T)-signed most "active" hits. It should be stressed, however, that all the combinatorial analogs comprising our previously constructed focused "drug-like" combinatorial library are somehow hypothetically "active" as defined by the additionally implemented biological activity range filter. For instance, as previously stated, it was found that 6-fluoroquinolone antibacterials are active as *M. tuberculosis* DNA gyrase inhibitors in the range of $0.0 \leq \text{MIC}\ [\mu g/mL] \leq 1.0$ or 2.0. Let's consider that one needs to extract those highly "active" 6-fluoroquinolones with predicted MIC ≤ 0.05 μg/mL. Therefore, the dock solutions which QSAR predicted biological activity is less or equal then 0.05 μg/mL will be signed as (T) and used in the end phase of the hit(s) identification/selection process.

Finally, all (T)-signed dock poses isolated within the last level of the Boolean-based clustering method, are further subjected to an additional substructural assessment of the combinatorially attached fragments. This selection procedure resides in the determination of the most frequently appearing fragments at the pre-defined variable scaffold positions followed by a subsequent comparative SAR analysis for exposing novel possible guidelines utilizing a well-established SAR for known drugs. The combinatorial hits thus selected are now ready for synthesis, purification, and further *in vitro* biological activity verification.

CASE STUDY: *IN SILICO* DESIGN AND IDENTIFICATION OF NOVEL 6-FLUOROQUINOLONES AS POTENTIAL INHIBITORS AGAINST *MYCOBACTERIUM TUBERCULOSIS* DNA GYRASE

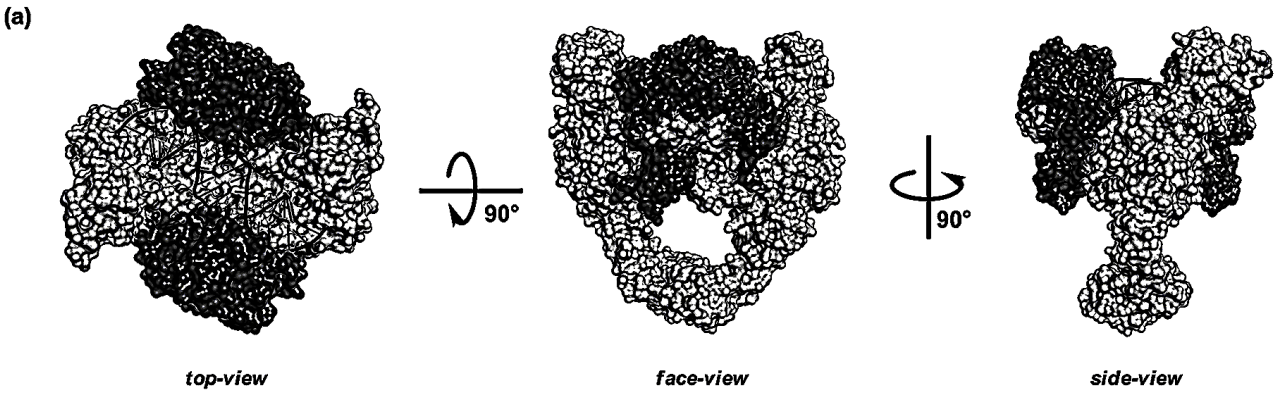
The integrated *in silico* screening protocol elaborated in the previous section (Figure 2) could generally be applied for the design and identification of novel drug analogs, irrespectively of the drugs class that are derived from as well as the *in silico* methodologies employed. However, to demonstrate its practical implementation, we opted to illustrate a specific example related to the design and identification of novel 6-fluoroquinolone (6-FQ) antibacterials as potential inhibitors against *M. tuberculosis* DNA gyrase enzyme.

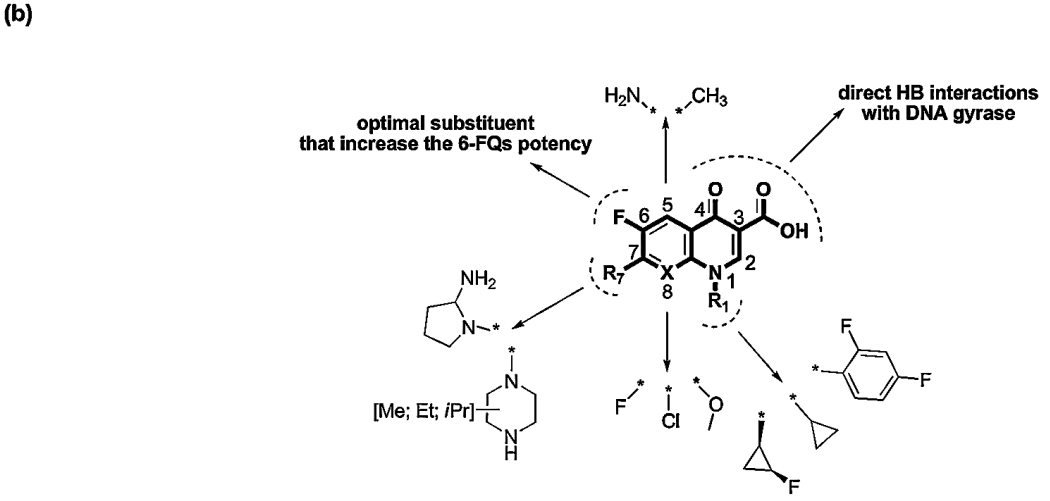
M. tuberculosis DNA Gyrase as a Therapeutic Target and 6-Fluoroquinolone Antibacterials

It is widely known that in all living organisms, the correct spatial DNA topology is of crucial importance for the proper regulation of the DNA processing including all the basic biological processes, control of the gene replication and transcription, as well as the DNA segregation (Wasserman & Cozzarelli 1986; Liu et al., 2009). In order to be spatially correct, the DNA topology is fundamentally maintained by a specific family of enzymes broadly known as the DNA topoisomerases (Sissi & Palumbo, 2010); this family of superior molecular nanomachines is substantially involved in the maintenance of two vital sub-cellular processes such as the DNA single-strand breaks (controlled by the so-called type I enzymes) and DNA double-strand breaks (controlled by the so-called type II enzymes).

In contrast to the other, mainly higher organisms, which possess multiple topoisomerase enzymes involved in various essential sub-cellular functions, the bacterial organisms usually possess two general types of topoisomerases (Levine et al., 1998) - DNA gyrase enzyme (responsible for the unwinding of the bacterial DNA during the DNA replication phase) and topoisomerase IV (a DNA gyrase paralogous form involved in the DNA decatenation process). However, unlike other bacterial species, *M. tuberculosis* is an unusual and unique bacterial organism, since it possesses only one type II topoisomerase - the DNA gyrase with a specific simultaneous functional role of topoisomerase IV (Cole et al., 1998; Collin et al., 2011; Bouige et al., 2013). Structurally, the mycobacterial DNA gyrase (Figure 6a) is comprised of two cardinal subunits GyrA and GyrB (relevant to the ParC and ParE subunits in topoisomerase IV) that together assemble a functional heart-shaped heterotetrameric complex A_2B_2 (C_2E_2 in topoisomerase IV). The structural and biochemical studies performed so far revealed that the GyrA subunit is responsible for the mycobacterial DNA breakage/reunion catalytic process (i.e., DNA replication and elongation), while the correct spatial topology of the mycobacterial DNA is maintained by the GyrB subunit (Levine et al., 1998; Cole et al., 1998). This catalytic process in *Mycobacteria* was recognized as a promising targeting mechanism for efficient tuberculosis chemotherapy (Ferrero et al., 1994; Maxwell, 1997; Collin et al., 2011).

Figure 6. (a) Three-dimensional structural views of the mycobacterial DNA gyrase enzyme; the individual DNA gyrase parts are colored differently – GyrA subunit (in light gray), GyrB subunit (in dark gray), and the mycobacterial DNA molecule (in black); (b) The current SAR knowledge of the 6- FQ antibacterials





The quinolone antibacterials (e.g., Nalidix acid and its structurally derived analogs) were found as the sole and most efficient DNA gyrase inhibitors so far. Among them, the 6-FQ class (quinolone's structural congeners that contain a fluorine atom attached at the 6 position of the main quinolone core) are probably one of the most utilized antitubercular agents in the clinical practice (Figure 6b). Their unique mode of action is grounded on the inhibition of the essential DNA gyrase mechanisms (e.g., supercoiling and relaxation of the double-stranded DNA), followed by instant formation of an irreversible covalent complex between the GyrA subunit, 5'-end of the mycobacterial DNA, and the 6-FQ inhibitor itself (Gellert et al., 1977; Sugino et al., 1977). It was found that the complex thus formed is highly stable and cytotoxic for the *Mycobacteria*, leading to an irrecoverable distortion of the entire DNA topology, failures in the DNA processing mechanisms, and finally bacterial cell destruction (Drlica & Malik, 2003).

The current SAR knowledge of 6-FQ antibacterials (Tillotson, 1996) explicitly shows the crucial 6-FQs scaffold positions, which structural alteration could significantly enhance or reduce the antimycobacterial activity (Figure 6b). As demonstrated, the main quinolone moiety (1,4-dihydro-4-oxo-3-pyridinecarboxylic acid) is of significant importance for the antimycobacterial activity. At the position 1, the substitution of a cyclopropyl group was found as the optimal one, however, various bulky substituents could also be attached at this position, which could enhance not only the lipophilic profile of the entire drug, but could also increase its metabolic stability after possible oral administration. On the other hand, the positions 5 and 8 could successfully be substituted with some small substituents (e.g., $-NH_2$ or CH_3 at position 5, i.e., -F, -Cl, or $-OCH_3$ at position 8); these structural alterations could lead to an increased biological activity. The position 6 of the main scaffold is commonly reserved for a fluorine atom, which was found as the optimal one leading to an increased potency, however, it could also be replaced by some small substituents. Three positions (3, 4, and 7) on the main quinolone core were found as of exceptional importance for the 6-FQs antimycobacterial activity. The mechanistic studies revealed that the substituents attached at these positions could establish direct interaction not only with the enzyme (GyrA subunit), but also with the DNA molecule (Laponogov et al., 2009; Laponogov et al., 2010). Thus, the positions 3 and 4 should be occupied by a carboxyl and carbonyl group, respectively, as they lead to a significant enhancement of the 6-FQs potency through establishing a direct hydrogen-bonding interaction with the surrounding amino acid residues of the enzyme (Tillotson, 1996). On the other hand, the position 7 was recognized as a key attachment point for the biological activity of 6-FQ antibacterials, and aminopyrrolidinyl- or piperazinyl-like substituents were found as the most important, leading to an additional stabilization of the 6-FQ in the complex formed between the DNA gyrase enzyme and the mycobacterial DNA.

However, as a consequence of the considerable structural divergences between the available gyrase/topoisomerase crystal structures (Laponogov et al., 2009; Laponogov et al., 2010; Wohlkonig et al., 2010; Bax et al., 2010), the explicit protein-ligand interactions mainly remain unclear, and regrettably the 6-FQs binding mechanism is still a subject of speculation. Moreover, some recently confirmed quinolone-caused amino acids alterations largely located at the GyrA α2 and α3 helices delineating a key part of the quinolone-binding pocket (QBP), were found as one of the major determinants of the ineffectiveness of the current 6-FQs in the chemotherapy of tuberculosis (Johnson et al., 2006; Shi et al., 2006; Matrat et al., 2006; Groll et al., 2009). All these issues, represent a major challenge towards the structural optimization of the existing 6-FQs or even development of novel, more effective 6-FQ antituberculars, and nowadays, the integrated *in silico* drug design approaches (Figure 2) have proven to be skillfully applied in achieving these goals.

The structural information used in this study including the description of datasets of 6-FQs, i.e., DNA gyrase/topoisomerase crystal structures as well as all the methodologies employed are broadly elaborated in our previous publications (Minovski & Šolmajer, 2010; Minovski et al., 2011; Minovski et al., 2012, Minovski et al., 2013), and therefore we give here just a short summary.

