SUPPLEMENTARY MATERIAL:

**PASS (Prediction of Activity Spectra for Substances) Method**

In PASS biological activities are described qualitatively (active or inactive). Any property of chemical compounds, which is determined by their structural peculiarities, can be used for prediction by PASS. It is clear, that the applicability of PASS is broader than the prediction of biological activity spectra.

The MNA descriptors (both for predicting the activity spectrum of the compound and for adding substances to SAR Base) are generated only if molecular structure corresponds to the following criteria:

• each of the atoms in a molecule must be presented by atom symbol from the periodic table. Symbols of unspecified atom A, Q, \*, or R group labels are not allowed;

• each of the bonds in a molecule must be covalent bond presented by single, double or triple bond types only;

• molecular structure must include three or more carbon atoms;

• molecular structure must include only one component. Single atom parts like HCl, Cl-, OH-, Na+, etc., (hydrogen atoms do not take into account) are excluded from MNA descriptors generation;

• structure (main part, see the previous sentence) must be uncharged;

• the absolute molecular weight of a substance must be less than 1250 Da.

A relevant error will be generated whenever a molecular structure does not correspond to these criteria or input data contains some other errors.

The PASS estimations of biological activity spectra of new compounds are based on the Structure-Activity Relationships knowledge-base (SAR Base), which accumulates the results of the training set analysis. SAR Base includes a biological activities dictionary and MNA descriptors dictionary, data and knowledge on the "structure – biological activity" relationships, the database of structure of compounds from the training set with their biological activity spectra. The structure of the compound represented as a set of MNA descriptors.

Algorithm of activity spectrum prediction description may be done easy using well known Bayesian approach. For the chemical compound $C$, which molecular structure is represented by the set $\{D\_{1}, ..., D\_{m}\}$ of $m$ MNA descriptors, estimate the probability $P(A|C)$ that the compound $C$ has an activity $A$. According to the Bayes formula:

$$P(A|C) = \frac{P(A)P(C|A)}{P(C)} $$

where $P(A)$ is the activity $A$ *prior* probability; $P(C|A)$ is the conditional probability of compound $C$ providing that it has activity $A$; $P(C)$ is the compound $C$ prior probability.

On the assumption that descriptors $D\_{1}, ..., D\_{m}$ are independent one can write the probability $P(C|A)$ as the product of conditional probabilities for particular descriptors:

$$P(C|A) ≅ P(D\_{1}, ..., D\_{m}|A) = \prod\_{i=1}^{m}P( D\_{i}|A)$$

This expression is approximately true since the MNA descriptors are a fortiori dependent due to their generation method. However, we have not acceptable alternatives, and we cannot forget about the approximation of obtained formulas.

After simple transformations, we obtain the expression for the log-likelihood ratio of the conditional probabilities $P(A|C)$ for activity $A$ and $P(B|C)$ for activity $B$ as:

$$ln\left[\frac{P\left(A|C\right)}{P\left(B|C\right)}\right]≅ln\left[\frac{P\left(A\right)}{P\left(B\right)}\right]+\sum\_{i=1}^{m}\left\{ln\left[\frac{P\left(A|D\_{i}\right)}{P\left(B|D\_{i}\right)}\right]-ln\left[\frac{P\left(A\right)}{P\left(B\right)}\right]\right\}$$

In a particular case, where $B$ is the lack of activity $A$, we obtain:

$$ln\left[\frac{P\left(A|C\right)}{1-P\left(A|C\right)}\right]≅ln\left[\frac{P\left(A\right)}{1-P\left(A\right)}\right]+\sum\_{i=1}^{m}\left\{ln\left[\frac{P\left(A|D\_{i}\right)}{1-P\left(A|D\_{i}\right)}\right]-ln\left[\frac{P\left(A\right)}{1-P\left(A\right)}\right]\right\}$$

The implication of the expression is quite clear: the logarithm of the posterior likelihood ratio is the sum of the logarithm of the a priori likelihood ratio and the sum of individual descriptors contributions. And, if the activity is not dependent on this descriptor, then $P(A|D\_{i})=P(A)$ and the such descriptor has no affect on the results and its contribution to the sum is zero. This is the classical result of the probabilistic approach. But, apart from the already marked proximity, this result has another significant, well-known disadvantage: the contribution of some descriptors is too large and suppresses all other terms of the sum for which the conditional probability of activity is too close to 0 or 1, when its present in the structure. The most pronounced effect could be in the situation, when for probabilities $P\left(A|D\_{i}\right)$ used frequency estimates based on analysis of the training set and the values 0 and 1 are the rule rather than the exception.

