CHAPTER 6

Probabilistic Approaches in Activity Prediction

DMITRY FILIMONOV AND VLADIMIR POROIKOV

Institute of Biomedical Chemistry of Russian Academy of Medical Sciences
10, Pogodinskaya Str., Moscow, 119121, Russia

6.1 Introduction

Biological activity has the probabilistic nature, and the most appropriate approaches in activity prediction are based on the theory of probability. The statistical nature of maximum likelihood method and Bayesian approach is well recognized, but many other methods (multiple regression, factor analysis, pattern recognition methods such as linear discriminant analysis, linear learning machine, support vector machines *etc*.)[1-3] can also be considered as probabilistic ones.[4,5] An informational search in PubMed Central with the queries "(probabilistic approach) OR (probabilistic method)" or "(statistical approach) OR (statistical method)", will find 3,477 documents or 180,475 documents, respectively. It is impossible to analyze all these publications, particularly taking into account that, despite of the presence of this term in their titles many of them are not really probabilistic (see, for instance, refs 6−20). We propose the following definition of probabilistic approaches: "The methods that use probabilities as an essential part of the algorithm, and/or for which the results of application are presented as probability estimates". Thus, many approaches that do not correspond strictly to the definition, are not considered in this chapter.

Since data on general dose-response relationships are not available in many cases, biological activity is often represented by a single quantitative or even qualitative characteristic. Therefore, many training sets are created with activity data presented in such mode. These probabilistic ligand-based drug design methods are further used for virtual screening. Existing training sets are not ideal, not just due to the simplified definition of biological activity, but also because (i) no one activity is represented by all relevant chemical classes and (ii) no one compound has been tested against all kinds of biological activity. So, the probabilistic character of biological activity is caused not only by experimental errors of its determination but also by incompleteness of available information.

Typically, virtual screening methods are used to select hits with a single required activity,[21-24] while the final aim of pharmaceutical R & D is to identify safety and potent leads and drug-candidates.[25-28] To overcome this problem, the authors have developed a method for prediction of many kinds of biological activity simultaneously based on the structural formula of chemical compound, which is realized in the computer program PASS (Prediction of Activity Spectra for Substances).[29,30] PASS provides the means for evaluation of general biological activity profile at the early stages of R & D, and thus its prediction can be used as a basis for the selection of compounds with the required kinds of biological activity but without unwanted ones.[31,32]

In this chapter we overview some probabilistic methods used for biological activity prediction, paying particular attention to the problems of creation of the training and evaluation sets, validation of (Q)SAR models, estimation of prediction accuracy, interpretation of the prediction results and their application in virtual screening.

In: Chemoinformatics Approaches to Virtual Screening. Eds. Alexandre Varnek and Alex Tropsha. RSC Publishing, 2008, pp. 182 – 216.

6.2 Biological Activity

Biological activity is the result of chemical compound's interaction with biological objects. It depends on the characteristics of (i) compound (structure of molecule and its physical-chemical properties), (ii) biological object (kind, sex, age, *etc*.), (iii) way of exposure (route of administration, dosage), (iv) peculiarities of the experimental terms and conditions.

The major paradigm of the twentieth century was based on the concept "one disease – one target",[27,33] therefore, at first chemical compounds were tested against the targeted activity, and only for those leads that passed through this "filter" was a more general biological activity profile was estimated. Currently, it is recognized that most pharmaceutical agents interact with several or even many targets in the organism, and thus their selectivity is rather relative. For example, by analysis of the available literature one may find that biological activity of caffeine (CAS No. 58-08-2) is described by the terms related to the following:

- ten pharmacotherapeutic effects (analeptic, antihypertensive, antihypotensive, cardiotonic, diuretic, immunosuppressant, psychostimulant, respiratory analeptic, saluretic, spasmolytic);
- 18 biochemical mechanisms of action (ATP diphosphatase inhibitor, adenosine deaminase inhibitor, cyclic AMP phosphodiesterase inhibitor, cytochrome P450 inhibitor, "dATP(dGTP)-DNA purinetransferase inhibitor, glycogen (starch) synthase inhibitor, guanylate cyclase inhibitor, hydroxyacylglutathione hydrolase inhibitor, lactoylglutathione lyase inhibitor, nucleotide metabolism regulator, P-glycoprotein inhibitor, phosphatidylinositol kinase inhibitor, phosphodiesterase inhibitor, phosphorylase inhibitor, purine nucleosidase inhibitor, thymidine kinase inhibitor, urate oxidase inhibitor, xanthine-like agent);
- nine adverse/toxic effects [arrhythmogenic, spasmogenic, convulsant, non mutagenic (salmonella), embryotoxic, teratogen, carcinogenic, carcinogenic (group 3), toxic];
- 16 metabolic terms (CYP1 substrate, CYP1A inhibitor, CYP1A substrate, CYP1A1 substrate, CYP1A2 inhibitor, CYP1A2 substrate, CYP2 substrate, CYP2B substrate, CYP2B1 substrate, CYP2B2 substrate, CYP2E substrate, CYP2E1 substrate, CYP3A substrate, CYP3A1 substrate, CYP3A4 substrate, CYP3A5 substrate).

Some apparent contradictions in terms representing the biological activity of caffeine can be explained either by its opposite effects in different doses or by peculiarities of experimental terms and conditions in the appropriate studies. A similar picture can be observed also for most well-known pharmaceuticals.

On the other hand, even acting on the same target, different chemical compounds can bind to them in different modes.[34] Therefore, any individual chemical structure exhibits many biological activities, and *vice versa* a particular biological activity can be caused by many different chemical structures.[35,36]

Biological activity is tested both *in vivo* and *in vitro*. In the past 20 years, due to advances in preparative and measuring techniques, a significant part of assays is the testing of ligand binding to the macromolecular target *in vitro*. It is necessary to keep in mind that such binding can occur not with the site of macromolecule that is responsible for its biological activity or for suppressing of this biological activity. As a result, many ligands found in high-throughput assays may appear to be nonspecific or "promiscuous" inhibitors.[37] Moreover, binding is not a sufficient condition for ensuring that a beneficial function will ensue in the cell or in the organism as a whole.[38] After the deciphering of the human genome and first results in postgenomic studies it became obvious that many diseases have a complex etiology,[27] while drug action on a certain target often leads to activation/inhibition of other elements in the appropriate regulatory network. As a consequence of negative feedback, expected pharmacotherapeutic action may be significantly decreased or even completely suppressed.[39] Therefore, specially designed multi-targeted drugs may have certain advantages over single-targeted medicines.[33]

Since the final purpose of pharmaceutical studies to find hits & leads with the required, but without unwanted, properties the virtual screening should provide the estimation of general biological activity profile because such experimental studies are highly expensive and time-consuming.

We proposed the biological activity spectrum of a substance concept, which seems to be a fundamental basis for description of biologically active substances.[29,30,32,40-43] The "biological activity spectrum" of a substance is the set of different kinds of biological activity, which reflect the results of chemical substance's interaction with various biological entities. This more general concept was introduced earlier than "biospectra"[44,45] or other "activity spectra".[46] Biological activity is defined qualitatively ("yes" /"none"), suggesting that the "biological activity spectrum" represents the "intrinsic" property of a substance, depending only on its structure and physicochemical

characteristics. Certainly, this is a simplified definition because the exhibition of biological activity depends on the presence and state of the corresponding targets and experimental conditions (object, route of administration, dose, *etc.).* However, such approximation provides a possibility for combining of information from many different sources, which is necessary because no one particular publication represents comprehensively different aspects of biological action of a compound. For example, to collect information on the biological activity profile of caffeine discussed above, an extensive information search was performed of the available literature and databases.

6.2.1 Dose-Effect Relationships

In the most general form the description of biological activity of a certain chemical compound can be represented as a probability of occurrence of a certain biological response, depending on the experimental conditions (object, its state, means of exposure) and "dosage" of the compound ("dosage" can be represented in many different ways, in particular a single *per os* administration or fixed amount of a substance): $\Pr(\text{Doze},\text{Test})$. Under the fixed experimental conditions one obtains a simple relationship "dose-effect": $\Pr(\text{Doze}) = P(D)$. It must be stressed that *P(D)* is the probability of occurrence of a certain effect, which depends on a dose *D* as a parameter.

According to the recommendations,[47] in quantitative measurements of biological activity drug action is expressed in terms of the effect, E, produced when an agonist, A, is applied at a concentration $[A]$. The relationship between E and $[A]$ can be often described empirically by the Hill's equation,[48,49] which has the form:

$$\frac{E}{E_{\max}} = \frac{[A]^{nH}}{[A]^{nH} + [A]_{50}^{nH}}, \qquad (6.1)$$

where $E_{\max}$ is the maximal action of A, nH is the Hill coefficient and $[A]_{50}$ is the concentration that produces an effect that is 50% of $E_{\max}$. In Figure 6.1 shows an example of effect-concentration relationships estimated according to the Hill equation (Equation 6.1). Clearly, if $[A]=[A]_{50}$, all curves pass through the point at which the effect is half of its maximal value.

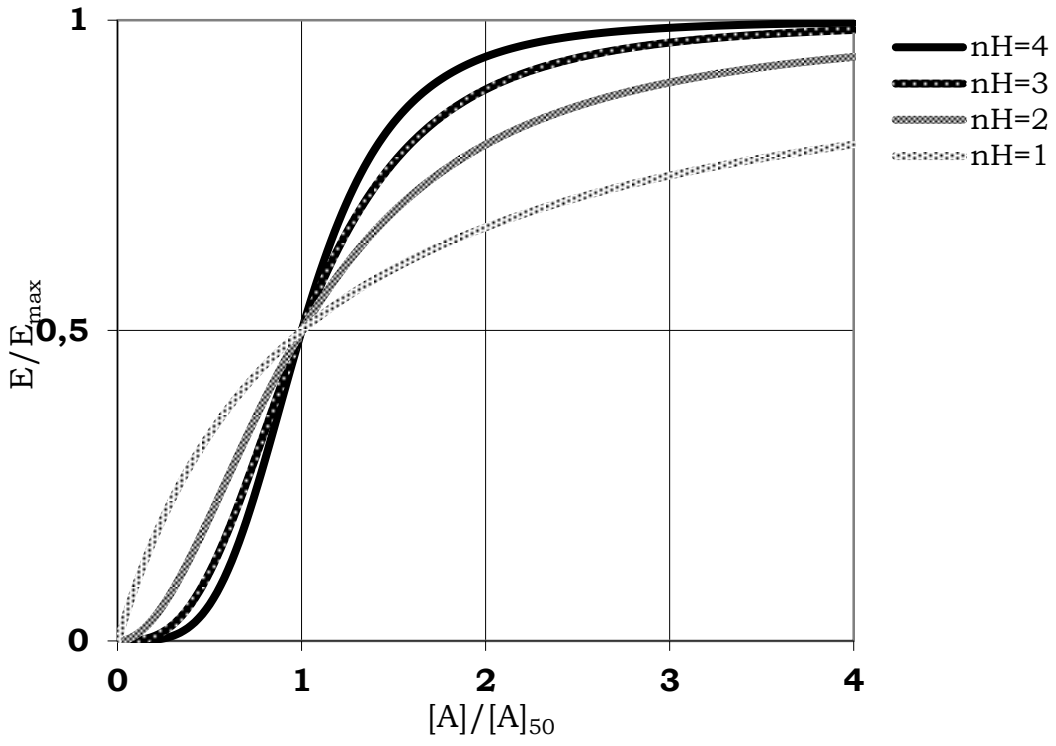
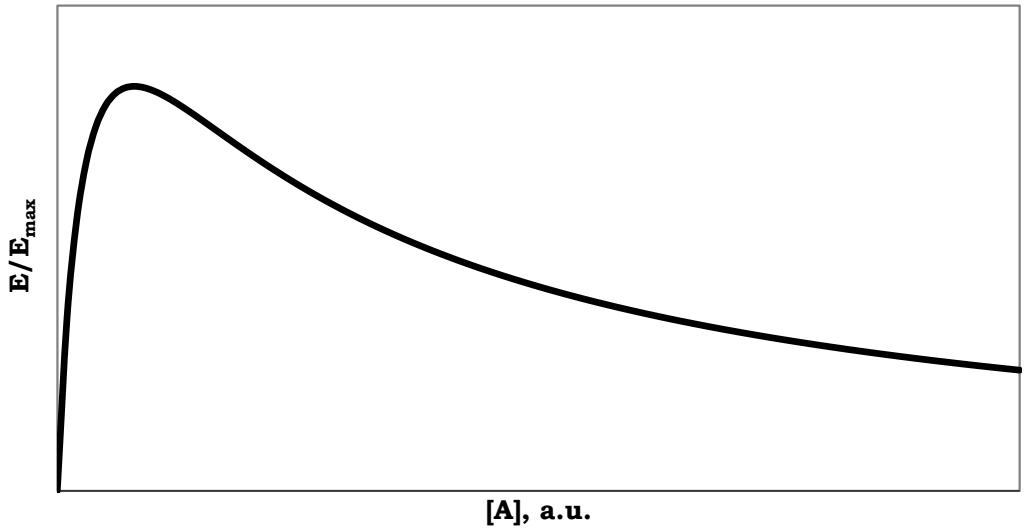