6-FQs Compound Libraries

Two compound libraries of structurally similar 6-FQs coming from different sources (named as *ExpLib* and *CombiLib*, respectively) were primarily utilized as a starting point for QSAR predictive modeling, and later on to perform structure-based VS experiments for identification and selection of novel 6-FQ antitubercular agents.

The *ExpLib* library comprised of 145 structurally similar 6-FQs with experimentally measured biological activity values expressed as minimal inhibitory concentrations – MIC_{exp} [μg/mL] was collected from an online structural data source (Division of AIDS Anti-HIV/OI/TB Therapeutic database[18]; Minovski et al., 2011). According to the 6-FQs activity information obtained within the framework of various functional and biochemical studies performed so far, one could easily distinguish between the active and inactive 6-FQs comprising our *ExpLib* library (Aubrey et al., 2006; Pantel et al., 2011; Pantel et al., 2012); therefore, 114 out of total 145 6-FQs could be determined as active inhibitors against *M. tuberculosis* DNA gyrase enzyme ($MIC_{exp} \leq 1.0$ μg/mL), while the rest of the *ExpLib* molecules (31 compounds with $MIC_{exp} > 1.0$ μg/mL) belong to the class of inactive 6-FQs.

On the other hand, the *CombiLib* library comprised of 1.101 mixed "drug-like" 6-FQs was assembled *in-house*, by employing a classical virtual reaction-based combinatorial enumeration approach (Minovski & Šolmajer, 2010; Minovski et al., 2012). At the beginning, the original synthetic pathways of three most frequently used 6-FQs in tuberculosis chemotherapy, i.e., ciprofloxacin (CIP), moxifloxacin (MOX), and ofloxacin (OFL), were exploited as templates for construction of their virtual combinatorial synthetic schemes (Schwalbe et al., 2000; Martel et al., 1997; Serradel et al., 1983), while a commercial pre-filtered "*Ro3*" fragments library comprised of 6.995 building-blocks (Key Organics Bionet Fragment Library "Rule of 3"[19]) was utilized as a reactants subset for performing SAR-based fragmental permutations at the pre-defined 6-FQs variable scaffold positions (Figure 6b) for each virtual synthetic pathway separately (R_1 and R_7 substructural modifications for CIP and MOX, while only R_7 substructural modifications for OFL analogs). According to the current SAR knowledge of 6-FQs (Figure 6b), the position R_1 of the CIP and MOX scaffold was targeted by primary amines (e.g., R_1-NH_2), while the position R_7 in all three cases (CIP, MOX, and OFL) was targeted by secondary amines (e.g., R_1, R_2-NH). Moreover, although contrary to the current SAR recommendations, an additional structural diversity was introduced by attaching non-amino fragments (e.g., R-*) at the position R_7 in all three cases (CIP, MOX, and OFL). This resulted in obtaining a total of six reaction-based enumeration definitions (for more information check Minovski et al., 2012) for generation of six different 6-FQs virtual combinatorial libraries (Table 2).

As demonstrated in Table 2, the performed virtual reaction-based combinatorial enumerations yielded a general virtual combinatorial library comprised of total 53.871 compounds (CIP, MOX, and OFL structural analogs). Although, the resulting library could generally be considered as a small sub-space within

Table 2. Virtual combinatorial definitions and substructural fragments used for construction of six different virtual combinatorial libraries of 6-FQ structural analogs

ID	6-FQs	Virtual Combinatorial Definition		Substructural Fragments		$\rho_{tot} = N_i \times M_j$
		R_1	R_7	R_1 (N_i)	R_7 (M_j)	
1.	CIP-N_i-M_j	R_1-NH_2	R_2, R_3-NH	116	106	12.296
2.	CIP'-N_i-M_j	R_1-NH_2	R_2-*	115	191	21.965
3.	MOX-N_i-M_j	R_1-NH_2	R_2, R_3-NH	115	74	8.510
4.	MOX'-N_i-M_j	R_1-NH_2	R_2-*	73	147	10.731
5.	OFL-M_j	N/A	R_1, R_2-NH	N/A	180	180
6.	OFL'-M_j	N/A	R_1-*	N/A	189	189
Sum						**53.871**

(Minovski et al., 2012).

the available chemical space (Bohacek et al., 1996), one should be aware about the very low probability that each compound in such a virtual compound library possess desired "drug-like" properties (Walters et al., 1999; Hann et al., 2001; Muegge et al., 2003). Therefore, in order to distill only those combinatorially generated 6-FQs delineating the "drug-like" chemical sub-space, the entire virtual combinatorial library was subjected to a robust druggability properties assessment. Assuming that our generated 6-FQs would be administered *per os*, we constructed a combined Lipinski-Veber "drug-likeness" filtering tool for automated isolation of the desired sub-space of "drug-like" 6-FQs (Minovski et al., 2012). This filtering strategy produced a list of 1.101 "drug-like" virtual combinatorial 6-FQ analogs, which were further used as an external dataset for prediction of their unknown biological activity values ($MIC_{pred\text{-}combi}$ [μg/mL]) by employing a pre-constructed predictive QSAR model (Minovski et al., 2011, Minovski et al. 2012).

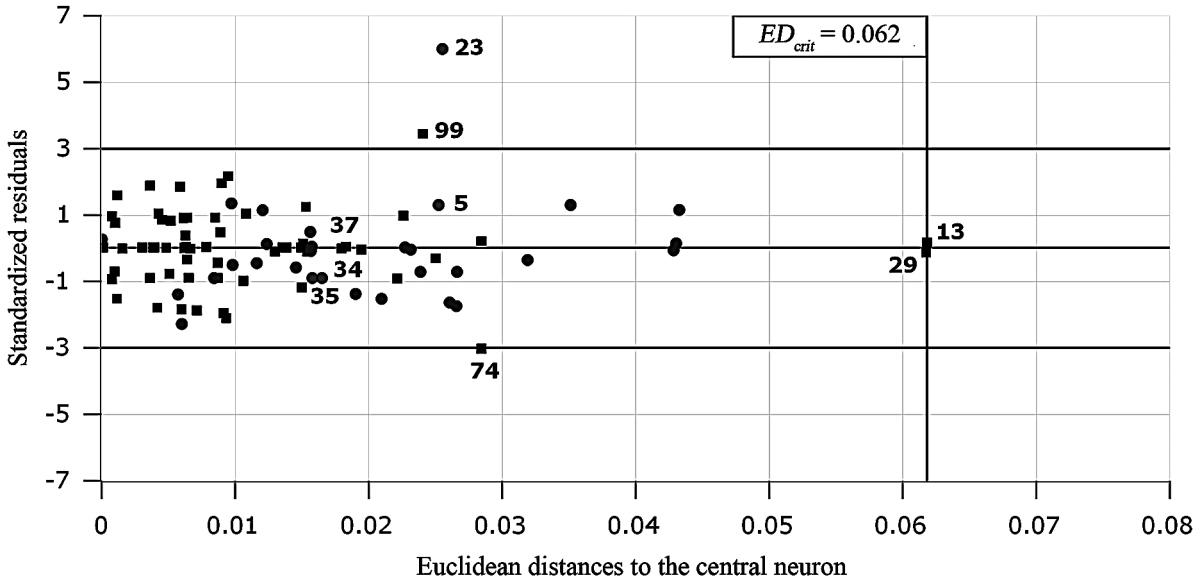
Development of an ANNs-Based Predictive QSAR Model

Following the integrated *in silico* screening protocol described previously (Figure 2), non-linear artificial neural networks (ANNs)-based predictive QSAR modeling was performed on the *ExpLib* library of 145 6-FQs using a comprehensive set of approximately 600 calculated 2D molecular descriptors (Minovski et al., 2011). The heuristic algorithm that is a step-wise selection procedure was initially used for descriptors pool reduction to a sub-pool comprised of up to 10 significant parameters, while its further reduction was achieved by construction of an inter-correlation descriptor matrix and subsequent elimination of those descriptors for which the inter-correlation coefficient is $R^2(P_i, P_m) \leq 0.40$. Here, a proposed upper R^2 value of 0.40 was used in order to avoid a possible chance correlation during the modeling (Topliss, 1983). A total of 7 molecular descriptors retained at the end of the parameters reduction/selection procedure, which further served as independent input variables for the ANNs model development. Kohonen artificial neural networks (KANNs) were employed for division of the compound's data (Kohonen, 1982; Novič & Zupan, 1995; Zupan & Gasteiger, 1999) on a training set (115 compounds) and an external validation set (30 compounds), while counter-propagation artificial neural networks (CP ANNs) were used for the purpose of modeling (Zupan et al., 1995; Zupan et al., 1997; Zupan & Gasteiger, 1999).

As stated previously, the modeling was performed solely on the training set objects, where various network architectures and number of learning epochs were extensively evaluated in the search for the most optimal CP ANN predictive model ($R_{tr} = 0.96$). The model was internally validated by using a CV LOO procedure ($R_{tr\text{-}cv} = 0.62$) as well as externally validated for its predictive performances using the previously excluded validation set objects ($Q_{ext} = 0.84$). Furthermore, the applicability domain (AD) of thus established and validated CP ANN predictive model was assessed by the *minimum Euclidean distance space* (MEDS) approach (Minovski, Župerl et al., 2013) – an efficient distance-based AD estimation method reflecting the reliability of the established CP ANN predictive model through utilization of the Euclidean distance (ED) metric in the model's structure-representation vector space (Figure 7).

As demonstrated in Figure 7, no external validation set compounds could be identified as outliers according to critical ED value ($ED_{crit} = 0.062$), i.e., a valuable information indicating that no significant structural differences exist between the investigated 6-FQs (structure-representation vector space). On the other hand, two training set objects (ID = 74 and 99) are somehow wrongly predicted by the model as demonstrated by their calculated standardized residual values, which are slightly above the $\pm 3\sigma$ boundaries. Among the external validation set objects, only one compound (ID = 23) could be determined as an extreme point with calculated standardized residual value around 6.0, i.e., a very poor biological activity prediction ($pMIC_{exp[ID=23]} = 0.7404$, $pMIC_{pred[ID=23]} = 0.2334$). The structural analysis

Figure 7. Graphical representation of the CP ANN predictive model's applicability domain assessed by the MEDS-based AD estimation method; the AD boundaries are defined by the training set object (ID = 13) with maximal ED to the central neuron (ED_{crit} = 0.062) and ±3σ units for the calculated standardized residuals (predictability of the model), respectively. The training set objects (115 compounds) are represented as solid dark-gray rectangles, while the external validation set objects (30 compounds) as solid light-gray circles.



of this compound shows that it belongs to a group of unconventional 6-FQs regarding the main scaffold (1,8-naphthyridine instead of quinoline moiety). In addition to this one, five more *ExpLib* compounds share the same structural feature, of which two (ID = 29 and 37) are training set objects, while three of them (ID = 5, 34, and 35) belong to the external validation set, and apparently all of them have acceptably predicted biological activity values, as they are situated within the boundaries of ±3σ units (Figure 7). However, (ID = 23) is the only *ExpLib* compound that has attached (*S*)-3-aminopirrolydinil group at the position R_7 of the naphthyridine ring and no similar compounds exist in the training set. These findings explicitly pinpoint to a certain degree of weakness of the established CP ANN model to correctly predict the biological activity values for naphthyridine-like 6-FQs with (*S*)-3-aminopirrolydinil fragments attached at the R_7 position. Anyhow, the rest of the external validation set objects are situated within the boundaries of the model's AD – a result that clearly confirms the reliability of the constructed CP ANN predictive model for its further utilization as an efficient tool for evaluation of the biological activities for novel, not yet synthesized 6-FQs.

