Supplement

to the paper of Alexey Zakharov, Alexey Lagunin, Dmitry Filimonov and Vladimir Poroikov “Quantitative prediction of antitarget interaction profiles for chemical compounds”

1. QSAR Modelling Using Quantitative Neighborhoods of Atoms (QNA) Descriptors by GUSAR Program

QNA descriptors are calculated based on the connectivity matrix (**C**) and the standard values of ionization potential (IP) and electron affinity (EA) of atoms in a molecule. For any given atom i QNA descriptors are calculated as follow:

$$P_i = B_i \sum_k \left(\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)\right)_{ik} B_k \qquad \textbf{\textit{(2)}}$$

$$Q_i = B_i \sum_k \left(\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)\right)_{ik} B_k A_k\,, \qquad \textbf{\textit{(3)}}$$

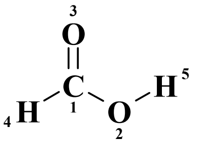
where $A_k = \tfrac{1}{2}(IP_k + EA_k)$, $B_k = (IP_k - EA_k)^{-\frac{1}{2}}$. The values of EA and IP collected from many different sources and used in this work are represented in Table 1. Though the value $\mu P - Q$ can be considered by convention as the partial atomic charge, where μ is the chemical potential, in general, the *P* and *Q* values are not the estimate of partial atomic charges, hardness or so forth.

Table 1. Electron affinity (EA) and first ionization potential (IP), electron volts, and atomic radius (AR), angstroms.

Atom	EA	IP	AR	Atom	EA	IP	AR	Atom	EA	IP	AR
H	0.75	13.60	0.46	**Kr**	-0.42	14.00	1.98	**Lu**	0.20	6.15	1.75
He	0.08	24.59	1.22	**Rb**	0.49	4.18	2.48	**Hf**	0.33	7.50	1.59
Li	0.62	5.39	1.55	**Sr**	-0.15	5.69	2.15	**Ta**	0.40	7.89	1.46
Be	-0.20	9.32	1.13	**Y**	0.31	6.22	1.81	**W**	0.67	7.98	1.40
B	0.28	8.30	0.91	**Zr**	0.33	6.84	1.60	**Re**	0.23	7.88	1.37
C	1.26	11.26	0.77	**Nb**	0.51	6.88	1.45	**Os**	1.44	8.73	1.35
N	0.44	14.53	0.71	**Mo**	0.68	7.09	1.39	**Ir**	1.57	9.10	1.35
O	1.46	13.62	0.73	**Tc**	0.54	7.23	1.36	**Pt**	1.10	8.96	1.38
F	3.45	17.42	0.71	**Ru**	1.10	7.37	1.34	**Au**	1.25	9.23	1.44
Ne	0.00	21.57	1.60	**Rh**	1.14	7.46	1.34	**Hg**	-0.19	10.44	1.57
Na	0.55	5.14	1.87	**Pd**	1.11	8.34	1.37	**Tl**	0.31	6.11	1.71
Mg	-0.31	7.64	1.60	**Ag**	1.22	7.58	1.44	**Pb**	1.39	7.42	1.75
Al	0.30	5.99	1.43	**Cd**	-0.43	8.99	1.56	**Bi**	0.97	7.29	1.82
Si	1.39	8.15	1.34	**In**	0.31	5.79	1.66	**Po**	1.97	8.42	1.56
P	0.75	10.49	1.30	**Sn**	1.39	7.34	1.58	**At**	2.90	9.20	1.48
S	2.00	10.36	1.04	**Sb**	0.90	8.64	1.61	**Rn**	-0.15	10.75	2.27
Cl	3.61	12.97	0.99	**Te**	1.97	9.01	1.70	**Fr**	0.48	3.98	2.80
Ar	-0.37	15.76	1.92	**I**	3.23	10.45	1.53	**Ra**	-0.15	5.28	2.35
K	0.50	4.34	2.36	**Xe**	-0.25	12.13	2.18	**Ac**	0.80	5.20	2.03
Ca	-0.19	6.11	1.97	**Cs**	0.47	3.89	2.66	**Th**	0.80	6.10	1.80
Sc	0.19	6.56	1.64	**Ba**	-0.15	5.21	2.23	**Pa**	0.84	6.00	1.62
Ti	0.33	6.82	1.46	**La**	0.30	5.59	1.87	**U**	0.82	6.19	1.53
V	0.53	6.74	1.34	**Ce**	0.25	5.54	1.83	**Np**	0.82	6.20	1.50
Cr	0.67	6.77	1.27	**Pr**	0.20	5.47	1.83	**Pu**	0.84	6.06	1.62
Mn	-0.17	7.43	1.30	**Nd**	0.20	5.53	1.82	**Am**	0.85	6.00	1.70
Fe	0.50	7.90	1.26	**Pm**	0.20	5.58	1.81	**Cm**	0.85	6.09	1.55
Co	0.66	7.86	1.25	**Sm**	0.20	5.64	1.80	**Bk**	0.82	6.23	1.49
Ni	1.16	7.64	1.24	**Eu**	0.20	5.67	2.04	**Cf**	0.84	6.27	1.42
Cu	1.23	7.72	1.28	**Gd**	0.20	6.15	1.80	**Es**	0.86	6.47	1.43
Zn	-0.44	9.39	1.39	**Tb**	0.20	5.86	1.78	**Fm**	0.86	6.60	1.38
Ga	0.30	6.00	1.39	**Dy**	0.20	5.94	1.77	**Md**	0.83	6.68	1.38
Ge	1.39	7.90	1.39	**Ho**	0.20	6.02	1.78	**No**	0.79	6.58	1.47