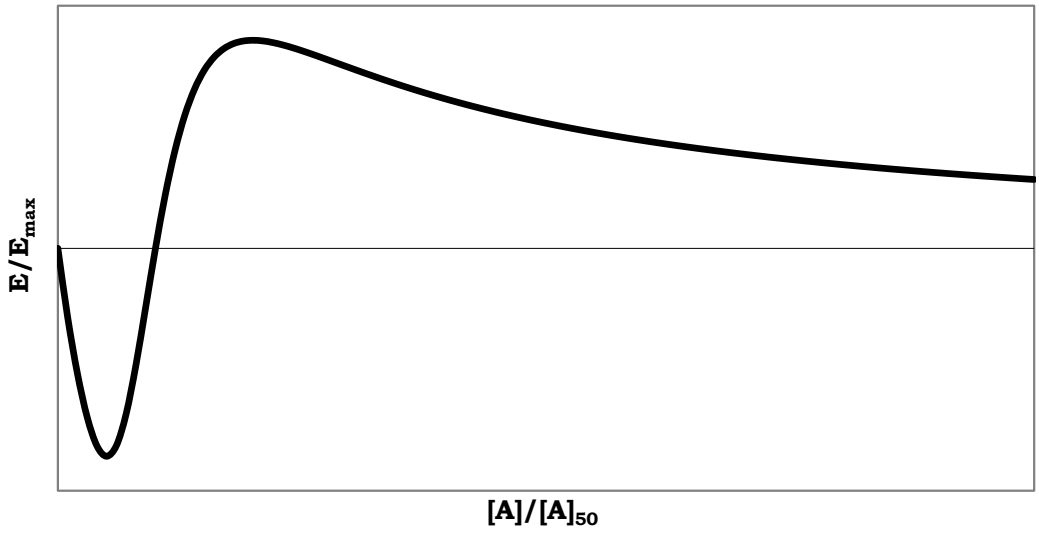
Figure 6.1 Relative values of effect depending on relative agonist concentration calculated according to the Hill equation. nH are the different values of Hill coefficient.

Unfortunately, Hill's equation (Equation 6.1) is only a convenient mathematical idealization, which can be realized for ligand binding to the pure isolated receptor *in vitro*. In an intact biological object a ligand interacts with several or even many different macromolecules,[50] and the final biological effect may dramatically differ from the simple relationship presented in Figure 6.1. For example, if some effect may be caused by two mechanisms, and a ligand interacts with the appropriate receptors, both activating and inhibiting them, then either activating or suppressing of the effect E can be observed depending of the concentration $[A]$ of the ligand (Figures 6.2 and 6.3).



$$\frac{E}{E_{\max}} = \frac{[A]}{[A] + [A]_{50}} - \frac{[A]}{[A] + 4[A]_{50}}$$

Figure 6.2 Relative effect vs. concentration of agonist provided the agonist simultaneously acts on another target as a weak antagonist.



$$\frac{E}{E_{\max}} = \frac{[A]^4}{[A]^4 + [A]_{50}^4} - \frac{[A]}{[A] + [A]_{50}}$$

Figure 6.3 Example of relative effect dependence on the agonist's concentration provided the agonist acts on another target as antagonist with equal semi-effective concentrations and different Hill coefficients.

In experimental testing of toxicity the results are presented by the numbers of surviving (n) and dying (m) biological objects within the fixed period of time under the fixed doses of acting substance *D*. The conditional probability $P(m,n \mid D)$ of certain numbers m and n at the certain D corresponds to the Bernoulli distribution:

$$P(m,n \mid D) = (m+n)!\frac{P(D)^m (1-P(D))^n}{m!n!}, \qquad (6.2)$$

where *P(D)* is the probability of death of a biological object at the obtained dosage *D*. Based on the experimental results, *P(D)* can be estimated only approximately by calculation of parameters for a definite parameter relationship *P(D)*, for instance Equation (6.1).

Usually, the dose-effect relationship $P(D)$ is simplified to the single quantitative or even qualitative characteristic. For example, for a certain level of probability q it is possible to determine an appropriate characteristic dose (quantile) $D_q = Arg\{P(D) = q\}$. Most often, the ED_{50} for $q = 0.5$ values are used, but $q = 0.16$, $q = 0.75$, $q = 0.84$ are also considered sometimes. However, in case of non-monotonic dependence $P(D)$ the value D_q can be ambiguous or even not exist if $P(D) < q$; for instance if (i) the part of population is resistant to the acting compounds and (ii) the suggested threshold *q* exceeds the fraction of the responsible part of population.

In accordance with the probabilistic nature of the biological activity concept, the most relevant methods for prediction of activity are those based on probabilistic theory and mathematical statistics, and the purpose of prediction is the complete relationship $\Pr(\text{Doze}) = P(D)$.

Unfortunately, in practice, the application of such approaches is strongly limited by the available experimental data, which in most cases are presented by semi-effective doses and even by qualitative characteristics "active/inactive".[51,52]

6.2.2 Experimental Data

The determination of biological activity is always associated with some experimental errors, which may be caused by variability of biological objects, inaccuracy of measurements due to the limited precision of the used equipment, inaccuracy of the personnel doing manual and mental work.

If the experimental measurements have been repeated several times, the resultant data are presented as average values and standard deviations (SDs) of the measurements. In many cases numerical data in the literature and, particularly, in databases are presented without SDs even in cases where such values could be calculated on the basis of primary data. Also, the results of testing in high-throughput assays for inactive compounds typically mean that the compound does not cause the studied effect at a certain threshold, e.g., at 10 μM, 1 μM, etc.[52]

Experimental errors associated with human error may be introduced both in experimental procedures *(e.g.*, inaccuracies of sample preparation) and in theoretical analysis of the study results *(e.g.*, errors in data drawing in publications, errors during the input of data into a computer).

As was concluded by Christoph Helma *et al.*:[53]

After summarizing our experiences with the quality assurance of chemical data in predictive toxicology, we conclude that the currently available databases and computational chemistry programs are too faulty to be trusted without further inspection. The development of reliable quality control procedures definitely needs more discussion, exchange of experience, and research activity. In this sense, we hope that we will raise some awareness in regard to data quality issues and quality assurance in predictive toxicology.

The necessity of quality control for chemical structures, particularly when the data are aggregated from different sources, was recently emphasized in another publication.[54]

However, the main source of scattering in experimental data is certainly determined by the variability of biological response. As was shown by comparison of results obtained in rodent carcinogenicity experiments, the concordance between the results taken from general literature and the results obtained from US National Toxicology

Program is only about 57%.[55] Therefore, the reproducibility of biological assays may be quite poor. It is well known that LD_{50} values for rodents obtained in different laboratories may vary significantly *(e.g.*, in LD_{50} studies performed by eleven laboratories to standardize a type A botulinum toxin assay for accessing the toxin in food contaminations, up to a ten-fold difference in results was shown).[56]

Notably, in actual practice training sets are not ideal: in addition to a simplified definition and high variability of biological activity they do not contain all chemical classes relevant to a particular biological activity, and information about all kinds of biological activity that can be revealed by a particular compound is always incomplete (no one compound is tested against all kinds of biological activity, and there is no one activity for which all possible ligands are known). Consequently, the probabilistic character of biological activity is caused not only by experimental errors of its determination but also by the incompleteness of available information.

6.3 Probabilistic Ligand-Based Virtual Screening Methods

Virtual screening methods are based on the modeling of the biological phenomenon of molecular recognition, either by the principle of complementarity or by the principle of similarity.[57]

Probabilistic ligand-based virtual screening methods look rather simple and fast; however, for their successful application it is necessary to have a training set of compounds with known activity. Probabilistic methods are based on the achievements of machine learning and have a long history, starting from pattern recognition methods.[4,5,58-63] Especially for the purposes of drug design, probabilistic methods were developed by Golender and Rozenblit,[64] and realized later in the expert system OREX.[65] In Section 6.4 we describe in detail the probabilistic method developed by our team, and to which the methods[7,66-74] and binary QSAR[51,52,75,76] are rather close in basic characteristics.

An important component of probabilistic ligand-based virtual screening methods is the design of the training set, which is the set of ligands available or selected to develop the virtual screening system.[77-81] The selection of this set and its usage strongly influence the overall performance of the final system.[82,83] Also, it is necessary to use the appropriate evaluation of prediction accuracy and reliability, and the representation and interpretation of biological activity prediction results is very important. Based on the probabilistic approach, it is possible to solve all these problems.

6.3.1 Preparation of Training Sets

Training sets should be representative for the compounds to be classified by the ligand-based virtual screening system.[83] Virtual screening is usually performed on a set containing a large number of ligands with a high diversity of molecular structure. For successful results, the diversity of structures from the training set must be comparable to those from the corresponding set used for virtual screening. As a rule, any training set must include sufficient active compounds as well as inactive ones.

It seems obvious that an "ideal" training set must include all tested active and inactive compounds. However, in practice it is necessary to be very careful during the design of training set because "a data set consisting of database chemical drawings and HTS assay measurements may be very misleading".[52]

There exist some other peculiarities, for instance every compound in the MDDR database (MDL® Drug Data Report[84]) has one or several records in the field "activity class", indicating that the compound is related to a certain therapeutic area. However, because of "umbrella patents", not each substance in MDDR was actually tested in biological assays. Those substances for which biological activity was studied in detail are called "principal compounds", and they have some records in the field "Action", such as experimental data on activity, LD_{50}, IC_{50}, K_i, *etc.* There are some publications, in which the training set is prepared on the basis of the MDDR database but this peculiarity is not taken into account. [7,73,74,85-87] In these publications, for each ligand from the training sets that was actually tested in biological assays there are several structurally similar molecules for which biological activity was assigned with the purpose of umbrella patenting. Therefore, unsurprisingly, structure similarity methods studied in these publications were shown to be rather successful during the validation.

In a well-designed training set the structural diversity must be as uniform as possible. It is very difficult to control such uniformity; however, the presence of closely similar compounds series in the set could (and have to) be checked, to avoid degeneracy.

In general, any ligand-based virtual screening method is based on direct or generalized similarity between the screened compound and compounds from the training set. Therefore, if such similarity is absent at all, no reasonable prediction of screened compound's properties can be made by using this training set.

6.3.2 Creation of Evaluation Sets

There are two fundamental problems in ligand-based virtual screening systems development: model selection and performance estimation. Almost invariably, all ligand-based methods have one or more adjustable parameters. To select the "optimal" parameter(s) or model for a given classification problem, it is necessary to utilize the independent evaluation set that was not used in the training procedure. Once the predictive system is developed, to estimate its performance, one must utilize the test set that was not used during the development process. To obtain the precise estimation of system performance, the test set must be large, ideally infinite. However, for a good choice of a model or its parameter(s), the number of compounds in training and evaluation sets must also be large. For theoretical analysis one can subdivide all available data into two (training and test) or three (training, evaluation and test) sets, which have to be approximately equal in size. However, to develop the actual working virtual screening system one must used all available data for the training; therefore, nothing remains for the evaluation and test sets. To overcome this contradiction, the most suitable methods for construction of evaluation (test) sets are K-Fold Cross-Validation (KF CV) and Leave-One-Out Cross-Validation (LOO CV).[87-90]

To perform KF CV a K-fold partition of the data set is created. For each from K experiments, K-l folds are used for training and the remaining one for testing. The true error estimate is obtained as the average of the separate K estimates. LOO CV is the degenerated case of KF CV, where K is chosen as the total number of examples. For a data set with *N* examples, perform *N* experiments, for each experiment use *N*-1 examples for training and the remaining one example for testing. The true error is estimated as the average error value on test examples – on all existing examples. Vapnik[4] proved several theorems, which stated unbiasedness and consistency of LOO CV estimation, if LOO CV is carefully performed: no information about the excluded compound is used for training and tuning the system based on a residual part of data set. Unfortunately, in the general case the computational time for LOO CV or even for KF CV will be very large due to the large number of sequential experiments. Fortunately, the probabilistic approaches usually have a small or zero number of tuned parameters and the LOO CV procedure can be performed quite easily. Thus, all available data can be used both for training and for evaluation of probabilistic ligand-based virtual screening systems. Earlier we have shown[91] that LOO CV provides a more rigorous accuracy estimation than the repeated many times 2-Fold (or jack-knife) CV.