Finally, the established CP ANN predictive model was used for prediction of the biological activity values ($MIC_{pred\text{-}combi}$ [μg/mL]) for our previously generated 1.101 "drug-like" 6-FQ combinatorial analogs (Minovski et al., 2012). The predicted biological activities for these compounds were in the range between $0.0021 \leq MIC_{pred\text{-}combi}$ [μg/mL] ≤ 6.3726, i.e., a mix of highly active 6-FQs, but also totally inactive representatives according to the *in vitro* inhibition assays performed so far (Aubrey et al., 2006; Pantel et al., 2011; Pantel et al., 2012). Taking this information into account, we defined a so-called global hypothetical activity (GHA) range $0.00 \leq MIC_{pred\text{-}combi}$ [μg/mL] ≤ 1.00, which served as

an efficient activity-based filtering tool for our 1.101 *CombiLib* compounds. This filtering procedure, resulted in selection of a subset of 427 "drug-like", but also hypothetically "active" 6-FQ combinatorial analogs, which were further used in structure-based calculations for final identification and selection of novel 6-FQs as potential *M. tuberculosis* DNA gyrase inhibitors.

Construction of *M. tuberculosis* DNA Gyrase Protein Homology Models for Molecular Docking Calculations

Although, various separate topoisomerase IIA subunits originating from different bacterial species are solved recently by X-ray crystallography (Laponogov et al., 2009; Laponogov et al., 2010; Wohlkonig et al., 2010; Bax et al., 2010; Tretter et al., 2010; Darmon et al., 2012; Bouige et al., 2013), unfortunately the entire crystal structure of the *M. tuberculosis* DNA gyrase holoenzyme in complex with the DNA molecule and an intercalated 6-FQ ligand still remains an indecipherable issue (Collin et al., 2011). Therefore, taking into consideration the recently proposed 6-FQs-topoisomerase binding mechanisms (Laponogov et al., 2010; Wohlkonig et al., 2010; Bax et al., 2010) as well as the structural and functional similarity between the both topoisomerase IIA paralogous forms (DNA gyrase and topoisomerase IV), we constructed three *M. tuberculosis* DNA gyrase protein homology models (Minovski et al., 2013).

The available topoisomerase IIA-DNA-6-FQ crystal structure complexes originating from three different bacterial species (*Streptococcus pneumoniae* topo IV-DNA-levofloxacin, PDB ID: 3K9F; *Acinetobacter baumannii* topo IV-DNA-moxifloxacin, PDB ID: 2XKK; and *Staphylococcus aureus* DNA gyrase-DNA-ciprofloxacin, PDB ID: 2XCT) and recently determined crystal structures of the separate *M. tuberculosis* DNA gyrase subunits - GyrA breakage-reunion domain, GyrA-BRD (PDB ID: 3IFZ) and GyrB-Toprim domain (PDB ID: 3M4I) were utilized for protein sequence alignment, while the proposed key amino acid sequences covering the QBP in *M. tuberculosis* DNA gyrase (Piton et al., 2010) were exploited for structural interchange of the original QBP sequences. It should be stressed that during the modeling, the nascent conformations of the DNA molecule, the intercalated ligands (levofloxacin [LFX], moxifloxacin [MOX], and ciprofloxacin [CIP], respectively), as well as the contributing co-factors present in the QBPs (e.g., water molecule(s) and/or Mg^{2+} ion(s)) were left intact. Moreover, the experimentally determined spatial coordinates of the co-crystallized 6-FQ conformations (LFX, MOX, and CIP, respectively) were used to define the QBP (a screening area with cavity radius of 12.5 Å) for each constructed homology model (for more details check Minovski et al., 2013).

The *M. tuberculosis* DNA gyrase protein homology models thus assembled and prepared (named as $3K9F_{mod}$, $2XKK_{mod}$, and $2XCT_{mod}$, respectively) which QBP emulates the one present in the wild-type enzyme (PDB IDs: 3IFZ and 3M4I), were further used as starting points to perform molecular docking calculations.

Molecular Docking Calculations and Protein Homology Models Validation

The molecular docking calculations on both 6-FQ compound libraries (*ExpLib* and *CombiLib*), within the previously defined QBP of each *M. tuberculosis* DNA gyrase protein homology model ($3K9F_{mod}$, $2XKK_{mod}$, and $2XCT_{mod}$, respectively), were performed by using the GOLD docking suite[20] (Jones et al., 1997). The entire docking procedure (including all the GA settings and required technical parameters) for each protein homology model was accomplished as described in our previous work (Minovski et al., 2013), while the GOLDScore Fitness (GSF) function was used for estimation of the 6-FQs binding affinity.

At the beginning, the quality of the constructed protein homology models was confirmed by an initial re-docking validation run (ligand reproduction assessment) on the co-crystallized ligand conformations (LFX, MOX, and CIP, respectively) present within the QBP of each protein homology model separately, followed by heavy-atoms RMSD (Å) comparison between the experimental ligand conformations and their corresponding GOLD-derived ligand poses (Figure 8a and Table 3).

As demonstrated in Table 3, a total of three dock poses were computed for each co-crystallized ligand. Except $2XKK_{mod}$ model for which the initial validation failed according to the calculated RMSD (Å) values for MOX-reproduced dock poses (RMSD > 2.0 Å; an apparently poor model), the rest two homology models ($3K9F_{mod}$ and $2XCT_{mod}$) successfully reproduced their experimental ligand conformations (LFX and CIP, respectively) with calculated RMSD values significantly below 2.0 Å (Verdonk et al., 2003). These preliminary validation results suggest the following order of quality of our constructed *M. tuberculosis* DNA gyrase protein homology models ($2XKK_{mod} < 2XCT_{mod} < 3K9F_{mod}$), which result could also be visually determined (Figure 8a).

However, to thoroughly assess their quality as well as reliability as efficient VS filters, a second validation experiment of their discriminatory performances (an investigation of the models capability to correctly discriminate between known active and inactive compounds) was carried out as described previously (Hevener et al., 2009). For that purposes, the *ExpLib* library of 145 6-FQs was first docked within the QBP of each protein homology model ($3K9F_{mod}$, $2XKK_{mod}$, and $2XCT_{mod}$, respectively) by using the same GA settings, and afterwards only the subset of 114 active top-scored dock poses ($MIC_{exp} \leq 1.0$ μg/mL) was enriched with a multiconformer library of total 13.990 artificial decoy molecules randomly selected from the Asinex Elite Library[21] (Minovski et al., 2013). Thus prepared, the library of active/decoy molecules was subjected to a robust similarity screening against the co-crystallized ligand conformations (LFX, MOX, and CIP, respectively) used as queries (Huang et al., 2006). vROCS tool[22] was used to evaluate the VS discriminatory performances of the constructed protein homology models (ROC-AUC and early enrichment parameters at 0.5, 1.0, and 2.0% retrieved within ±95% confidence interval) directly from the generated ROC curves (Figure 8b and Table 4).

Table 4 summarizes the statistical parameters carrying the discriminatory performances of our constructed *M. tuberculosis* DNA gyrase protein homology models as efficient VS filters calculated directly from the generated ROC curves where all the comparisons are performed relative to the $3K9F_{mod}$ model (Figure 8b). As displayed here, according to the highest calculated ROC-AUC value the best VS performances could be assigned to $3K9F_{mod}$ model ($ROC\text{-}AUC_{[3K9Fmod]} = 0.844$), $2XCT_{mod}$ has somehow moderate VS performances ($ROC\text{-}AUC_{[2XCTmod]} = 0.826$), while $2XKK_{mod}$ could be identified as a poorest one ($ROC\text{-}AUC_{[2XKKmod]} = 0.815$). These findings are also supported by the calculated *p*-values – a probability for an investigated model to give a better outcome then the reference one (in our case $3K9F_{mod}$) under assumption that the null hypothesis is true. As demonstrated, the calculated ROC-AUC *p*-values for $2XKK_{mod}$ and $2XCT_{mod}$ models both tend toward 0 ($p\text{-value}_{[2XKKmod]} = 0.192$ and $p\text{-value}_{[2XCTmod]} = 0.299$, respectively), i.e., the probability for these two models to give a better outcome (identification of more active molecules than the reference model) is practically very low and this result is not due to a random chance (p-value < 0.5; the null hypothesis could be rejected). Similarly, the probability for $2XKK_{mod}$ to give better outcome than $2XCT_{mod}$ is very low as well ($p\text{-value}_{[2XCTmod]} > p\text{-value}_{[2XKKmod]}$), which results are additionally grounded by the estimated early enrichment factors (calculated at 0.5, 1.0, and 2.0% in the boundaries of ±95% confidence interval) and their corresponding probability values. Moreover, these results are also congruent with the initial re-docking validation results (Figure 8a and Table 3), allowing us to confirm our previously established quality order of our protein homology models (2XK-

Figure 8. Validation of the constructed M. tuberculosis DNA gyrase protein homology models. (a) Re-docking of the co-crystallized ligand conformations (LFX, MOX, and CIP, respectively) within the QBP of each protein homology model (3K9F$_{mod}$, 2XKK$_{mod}$, and 2XCT$_{mod}$), separately. The experimental ligand conformations and the corresponding GOLD-calculated ligand poses are depicted in dark and light gray, respectively (stick representation), the water molecule(s) are represented as black sphere(s), while Mg^{2+} ion as a light gray sphere. (b) Protein homology models discriminatory performances for correct identification of active and inactive 6-FQs assessed by ROC methodology. The diagonal line (line of no discrimination) corresponds to the randomly distributed data (RDD = 0.5).

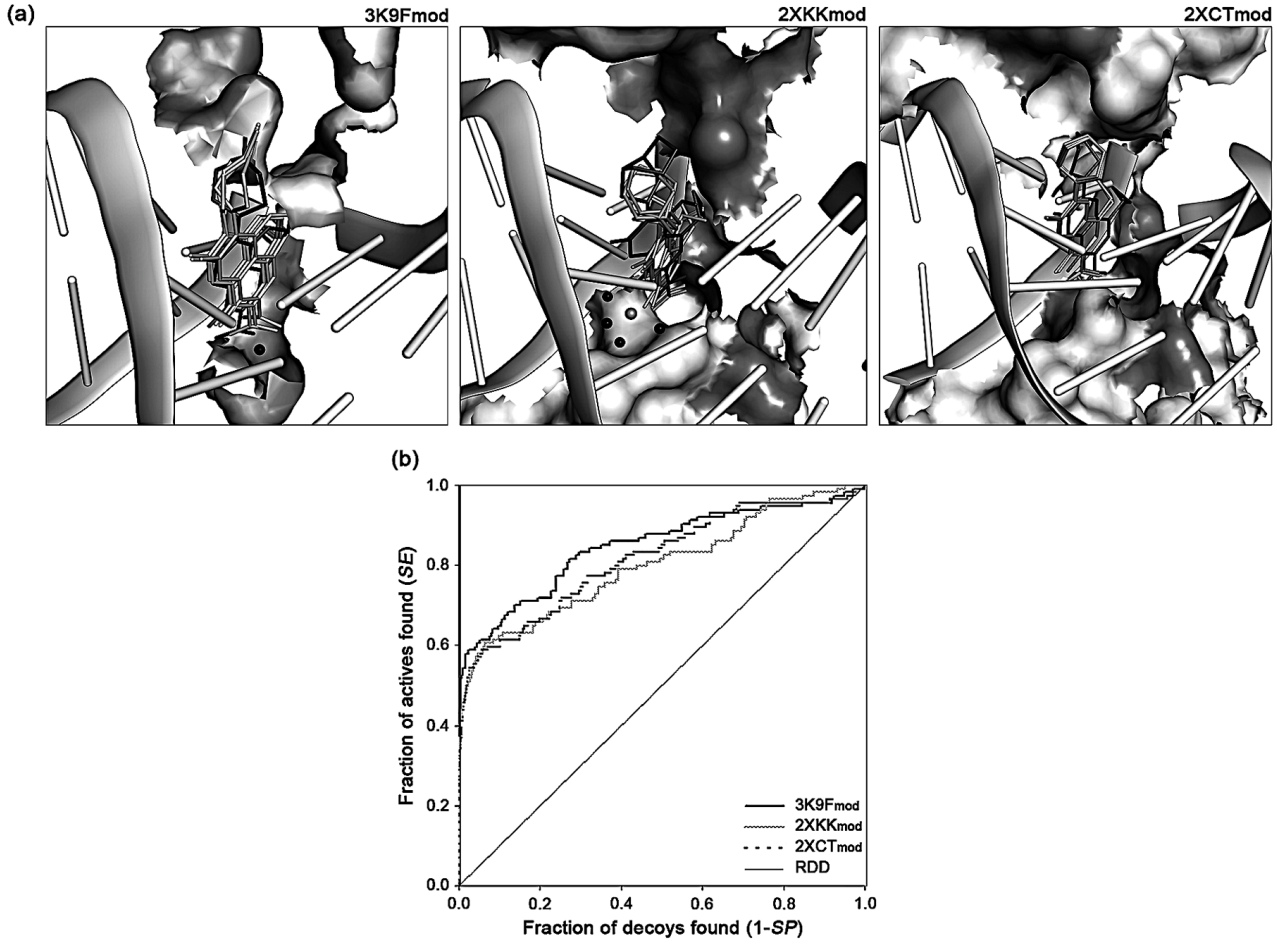


Table 3. Ligand reproduction assessment for initial validation of the quality of the assembled M. tuberculosis DNA gyrase protein homology models (3K9F$_{mod}$, 2XKK$_{mod}$, and 2XCT$_{mod}$, respectively) performed by heavy-atoms RMSD (Å) comparison between each experimental ligand conformation (LFX, MOX, and CIP, separately) and their corresponding GOLD-computed dock poses