As	0.80	9.79	1.48	**Er**	0.20	6.11	1.76	**Lr**	0.85	6.69	1.30
Se	2.02	9.75	1.17	**Tm**	0.20	6.18	1.75	**Db**	0.46	6.43	1.14
Br	3.45	11.81	1.14	**Yb**	0.20	6.25	1.94	**Jl**	0.50	6.78	1.01

QNA describes each of the atoms in a molecule, and, at the same time, each of the *P* and *Q* values depends on the whole composition and structure of a molecule (Figure 1).



a)

$$\mathbf{C} = \begin{vmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

b)

$$\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right) = \begin{vmatrix} 1.40 & -0.59 & -0.57 & -0.57 & 0.14 \\ -0.59 & 1.27 & 0.14 & 0.14 & -0.54 \\ -0.57 & 0.14 & 1.13 & 0.13 & -0.02 \\ -0.57 & 0.14 & 0.13 & 1.13 & -0.02 \\ 0.14 & -0.54 & -0.02 & -0.02 & 1.13 \end{vmatrix}$$

c)

	EA	IP	*A*	*B*	P	Q
C	1.263	11.26	6.262	0.316	-0.00218	-0.1820
O	1.461	13.62	7.541	0.287	0.02944	0.3019
O	1.461	13.62	7.541	0.287	0.06199	0.5297
H	0.754	13.60	7.177	0.279	0.05812	0.4706
H	0.754	13.60	7.177	0.279	0.05304	0.3533

d)

Figure 1. Example of QNA description of *formic acid*: (a) structure diagram; (b) connectivity matrix; (c) exponent of the connectivity matrix; (d) electron affinities, ionization potentials, variables of equations (2) and (3), *P* and *Q* values for each atom of *formic acid* molecule.

From Figure 1c it is clear that any atom influences the others, though the influence decreases with the increase of the distance between them; *e.g.* components of matrix $\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)$ for atom 1 (C) are: 1.40 for atom 1 itself, -0.59 for its immediate neighbor atom 2 (O), -0.57 for atoms 3 (O) and 4 (H), and 0.14 for atom 5 (H).

The algorithm of QNA descriptor calculation is quite simple due to uselessness of the matrix $\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)$ itself, only product of $\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)$ by vector is necessary, and that matrix $\mathbf{C}$ consists of 0 and 1 only.

QNA describe each particular atom of a molecule, at the same time, each *P* or *Q* value depends on the total molecule composition and structure. To use this molecule structure representation in equation (1) we have proposed to calculate each $f_i(S)$ function of the structure of a molecule as the average value of $g_i(P,Q)$ function of *P* and *Q* variables for those m atoms in a molecule, which have two or more immediate neighbors:

$$f_i(S) = \tfrac{1}{m}\sum_k g_i(P_k,Q_k) \tag{4}$$

After substitution of expression (4) into the equation (1) and interchange of summations we found:

$$y_{pred} = a_0 + \sum_i a_i \tfrac{1}{m}\sum_k g_i(P_k,Q_k) = \tfrac{1}{m}\sum_k \left(a_0 + \sum_i a_i g_i(P_k,Q_k)\right) \tag{5}$$

According to the equation (5) the estimate y_{pred} for molecule can be interpreted as an average of values predicted for particular atoms in a molecule. Formally, QNA descriptors represent a molecule structure by two descriptors only (*P* and *Q*), in contrast to numerous traditional descriptors used in QSAR.