6.3.3 Mathematical Approaches

Many different methods can be applied to virtual screening, and such methods are described in other chapters of this book and/or in the *Handbooks of Che-minformatics.*[3] Here we discuss the methods based on a probabilistic approach. Unfortunately, there are many publications in which the "probabilistic" or "statistical" approach items are farfetched. The Binary Kernel Discrimination[8-10,17,20] and the Bayesian Machine Learning Models[6] are actually special cases of Artificial Neural Networks; whereas the Probabilistic Neural Networks[14-16] are really similarity-based methods, which do not take into account the results of well-developed nonparametric regression methods.[92]

In virtual screening of the chemical structures set called the Screening Set (SS) for each compound $C \in SS$ any proposed method *P* should give the estimate *P(C),* which, being compared with a certain criterion, provides the basis for decision about the advisability of further testing of the chemical compound *C.* In other words, it is necessary to recognize whether compound C belongs to the class of compounds in which we are interested in, *i.e.,* to solve the task of pattern recognition (PR), which is a typical problem of Machine Learning (ML). There are a lot of publications, monographs and specialized journals devoted to the problems of ML and PR; machine learning approaches are widely used in cheminformatics (see, for example, refs. 11,67,69-71,73,87,93,94). Notably, the fundamentals of machine learning were developed much earlier than the informational technologies (IT) became

widely introduced. For example, Nilsson[61] noted, referring to Kanal,[59] that the engineers rediscover for themselves well-known methods of statistics. Later, in machine learning these methods were discovered for a second time, and now the same situation is observed in cheminformatics: methods well known to engineers and IT specialists are rediscovered once again. Mathematically, the estimate *P(C)* in many cases can be represented as:[1, 61]

$$P(C) = \sum_i a_i f_i(C), \qquad (6.3)$$

where $f_i(C)$ are the different functions of chemical structure of compound C, independent from the coefficients a_i. Various methods differ in the values of estimates $P(C)$, in the choice of functions $f_i(C)$, and in approaches that are used to determine the coefficients a_i. Without restriction of generality, let us suggest that the estimate $P(C)$ is a real quantity, and decision about advisability of further testing of chemical compound C is taken if $P(C) > \theta$, where θ is a threshold value. If the functions $f_i(C)$ represent physical-chemical parameters or other quantitative characteristics of molecular structure and/or every possible function of these characteristics, and coefficients a_i are determined on the basis of regression, PLS, SVM etc., then the estimate $P(C)$ is the results of QSAR method. If, at the same time, $f_i(C)$ are determined as a measure of similarity of structure of molecule C with another molecule C_i from the training set, it is a QSAR method based on similarity. If the functions $f_i(C)$ possess only the values 0 and 1, and coefficients a_i are determined on the basis of probabilistic approach, it is the method described in this Chapter.

It is widely accepted that probabilistic approach was first developed and applied in expert systems MYCIN[95,96] and PROSPECTOR.[97] In these expert systems the likelihood estimates are calculated for several competitive hypothesis *H* on the basis of available evidences *E*. In expert system MYCIN each hypothesis was estimated by a confidence factor $CF(H \mid E_1, E_2, ...)$ as a difference of estimates for measure of belief $MB(H \mid E_1, E_2, ...)$ and measure of distrust $MD(H \mid E_1, E_2, ...)$:

$$CF(H \mid E_1, E_2, ...) = MB(H \mid E_1, E_2, ...) - MD(H \mid E_1, E_2, ...), \qquad (6.4)$$

where MB and MD were calculated by aggregation of values for separate evidences E_i $MB(H \mid E_i)$ and $MD(H \mid E_i)$ according to the theory of probability rules. In fact, these aggregation rules are piecewise-linear approximations of simple formula:

$$CF(H \mid E_1, E_2, ..., E_m) = \frac{CF(H \mid E_1, E_2, ..., E_{m-1}) + CF(H \mid E_m)}{1 + CF(H \mid E_1, E_2, ..., E_{m-1}) CF(H \mid E_m)} \qquad (6.5)$$

These equations follow directly from the approach, which is very popular in recent times in Machine Learning, Data Mining, Text Mining and Knowledge Data Discovery, bioinformatics and cheminformatics, and called "naive Bayes classifier".[7,63,66,68,98,99] Such approach was applied for virtual screening by Labute and Gao,[51,52,75,76] and other researchers[67,69-71, 73], and also by the authors of this Chapter.[91,100-104]

When applied to virtual screening the naive Bayes classifier consists in the following.

Let a molecular structure of compound C to be represented by the set of descriptors $\{D_1, D_2, ..., D_m\}$, and the probability that it belongs to a given class A is estimated by a conditional probability $P(A \mid C) = P(A \mid D_1, D_2, ..., D_m)$.

Using Bayes' theorem, we write:

$$P(A \mid D_1, D_2, ..., D_m) = \frac{P(A) \cdot P(D_1, D_2, ..., D_m \mid A)}{P(D_1, D_2, ..., D_m)}, \qquad (6.6)$$

where $P(D_1,D_2,...,D_m \mid A)$ is the conditional probability of the descriptors set $\{D_1,D_2,...,D_m\}$ occurrence in a compound C from class A; $P(A)$ is the class A *prior probability*, $P(D_1,D_2,...,D_m)$ is the descriptors set $\{D_1,D_2,...,D_m\}$ *prior probability*. The "naïve" conditional independence assumptions mean that each descriptor D_i is conditionally independent of every other descriptor D_j for $j \neq i$. This means that:

$$P(D_1,D_2,...,D_m \mid A) \cong P(D_1 \mid A)P(D_2 \mid A)...P(D_m \mid A) = \prod_{i=1}^{m} P(D_i \mid A) \qquad (6.7)$$

As a result, the log-likelihood ratio of the conditional probability $P(A \mid D_1,D_2,...,D_m)$ of the class A and $P(\neg A \mid D_1,D_2,...,D_m)$ of its complement $\neg A$ can be expressed as:

$$ln\left[\frac{P(A \mid D_1,D_2,...,D_m)}{P(\neg A \mid D_1,D_2,...,D_m)}\right] = ln\left[\frac{P(A)}{P(\neg A)}\right] + \sum_i ln\left[\frac{P(D_i \mid A)}{P(D_i \mid \neg A)}\right] \qquad (6.8)$$

Taking into account that $P(\neg A \mid D_1,D_2,...,D_m) = 1 - P(A \mid D_1,D_2,...,D_m)$ and using Bayes' theorem for ratio $P(D_i \mid A) / P(D_i \mid \neg A)$, we find:

$$ln\left[\frac{P(A \mid D_1,D_2,...,D_m)}{P(\neg A \mid D_1,D_2,...,D_m)}\right] = ln\left[\frac{P(A)}{1-P(A)}\right] + \sum_i \left\{ln\left[\frac{P(A \mid D_i)}{1-P(A \mid D_i)}\right] - ln\left[\frac{P(A)}{1-P(A)}\right]\right\} \qquad (6.9)$$

In terms of the general formula (6.3) $P(C) = \sum_i a_i f_i(C)$ we can write:

$$P(C) = ln\left[\frac{P(A \mid D_1,D_2,...,D_m)}{P(\neg A \mid D_1,D_2,...,D_m)}\right], \qquad (6.10.a)$$

$$a_0 = ln\left[\frac{P(A)}{1-P(A)}\right], \quad f_0(C) \equiv 1, \qquad (6.10.b)$$

$$a_i = \sum_i \left\{ln\left[\frac{P(A \mid D_i)}{1-P(A \mid D_i)}\right] - ln\left[\frac{P(A)}{1-P(A)}\right]\right\}, \qquad (6.10.c)$$

$$f_i(C) = 1 \text{ if } D_i \in \{D_1,D_2,...,D_m\} \text{ and } f_i(C) = 0 \text{ if } D_i \notin \{D_1,D_2,...,D_m\}. \qquad (6.10.d)$$

Clearly, the constant a_0 can be included into threshold value θ, so that the function $f_0(C) \equiv 1$ is not necessary. We must stress that in such form the probabilistic approach has no tuned parameters at all. Some tuning of naive Bayes classifier can be performed by selection of the molecular structure descriptors (or $f_i(C)$) set. This is a wonderful feature in contrast to QSAR methods, especially to Artificial Neural Networks.

The describing functions $f_i(C) = 1$ if $D_i \in \{D_1,D_2,...,D_m\}$ (and $f_i(C) = 0$ otherwise) can be constructed on the basis of very wide approaches. In our investigations we use as descriptor sets $\{D_1,D_2,...,D_m\}$ substructure fragment descriptors (see below). For quantitative parameters the describing function $f_i(C)$ can be equal to 1 if molecule parameter(s) x_j satisfies the same condition k, *e.g.*, if value of x_j belongs to some interval or multidimensional x_j belongs to some region in appropriate space, and so on. Like this, naive Bayes approach was proposed and developed by Labute and Gao.[51,52,74,75,76,105]

The naive Bayes approach has several well known difficulties. The conditional independence of descriptors of a molecule structure is not true as a rule. The probability $P(A \mid D_i)$ estimations can be close or even equal to 0 or 1 and in such case coefficients a_i become too large or infinite. To overcome this problem, we have substituted the logarithms of probabilities ratios $ln[P(A \mid D_i)/(1-P(A \mid D_i))]$ for $ArcSin(2P(A \mid D_i)-1)$. The $ArcSin(2P(A \mid D_i)-1)$ shape coincides with the shape of $ln[P(A \mid D_i)/(1-P(A \mid D_i))]$ for almost all values of $P(A \mid D_i)$, but $ArcSin(2P(A \mid D_i)-1)$ values are bounded by the values $\pm\pi/2$.

Interestingly, the naive Bayes approach is "too simple", but as a rule it provides high accuracy of recognition[7,63,68].

6.3.4 Evaluation of Prediction Accuracy

When a classifier that provides the estimation of *P(C)* is constructed, its performance must be estimated. The most important estimation is of the prediction accuracy. To do this, an evaluation set (test set or validation set - see Section 6.3.2) must be used. The evaluation set (ES) must be relevant and include both type of examples - positive and negative ("active" and "inactive" compounds). For all compounds $C \in$ES estimations *P(C)* are calculated, and obtained values are analyzed using knowledge about the "true" classification of compounds in ES. Figure 6.4 shows the main features of this task.