Model	**3K9F$_{mod}$ [LFX]**			**2XKK$_{mod}$ [MOX]**			**2XCT$_{mod}$ [CIP]**		
Dock pose	pose 1	pose 2	pose 3	pose 1	pose 2	pose 3	pose 1	pose 2	pose 3
RMSD (Å)	1.056	1.121	1.027	2.553	2.652	2.488	1.249	1.311	1.340

Table 4. The statistical parameters describing the VS performances of our constructed M. tuberculosis DNA gyrase protein homology models, calculated directly from the obtained ROC curves: ROC-AUC (area under the ROC curve), EF (enrichment factor), and p-values (the probability for a model to retrieve better outcome then the reference one, assuming that the null hypothesis is true). All comparisons are performed relative to the $3K9F_{mod}$ model used as reference.

Model	$3K9F_{mod}$	$2XKK_{mod}$	*p*-Value	$2XCT_{mod}$	*p*-value
ROC-AUC	0.844 [0.795, 0.891]	0.815 [0.769, 0.864]	0.192	0.826 [0.774, 0.870]	0.299
EF (0.5%)	85.464 [66.67, 105.60]	68.829 [50.88, 89.23]	0.108	77.032 [57.14, 97.48]	0.273
EF (1.0%)	52.046 [45.50, 61.46]	43.473 [34.62, 52.73]	0.091	46.779 [36.84, 56.67]	0.215
EF (2.0%)	27.537 [22.80, 32.66]	24.786 [20.08, 29.31]	0.203	27.365 [23.61, 32.00]	0.479

$K_{mod} < 2XCT_{mod} < 3K9F_{mod}$). Based on both validation experiments, the $3K9F_{mod}$ could be recognized as a model with by far the best VS discriminatory performances, and therefore it was selected for further utilization as a robust *in silico* filtering device for VS of our novel 6-FQ combinatorial analogs.

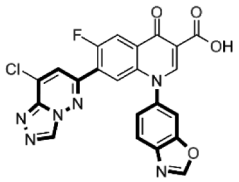
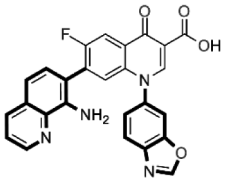
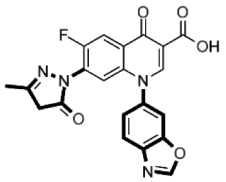
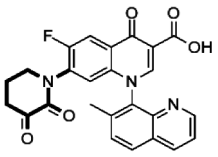
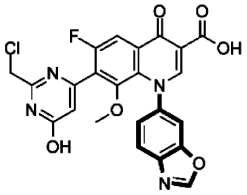
Finally, the *CombiLib* library of 427 "drug-like" 6-FQ combinatorial analogs was docked into QBP of the selected *M. tuberculosis* DNA gyrase protein homology model ($3K9F_{mod}$), while the experimental conformation of the co-crystallized LFX ligand present within the QBP was used as a template structure against which an *in-depth* post-docking analysis of the GOLD-computed docking solutions was performed (Minovski et al., 2013).

Boolean-Based Clustering for Identification and Selection of Novel 6-FQ Hits

As described previously, the generated docking solutions are usually stored in a form of extended virtual compound libraries where all computed dock poses per ligand are organized as small clusters. Therefore, to thoroughly assess all the generated docking solutions for our 427 "drug-like" *CombiLib* 6-FQ analogs (a total of 1281 poses arranged into 427 clusters, i.e., 3 poses per cluster) as well as to identify and select the most promising 6-FQ hit candidates, each calculated dock pose was subjected to a detailed post-docking analysis based on their visual examination using the available structural data.

Following the Boolean-based (T/F) clustering method described in the previous section, the post-docking analysis was carried out in three consecutively-coupled levels: geometric properties assessment, score-based clustering, and activity-based clustering (Minovski et al., 2013). Within the scope of the first level of the performed post-docking analysis (geometric properties assessment), a total of 162 (T)-signed 6-FQ analogs were initially identified as geometrically well positioned relative to the experimental co-crystallized LFX conformation. The structural analysis of these compounds clearly reflected the capability of the selected protein homology model $3K9F_{mod}$ to correctly identify 6-FQ structural analogs that belong to different structural classes (CIP, MOX, and OFL *CombiLib* compounds). All (T)-signed 6-FQ combinatorial analogs thus selected were extracted as a separate cluster and used in the second level of the Boolean-based (T/F) post-docking analysis by taking into consideration their estimated GSF function (score-based clustering). Implementing a pre-defined GSF threshold of ($GSF \geq 80$), a total of 92 top-scored clusters were determined and accordingly a total of 92 (T)-signed top-scored 6-FQ combinatorial analogs were selected. Similarly, the selected docking solutions were distilled as a separate cluster and

Table 5. Some promising CombiLib 6-FQ representatives identified by using the Boolean-based (T/F) post-docking analysis; the most frequently occurring substructural fragments at R_1 and R_7 position are represented in bold.

ID	Code	*CombiLib* Hits	R_1	R_7	$MIC_{pred\text{-}combi}$ [μg/mL]	GSF
1.	CIP'-**028-059**		**028**	**059**	0.0499	83.55
2.	CIP'-**028-073**		**028**	**073**	0.0021	88.89
3.	CIP-**028-102**		**028**	**102**	0.013	80.87
4.	CIP-049-**096**		049	**096**	0.0329	86.57
5.	MOX'-**016**-137		**016**	137	0.03	89.89

used as an input in the final level of the Boolean-based (T/F) post-docking analysis (activity-based clustering) by taking into account their QSAR-predicted biological activity values ($MIC_{pred\text{-}combi}$ [μg/mL]). It should be stressed once again that all 6-FQ analogs comprising our *CombiLib* library are somehow hypothetically active as inhibitors of *M. tuberculosis* DNA gyrase enzyme according to the previously implemented GHA range ($0.00 \leq MIC_{pred\text{-}combi}$ [μg/mL] ≤ 1.00). Nevertheless, in order to extract the most "active" 6-FQ hits, the cluster of (T)-signed highly-scored hits assembled at the end of the previous level, was subjected to an additional activity-based filtering routine ($MIC_{pred\text{-}combi} \leq$ 0.05 μg/mL). In this procedure, a total of 48 (T)-signed combinatorial hits were identified as most "active", mainly CIP and MOX structural analogs of which the most promising are listed in Table 5, while no OFL combinatorial compounds were identified at the end of the Boolean-based (T/F) post-docking analysis.

The substructural examination of the building-blocks attached at the pre-defined variable scaffold positions of the isolated 6-FQ combinatorial hits revealed the most frequently occurring fragments (Table 5). Namely, at R_1 position benzo[*d*]oxazole was found as the most frequently occurring one (fragment **028** in CIP- and CIP'-analogs, which is relevant to the fragment **016** in MOX'-analogs). On the other hand, four fragments were identified as the most frequently appearing at R_7 position including 8-chloro-6-methyl-[1,2,4]triazolo[4,3-*b*]pyridazine (fragment **059**), 7-methylquinolone-8-amine (fragment **073**), 1-methylpiperidine-2,3-dione (fragment **096**), and 1,3-dimethyl-1*H*-pyrazol-5(4*H*)-one (fragment **102**). From a medicinal chemistry perspective, these fragments are mainly small aromatic N-heterocyclic systems with molecular weight between 98 and 155 g/mol, which contain more than two HBA atoms on average – an important SAR feature that undoubtedly increases the probability of establishing HB interactions between 6-FQs and surrounding amino acid residues of the GyrA/GyrB subunit of the enzyme.

CONCLUSION

In the current era of enormous technological advancements, we have witnessed a rapid breakthrough in the modern design and optimization of novel drug candidates. Nowadays, the *in silico* drug design methods are indeed an irreplaceable supplement to the experimentally grounded approaches not only in the early hit(s) identification and hit-to-lead stages, but also in the late points of the drug discovery pipeline. From the profusion of *in silico* drug discovery methods currently available, various two- and three-dimensional approaches (e.g., QSAR, ligand-based, structure-based, etc.) became a cornerstone of the modern drug discovery for rapid and efficient development of novel successful drug candidates. Unfortunately, it was found that their individual utilization as *in silico* ligand filters could frequently result in a very small number of newly identified or *de novo* developed drug candidates, and therefore, more and more efforts are currently devoted to a skillful, rational, and knowledge-based development of integrated *in silico* drug discovery platforms.

In this chapter, we illustrated such an integrated *in silico* screening platform for fast and efficient identification of novel drug candidates from a large number of possibilities. A variety of well-established *in silico* drug discovery methods were covered and their specific assemblage into an efficient screening integration for identification and selection of novel hit candidates was thoroughly reviewed. Finally, we illustrated its practical application in a case study related to the design and identification of novel 6-FQ antibacterials as potential *M. tuberculosis* DNA gyrase inhibitors and proposed some new SAR guidelines. In conclusion, we believe that the future of the modern drug development resides in the implementation of such integrated *in silico* screening schemes, which could aid not only the development of novel and potent drugs, but could also introduce some novel standards in the on-going drug discovery programs.

REFERENCES

Abuhamdah, S., Habash, M., & Taha, M. O. (2013). Elaborate ligand-based modeling coupled with QSAR analysis and *in silico* screening reveal new potent acetylcholinesterase inhibitors. *Journal of Computer-Aided Molecular Design*, *27*(12), 1075–1092. doi:10.1007/s10822-013-9699-6 PMID:24338032

Agrafiotis, D. K. (2000). Multiobjective optimization of combinatorial libraries. *Molecular Diversity*, *5*(4), 209–230. doi:10.1023/A:1021320124615 PMID:12549673

Agrafiotis, D. K., Lobanov, V. S., & Salemme, F. R. (2002). Combinatorial informatics in the post-genomic era. *Nature Reviews. Drug Discovery*, *1*(5), 337–346. doi:10.1038/nrd791 PMID:12120409

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & Biology*, *10*(9), 787–797. doi:10.1016/j.chembiol.2003.09.002 PMID:14522049

Andersson, P. M., Sjöström, M., Wold, S., & Lundstedt, T. (2001). Strategies for subset selection of parts of an in-house chemical library. *Journal of Chemometrics*, *15*(4), 353–369. doi:10.1002/cem.671

Aronov, A. M. (2002). Design of virtual combinatorial libraries. In L. Bellavance English (Ed.), *Methods in molecular biology, combinatorial library: Methods and protocols* (Vol. 201, pp. 267–276). Totowa, NJ: Humana Press; doi:10.1385/1-59259-285-6:267

Aubrey, A., Veziris, N., Cambau, E., Truffot-Pernot, C., Jarlier, V., & Fisher, L. M. (2006). Novel gyrase mutations in quinolone-resistant and –hypersusceptible clinical isolates of *Mycobacterium tuberculosis*: Functional analysis of mutant enzymes. *Antimicrobial Agents and Chemotherapy*, *50*(1), 104–112. doi:10.1128/AAC.50.1.104-112.2006 PMID:16377674

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Reviews. Drug Discovery*, *1*(11), 882–894. doi:10.1038/nrd941 PMID:12415248

Bajot, F. (2006). The use of QSAR and computational methods in drug design. In T. Puzyn, J. Leszczynski, & M. T. D. Cronin (Eds.), *Recent advances in QSAR studies: Methods and applications* (pp. 261–282). Springer Dordreht.