Since *P* and *Q* values have different scales (standard deviations are 0.023 and 0.208, respectively), we made the normalization to optimize a family of functions $g_i(P,Q)$.

Normalization has been performed by calculation of the average values (E_P and E_Q), standard deviations (D_P and D_Q) and correlation between *P* and *Q* values (R_{PQ}):

$$P' = \frac{P - E_P}{D_P}, \qquad Q' = \frac{Q - E_Q}{D_Q} \tag{6}$$

$$U = \frac{P' + Q'}{\sqrt{2(1 + R_{PQ})}}, \qquad V = \frac{P' - Q'}{\sqrt{2(1 - R_{PQ})}} \tag{7}$$

The orthonormal U and V have zero mean, unit variance and they are uncorrelated.

Chebyshev polynomials were chosen as the family of functions $g_i(P,Q)$, and the orthonormal U and V values have been additionally transformed by using hyperbolic tangent, so, the "normalized QNA" vary from -1 to 1. After this, the functions $g_i(P,Q)$ in equation (5) are represented using Chebyshev polynomials as:

$$g_i(P,Q) = T_{uv}(P,Q) = \mathrm{Cos}(u * \mathrm{ArcCos}(\mathrm{TanH}(U))) * \mathrm{Cos}(v * \mathrm{ArcCos}(\mathrm{TanH}(V))), \tag{8}$$

where the integer numbers u , v = 0, 1, 2, ... define 2-dimesional Chebyshev polynomial degree. The final equation for estimate y_{pred} using QNA descriptors is:

$$y_{pred} = \tfrac{1}{m}\sum_k \left(a_0 + \sum_{uv} a_{uv} T_{uv}(P_k, Q_k)\right) = a_0 + \sum_{uv} a_{uv} T_{uv}, \qquad T_{uv} = \tfrac{1}{m}\sum_k T_{uv}(P_k, Q_k). \tag{9}$$

QNA descriptors and their polynomial transformations (6)-(8) do not provide information on the shape and volume of a molecule although this information may be important for determination of the structure-activity relationships. Therefore, these parameters were added to QNA descriptors. Topological length of a molecule was calculated as the maximal distance between any two atoms and the volume of a molecule – as the sum of each atom's volume, $\frac{4}{3}\pi R^3$, where R is the atomic radius (see Table 1 (AR)).

The Chebyshev polynomials are arranged in ascending order of their degrees $u + v$. For $u + v$ =1 they are $T_{1,0}$, $T_{0,1}$; for $u + v$ =2 they are $T_{2,0}$, $T_{1,1}$, $T_{0,2}$; for $u + v$ =3 they are $T_{3,0}$, $T_{2,1}$, $T_{1,2}$, $T_{0,3}$, etc. The first, second and third power of topological length and volume of a molecule were used. The number of initial variables equals to the number of Chebyshev polynomials plus the number of the first, second and third power of topological length and volume of a molecule.

The number of initial variables depends on the number of compounds in the training set. If the number of compounds in the training set less than 25, then the initial variables are 24. If the number of compounds in the training set varied from 25 to 100, then the initial variables depends on the following equation:

$$A = (Ln(B) \times 18.755) - 36.37,$$

where *A* – initial variables and *B* – the number of compounds in the training set. If the number of compounds in the training set varied from 100 to 2000, then the number of initial variables equal to one-half of the number of compounds in the training set. If the number of compounds in the training set exceeds 2000, then the initial variables are 1000.

GUSAR algorithm uses three procedures describe below for generation of different QSAR models based on QNA descriptors.

1) U and V values from (8) are multiplied by value, which are randomly chosen from 0.98 to 1.09.

2) Coefficient before matrix $\mathbf{C}$ is randomly chosen from -0.1 to -0.9.

3) Calculation of QNA descriptors is randomly chosen from two cases. First case is when *P* and *Q* values are calculated for all atom types. Second case is when *P* and *Q* values are calculated for each atom, which has connection with more than one carbon.