Let us suggest that for compounds in ES we have values of some targeted molecular property. "Expert" divides ES into two parts: positive and negative examples. Using a constructed estimator we calculate *P(C)* values and, selecting the threshold value, divide ES into two other parts: predicted positive if $P(C) > \theta$ and predicted negative if $P(C) < \theta$. We compare prediction results with known data and calculate four numbers: *TP* is the number of true positives, *FP* is the number of false positives, *TN* is the number of true negatives, and *FN* is the number of false negatives (Figure 6.4)

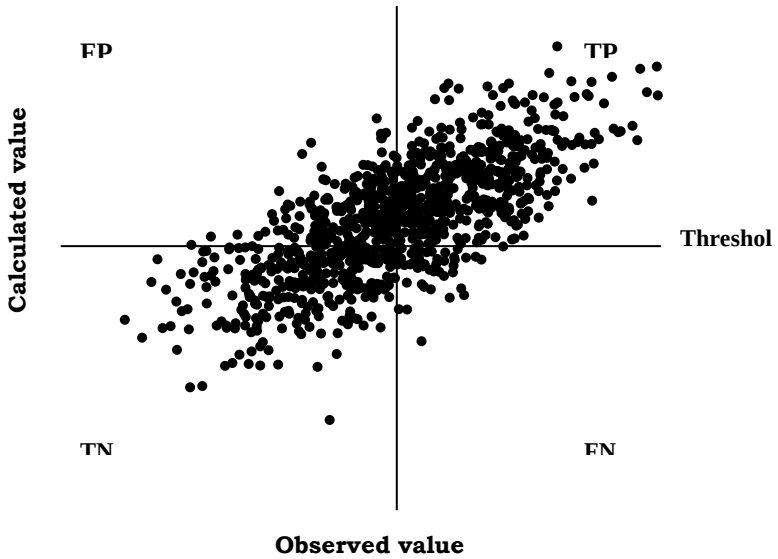


Figure 6.4 An artificially generated relationship between observed and calculated values of effect is shown as points with binomial distribution. Compounds are divided by the vertical line into actives and inactives according to the experimental values and by the horizontal line into predicted actives and inactives, at the selected threshold value. Compounds that fall into the appropriate

quadrants are classified based on the test as "True Positives" (TP), "True Negatives" (TN), "False Positives" (FP), and "False Negatives" (FN).

It is important to keep in mind that the situation illustrated in Figure 6.4 is a common case and it has symmetry in relation to errors: errors can be both in estimations *P(C)* and in experimental values. The result like that shown in Figure 6.4 occurs, even if the classifier is ideally true but experimental values are known with finite accuracy.

For pattern recognition or classification, usually, the following characteristics of recognition accuracy are used (see, for example, refs. 66,106-108):

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy (Concordance)} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Predictive value positive} = \frac{TP}{TP + FP}$$

$$\text{Predictive value negative} = \frac{TN}{TN + FN}$$

$$\text{False Negative Rate} = \frac{FN}{TP + FN}$$

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$

and others; each of them has some disadvantages. To minimize the disadvantages, Youden's index was proposed in 1950.[109] Youden's index summarizes the test accuracy into a single numeric value, Sensitivity + Specificity - 1, or:

$$YI = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 = \frac{TP \cdot TN - FP \cdot FN}{(TP + FN) \cdot (TN + FP)} \qquad (6.11)$$

The recognition accuracy estimation described above faces one very important problem: what is the best choice for the threshold value θ? To solve this problem, statistical decision theory is used.[110-113] The basis for this is an analysis of the so-called the Received Operating Characteristic (ROC) curve. By tradition, ROC is plotted as a function of true positive rate $TP/(TP + FN)$ (or sensitivity) *versus* false positive rate $FP/(TN + FP)$) (or 1-Specificity) for all possible threshold values θ. Figure 6.5 presents an example of such a ROC curve for the results obtained with our computer program PASS in predicting antineoplastic activity.

Estimation of the optimal threshold value is provided by minimizing a risk function, which depends on *a priori* probabilities of positive and negative examples and loss values for all four *(TP, FP, TN* and *FN)* possible results. If *a priori* probabilities or losses are not known, the optimal choice is *MiniMax* (*Mimi*mizing the *Max*imum possible loss) according to which the optimal threshold value must satisfy the condition "Sensitivity = Specificity". Another choice may be the maximum of Youden's index.

In any case, this approach uses several additional assumptions. For this reason in the last time in ML the recognition accuracy criterion of the Area Under the ROC Curve *(AUC)*, which is free of additional assumptions, becomes very popular.[7,63,68-71,106-108,112-116] Mathematically, *AUC* equals the probability that the estimation $P(C)$ assigns the higher value to a randomly drawn positive example C_+ than to the randomly drawn negative example C_-:

$$AUC(P) = PROBABILITY\{P(C_+) > P(C_-)\} \qquad (6.12)$$

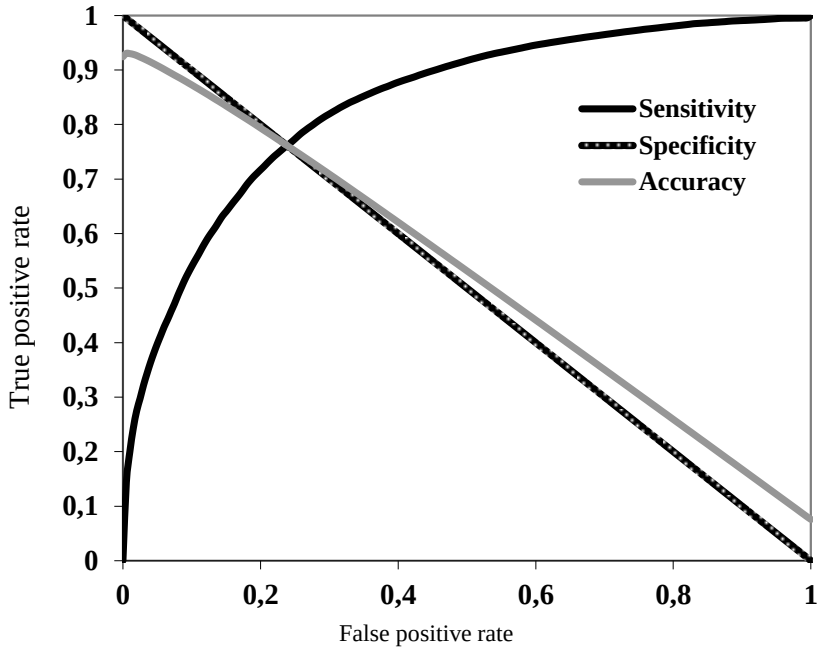


Figure 6.5 Relationships between the Sensitivity [TP/(TP + FN)] (shown by the curve), Specificity [TN/(TN + FP)] and accuracy (concordance) [(TP+TN)/(TP + FP + TN + FN)], as functions of False Positive Rate = [FP/(TN + FP)]. The estimations were obtained by PASS 2007 in leave-one-out cross-validation procedure for antineoplastic activity.

In our papers[91,117] we have used the Invariant Accuracy of Prediction (IAP) criterion, which exactly coincides with AUC , and it is calculated as:

$$IAP = \frac{NumberOf\{P(C_+) > P(C_-)\}}{NumberOf(C_-) \cdot NumberOf(C_+)} \tag{6.13}$$

In our computer program PASS (Section 6.4) we also use the Invariant Error of Prediction (IEP) criterion: $IEP \equiv 1 - IAP$.

Computationally, it is more convenient to calculate the estimate of prediction accuracy on ES as an Invariant Accuracy *{IA)*, which equals $2AUC(P) - 1$ and can be calculated as a result of comparison of estimates $P(C_+)$ for positive and $P(C_-)$ for negative examples through all pairs (each positive example and each negative example) in a form:

$$IA = \frac{\sum_{VS} Sgn(P(C_+) - P(C_-))}{N_+ N_-}, \quad Sgn(z) = \begin{cases} -1, & z < 0 \\ 0, & z = 0 \\ +1, & z > 0 \end{cases} \tag{6.14}$$

which is the difference of numbers of cases of true $P(C_+) > P(C_-)$ and false $P(C_+) < P(C_-)$ divisions of pairs of positive and negative examples, divided on the number of all pairs $N_+ N_-$. These are the following general cases:

- If all objects are predicted with the same value $P(C)$, then $IA = 0$.
- If the prediction is random and the estimates $P(C_+) > P(C_-)$ and $P(C_+) < P(C_-)$ have equal probabilities, then $IA = 0$ on average of probability.
- If all outcomes $P(C_+) > P(C_-)$ or $P(C_+) < P(C_-)$, then $IA = 1$ or $IA = -1$, respectively.

If inaccuracy of division of ES onto two classes exists, then:

$$IA = A \cdot \left(1 - \frac{m_-}{N_-} - \frac{m_+}{N_+} \right)$$

where $A \leq 1$ is the potential accuracy of the method, m_- is the number of compounds mistakenly described as negative examples (not found yet or not studied positive examples), and m_+ is the number of compounds mistakenly described as positive examples, for instance due to the errors in data used for creation of the sets, mistakes of personnel, *etc.* With:

$$IA = A \cdot \left(1 - \frac{m_-}{N_-} - \frac{m_+}{N_+} \right)$$

it is possible to compare the accuracy of several classifiers using ES with "errors of the teacher" correctly.

The IA (IAP, AUC)) criterion gives a robust estimation of general classifiers performance, but in the case of virtual screening to find several ligands at a top of ranked compounds list, the minimal number of decoys may be more important.[116,118] For this purpose, Enrichment Factor,[7,115,119-122] analysis of the robust initial enhancement (RIE)[116,118] and Boltzmann-Enhanced Discrimination ROC (BEDROC)[116] criteria were proposed.

6.3.5 Single-Targeted vs. Multi-Targeted Virtual Screening

Most existing virtual screening methods have been developed to be used for selection of hits with a single targeted activity.[22-24] However, most discovered pharmaceutical agents have several or even many kinds of biological activity. Some of these biological activities represent adverse/toxic effects, some others can be considered as a reason for utilization of known medicines according to new indications, which is called repositioning of drugs.[123-126]

Both new pharmacotherapeutics and adverse/toxic effect can be discovered on the basis of computer predictions with probabilistic methods. Different methods can be applied either sequentially or simultaneously. Early attempts to predict many kinds of biological activity simultaneously using such an approach were performed by Avidon and co-authors,[127] Golender and Rozenblit,[64,65] and Vassiliev and co-authors.[128]

Since the early 1990s, the authors have been developing the computer program PASS, which predicts many kinds of biological activity based on the structural formula of a compound.[29,30,32,40,41,43,100] This program, the present version of which predicts over 3000 kinds of biological activity with a mean accuracy of about 95%, is described in more detail below. Different PASS applications in virtual screening of multi-targeted ligands have been presented in several publications.[100-104,129]

The Prous Institute for Biomedical Research[130] is developing a computational method based on a wide range of molecular descriptors and binding profiles, called BioEpisteme®, which is claimed to facilitate the discovery of new medicines and new uses for existing drugs. Pre-requisites of the BioEpisteme approach are quite close to the PASS concept: "A drug may interact with multiple targets and produce more than one therapeutic response and/or adverse effect." Unfortunately, we could not find a detailed description of the method used in BioEpisteme in the available literature - only the very general scheme presented on the web-site.[130] Recently, the number of different molecular mechanisms covered by BioEpisteme was reported to be about 400.[131]

Quantum Pharmaceuticals[132] recently proposed a new method for toxicity prediction based on computation of small molecules' affinity to about 500 human proteins. The analysis of binding profiles for about 1000 known pharmaceutical agents led to establishment of a relation between the toxicological properties of a molecule and its activity against the selected representatives of approximately 50 protein families. This activity profile was further used as a "natural" set of descriptors for various toxicological endpoints predictions, including human-MRDD, human-MRTD, human-TDLo, mouse-LD_{50} (oral, intravenous, subcutaneous), rat-LD_{50} (oral, intravenous, subcutaneous, intraperitoneal), *etc.*[46]

Thus, probabilistic biological activity prediction methods can be used for both estimation of adverse/toxic effects in molecules under study and for finding the multi-targeted ligands, which might "yield drugs of superior clinical value compared with monotargeted formulations".[33]

6.4 PASS Approach

The computer program PASS was designed to predict many kinds of biological activity simultaneously based on the structural formulae of chemical compounds. Thus, PASS may estimate the biological activity profiles for virtual molecules, prior to their chemical synthesis and biological testing.

6.4.1 Biological Activities Predicted by PASS

The latest version of PASS (2007) predicts 3300 kinds of biological activity with a mean prediction accuracy of about 95%. PASS could predict about 1000 kinds of biological activity in 2004,[32] only 541 activities in 1998,[133] and 114 activities in 1996.[30]

The default list of predictable biological activities currently includes 374 pharmacotherapeutic effects *(e.g.*, antihypertensive, hepatoprotectant, nootropic, *etc.),* 2755 mechanisms of action, *(e.g.,* 5-hydroxytryptamine antagonist, acetylcholine Ml receptor agonist, cyclooxygenase inhibitor, *etc.),* 50 adverse and toxic effects *(e.g.*, carcinogenic, mutagenic, hematotoxic, *etc.)* and 121 metabolic terms *(e.g.,* CYP1A inducer, CYP1A1 inhibitor, CYP3A4 substrate, *etc.).* Information about novel activities and new compounds can be straightforwardly included into PASS.