Barril, X. (2012). Druggability predictions: Methods, limitations, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(4), 327–338. doi:10.1002/wcms.1134

Bartfai, T., & Lees, G. V. (2006). The business basics (general). In T. Bartfai & G. V. Lees (Eds.), *Drug discovery: From bedside to Wall Street* (pp. 183–192). Oxford, UK: Elsevier Academic Press.

Bax, B. D., Chan, P. F., Eggleston, D. S., Fosberry, A., Gentry, D. R., & Gorec, F. et al. (2010). Type IIA topoisomerase inhibition by a new class of antibacterial agents. *Nature*, *466*(7309), 935–940. doi:10.1038/nature09197 PMID:20686482

Bhal, S. K., Kassam, K., Peirson, I. G., & Pearl, G. M. (2007). The rule of five revisited: Applying logD in place of logP in drug-likeness filters. *Molecular Pharmaceutics*, *4*(4), 556–560. doi:10.1021/mp0700209 PMID:17530776

Bharath, E. N., Manjula, S. N., & Vijaychand, A. (2011). *In Silico* drug design – Tool for overcoming the innovation deficit in the drug discovery process. *International Journal of Pharmacy and Pharmaceutical Sciences*, *3*(2), 8–12.

Bleicher, K. H., Böhm, H.-J., Müller, K., & Alanine, A. I. (2003). Hit and lead generation: Beyond high-throughput screening. *Nature Reviews. Drug Discovery*, *2*(5), 369–378. doi:10.1038/nrd1086 PMID:12750740

Bohacek, R. S., mcMartin, C., & Guida, W. C. (1996). The art and practice of structure-based drug design. *Medicinal Research Reviews*, *16*(1), 3–50. doi:10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6 PMID:8788213

Borchardt, R. T. (2004). Scientific, educational and communication issues associated with integrating and applying drug-like properties in drug discovery. In R. Borchardt, E. Kerns, C. Lipinski, D. Thakker, & B. Wang (Eds.), *Pharmacological profiling in drug discovery for lead selection* (pp. 451–466). Arlington, VA: AAPS Press.

Bouige, A., Darmon, A., Piton, J., Roue, M., Petrella, S., & Capton, E. et al. (2013). *Mycobacterium tuberculosis* DNA gyrase possesses two functional GyrA-boxes. *The Biochemical Journal*, *455*(3), 285–294. doi:10.1042/BJ20130430 PMID:23869946

Bouvier, G., Evrard-Todeschi, N., Girault, J.-P., & Bertho, G. (2010). Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics (Oxford, England)*, *26*(1), 53–60. doi:10.1093/bioinformatics/btp623 PMID:19910307

Burden, F. R., & Winkler, D. A. (1999). New QSAR method applied to structure-activity mapping and combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, *39*(2), 236–242. doi:10.1021/ci980070d

Cavazzani, P., & Rajagopal, N. A. (2010). Market analysis of growing competition in pharmaceutical industry. *International Journal of Business Competition and Growth*, *1*(1), 31–44. doi:10.1504/IJBCG.2010.032827

Chabala, J. C. (1995). Solid-phase combinatorial chemistry and novel tagging methods for identifying leads. *Current Opinion in Biotechnology*, *6*(6), 632–639. doi:10.1016/0958-1669(95)80104-9 PMID:8527833

Chen, C. Y. (2013). A novel integrated framework and improved methodology of computer-aided drug design. *Current Topics in Medicinal Chemistry*, *13*(9), 965–988. doi:10.2174/1568026611313090002 PMID:23651478

Cheng, T., Li, Q., Zhou, Z., Wang, Y., & Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: A problem-centric review. *The AAPS Journal*, *14*(1), 133–141. doi:10.1208/s12248-012-9322-0 PMID:22281989

Chirico, N., & Gramatica, P. (2011). Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *Journal of Chemical Information and Modeling*, *51*(9), 2320–2335. doi:10.1021/ci200211n PMID:21800825

Clark, D. E. (2003). In silico prediction of blood-brain barrier permeation. *Drug Discovery Today*, *8*(20), 927–933. doi:10.1016/S1359-6446(03)02827-7 PMID:14554156

Coe, D. M., & Storer, R. (1999). Solution-phase combinatorial chemistry. *Molecular Diversity*, *4*(1), 31–38. doi:10.1023/A:1009694409264 PMID:10320987

Coi, A., & Bianucci, A. M. (2013). Combining structure- and ligand-based approaches for studies of interactions between different conformations of hERG K^+ channel pore and known ligands. *Journal of Molecular Graphics & Modelling*, *46*, 93–104. doi:10.1016/j.jmgm.2013.10.001 PMID:24185260

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., & Harris, D. et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, *393*(6685), 537–544. doi:10.1038/31159 PMID:9634230

Collin, F., Karkare, S., & Maxwell, A. (2011). Exploiting bacterial DNA gyrase as a drug target: Current state and perspectives. *Applied Microbiology and Biotechnology*, *92*(3), 479–497. doi:10.1007/s00253-011-3557-z PMID:21904817

Congreve, M., Carr, R., Murray, C. W., & Jhoti, H. (2003). A "rule of three" for fragment-based lead discovery. *Drug Discovery Today*, *8*(19), 876–877. doi:10.1016/S1359-6446(03)02831-9 PMID:14554012

Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the definition of the Q^2 parameter for QSAR validation. *Journal of Chemical Information and Modeling*, *49*(7), 1669–1678. doi:10.1021/ci900115y PMID:19527034

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of Chemometrics*, *24*(3-4), 194–201. doi:10.1002/cem.1290

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, *110*(18), 5959–5967. doi:10.1021/ja00226a005 PMID:22148765

Cruz-Monteagudo, M., Borges, F., Cordeiro, M. N. D. S., Fajin, J. L. C., Morell, C., & Ruiz, R. M. et al. (2008). Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. *Journal of Combinatorial Chemistry*, *10*(6), 897–913. doi:10.1021/cc800115y PMID:18855460

Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M., & Chen, H. (2013). Chemical predictive modeling to improve compound quality. *Nature Reviews. Drug Discovery*, *12*(12), 948–962. doi:10.1038/nrd4128 PMID:24287782

Darmon, A., Piton, J., Roué, M., Petrella, S., Aubrey, A., & Mayer, C. (2012). Purification, crystallization and preliminary X-ray crystallographic studies of the *Mycobacterium tuberculosis* DNA gyrase CTD. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *68*(Pt 2), 178–180. doi:10.1107/S1744309111051888 PMID:22297993

Deng, Z.-L., Du, C.-X., Li, X., Hu, B., Kuang, Z.-K., & Wang, R. et al. (2013). Exploring the biologically relevant chemical space for drug discovery. *Journal of Chemical Information and Modeling*, *53*(11), 2820–2828. doi:10.1021/ci400432a PMID:24125686

Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., & Stallings, W. C. et al. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of Medicinal Chemistry*, *45*(11), 2213–2221. doi:10.1021/jm010548w PMID:12014959

Drlica, K., & Malik, M. (2003). Fluoroquinolones: Action and resistance. *Current Topics in Medicinal Chemistry*, *3*(3), 249–282. doi:10.2174/1568026033452537 PMID:12570763

Drwal, M. N., & Griffith, R. (2013). Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today. Technologies*, *10*(3), e395–e401. doi:10.1016/j.ddtec.2013.02.002 PMID:24050136

Ekins, S., Crumb, W. J., Sarazan, R. D., Wikel, J. H., & Wrighton, S. A. (2002). Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *The Journal of Pharmacology and Experimental Therapeutics*, *301*(2), 427–434. doi:10.1124/jpet.301.2.427 PMID:11961040

Ekins, S., Mestres, J., & Testa, B. (2007). *In silico* pharmacology for drug discovery: Methods for virtual ligand screening and profiling. *British Journal of Pharmacology*, *152*(1), 9–20. doi:10.1038/sj.bjp.0707305 PMID:17549047

Ellingson, S. R., & Baudry, J. (2012). High-throughput virtual molecular docking with AutoDockCloud. *Concurrency and Computation*, *26*(4), 907–916. doi:10.1002/cpe.2926

Ellingson, S. R., Dakshanamurthy, S., Brown, M., Smith, J. C., & Baudry, J. (2013). Accelerating virtual high-throughput ligand docking: Current technology and case study on a petascale supercomputer. *Concurrency and Computation*. doi:10.1002/cpe PMID:24729746

Ellingson, S. R., Smith, J. C., & Baudry, J. (2013). VinaMPI: Facilitating multiple receptor high-throughput virtual docking on high-performance computers. *Journal of Computational Chemistry*, *34*(25), 2212–2221. doi:10.1002/jcc.23367 PMID:23813626

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives*, *111*(10), 1361–1375. doi:10.1289/ehp.5758 PMID:12896860

Esposito, E. X., Hopfinger, A. J., & Madura, J. D. (2004). Methods for applying the quantitative structure-activity relationship paradigm. In J. Bajorath (Ed.), *Methods in molecular biology, chemoinformatics: Concepts, methods, and tools for drug discovery* (Vol. 275, pp. 131–213). Totowa, NJ: Humana Press; doi:10.1385/1-59259-802-1:131

Ferrero, L., Cameron, B., Manse, B., Lagneaux, D., Crouzet, J., Famechon, A., & Blanche, F. (1994). Cloning and primary structure of Staphylococcus aureus DNA topoisomerase IV: A primary target of fluoroquinolones. *Molecular Microbiology*, *13*(4), 641–653. doi:10.1111/j.1365-2958.1994.tb00458.x PMID:7997176

Free, S. J., & Wilson, J. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, *7*(4), 395–399. doi:10.1021/jm00334a001 PMID:14221113

Gedeck, P., & Lewis, R. A. (2008). Exploiting QSAR models in lead optimization. *Current Opinion in Drug Discovery & Development*, *11*(4), 569–575. PMID:18600573

Gellert, M., Mizuuchi, K., O'Dea, M. H., Itoh, T., & Tomizawa, J.-I. (1977). Nalidixic acid resistance: A second genetic character involved in DNA gyrase activity. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 4772–4776. doi:10.1073/pnas.74.11.4772 PMID:337300

Ghosh, S., Nie, A., An, J., & Huang, Z. (2006). Structure-based virtual screening of chemical libraries for drug discovery. *Current Opinion in Chemical Biology*, *10*(3), 194–202. doi:10.1016/j.cbpa.2006.04.002 PMID:16675286

Gohlke, H., & Klebe, G. (2001). Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology*, *11*(2), 231–235. doi:10.1016/S0959-440X(00)00195-0 PMID:11297933

Golbraikh, A., & Tropsha, A. (2002). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*, *16*(5-6), 357–369. doi:10.1023/A:1020869118689 PMID:12489684

Gozalbes, R., Mosulén, S., Ortí, L., Rodríguez-Díaz, J., Carbajo, R. J., Melnyk, P., & Pineda-Lucena, A. (2013). Hit identification of novel heparanase inhibitors by structure- and ligand-based approaches. *Bioorganic & Medicinal Chemistry*, *21*(7), 1944–1951. doi:10.1016/j.bmc.2013.01.033 PMID:23415087

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, *331*(1), 281–299. doi:10.1016/S0022-2836(03)00670-3 PMID:12875852