To overcome this problem one can offer many different approaches and they were tested in the PASS development. The best result was obtained by using of so-called Fischer $ArcSin\left(2p-1\right)$ conversion instead of $ln\left[p/\left(1-p\right)\right]$: its shape coincides with the shape of $ln\left[p/\left(1-p\right)\right]$ for almost all values of $p$, but $ArcSin\left(2-1\right)$ values are bounded by the values ±π/2. The accuracy of prediction also improved after changing the sum of descriptor contributions by the their average value, that apparently compensates for the assumption of descriptors independence. Logarithm of the a priori likelihood ratio has little information about a specific predicted organic compound and can be omitted.

Bayesian approach described above explains why PASS prediction algorithm based on the following specific statistics: on the basis of a molecular structure represented by the set $\{D\_{1}, ..., D\_{m}\}$ of $m$ MNA descriptors, the $B\_{k}$ values are calculated for each activity $A\_{k}$:

$$B\_{k}=\frac{S\_{k}-S\_{0k}}{1-S\_{k}∙S\_{0k}}$$

$$S\_{k}=Sin\left[\sum\_{i}^{}ArcSin\left(2P\left(A\_{k}|D\_{i}\right)-1\right)/m\right]$$

$$S\_{0k}=2P\left(A\_{k}\right)-1$$

For each kind of activity, if for all descriptors of molecule $P\left(A\_{k}|D\_{i}\right)=1$, then $B\_{k}=1$; if for all descriptors of molecule $P\left(A\_{k}|D\_{i}\right)=0$, then $B\_{k}=-1$; if the relationship between descriptors of molecule and activity $A\_{k}$ does not exist and $P\left(A\_{k}|D\_{i}\right)≈P\left(A\_{k}\right)$, then $B\_{k}≈0$.

The PASS prediction algorithm uses the following data on the "structure-activity" relationships:

$N$ is the total number of compounds in the SAR Base;

$N\_{i}$ is the number of compounds contained descriptor $D\_{i}$ in the structure description;

$N\_{k}$ is the number of compounds contained the activity $A\_{k}$ in the activity spectrum;

$N\_{ik}$ is the number of compounds contained both the activity $A\_{k}$ and the descriptor $A\_{k}$.

The simplest frequency estimations of probabilities $P\left(A\_{k}\right)$ и $P\left(A\_{k}|D\_{i}\right)$ are given by:

$P\left(A\_{k}\right)=\frac{N\_{k}}{N}$*,* $P\left(A\_{k}|D\_{i}\right)=\frac{N\_{ik}}{N\_{i}}$

Estimation of PASS prediction accuracy and dependency required for calculation probabilities $Pa$ and $Pi$ on the basis of $B$ statistics, are the end result of the training procedure, which consists in the following. According to the SAR Base, formed on the basis of the training set, for each kind of activity $A\_{k}$, for each $N\_{k}$ active, and for each $N-N\_{k}$ inactive compound, $B$ statistics values are calculated. Calculations are carried out in the Leave-One-Out Cross-Validation (LOO CV), i.e., after the "exclusion" of the compound from SAR Base, for what is enough to not include it in sum. Smooth estimations of the distribution functions $Pa(B)$ and $Pi(B)$ are based on the obtained sets of $B$ statistics.

The probabilities $Pa$ and $Pi$ 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 $Pa$ and $Pi$ are equivalent and can be used for understanding the results of prediction. It allows constructing different critera for analyzing the results of prediction corresponding to the solution of specific practical problems.

An important feature of the PASS prediction algorithm is its robustness to the imperfection of information on the structure and biological activity spectrum of organic compounds in the training set. In a special study was showed that halving the actual well-known information about the structure or activity of organic compounds in the training set only slightly reduces the accuracy of prediction in cross-validation. There's demonstrated that the accuracy base on a LOO CV is even more rigid than for cross-validation.