In PASS *biological activities* are described qualitatively ("active" or "inactive"). Qualitative presentation allows integrating information concerning compounds tested under different terms and conditions and collected from many different sources, as in the general PASS training set. Any property of chemical compounds that is determined by their structural peculiarities can be used for prediction by PASS. Clearly, the applicability of PASS is broader than the prediction of biological activity spectra. For example, we used this approach to predict drug-likeness[134] and the biotransformation of drug-like compounds.[135]

6.4.2 Chemical Structure Description in PASS

The 2D structural formulae of compounds were chosen as the basis for *description of chemical structure* because this is the only information available in the early stage of research. Plenty of characteristics of chemical compounds can be calculated on the basis of structural formulae.[3,67,136-139] Earlier[29] we applied the Substructure Superposition Fragment Notation (SSFN) codes.[140] But SSFN, like many other structural descriptors, reflects rather abstraction of chemical structure by the human mind than the nature of the biological activity revealed by chemicals. The Multilevel Neighborhoods of Atoms (MNA) descriptors[91,141,142] have certain advantages over SSFN. These descriptors are based on the molecular structure representation, which includes the hydrogens according to the valences and partial charges of present atoms and does not specify the types of bonds. MNA descriptors are generated as a recursively defined sequence:

- zero-level MNA descriptor for each atom is the mark A of the atom itself;
- any next-level MNA descriptor for the atom is the sub-structure notation $A(D_1D_2...D_i...)$,

where D_i is the previous-level MNA descriptor for i –th immediate neighbours of the atom A.

The mark of atom may include not only the atomic type but also any additional information about the atom. In particular, if the atom is not included into the ring, it is marked by "-". The neighbour descriptors $D_1D_2...D_i...$ are arranged in uniquely, *e.g.,* in lexicographic order. Iterative process of MNA descriptors generation can be continued, covering first, second, *etc.* neighborhoods of each atom. MNA descriptors have a more general background than the descriptors,[67,137] which look like MNA.

The molecular structure is represented by the set of unique MNA descriptors of the 1st and 2nd levels. The substances are considered to be *equivalent* in PASS if they have the same set of MNA descriptors. Since MNA

descriptors do not represent the stereochemical peculiarities of a molecule, substances whose structures differ only stereochemically are formally considered as equivalent.

6.4.3 SAR Base

The PASS estimations of biological activity spectra of new compounds are based on the Structure-Activity Relationships data and knowledge-base (SAR Base), which accumulates the results of the training set analysis. The in-house developed general PASS training set currently (December 2007) includes about 117000 known biologically active substances (drugs, drug-candidates, leads, and toxic compounds). Since new information about biologically active compounds is discovered regularly, we perform a special informational search and analyze the new information, which is further used for updating and correcting the PASS training set.

6.4.4 Algorithm of Activity Spectrum Estimation

The algorithm of activity spectrum estimation is based on the above-mentioned Bayesian approach, but differs in several details. For each kind of activity A_k, which can be predicted by PASS, on the basis of a molecule's structure represented by the set of MNA descriptors { $D_1, D_2, ..., D_m$ } the following values are calculated:

$$S_{0k} = 2P(A_k) - 1, \tag{6.15.1}$$

$$S_k = \mathrm{Sin}\left[\frac{1}{m}\sum \mathrm{ArcSin}\left(2P(A_k \mid D_i) - 1\right)\right], \tag{6.15.2}$$

$$B_k = \frac{S_k - S_{0k}}{1 - S_k S_{0k}}, \tag{6.15.3}$$

where $P(A_k)$ is *a priori* probability to find a compound with activity of kind A_k; $P(A_k \mid D_i)$ is a conditional probability of activity of kind A_k if the descriptor D_i is present in a set of molecule's descriptors. For each kind of activity, if for all descriptors of molecule $P(A_k \mid D_i) = 1$, then $B_k = 1$; if for all descriptors of molecule $P(A_k \mid D_i) = 0$, then $B_k = -1$; if the relationship between descriptors of molecule and activity A_k does not exist and $P(A_k) \approx P(A_k \mid D_i)$, then $B_k \approx 0$.

The simplest frequency estimations of probabilities $P(A_k)$, $P(A_k \mid D_i)$ are given by:

$$P(A_k) = \frac{N_k}{N}, \qquad P(A_k \mid D_i) = \frac{N_{ik}}{N_i}, \tag{6.16}$$

where N is the total number of compounds in the SAR Base; N_k is the number of compounds containing the activity A_k in the activity spectrum; N_i is the number of compounds containing descriptor D_i in the structure description; N_{ik} is the number of compounds containing both the activity A_k and the descriptor D_i.

In PASS version 1.703 and later the estimations of probabilities $P(A_k)$, $P(A_k \mid D_i)$ are calculated as:

$$P(A_k) = \frac{\sum_n f_n(A_k) \sum_i g_n(D_i)}{\sum_n \sum_i g_n(D_i)}, \quad (6.17.a)$$

$$P(A_k \mid D_i) = \frac{\sum_n f_n(A_k) g_n(D_i)}{\sum_n g_n(D_i)}, \quad (6.17.b)$$

where $f_n(A_k)$ is the generic function of compound n belonging to a set of compounds containing the activity A_k in the activity spectrum, $f_n(A_k)$ is equal to 0 or 1; $g_n(D_i)$ is the measure of compound n belonging to the set of compounds containing descriptor D_i in the structure description, now $g_n(D_i)$ is equal to 0 or $\frac{1}{m_n}$, where m_n is the number of descriptors for the molecule n, and $\sum_i g_n(D_i) \equiv 1$ in this case.

The estimations Equations (6.17a, b) of probabilities $P(A_k)$, $P(A_k \mid D_i)$ not only increase the algorithm's prediction accuracy, but also open up new possibilities. For example, function $f_n(A_k)$ in the range [0, 1] can be considered as a measure of molecule n belonging to a fuzzy set of molecules that reveal activity A_k. The descriptor weight $g_n(D_i)$ can be considered in the same manner, and then the molecule structure descriptors can be of arbitrary nature, *e.g.*, such as in the refs. 51 and 52.

The main purpose of PASS is the prediction of activity spectra for new, possibly not yet synthesized compounds. Therefore, the general principle of the PASS algorithm is the exclusion from SAR Base of substances that is equivalent to the substance under prediction. So, if molecule n is equivalent to the molecule under prediction then this substance is excluded from sums in (Equations 6.17a,b).

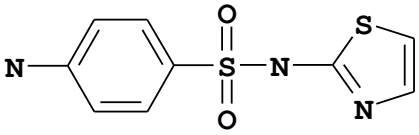
To obtain the qualitative ("Yes/No") results of prediction, it is necessary to define the threshold B_k values for each kind of activity A_k. On the basis of statistical decision theory (Section 6.3.4) it is possible using the risk functions minimization, but nobody can *a priori* determine such functions for all kinds of activity and for all possible real-world problems. Therefore the *predicted activity spectrum* is presented in PASS by the list of activities with probabilities "to be active" P_a and "to be inactive" P_i calculated for each activity. The list is arranged in descending order of $P_a - P_i$; thus, the more probable activities are at the top of the list. The list can be shortened at any desirable cutoff value, but $P_a > P_i$ is used by default. If the user chooses a rather high value of P_a as a cutoff for selection of probable activities, the chance to confirm the predicted activities by the experiment is high too, but many activities will be lost. For instance, if P_a >80% is used as a threshold, about 80% of real activities will be lost; for P_a >70%, the portion of lost activities is 70%, etc.

An example of prediction results for sulfathiazole is shown in Figure 6.6. This substance was found in SAR Base and was excluded from the SAR Base on prediction of its activity spectrum. The known (contained in SAR Base of PASS version 2007) activity spectrum includes the following activities: antibacterial, antibiotic, dihydropteroate synthase inhibitor, iodide peroxidase inhibitor. In Figure 6.6 the predicted activity spectrum includes 65 of 374 pharmacological effects, 176 of 2755 molecular mechanisms, 7 of 50 side effects and toxicity, 11 of 121 metabolism terms at default $P_a > P_i$ cutting points. All activities included in SAR Base are predicted with $P_a > P_i$. The activity of as a dihydropteroate synthase inhibitor is the second among the 176 predicted molecular mechanisms.

The probabilities P_a and P_i are functions of the initial estimation B_k defined by the equations:

$$FA_k(P_a) = B_k, FI_k(P_i) = B_k, \quad (6.18)$$

where the functions FA_k, FI_k are obtained as the final result of the *training procedure* which consists in the following.



```
>  <PASS_MNA_COUNT>
   32

>  <PASS_KNOWN_ACTIVITIES>
              Antibacterial
              Antibiotic
              Dihydropteroate synthase inhibitor
              Iodide peroxidase inhibitor

>  <PASS_RESULT_COUNT>
   65 of   374 Possible Pharmacological Effects at Pa > Pi
  176 of  2755 Possible Molecular Mechanisms at Pa > Pi
    7 of    50 Possible Side Effects and Toxicity at Pa > Pi
   11 of   121 Possible Metabolism at Pa > Pi

>  <PASS_EFFECTS>
0.886  0.004  Antiobesity
0.769  0.004  Antidiabetic
0.766  0.008  Antieczematic atopic
0.738  0.010  Antiprotozoal (Toxoplasma)
0.752  0.027  Antineoplastic (colorectal cancer)
0.727  0.002  Antiprotozoal (Coccidial)
0.651  0.043  Antineoplastic (brain cancer)
0.601  0.072  Antinephritic
0.601  0.091  Antiviral (Arbovirus)
0.578  0.083  Antineoplastic (lymphocytic leukemia)
0.578  0.083  Antineoplastic (non-Hodgkin's lymphoma)
0.418  0.005  Hypoglycemic
0.484  0.093  Allergic conjunctivitis treatment
0.408  0.019  Diuretic inhibitor
0.395  0.016  Antibacterial
0.421  0.043  Hematopoietic inhibitor
     ...
0.253  0.059  Antiprotozoal (Trichomonas)
0.209  0.021  Antibiotic
0.267  0.093  Anticoagulant
     ...
0.008  0.005  Histone acetylation inducer

>  <PASS_MECHANISMS>
0.732  0.004  Para amino benzoic acid antagonist
0.675  0.004  Dihydropteroate synthase inhibitor
0.661  0.028  Chloride peroxidase inhibitor
0.592  0.025  5 Hydroxytryptamine 6 agonist
0.591  0.062  Phthalate 4,5-dioxygenase inhibitor
    ...
0.265  0.227  Pterin deaminase inhibitor
0.138  0.100  Iodide peroxidase inhibitor
0.166  0.129  Cathepsin H inhibitor
    ...
0.141  0.140  3-Hydroxybenzoate 4-monooxygenase inhibitor

>  <PASS_TOXICITY>
0.555  0.112  Hematotoxic
0.442  0.139  Hepatotoxic
0.392  0.135  Nephrotoxic
0.275  0.066  Carcinogenic, female rats
0.205  0.114  Carcinogenic, female mice
0.341  0.269  Torsades de pointes
0.162  0.123  Carcinogenic

...
```

Figure 6.6 Structure of sulfathiazole and part of its predicted activity spectrum. Activities contained in the SAR Base of PASS version 2007 are marked in bold.