Gribbon, P., & Sewing, A. (2005). High-throughput drug discovery: What can we expect from HTS? *Drug Discovery Today*, *10*(1), 17–22. doi:10.1016/S1359-6446(04)03275-1 PMID:15676295

Groll, A. V., Martin, A., Jureen, P., Hoffner, S., Vandamme, P., & Portaels, F. et al. (2009). Fluoroquinolone resistance in *Mycobacterium tuberculosis* and mutations on *gyrA* and *gyrB*. *Antimicrobial Agents and Chemotherapy*, *53*(10), 4498–4500. doi:10.1128/AAC.00287-09 PMID:19687244

Grzybowski, B. A., Ishchenko, A. V., Kim, C.-Y., Topalov, G., Chapman, R., & Christianson, D. W. et al. (2002). Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(3), 1270–1273. doi:10.1073/pnas.032673399 PMID:11818565

Gussio, P., Pattabiraman, N., Zaharevitz, D. W., Kellogg, G. E., Topol, I. A., & Rice, W. G. et al. (1996). All-atom models for the non-nucleoside binding site of HIV-1 reverse transcriptase complexed with inhibitors: A 3D QSAR approach. *Journal of Medicinal Chemistry*, *39*(8), 1645–1650. doi:10.1021/jm9508088 PMID:8648604

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, *47*(4), 409–443. doi:10.1002/prot.10115 PMID:12001221

Hann, M. M., Leach, A. R., & Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of Chemical Information and Computer Sciences*, *41*(3), 856–864. doi:10.1021/ci000403i PMID:11410068

Hann, M. M., & Oprea, T. I. (2004). Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology*, *8*(3), 255–263. doi:10.1016/j.cbpa.2004.04.003 PMID:15183323

Hansch, C. J., & Fujita, T. (1964). ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626. doi:10.1021/ja01062a035

Harris, C. J., Hill, R. D., Sheppard, D. W., Slater, M. J., & Stouten, P. F. W. (2011). The design and application of target-focused compound libraries. *Combinatorial Chemistry & High Throughput Screening*, *14*(6), 521–531. doi:10.2174/138620711795767802 PMID:21521154

Hecker, E. A., Duraiswami, C., Andrea, T. A., & Diller, D. J. (2002). Use of ctalyst pharmacophore models for screening of large combinatorial libraries. *Journal of Chemical Information and Modeling*, *42*(5), 1204–1211. doi:10.1021/ci020368a PMID:12377010

Hevener, K. E., Zhao, W., Ball, D. M., Babaoglu, K., Qi, J., White, S. W., & Lee, R. E. (2009). Validation of molecular docking programs for virtual screening against dihydropteroate synthase. *Journal of Chemical Information and Modeling*, *49*(2), 444–460. doi:10.1021/ci800293n PMID:19434845

Hopkins, A. L. (2008). Pharmacological space. In C. G. Wermuth (Ed.), *The practice of medicinal chemistry* (3rd ed.; pp. 521–532). Amsterdam, Netherlands: Academic Press/Elsevier. doi:10.1016/B978-0-12-374194-3.00025-1

Hsu, C. –H., Lin, C. –Y., Ouyang, M., & Guo, Y. K. (2013). Biocloud: Cloud computing for biological, genomics, and drug design. *BioMed Research International, 2013*(Article ID 909470), 1-3. doi: 10.1155/2013/909470

Hu, S., Yu, H., Zhao, L., Liang, A., Liu, Y., & Zhang, H. (2013). Molecular docking and 3D-QSAR studies on checkpoint kinase 1 inhibitors. *Medicinal Chemistry Research*, *22*(10), 4992–5013. doi:10.1007/s00044-013-0471-1

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, *49*(23), 6789–6801. doi:10.1021/jm0608356 PMID:17154509

Huang, S.-Y., Grinter, S. Z., & Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions. *Physical Chemistry Chemical Physics*, *12*(40), 12899–12908. doi:10.1039/c0cp00151a PMID:20730182

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x PMID:21091654

Jahn, A., Rosenbaum, L., Hinselmann, G., & Zell, A. (2011). 4D flexible atom-pairs: An efficient probabilistic conformation space comparison for ligand-based virtual screening. *Journal of Cheminformatics*, *3*(23), 1–17. doi:10.1186/1758-2946-3-23 PMID:21214931

Jain, A. N. (2006). Scoring functions for protein-ligand docking. *Current Protein & Peptide Science*, *7*(5), 407–420. doi:10.2174/138920306778559395 PMID:17073693

Johnson, R., Streicher, E. M., Louw, G. E., Warren, R. M., van Helden, P. D., & Victor, T. C. (2006). Drug resistance in *Mycobacterium tuberculosis. Current Issues in Molecular Biology*, *8*(2), 97–111. PMID:16878362

Jones, G., Willet, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, *267*(3), 727–748. doi:10.1006/jmbi.1996.0897 PMID:9126849

Kamaria, P., & Kawathekar, N. (2014). Ligand-based 3D-QSAR analysis and virtual screening in exploration of new scaffolds as *Plasmodium falciparum* glutathione reductase inhibitors. *Medicinal Chemistry Research*, *23*(1), 25–33. doi:10.1007/s00044-013-0603-7

Katritzky, A. R., Karelson, M., & Lobanov, V. S. (1997). QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure and Applied Chemistry*, *69*(2), 245–248. doi:10.1351/pac199769020245

Kenny, B. A., Bushfield, M., Parry-Smith, D. J., Fogarty, S., & Treherne, J. M. (1998). The application of high-throughput screening to novel lead discovery. In E. Jucker (Ed.), *Progress in drug research* (pp. 245–269). Basel, Switzerland: Birkhäuser Verlag; doi:10.1007/978-3-0348-8845-5_7

Khalaf, R. A., Sheikha, G. A., Bustanji, Y., & Taha, M. O. (2010). Discovery of new cholesteryl ester transfer protein inhibitors via ligand-based pharmacophore modeling and QSAR analysis followed by synthetic exploration. *European Journal of Medicinal Chemistry*, *45*(4), 1598–1617. doi:10.1016/j.ejmech.2009.12.070 PMID:20116902

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews. Drug Discovery*, *3*(11), 935–949. doi:10.1038/nrd1549 PMID:15520816

Klebe, G. (2006). Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discovery Today*, *11*(13-14), 580–594. doi:10.1016/j.drudis.2006.05.012 PMID:16793526

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. doi:10.1007/BF00337288

Kroemer, R. T. (2007). Structure-based drug design: Docking and scoring. *Current Protein & Peptide Science*, *8*(4), 312–328. doi:10.2174/138920307781369382 PMID:17696866

Krovat, E. M., Fruhwirth, K. H., & Langer, T. (2005). Pharmacophore identification, *in silico* screening, and virtual library design for inhibitors of the human factor Xa. *Journal of Chemical Information and Modeling*, *45*(1), 146–159. doi:10.1021/ci049778k PMID:15667140

Kuhn, T., Willighagen, E. L., Zielesny, A., & Steinbeck, C. (2010). CDK-Taverna: An open workflow environment for cheminformatics. *BMC Bioinformatics*, *11*(1), 159. doi:10.1186/1471-2105-11-159 PMID:20346188

Kuz'min, V. E., Artemenko, A. G., Muratov, E. N., Volineckaya, I. L., Makarov, V. A., & Riabova, O. B. et al. (2007). Quantitative structure-activity relationship studies of [(biphenoxyl)propyl]isoxazole derivatives. Inhibitors of human rhinovirus 2 replication. *Journal of Medicinal Chemistry*, *50*(17), 4205–4213. doi:10.1021/jm0704806 PMID:17665898

Lahana, R. (1999). How many leads from HTS? *Drug Discovery Today*, *4*(10), 447–448. doi:10.1016/S1359-6446(99)01393-8 PMID:10481138

Langer, T., & Hoffmann, R. D. (2001). Virtual screening: An effective tool for lead structure discovery? *Current Pharmaceutical Design*, *7*(7), 509–527. doi:10.2174/1381612013397861 PMID:11375766

Laponogov, I., Pan, X.-S., Veselkov, D. A., McAuley, K. E., Fisher, L. M., & Sanderson, M. R. (2010). Structural basis of gate-DNA breakage and resealing by type II topoisomerases. *PLoS ONE*, *5*(6), e11338. doi:10.1371/journal.pone.0011338 PMID:20596531

Laponogov, I., Sohi, M. K., Veselkov, D. A., Pan, X.-S., Sawhney, R., & Thompson, A. W. et al. (2009). Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. *Nature Structural & Molecular Biology*, *16*(6), 667–669. doi:10.1038/nsmb.1604 PMID:19448616

Leach, A. R., Bradshaw, J., Green, D. V. S., Hann, M. M., & Delany, J. J. (1999). Implementation of a system for reagent selection and library enumeration. *Journal of Chemical Information and Computer Sciences*, *39*(6), 1161–1172. doi:10.1021/ci9904259 PMID:10614028

Lee, K.-O., Park, H.-J., Kim, Y.-H., Seo, S.-Y., Lee, Y.-S., & Moon, S.-H. et al. (2004). CoMFA and CoMSIA 3D QSAR studies on primarane cyclooxygenase-2 (COX-2) inhibitors. *Archives of Pharmacal Research*, *27*(5), 467–470. doi:10.1007/BF02980117 PMID:15202549

Leland, B. A., Christie, B. D., Nourse, J. G., Grier, D. L., Carhart, R. E., & Maffett, T. et al. (1997). Managing the combinatorial explosion. *Journal of Chemical Information and Computer Sciences*, *37*(1), 62–70. doi:10.1021/ci960088t

Levine, C., Hiasa, H., & Marians, K. J. (1998). DNA gyrase and topoisomerase IV: Biochemical activities, physiological roles during chromosome replication, and drug sensitivities. *Biochimica et Biophysica Acta*, *1400*(1-3), 29–43. doi:10.1016/S0167-4781(98)00126-2 PMID:9748489

Lewis, R. A. (2005). A general method for exploiting QSAR models in lead optimization. *Journal of Medicinal Chemistry*, *48*(5), 1638–1648. doi:10.1021/jm049228d PMID:15743205

Lill, M. A. (2007). Multi-dimensional QSAR in drug discovery. *Drug Discovery Today*, *12*(23-24), 1013–1017. doi:10.1016/j.drudis.2007.08.004 PMID:18061879

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *46*(1-3), 3–26. doi:10.1016/S0169-409X(00)00129-0 PMID:11259830

Liu, D. X., Jiang, H. L., Chen, K. X., & Ji, R. Y. (1998). A new approach to design virtual combinatorial library with genetic algorithm based on 3D grid property. *Journal of Chemical Information and Computer Sciences*, *38*(2), 233–242. doi:10.1021/ci970086o

Liu, X., Vogt, I., Haque, T., & Campillos, M. (2013). HitPick: A web server for hit identification and target prediction of chemical screening. *Bioinformatics (Oxford, England)*, *29*(15), 1910–1912. doi:10.1093/bioinformatics/btt303 PMID:23716196

Liu, Z., Diebler, R. W., Chan, H. S., & Zechiedrich, L. (2009). The why and how of DNA unlinking. *Nucleic Acids Research*, *37*(3), 661–671. doi:10.1093/nar/gkp041 PMID:19240147

Lobanov, V. S., & Agrafiotis, D. K. (2002). Scalable method for the construction and analysis of virtual combinatorial libraries. *Combinatorial Chemistry & High Throughput Screening*, *5*(2), 167–178. doi:10.2174/1386207024607392 PMID:11966425

Lobell, M., Molnar, L., & Keseru, G. M. (2003). Recent advances in the prediction of blood-brain partitioning from molecular structure. *Journal of Pharmaceutical Sciences*, *92*(2), 360–370. doi:10.1002/jps.10282 PMID:12532385

Lybrand, T. P. (1995). Ligand-protein docking and rational drug design. *Current Opinion in Structural Biology*, *5*(2), 224–228. doi:10.1016/0959-440X(95)80080-8 PMID:7648325

Manallack, D. T., Pitt, W. R., Gancia, E., Montana, J. G., Livingstone, D. J., Ford, M. G., & Whitley, D. C. (2002). Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *Journal of Chemical Information and Computer Sciences*, *42*(5), 1256–1262. doi:10.1021/ci020267c PMID:12377017

Marcou, G., & Rognan, D. (2007). Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling*, *47*(1), 195–207. doi:10.1021/ci600342e PMID:17238265

Martel, A. M., Leeson, P. A., & Castañer, J. (1997). BAY-12-8039. Fluoroquinolone antibacterial. *Drugs of the Future*, *22*(2), 109–113.