For each kind of activity and each MNA descriptor the estimations of probabilities $P(A_k)$, $P(A_k \mid D_i)$ are calculated by Equations (6.17a,b). For each kind of activity A_k, for each p of N_k active, and for each q of $N - N_k$ inactive compound in SAR Base, after excluding this compound, the estimates B_{kp} and B_{kq} are calculated. The N_k estimates of B_{kp} for active compounds are sorted in the ascending order; the $N - N_k$ estimates of B_{kq} for inactive compounds are sorted in the descending order. The functions FA_k, FI_k are calculated as conditional expectations:

$$FA_k(F) = \sum_{p=1}^{N_k} C_{N_k-1}^{p-1} F^{p-1} (1-F)^{N_k-p} B_{kp}, \qquad (6.19.a)$$

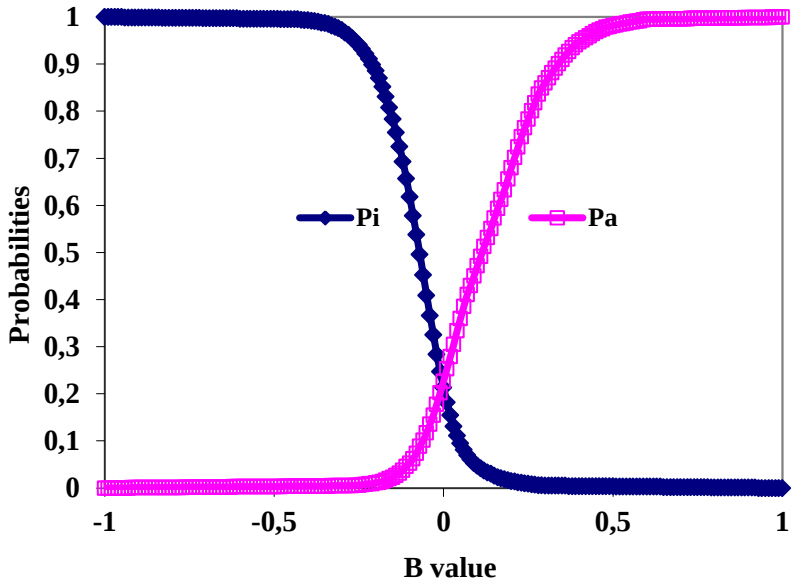
$$FI_k(F) = \sum_{q=1}^{N-N_k} C_{N-N_k-1}^{q-1} F^{q-1} (1-F)^{N-N_k-q} B_{kq}, \qquad (6.19.b)$$

where $C_n^m F^m (1-F)^n$ is the binomial distribution, $C_n^m = \dfrac{n!}{m!(n-m)!}$ is the binomial coefficient, F is in the range [0, 1]. Clearly, FA_k and FI_k are estimations of the quantile functions of the probability distributions of the estimations B_{kp} and B_{kq}. Thus, the probabilities P_a and P_i are both the measures of belonging to subsets of "active" and "inactive" compounds and the probabilities of the 1st and 2nd kinds of prediction error, respectively. These two interpretations of the probabilities P_a and P_i are equivalent and can be used in understanding the results of prediction.

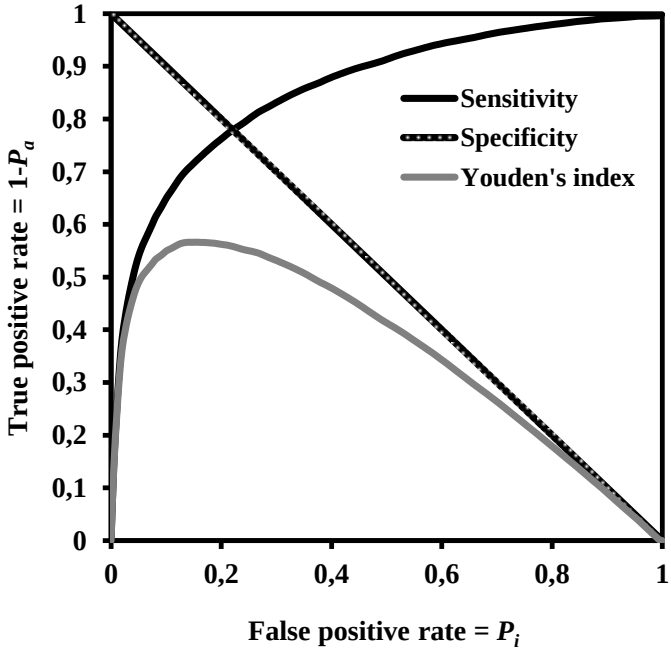
In Figure 6.7 shows an example of probabilities $P_a(B)$ and $P_i(B)$ estimation as functions of *B* value, and in terms of Sensitivity, Specificity and Youden's index, for antihypertensive activity in the SAR Base of PASS version 2007.

Leave one out cross-validation for 3300 kinds of biological activity and 117332 substances provides the estimate of PASS prediction accuracy during the training procedure. The average accuracy of prediction is about 94.7% according to the LOO CV estimation, while that for particular kinds of activity varies from 65% [System lupus erythematosus treatment, Immunomodulator (HIV)] to 99.9%) (Allergic rhinitis treatment, histone acetylation inducer). The estimated accuracy of prediction for all kinds of biological activity predicted by PASS is presented at the web site.[143]

The accuracy of PASS predictions depends on several factors, of which the quality of the training set seems to be the most important (Section 6.3.1). A perfect training set should include comprehensive information about biological activities known or possible for each compound. In other words, the whole *biological activity spectrum* should be thoroughly investigated for each compound included into the PASS training set. Actually, no database exists with information about biologically active compounds tested against each kind of biological activity. Therefore, information concerning known biological activities for any compound is always incomplete. We investigated the influence of the information's incompleteness on the prediction accuracy for new compounds. About 20000 "principal compounds" from the MDDR database (Section 6.3.1) were used to create the heterogeneous training and evaluation sets. At random, 20, 40, 60, 80% of information were excluded from the training set. Either structural data or biological activity data were removed in two separate computer experiments. In both cases it was shown that even if up to 60% of information is excluded the results of prediction are still satisfactory.[91] Thus, despite the incompleteness of information in the training set, the method used in PASS is robust enough to get reasonable prediction results.



a)



b)

Figure 6.7 Estimations of probabilities $P_a(B)$ and $P_i(B)$ as functions of B value (a) and in terms of sensitivity, specificity and Youden's index (b). The curves are obtained for activity antihypertensive based on data presented in SAR Base PASS version 2007.

6.4.5 Interpretation of Prediction Results

Only activities with $P_a > P_i$ are considered as possible for a particular compound.

It is necessary to remember that probability P_a first of all reflects the similarity of molecule under prediction with the structures of molecules that are the most typical in a sub-set of "actives" in the training set. Therefore, usually, there is no direct correlation between the P_a values and quantitative characteristics of activities.

Even an active and potent compound, whose structure is not typical of the structures of "actives" from the training set, may obtain a low P_a value and even $P_a < P_i$ during the prediction. This is clear from the way the functions $P_a(B)$ and $P_i(B)$ are constructed: the values P_a for "actives" and P_i for "inactives" are distributed fully uniformly. Taking this into account, the following interpretation of prediction results is possible.

If, for instance, P_a equals to 0.9, then for 90% of "actives" from the training set the *B* values are less than for this compound, and only for 10% of "actives" is this value higher. If we decline the suggestion that this compound is active, we will make a wrong decision with probability 0.9.

If P_a is less than 0.5, but $P_a > P_i$ then for more than half of "actives" from the training set the *B* values are higher than for this compound. If we decline the suggestion that this compound is active, we will make a wrong decision with a probability of <0.5. In such a case the probability of confirming this kind of activity in the experiment is small, but there is a more than 50% chance that this structure has a high degree of novelty and may become a New Chemical Entity (NCE).

If the predicted biological activity spectrum is wide, the structure of the compound is quite simple, and does not contain peculiarities that are responsible for the selectivity of its biological action.

If it appears that the structure under prediction contains a few new MNA descriptors (in comparison with the descriptors from the compounds of the training set), then the structure has low similarity with any structure from the training set, and the results of prediction should be considered as very rough estimates.

Based on these criteria, one may choose which activities have to be tested for the studied compounds on the basis of a compromise between the novelty of pharmacological action and the risk of obtaining a negative result in experimental testing. Certainly, one will also take into account a particular interest in some kinds of activity, experimental facilities, *etc*.

6.4.6 Selection of the Most Prospective Compounds

A fundamental limitation must be kept in mind: any observation, estimation or calculation has only restricted accuracy. In absolutely all cases instead of the desirable unknown intrinsic *Real* value we have only:

$$Observation = Real + Noise$$

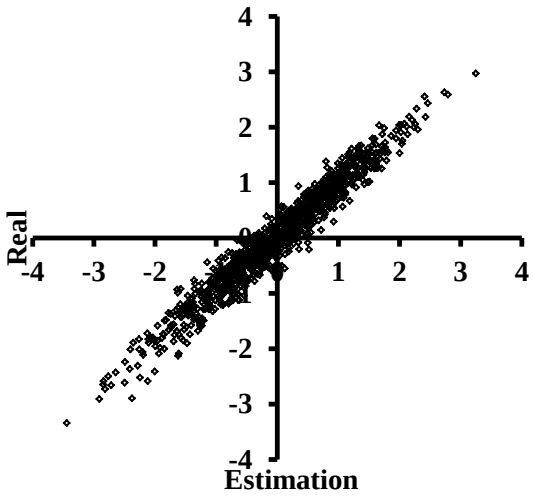
This is critically important for (virtual) screening especially. To highlight this, Figure 6.8 presents the generated data of 1000 points with binormal distribution and correlation coefficient square $R^2 = 0.95$ and $R^2 = 0.5$. Clearly, for $R^2 = 0.5$ the relationship looks like a weak tendency only. Figures 6.9-6.11 show the results of the selection of the 100 *Bests* (with the highest *Real* values) and the 100 *Winners* (with the highest *Estimation* values) among 1000000 "screened" examples. Clearly, only for $R^2 = 0.95$ is coincidence of the *Winners* and the *Bests* relatively good (about 60%), while for $R^2 = 0.5$ it is practically zero.

It is possible to perform a complete analysis of such relationships, but even the presented data provide enough evidence for the following conclusion: the method for (virtual) screening must be highly accurate, and/or many different virtual screening methods must be used in combination and/or the number of selected candidates must be sufficiently large at all stages of screening (in Figures 6.9 and 6.10, the number 100 is not "sufficiently large").[99,116,144,145]

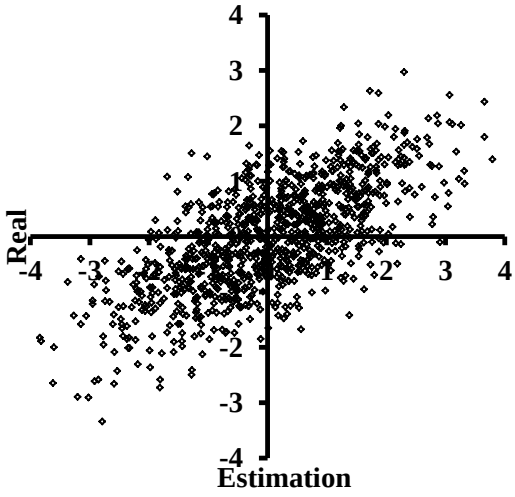
6.5 Conclusions

Since the predicted with PASS biological activity spectra contain the estimates of probabilities for the pharmacological main and side effects, molecular mechanisms of action and specific toxicity, the choice of the most prospective compounds from the available samples of chemical compounds can be realized on the basis of complex criteria. Both the presence of targeted biological effects with desirable mechanisms of action and the absence of unwanted adverse effects and toxicity have to be taken into account. In such studies, the search for leads with the required properties and their optimization to decrease the adverse and toxic effect, usually performed sequentially,

will be solved simultaneously. Moreover, it was shown that the algorithms used in PASS can be successfully applied for discrimination between the so-called drug-like and drug-unlike compounds,[134] which provides the possibility for extension of the applicability of the program by "filtering" in early stages chemical compounds, for which probability of becoming a drug is rather small.



a)



b)

Figure 6.8 An example of relationships between the available measured values and unavailable true values. 1000 points are presented, all values have a normal distribution. Error of measurement (calculation) corresponds to square of correlation coefficient $R^2 = 0.95$ (a) and $R^2 = 0.5$ (b).

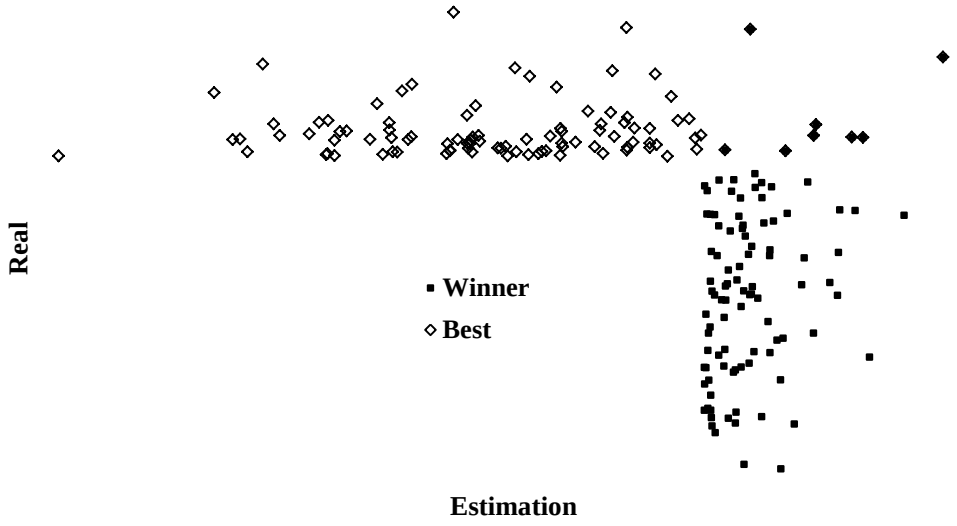


Figure 6.9 Example of relationship between the available measured (calculated) values and unavailable true values. The100 Winners and the100 Bests of 1000000 are presented. All compounds have a normal distribution, error of measurement (calculation) corresponds to the square of correlation coefficient $R^2 = 0.5$.

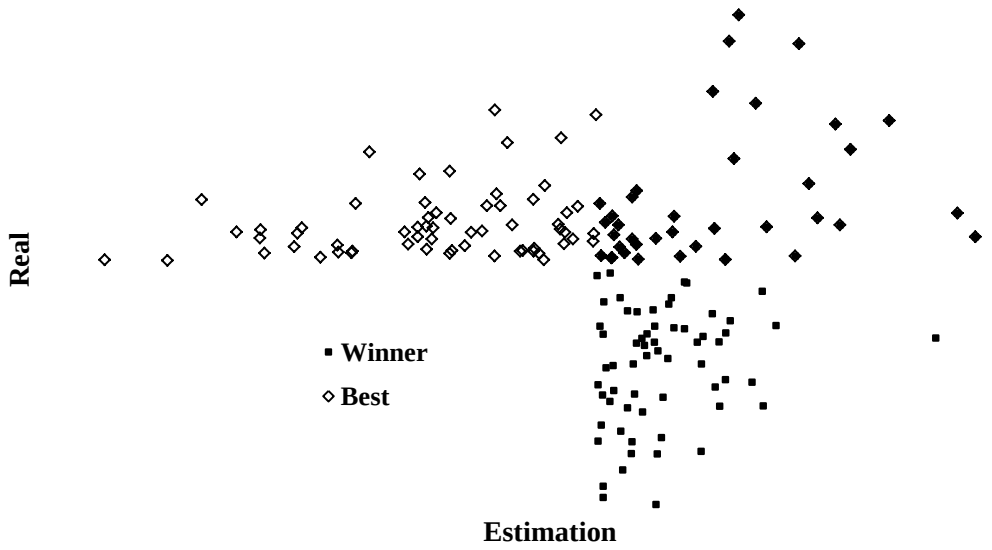


Figure 6.10 An example of relationship between the available measured (calculated) values and unavailable true values. The100 Winners and the100 Bests of 1,000,000 are presented, all values have a normal distribution, error of measurement (calculation) corresponds to the correlation coefficient $R^2 = 0.8$.

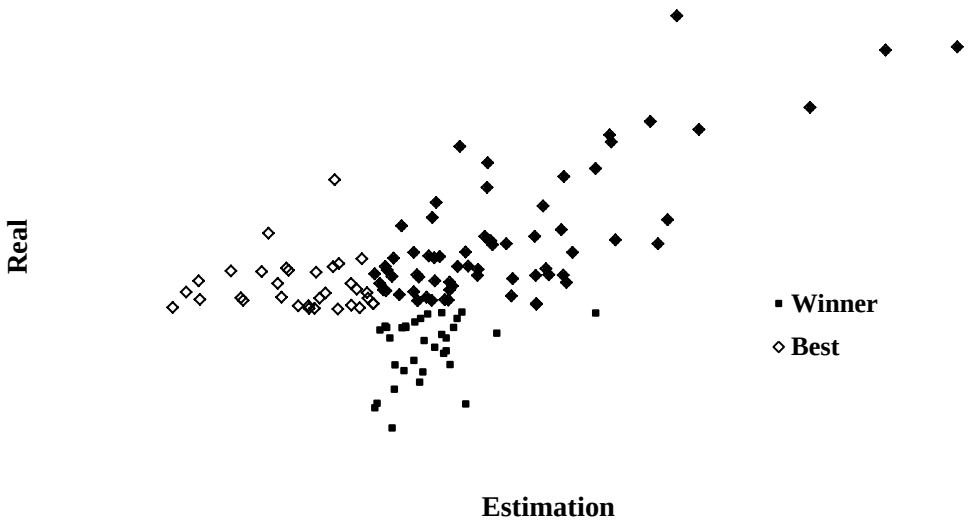


Figure 6.11 Example of relationship between the available measured (calculated) values and unavailable true values. The 100 Winners and the 100 Bests of 1 000 000 are presented. All values have a normal distribution; the error of measurement (calculation) corresponds to the correlation coefficient $R^2 = 0.95$.

The evolution of any molecule from hit to lead and from lead to drug-candidate typically is associated with the detailed evaluation of pharmacodynamics and pharmacokinetics of the compound. Using several different probabilistic methods for virtual screening together it might be possible to increase significantly the rate of promising substances in the selected subset.[101,103] A challenging task is to optimize simultaneously both pharmacodynamics and pharmacokinetics of lead compounds because it is very difficult to modify the appropriate molecular determinants that define the desired compound characteristics in a consistent manner. However, even this task might be solved using "an integrated software framework that monitors ligand (or library) alterations in the context of 'fitness landscape"".[26]

References

1. S. Wold, W. J. Dunn III, *J. Chem. Inf. Comput. Sci.*, 1983, **23**, 6-13.
2. D. Livingstone, *Data Analysis for Chemists. Applications to QSAR and Chemical Product Design*, Oxford, New York, Tokyo, Oxford University Press, 1995.
3. J. Gasteiger, ed., *Handbooks of Cheminformatics: From Data to Knowledge*, 4 Vols., Wiley-VCH, Weinheim, 2003.
4. V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*, New York, NY: Springer-Verlag, 1982.
5. V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.
6. D. Bahler, B. Stone, C. Wellington, D. W. Bristol, , *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 906-914.
7. E. O. Cannon, A. Amini, A. Bender, M. J. E. Sternberg, S. H. Muggleton, R. C. Glen and J. B. O. Mitchell, *J. Comput. Aided Mol. Des.*, 2007, **21**, 269–280.
8. J. W. Godden, J. R. Furr, J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 182-188.
9. J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.*, 2006, **46**, 1094-1097.
10. G. Harper, J. Bradshaw, J. C. Gittins, D. V. S. Green, A. R. Leach, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1295-1300.

11. C. Helma, T. Cramer, S. Kramer, L. De Raedt, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1402-1411.
12. R. D. King, S. Muggleton, R. A. Lewis, M. J. E. Sternberg, *Proc. Natd. Acad. Sci. USA*, 1992, **89**, 11322-11326.
13. R. D. King, S. H. Muggleton, A. Srinivasan, M. J. E. Sternberg, *Proc. Natd. Acad. Sci. USA*, 1996, **93**, 438-442.
14. P. D. Mosier, P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1460-1470.
15. T. Niwa, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 113 - 119.
16. T. Niwa, *J. Med. Chem.*, 2004, **47**, 2645-2650.
17. R. Rosipal, L. J. Trejo, B. Matthews, K. Wheeler in *Proceedings of 3rd International Symposium on PLS and Related Methods (PLS'03)*, Lisbon, Portugal, 2003, 249-260.
18. M. J. E. Sternberg, S. H. Muggleton, *QSAR Comb. Sci.*, 2003, **22**, 527-532.
19. W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 525-531.
20. D. J. Wilton, R. F. Harrison, P. Willett, *J. Chem. Inf. Model.*, 2006, **46**, 471-477.
21. B. Waszkowycz, T. D. J. Perkins, R. A. Sykes, J. Li, *IBM Systems Journal*, 2001, **40**, 2, 360-376.
22. P. D. Lyne, *Drug Discov. Today*, 2002, **7**, 1047-1055.
23. A. N. Jain, *Curr. Opin. Drug Discov. Devel.*, 2004, **7**, 396-403.
24. G. Klebe, *Drug Discov Today*, 2006, 11, 580-594.
25. T. I. Oprea, *Molecules*, 2002, **7**, 51–62.
26. T. I. Oprea, H. Matter, *Curr. Opin. Chem. Biol.*, 2004, **8**, 349–358.
27. H. Kubinyi, *Nat. Rev. Drug Discov.*, 2003, **2**, 665-668.
28. H. Kubinyi in *Computational Approaches to Structure Based Drug Design*, ed. R. M. Stroud, Royal Society of Chemistry, London, 2007, p. 24-45.
29. D. A. Filimonov, V. V Poroikov., E. I. Karaicheva et al., *Exper. Clin. Pharmacol. (Rus)*, 1995, **58**, 56-62.
30. V. V. Poroikov, D. A. Filimonov in *QSAR and Molecular Modelling Concepts, Computational Tools and Biological Applications*, ed. F. Sanz, J. Giraldo and F. Manaut, Prous Science Publishers, Barcelona, 1996, p. 49-50.
31. V. V. Poroikov, D. A. Filimonov, *J. Comput. Aid. Molec. Des.*, 2002, **16**, 819-824.
32. V. Poroikov, D. Filimonov in *Predictive Toxicology*, ed. C. Helma, Taylor & Francis, New-York, 2005, 459-478.
33. C. G. Wermuth, *Drug Discov. Today*, 2004, **9**, 826-827.
34. T. W. Schwartz, B. Holst, Trends Pharmacol. Sci., 2007, 28, 366-373.
35. A. C. R. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. B. O. Mitchell, C. Taroni, J. M. Thornton, *Structure*, 1998, **6**, 875-884.
36. M. H. V. Van Regenmortel, *J. Mol. Recognit.*, 1999, **12**, 1-2.
37. B. Y. Feng, A. Shelat, T. N. Doman, R. K. Guy, B. K. Shoichet, *Nat. Chem. Biol.*, 2005, **1**, 146-148.
38. M. H. V. Van Regenmortel, *J. Mol. Recognit.*, 2000, **13**, 1-4.
39. J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, J. Lankelma, *BioSystems*, 2006, **83**, 81-90.
40. V. V. Poroikov, D. A. Filimonov, A. P. Boudunova, *Automat. Document. Math. Linguist.*, Allerton Press, Inc., 1993, **27**, 40-43.
41. D. A. Filimonov, V. V. Poroikov in *Bioactive Compound Design: Possibilities for Industrial Use*, BIOS Scientific Publishers, Oxford, 1996, p.47-56.
42. V. Poroikov, D. Filimonov in *Rational Approaches to Drug Design*, ed. H.-D. Holtje, W. Sippl, Prous Science Publishers, Barcelona, 2001, p.403-407.
43. D. A. Filimonov, V. V. Poroikov, *Rus. Chem. J.*, 2006, **50**, 66-75.
44. A. F. Fliri, W. T. Loging, P. F. Thadeio, R. A. Volkmann, *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 2, 261–266.
45. A. F. Fliri, W. T. Loging, P. F. Thadeio, R. A. Volkmann, *J. Med. Chem.*, 2005, **48**, 22, 6918-6925.
46. P. O. Fedichev, A. A. Vinnik in *Fourth International Symposium Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2007)*, Moscow, 2007, p. 46.
47. R. R. Neubig et al., *Pharmacol. Rev.*, 2003, **55**, 597–606.
48. A. V. Hill, *Proc. Physiol. Soc.*, 1910, **40**, 4-7.
49. E. J. Ariens, *Molecular Pharmacology*, New York, Academic Press, 1964.
50. A. Macchiarulo, I. Nobeli, J. M. Thornton, *Nature biotechnology*, 2004, **22**, 8, 1039-1045.