Matrat, S., Veziris, N., Mayer, C., Jarlier, V., Truffot-Pernot, C., & Camuset, J. et al. (2006). Functional analysis of DNA gyrase mutant enzymes carrying mutations at position 88 in the A subunit in clinical strains of *Mycobacterium tuberculosis* resistant to fluoroquinolones. *Antimicrobial Agents and Chemotherapy*, *50*(12), 4170–4173. doi:10.1128/AAC.00944-06 PMID:17015625

Mavromoustakos, T., Durdagi, S., Koukoulitsa, C., Simcic, M., Papadopoulos, M. G., Hodoscek, M., & Golic Grdadolnik, S. (2011). Strategies in the rational drug design. *Current Medicinal Chemistry*, *18*(17), 2517–2530. doi:10.2174/092986711795933731 PMID:21568895

Maxwell, A. (1997). DNA gyrase as a drug target. *Trends in Microbiology*, *5*(3), 102–109. doi:10.1016/S0966-842X(96)10085-8 PMID:9080608

Minovski, N., Perdih, A., Novic, M., & Solmajer, T. (2013). Cluster-based molecular docking study for *in silico* identification of novel 6-fluoroquinolones as potential inhibitors against *Mycobacterium tuberculosis. Journal of Computational Chemistry*, *34*(9), 790–801. doi:10.1002/jcc.23205 PMID:23280926

Minovski, N., Perdih, A., & Solmajer, T. (2012). Combinatorially-generated library of 6-fluoroquinolone analogs as potential novel antitubercular agents: A chemometric and molecular modeling assessment. *Journal of Molecular Modeling*, *18*(5), 1735–1753. doi:10.1007/s00894-011-1179-0 PMID:21833830

Minovski, N., & Šolmajer, T. (2010). Chemometrical exploration of combinatorially generated drug-like space of 6-fluoroquinolone analogs: A QSAR study. *Acta Chimica Slovenica*, *57*(3), 529–540. PMID:24061797

Minovski, N., Vračko, M., & Šolmajer, T. (2011). Quantitative structure-activity relationship study of antitubercular fluoroquinolones. *Molecular Diversity*, *15*(2), 417–426. doi:10.1007/s11030-010-9238-5 PMID:20229318

Minovski, N., Župerl, Š., Drgan, V., & Novič, M. (2013). Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study. *Analytica Chimica Acta*, *759*, 28–42. doi:10.1016/j.aca.2012.11.002 PMID:23260674

Moal, I. H., Torchala, M., Bates, P. A., & Fernández-Recio, J. (2013). The scoring of poses in protein-protein docking: Current capabilities and future directions. *BMC Bioinformatics*, *14*, 286. doi:10.1186/1471-2105-14-286 PMID:24079540

Moldover, B., Solidar, A., Montgomery, C., Miziorko, H., Murphy, J., & Wyckoff, G. J. (2012). ChemVassa: A new method for identifying small molecule hits in drug discovery. *The Open Medicinal Chemistry Journal*, *6*(1), 29–34. doi:10.2174/1874104501206010029 PMID:23525139

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. In A. Kukol (Ed.), *Methods in molecular biology: Molecular modeling of proteins* (Vol. 443, pp. 365–382). Totowa, NJ: Humana Press; doi:10.1007/978-1-59745-177-2_19

Muegge, I. (2003). Selection criteria for drug-like molecules. *Medicinal Research Reviews*, *23*(3), 302–321. doi:10.1002/med.10041 PMID:12647312

Muegge, I., Heald, S. L., & Britelli, D. (2001). Simple selection criteria for drug-like chemical matter. *Journal of Medicinal Chemistry*, *44*(12), 1841–1846. doi:10.1021/jm015507e PMID:11384230

Musmuca, I., Caroli, A., Mai, A., Kaushik-Basu, N., Arora, P., & Ragno, R. (2010). Combining 3-D quantitative structure-activity relationship with ligand based and structure based alignment procedures for *in Silico* screening of new hepatitis C virus NS5B polymerase inhibitors. *Journal of Chemical Information and Modeling*, *50*(4), 662–676. doi:10.1021/ci9004749 PMID:20225870

Nevin, D. K., Peters, M. B., Carta, G., Fayne, D., & Lloyd, D. G. (2012). Integrated virtual screening for the identification of novel and selective peroxisome proliferator-activated receptor (PPAR) scaffolds. *Journal of Medicinal Chemistry*, *55*(11), 4978–4989. doi:10.1021/jm300068n PMID:22582973

Nicolaou, C. A., & Brown, N. (2013). Multi-objective optimization methods in drug design. *Drug Discovery Today. Technologies*, *10*(3), e427–e435. doi:10.1016/j.ddtec.2013.02.001 PMID:24050140

Novič, M., & Zupan, J. (1995). Investigation of infrared spectra-structure correlation using Kohonen and counterpropagation neural network. *Journal of Chemical Information and Computer Sciences*, *35*(3), 454–466. doi:10.1021/ci00025a013

O'Driscoll, C. (2004). A virtual space odyssey. In *Proceedings of Horizon Symposium, Charting Chemical Space: Finding New Tools to Explore Biology* (pp. 1-4). New York, NY: Nature Publishing Group. Retrieved from http://www.nature.com/horizon/chemicalspace/background/odyssey.html

OECD. (2004). *Principles for the validation for regulatory purposes of (quantitative) structure-activity relationship models*. Paris, France: OECD.

Oprea, T. I. (2000). Property distribution of drug-related chemical databases. *Journal of Computer-Aided Molecular Design*, *14*(3), 251–264. doi:10.1023/A:1008130001697 PMID:10756480

Pantel, A., Petrella, S., Matrat, S., Brossier, F., Bastian, S., & Reitter, D. et al. (2011). DNA gyrase inhibition assays are necessary to demonstrate fluoroquinolone resistance to *gyrb* mutations in *Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy*, *55*(10), 4524–4529. doi:10.1128/AAC.00707-11 PMID:21768507

Pantel, A., Petrella, S., Veziris, N., Brossier, F., Bastian, S., & Jarlier, V. et al. (2012). Extending the definition of the GyrB quinolone resistance-determining region in *Mycobacterium tuberculosis* DNA gyrase for assessing fluoroquiniolone resistance in *M. tuberculosis. Antimicrobial Agents and Chemotherapy*, *56*(4), 1990–1996. doi:10.1128/AAC.06272-11 PMID:22290942

Pardridge, W. M. (1995). Transport of small molecules through the blood-brain barrier: Biology and methodology. *Advanced Drug Delivery Reviews*, *15*(1-3), 5–36. doi:10.1016/0169-409X(95)00003-P

Piton, J., Petrella, S., Delarue, M., André-Leroux, G., Jarlier, V., Aubry, A., & Mayer, C. (2010). Structural insights into the quinolone resistance mechanism of *Mycobacterium tuberculosis* DNA gyrase. *PLoS ONE*, *5*(8), e12245. doi:10.1371/journal.pone.0012245 PMID:20805881

Ramesha, C. S. (2000). Comment: How many leads from HTS? *Drug Discovery Today*, *5*(2), 43–44. doi:10.1016/S1359-6446(99)01444-0 PMID:10652452

Rastelli, G., Degliesposti, G., Del Rio, A., & Sgobba, M. (2009). Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chemical Biology & Drug Design*, *73*(3), 283–286. doi:10.1111/j.1747-0285.2009.00780.x PMID:19207463

Rawlins, M. D. (2004). Cutting the cost of drug development? *Nature Reviews. Drug Discovery*, *3*(4), 360–364. doi:10.1038/nrd1347 PMID:15060531

Raymond, J.-L., van Deursen, R., Blum, L. C., & Ruddigkeit, L. (2010). Chemical space as a source for new drugs. *Medicinal Chemistry Communications*, *1*(1), 30–38. doi:10.1039/c0md00020e

Rishton, G. M. (1997). Reactive compounds and *in vitro* false positives in HTS. *Drug Discovery Today*, *2*(9), 382–384. doi:10.1016/S1359-6446(97)01083-0

Rishton, G. M. (2003). Non-leadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today*, *8*(2), 86–96. doi:10.1016/S1359644602025722 PMID:12565011

Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., & Das, R. N. (2013). Some case studies on application of „r^2_m“ metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data. *Journal of Computational Chemistry*, *34*(12), 1071–1082. doi:10.1002/jcc.23231 PMID:23299630

Sacan, A., Ekins, S., & Kortagere, S. (2012). Applications and limitations of in silico models in drug discovery. In R. S. Larson (Ed.), Bioinformatics and drug discovery, methods in molecular biology (pp. 87-124). New York, NY: Springer Science + Business Media. doi:10.1007/978-1-61779-965-5_6

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules (Basel, Switzerland)*, *17*(12), 4791–4810. doi:10.3390/molecules17054791 PMID:22534664

Salemme, F. R., Spurlino, J., & Bone, R. (1997). Serendipity meets precision: The integration of structure-based drug design and combinatorial chemistry for efficient drug discovery. *Structure (London, England)*, *5*(3), 319–324. doi:10.1016/S0969-2126(97)00189-5 PMID:9083110

Schlegel, B., Meier, R., Laggner, C., Schnell, D., Langer, T., & Seifert, R. et al. (2007). Generation of a homology model of the human histamine H3 receptor for ligand docking and pharmacophore-based screening. *Journal of Computer-Aided Molecular Design*, *21*(8), 437–453. doi:10.1007/s10822-007-9127-x PMID:17668276

Schwalbe, T., Kadzimirisz, D., & Jas, G. (2000). Synthesis of a library of ciprofloxacin analogues by means of sequential organic synthesis in microreactors. *QSAR & Combinatorial Science*, *24*(6), 758–768. doi:10.1002/qsar.200540005

Scior, T., Medina-Franco, J. L., Do, Q.-T., Martinez-Mayorga, K., Yunes Rojas, J. A., & Bernard, P. (2009). How to recognize and workaround pitfalls in QSAR studies: A critical review. *Current Medicinal Chemistry*, *16*(32), 4297–4313. doi:10.2174/092986709789578213 PMID:19754417

Seifert, M. H. J., & Lang, M. (2007). Essential factors for successful virtual screening. *Mini Reviews in Medicinal Chemistry*, *8*(1), 63–72. doi:10.2174/138955708783331540 PMID:18220986

Serradel, M. N., Blancafort, P., & Castañer, J. (1983). DL-8280. *Drugs of the Future*, *8*(5), 395.

Shaikh, S. A., Jain, T., Sandhu, G., Latha, N., & Jayaram, B. (2007). From drug target to leads – Sketching a physicochemical pathway for lead molecule design *In Silico. Current Pharmaceutical Design*, *13*(34), 3454–3470. doi:10.2174/138161207782794220 PMID:18220783

Shi, R., Zhang, J., Li, C., Kazumi, Y., & Sugawara, I. (2006). Emergence of ofloxacin resistance in *Mycobacterium tuberculosis* clinical isolates from China as determined by *gyrA* mutation analysis using denaturing high-pressure liquid chromatography and DNA sequencing. *Journal of Clinical Microbiology*, *44*(12), 4566–4568. doi:10.1128/JCM.01916-06 PMID:17035499

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*, *432*(7019), 862–865. doi:10.1038/nature03197 PMID:15602552

Sissi, C., & Palumbo, M. (2010). In front of and behind the replication fork: Bacterial type IIA topoisomerases. *Cellular and Molecular Life Sciences*, *67*(12), 2001–2024. doi:10.1007/s00018-010-0299-5 PMID:20165898

Stahura, F. L., & Bajorath, J. (2004). Virtual screening methods that complement HTS. *Combinatorial Chemistry & High Throughput Screening*, *7*(4), 259–269. doi:10.2174/1386207043328706 PMID:15200375

Sugino, A., Peebles, C. L., Kreuzer, K. N., & Cozzarelli, N. R. (1977). Mechanism of action of nalidixic acid: Purification of *Escherichia coli nalA* gene product and its relationship to DNA gyrase and a novel nicking-closing enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 4767–4771. doi:10.1073/pnas.74.11.4767 PMID:200930

Tan, W., Mei, H., Chao, L., Liu, T., Pan, X., Shu, M., & Yang, L. (2013). Combined QSAR and molecular docking studies on predicting P-glycoprotein inhibitors. *Journal of Computer-Aided Molecular Design*, *27*(12), 1067–1073. doi:10.1007/s10822-013-9697-8 PMID:24322389

Tian, S., Sun, H., Li, Y., Pan, P., Li, D., & Hou, T. (2013). Development and evaluation of an integrated virtual screening strategy by combining molecular docking and pharmacophore searching based on multiple protein structures. *Journal of Chemical Information and Modeling*, *53*(10), 2743–2756. doi:10.1021/ci400382r PMID:24010823

Tillotson, G. S. (1996). Quinolones: Structure-activity relationships and future predictions. *Journal of Medical Microbiology*, *44*(5), 320–324. doi:10.1099/00222615-44-5-320 PMID:8636945

Topliss, J. G. (1983). *Quantitative structure-activity relationships of drugs*. New York, NY: Academic Press.

Topliss, J. G., & Edwards, R. P. (1979). Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry*, *22*(10), 1238–1244. doi:10.1021/jm00196a017 PMID:513071

Tretter, E. M., Schoeffler, A. J., Weisfield, S. R., & Berger, J. M. (2010). Crystal structure of the DNA gyrase GyrA N-terminal domain from Mycobacterium tuberculosis. [PubMed]. *Proteins, Structure, Function, and Bioinformatics*, *78*(2), 492–495. doi:10.1002/prot.22600

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., & Bertrand, H.-O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of Medicinal Chemistry*, *48*(7), 2534–2547. doi:10.1021/jm049092j PMID:15801843

Truszkowski, A., Jayaseelan, K. V., Neumann, S., Willighagen, E. L., Zielesny, A., & Steinbeck, C. (2011). New developments on the chemoinformatics open workflow environment CDK-Taverna. *Journal of Cheminformatics*, *3*(1), 54. doi:10.1186/1758-2946-3-54 PMID:22166170

Ul-Haq, Z., Usmani, S., Shamshad, H., Mahmood, U., & Halim, S. A. (2013). A combined 3D-QSAR and docking studies for the *In-silico* prediction of HIB-protease inhibitors. *Chemistry Central Journal*, *7*(88), 1–12. doi:10.1186/1752-153X-7-88 PMID:23289739

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, *45*(12), 2615–2623. doi:10.1021/jm020017n PMID:12036371

Vedani, A., Dobler, M., & Lill, M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *Journal of Medicinal Chemistry*, *48*(11), 3700–3703. doi:10.1021/jm050185q PMID:15916421

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, *52*(4), 609–623. doi:10.1002/prot.10465 PMID:12910460

Vidal, D., Garcia-Serna, R., & Mestres, J. (2011). Ligand-based approaches to In Silico pharmacology. In J. Bajorath (Ed.), Methods in molecular biology: Chemoinformatics and computational chemical biology (vol. 672, pp. 489-502). Springer Science + Business Media, LLC. doi:10.1007/978-1-60761-839-3_19

Vilar, S., Karpiak, J., & Costanzi, S. (2009). Ligand and structure-based models for the prediction of ligand-receptor affinities and virtual screening: development and application to the β_2-adrenergic receptor. *Journal of Computational Chemistry*, *31*(4), 707–720. doi:10.1002/jcc PMID:19569204

Villoutreix, B. O., Eudes, R., & Miteva, M. A. (2009). Structure-based virtual ligand screening: Recent success stories. *Combinatorial Chemistry & High Throughput Screening*, *12*(10), 1000–1016. doi:10.2174/138620709789824682 PMID:20025565

Walters, W. P., Murcko, A. A., & Murcko, M. A. (1999). Recognizing molecules with drug-like properties. *Current Opinion in Chemical Biology*, *3*(4), 384–387. doi:10.1016/S1367-5931(99)80058-1 PMID:10419858

Wang, C., Bradley, P., & Baker, D. (2007). Protein-protein docking with backbone flexibility. *Journal of Molecular Biology*, *373*(2), 503–519. doi:10.1016/j.jmb.2007.07.050 PMID:17825317

Warr, W. A. (2012). Scientific workflow systems: Pipeline pilot and KNIME. *Journal of Computer-Aided Molecular Design*, *26*(7), 801–804. doi:10.1007/s10822-012-9577-7 PMID:22644661

Wasserman, S. A., & Cozzarelli, N. R. (1986). Biochemical topology: Applications to DNA recombination and replication. *Science*, *232*(4753), 951–960. doi:10.1126/science.3010458 PMID:3010458

Wieland, T. (1997). Combinatorics of combinatorial chemistry. *Journal of Mathematical Chemistry*, *21*(2), 141–157. doi:10.1023/A:1019166201637

Wohlkonig, A., Chan, P. F., Fosberry, A. P., Homes, P., Huang, J., & Kranz, M. et al. (2010). Structural basis of quinolone inhibition of type IIA topoisomerases and target-mediated resistance. *Nature Structural & Molecular Biology*, *17*(9), 1152–1153. doi:10.1038/nsmb.1892 PMID:20802486

Wolber, G., & Kosara, R. (2006). Pharmacophores from macromolecular complexes with LigandScout. In T. Langer & R. D. Hoffmann (Eds.), *Pharmacophores and pharmacophore searches* (pp. 131–150). New Weinheim, Germany: WILEY-VCH; doi:10.1002/3527609164.ch6

Wolber, G., & Langer, T. (2005). LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling*, *45*(1), 160–169. doi:10.1021/ci049885e PMID:15667141

Wold, S. (1991). Validation of QSAR's. *Quantitative Structure-Activity Relationship*, *10*(3), 191–193. doi:10.1002/qsar.19910100302

Wold, S., & Eriksson, L. (1995). Statistical validation of QSAR results. In Chemometrics methods in molecular design (pp. 309-318). Weinheim, Germany: Wiley-VCH.

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current Protein & Peptide Science*, *7*(3), 217–227. doi:10.2174/138920306777452312 PMID:16787261

Yang, J., Roy, A., & Zhang, Y. (2013). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)*, *29*(20), 2588–2595. doi:10.1093/bioinformatics/btt447 PMID:23975762

Yang, Y., Zhang, W., Cheng, J., Tang, Y., Peng, Y., & Li, Z. (2013). Pharmacophore, 3D-QSAR, AND Bayesian model analysis for ligands binding at the benzodiazepine site of $GABA_A$ receptors: The key roles of amino group and hydrophobic sites. *Chemical Biology & Drug Design*, *81*(5), 583–590. doi:10.1111/cbdd.12100 PMID:23279907

Young, S. S., Farmen, M. W., & Rusinko, A., III. (1996). Random versus rational: Which is better for general compound screening? *Network Science*. Retrieved from http://www.netsci.org/Science/Screening/feature09.html

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., & Tropsha, A. (2006). A novel automated lazy learning QSAR (ALL-QSAR) approach: Method Development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *Journal of Chemical Information and Modeling*, *46*(5), 1984–1995. doi:10.1021/ci060132x PMID:16995729

Zhang, Y., Yang, S., Jiao, Y., Liu, H., Yuan, H., & Lu, S. et al. (2013). An integrated virtual screening approach for VEGFR-2 inhibitors. *Journal of Chemical Information and Modeling*, *53*(12), 3163–3177. doi:10.1021/ci400429g PMID:24266594

Zheng, F., Zhan, M., Huang, X., Hammed, M. D. M. A., & Zhan, C.-G. (2014). Modeling *in vitro* inhibition of butyrylcholinesterase using molecular docking, multi-linear regression and artificial neural network approaches. *Bioorganic & Medicinal Chemistry*, *22*(1), 538–549. doi:10.1016/j.bmc.2013.10.053 PMID:24290065

Zupan, J., & Gasteiger, J. (1999). *Neural networks in chemistry and drug design* (2nd ed.). Weinheim, Germany: Wiley-VCH.

Zupan, J., Novič, M., & Gasteiger, J. (1995). Neural networks with counter-propagation learning strategy used for modelling. *Chemometrics and Intelligent Laboratory Systems*, *27*(2), 175–187. doi:10.1016/0169-7439(95)80022-2

Zupan, J., Novič, M., & Ruisánchez, I. (1997). Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemometrics and Intelligent Laboratory Systems*, *38*(1), 1–23. doi:10.1016/S0169-7439(97)00030-0

KEY TERMS AND DEFINITIONS

Antibacterial Agents: A common term for chemicals including synthetic or semi-synthetic drugs or other similar chemical entities that either kill or inhibit the bacterial growth.

Artificial Neural Networks: Computational non-linear modeling tools that mimic the structure and functions of the biological neural networks in the brain. During the learning process (training of the network), a complex relationship between the input and output data is established. They can be used either for numerical predictions of properties or pattern recognition purposes.

DNA Gyrase: An omnipresent molecular nanomachine from the type II DNA topoisomerase superfamily that is responsible for the unwinding of the DNA molecule during the DNA replication phase. The bacterial DNA gyrase is a well-established target of many antibacterials including nalidixic acid and their derivatives 6-fluoroquinolones.

Drug-Likeness: A qualitative measure based on a set of complex *in silico* calculable physico-chemical properties (e.g., molecular weight, logP, number of rotatable bonds, number of hydrogen bond donors and acceptors, and polar surface area) that determine whether an investigated compound is similar to the known drugs. It was found as a useful measure in the modern drug discovery and it is frequently used to filter out the so-called "drug-like" compounds from massive chemical libraries.

Protein Homology Modeling: Construction of a three-dimensional model of the "*target*" protein at atomic resolution from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein ("*template*").

Quantitative Structure-Activity Relationship: An approach designed to establish relationships between the chemical structure and biological activity (or other target property) of investigated compounds in a quantitative manner.

Virtual Combinatorial Library Design: Generation of a list of structurally similar molecules in a virtual (*in silico*) environment employing the principles of combinatorial chemistry where a set of reagents (substituents or building-blocks) are specifically attached at pre-defined scaffold positions on the main structure.

Virtual Screening: A computational methodology used in the modern drug discovery to search and identify those molecular entities (small molecules) from a chemical library which are most likely to bind to a drug target, usually protein receptor or enzyme.

ENDNOTES

[1] http://www.knime.org
[2] http://accelrys.com/products/pipeline-pilot
[3] http://sourceforge.net/apps/mediawiki/cdk/index/.php?title=Main_Page
[4] http://www.rdkit.org
[5] http://ggasoftware.com/opensource
[6] http://www.novamechanics.com/index.php
[7] https://github.com/knime-mpicbg/HCS-Tools/wiki
[8] http://www.seqan.de
[9] https://www.molecular-networks.com/pipelinepilot
[10] http://www.biosolveit.de/PipelinePilot
[11] http://www.cresset-group.com/cresset-pipeline-pilot-component-v2-0-0-release-notes
[12] http://cdktaverna.wordpress.com
[13] http://www.inhibox.com
[14] https://www.sciencecloud.com
[15] http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html
[16] http://accelrys.com/products/discovery-studio/pharmacophore-ligand-based-design.html
[17] http://www.inteligand.com/ligandscout
[18] http://chemdb.niaid.nih.gov
[19] http://www.keyorganics.ltd.uk
[20] https://www.ccdc.cam.ac.uk/Solutions/GoldSuite/pages/GoldSuite.aspx
[21] http://www.asinex.com
[22] http://www.eyesopen.com/rocs.