51. P. Labute in *Proceedings of the Pacific Symposium on Biocomputing'99 World Scientific*, ed. R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, K. Londerdale, New Jersey, 1999, p. 444-455.
52. P. Labute, S. Nilar, C. Wiiliams, *Comb. Chem. and HTS*, 2002, **5**, 135-145.
53. C. Helma, S. Kramer, B. Pfahringer, E. Gottmann, *Environ. Health Perspect.*, 2000, **108**, 1029-1033.
54. A. M. Richard, L. S. Gold, M. C. Nicklaus, *Cur. Opin. Drug Discov. Devel.*, 2006, **9**, 314-325.
55. E. Gottmann, S. Kramer, B. Pfahringer, C. Helma, *Environ. Health Perspect.*, 2001, 109, 509-64. V. Golender, A. Rozenblit, *Logical and combinatorial algorithms for drug design*, Letchforth, England: Research Studies Pr., 1983, 289 p.
56. D. Sesardic, K. McLellan, T. A. N. Ekong, D. R. Gaines. *Pharmacol. Toxicol.*, 1996, **78**, 283-288.
57. M. H. J. Seifert, K. Wolf, D. Vitt, *BioSilico*, 2003, **1**, 143-149.
58. G. Sebestyen, *Decision Making Processes in Pattern Recognition*, MacMillan, New York, 1962.
59. L. Kanal et al. in *Proceedings of National Electronics Conference*, Chicago, 1962, 279-295.
60. M. A. Aizerman, E. M. Braverman, L. I. Rozonoer, *Automation and Remote Control*, 1964, **25**, 821-837.
61. N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*, New York, St. Louis, San Francisco,Toronto, London, Sydney, McGraw-Hill Book, 1965.
62. V. Vapnik, A. Lerner, *Automation and Remote Control*, 1963, **24**, 774-780.
63. P. Domingos, M. Pazzani, *Machine Learning*, 1997, **29**, 103–130.
64. V. Golender, A. Rozenblit, *Logical and combinatorial algorithms for drug design*, Letchforth, England: Research Studies Pr., 1983.
65. N. Veretennikova, A. Skorova et al. in *Quantitative Structure-Activity Relationships in Environmental Sciences – VII, Proceedings of QSAR 96*, Elsinore, Denmark, 1996, SETAC Press, 115-131.
66. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environmental Health Perspectives*, 2003, **111**, 1361-1375.
67. A. Bender, H. Y. Mussa, R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170-178.
68. P. A. Flach, N. Lachiche, *Machine Learning*, 2004, **57**, 233–269.
69. A. E. Klon, M. Glick, J. W. Davies*J. Chem. Inf. Comput. Sci.*, 2004, **44**, 2216-2224.
70. A. E. Klon, M. Glick, M. Thoma, P. Acklin, J. W. Davies, *J. Med. Chem.*, 2004, **47**, 2743-2749.
71. A. E. Klon, M. Glick, J. W. Davies, *J. Med. Chem.*, 2004, **47**, 4356-4359.
72. U. Brefeld, T. Scheer in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.
73. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.*, 2004, **2**, 3256–3266.
74. Nidhi, M. Glick, J. W. Davies, J. L. Jenkins, *J. Chem. Inf. Model.*, 2006, **46** (3), 1124 -1133.
75. H. Gao, C. Williams, P. Labute, J. Bajorath, *J. Chem. Comput. Sci.*, 1999, **39**, 164-168.
76. H. Gao, M. S. Lajiness, J. Van Drie, *J. Mol. Graphics and Modelling*, 2002, **20**, 259-268.
77. W. J. Streich, S. Dove, R. Franke, *J. Med. Chem.*, 1980, **23**, 1452-1456.
78. S. Dove, W. J. Streich, R. Franke, *J. Med. Chem.*, 1980, **23**, 1456-1459.
79. A. Golbraikh, A. Tropsha, *J. Comp.-Aided Mol. Design*, 2002, **16**, 5, 357-369.
80. A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, A. Tropsha, *J. Comp.-Aided Mol. Design*, 2003, **17**, 2, 241-253.
81. C. Szantai-Kis, I. Kovesdi, G. Keri, L. Orfi, *Molecular Diversity*, 2003, **7**, 1, 37-43.
82. R.P.W. Duin, E. Pekalska in *Computer Recognition Systems (Proc. of 4th Int. Conf. on Computer Recognition Systems CORES'05)*, eds. M. Kurzynski, E. Puchala, M. Wozniak, A. Zolnierek, Advances in soft computing, Springer Verlag, Berlin, 2005, 27-42.
83. L. D. Hughes, D. S. Palmer, F. Nigsch, J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2008, **48**, 220-232.
84. http://www.mdl.com/
85. R. P. Sheridan, J. Shpungin, *J. Chem. Inf. Comput. Sci.*, 2004, **44** (2), 727 -740.
86. P. Willett, *Drug Discov. Today*, 2006, **11**, 1046-1053.
87. B. Chen, R. F. Harrison, G. Papadatos, P. Willett, D. J. Wood, X. Q. Lewell, P. Greenidge, N. Stiefl, *J. Comput. Aided Mol. Des.*, 2007, **21**, 53–62.
88. T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
89. U.M. Braga-Neto, E. R. Dougherty, *Bioinformatics*, 2004, **20**, 374-380.

90. A. Tropsha in *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules, EuroQSAR 2004* Proceedings, E. Aki (Sener), I. Yalcin eds., Turkey, 2004, 25-29.
91. V. V. Poroikov, D. A. Filimonov, Yu. V. Borodina, A. A. Lagunin, A. Kos, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349-1355.
92. W. Hardle, *Applied Nonparametric Regression*, Cambridge University Press, 1990.
93. J. R. Rose in *Handbooks of Cheminformatics: From Data to Knowledge*, ed. J. Gasteiger, 4 Vols., Wiley-VCH, Weinheim, 2003, 1082-1097.
94. T. Ohgaru, R. Shimizu, K. Okamoto, M. Kawase, Y. Shirakuni, R. Nishikiori, T. Takagi, *J. Chem. Inf. Model.*, 2008, *48*, 207-212.
95. E. H. Shortliff, *Computer-Based Medical Consultation: MYCIN*, Elsevier, New York, 1976.
96. W. Van Melle, A. C. Scott, J. S. Benett, M. A. Peairs, *The EMYCIN manual*, Technical Report, Heuristica Programming Project, Stanford University, 1981.
97. J. Gashnig in *Introductory Readings in Expert Systems*, D. Michie ed., Gordon and Breach Science Publishers, NY, 1982.
98. M. Vogt, J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.*, 2007, **47**, 39-46.
99. M. Vogt, J. Bajorath, *J. Chem. Inf. Model.*, 2007, **47**, 337-341.
100. V. Poroikov, D. Akimov, E. Shabelnikova, D. Filimonov, *SAR & QSAR Environ. Res*., 2001, **12**, 327-344.
101. V. V. Poroikov, D. A. Filimonov, W.-D. Ihlenfeldt, T. A. Gloriozova, A. A. Lagunin, Yu. V. Borodina, A. V. Stepanchikova, M. C. Nicklaus, *J. Chem. Inform. Comput. Sci.*, 2003, **43**, 228-236.
102. A. A. Lagunin, O. A. Gomazkov, D. A. Filimonov, T. A. Gureeva, E. A. Dilakyan, E. V. Kugaevskaya, Yu. E. Elisseeva, N. I. Solovyeva, V. V. Poroikov, *J. Med. Chem.*, 2003, **46**, 3326-3332.
103. Geronikaki A., Dearden J., Filimonov D., et al., *J. Med. Chem.*, 2004, **47**, 11, 2870-2876.
104. A. A. Geronikaki, A. A. Lagunin, D. I. Hadjipavlou-Litina, P. T. Elefteriou, D. A. Filimonov, V. V. Poroikov, I. Alam, A. K. Saxena, *J. Med. Chem.*, 2008. In press.
105. J. Hert, P. Willett, D. J. Wilton, *J. Chem. Inf. Model.*, 2006, **46**, 462-470.
106. J. Swets, *Science*, 1988, **240**, 1285-1293.
107. J. A. Swets, R. M. Dawes, J. Monahan, *Sci. Am.*, 2000, **283**, 82-87.
108. N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, *J. Med. Chem.*, 2005, **48**, 2534-2547.
109. W. J. Youden, *Cancer*, 1950, **3**, 32-35.
110. A. Wald, *Statistical Decision Functions*, John Wiley and Sons, New York, 1950.
111. D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.
112. T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 861-874.
113. T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 882-891.
114. A. P. Bradley, *Pattern Recognition*, 1997, **30**, 7, 1145-1159.
115. A. E. Cleves, A. N. Jain, *J. Med. Chem.*, 2006, **49**, 2921-2938.
116. J.-F. Truchon, C. I. Bayly, *J. Chem. Inf. Model.*, 2007, **47**, 488-508.
117. Yu. Borodina, D. Filimonov, V. Poroikov, *Quant. Struct.-Act. Relat.*, 1998, **17**, 459-464.
118. R. P. Sheridan, S. B. Singh, E. M. Fluder, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1395-1406.
119. R. N. Jorissen, M. K. Gilson, *J. Chem. Inf. Model.*, 2005, **45**, 549-561.
120. T. A. Pham, A. N. Jain, *J. Med. Chem.*, 2006, **49**, 5856-5868.
121. M. H. J. Seifert, *J. Chem. Inf. Model.*, 2006, **46**, 1456-1465.
122. J. Kirchmair, S. Ristic, K. Eder, P. Markt, G. Wolber, C. Laggner, T. Langer, *J. Chem. Inf. Model.*, 2007, **47** (6), 2182 -2196.
123. T. T. Ashburn, K. B. Thor, *Nat. Rev. Drug Discov.*, 2004, 3, 673-683.
124. Y. Y. Li, J. An, S. J. Jones, *Genome Inform.*, 2006, **7**, 239-247.
125. L. A. Tartaglia, *Expert. Opin. Investig. Drugs*, 2006, **15**, 1295-1298.
126. C. G. Wermuth, *Drug Discov. Today*, 2006, 11, 160-164.
127. V. V. Avidon, V. S. Arolovich, S. P. Kozlova, L. A. Piruzyan, *Chem. Pharm. J.* (Rus), 1978, No 5, 88-92.
128. A. N. Kochetkov, P. N. Vassiliev, A. G. Breslaukhov, in *Abstr. First All-Union Conf. Theoret. Org. Chem.*, Volgograd, 1991, part 2, 500.

129. V. Poroikov, D. Filimonov, A. Lagunin, T. Gloriozova, A. Zakharov, *SAR & QSAR Environ. Res*., 2007, **18**, 101-110.
130. http://www.prousresearch.com
131. J. Prous, D. Aragones, in *Abstracts of National American Society Meeting*, Boston, Aug. 19-23, 2007.
132. http://www.q-pharm.com/
133. T. A. Gloriozova, D. A. Filimonov, A. A. Lagunin, V. V. Poroikov, *Chem-Pharm J.* (Rus), 1998, **32**, 32-39.
134. S. Anzali, C. Barnickel, B. Cezanne, M. Krug, D. Filimonov, V. Poroikov, *J. Med. Chem.*, 2001, **44**, 2432-2437.
135. Yu. Borodina, A. Sadym, D. Filimonov, V. Blinova, A. Dmitriev, V. Poroikov, *J. Chem. Inform. Comput. Sci.*, 2003, **43**, 1636-1646.
136. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, WILEY-VCH, 2000.
137. L. Xing, R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 796-805.
138. W. Guba in *Predictive toxicology*, ed. Christoph Helma, Marcel Dekker, 2003, 11-35.
139. A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comp.-Aided Mol. Design*, 2005, **19**, 693-703.
140. V. V. Avidon, I. A. Pomerantsev, A. B. Rozenblit, V. E. Golender, *J. Chem. Inf. Comput. Sci.*, 1982, **22**, 207-214.
141. D. Filimonov, V. Poroikov, Yu. Borodina, T. Gloriozova, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 666-670.
142. A. Lagunin, A. Stepanchikova, D. Filimonov, V. Poroikov, *Bioinformatics*, 2000, **16**, 747-748.
143. http://www.ibmc.msk.ru/PASS
144. R. P. Sheridan, S. K. Kearsley, *Drug Discovery Today*, 2002, **7**(17), 903-911.
145. J. W. Raymond, M. Jalaie, M. P. Bradley, *J. Chem. Inf. Comput. Sci.*, 2004, **44**(2), 601-609.