All these procedures are used for creation of each QSAR model based on QNA descriptors. These procedures allow obtaining the different QNA models.

2. PASS (Prediction of Activity Spectra for Substances) Algorithm

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 Multilevel Neighbourhoods of Atoms (MNA) *descriptors* are based on the molecular structure representation, which includes the hydrogen atoms according to the valences and partial charges of other atoms and does not specify the types of bonds. MNA descriptors are generated as recursively defined sequence:

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

where $\boldsymbol{D_i}$ is the previous-level MNA descriptor for $\boldsymbol{i}$–th immediate neighbor’s of the atom $\boldsymbol{A}$.

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 neighbor descriptors $\boldsymbol{D_1D_2...D_i...}$ are arranged in unique manner, e.g., in lexicographic order. Iterative process of MNA descriptors generation can be continued covering first, second, etc. neighborhoods of each atom.

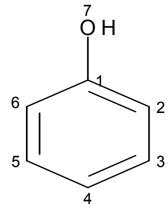
This process can be continued iteratively covering 2nd, 3rd, etc. neighbourhoods of the atom. We use 2nd level descriptors in the present version of PASS.

Multilevel Neighbourhoods of Atoms (MNA) structure descriptors of a molecule are generated on the basis of connection table (C) and table of atoms types (A) presented the substance. Connection table contains data on the covalent bonds in a molecule. Various bond types are not specified (topological approximation). All hydrogens based on valences and partial charges of atoms are taken into account. The types of atoms are specified according to the data presented in Table 1.

Table 1. Classification of different atom types used in calculation of descriptors

Class name	Elements
H	H
C	C
N	N
O	O
F	F
Si	Si
P	P
S	S
Cl	Cl
Ca	Ca
As	As
Se	Se
Br	Br
Li*	Li, Na
B*	B, Re
Mg*	Mg, Mn
Sn*	Sn, Pb
Te*	Te, Po
I*	I, At
Os*	Os, Ir
Sc*	Sc, Ti, Zr
Fe*	Fe, Hf, Ta
Co*	Co, Sb, W
Sr*	Sr, Ba, Ra
Pd*	Pd, Pt, Au
Be*	Be, Zn, Cd, Hg
K*	K, Rb, Cs, Fr
V*	V, Cr, Nb, Mo, Tc
Ni*	Ni, Cu, Ge, Ru, Rh, Ag, Bi
In*	In, La, Ce, Pr, Nd, Pm, Sm, Eu
Al*	Al, Ga, Y, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Tl
R*	R, He, Ne, Ar, Kr, Xe, Rn, Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr, Db, Jl

Example of structure presentation by zero-, first- and second-levels MNA descriptors for the phenol's molecule is shown in Figure below.



Atom	MNA/0	MNA/1	MNA/2
1	C	C(CC-O)	C(C(CC-H)C(CC-H)-O(C-H))
2	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
3	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
4	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
5	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
6	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
7	-O	-O(C-H)	-O(C(CC-O)-H(-O))
8	-H	-H(C)	-H(C(CC-H))
9	-H	-H(C)	-H(C(CC-H))
10	-H	-H(C)	-H(C(CC-H))
11	-H	-H(C)	-H(C(CC-H))
12	-H	-H(C)	-H(C(CC-H))
13	-H	-H(-O)	-H(-O(C-H))

MNA descriptors for phenol.
MNA/0, MNA/1,MNA/2 - zero, first and second levels
of MNA descriptors.

It is shown that usage of the 1st & 2nd levels MNA descriptors provides the best accuracy of prediction. MNA descriptors are generated for each substance. Unique integer identifier is assigned to each particular descriptor according to the descriptors' dictionary.

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, the substances whose structures differ only stereochemically, are formally considered as equivalent.

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. The in-house developed PASS training set includes more than 260,000 known biologically active substances (drugs, drug-candidates, leads, and toxic compounds). Since new information about biologically active compounds is discovered regularly, we perform the special informational search and analyse the new information, which is further used for updating and correcting the PASS training set.

MNA descriptors $\{D_1, \ldots, D_m\}$ for each kind of activity A_k the following B_k values are calculated:

$B_k = (S_k - S_{0k})/(1 - S_k \bullet S_{0k})$,

$S_k = Sin[\sum_i ArcSin(2P(A_k|D_i) - 1)/m]$,

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

where $P(A_k|D_i)$ is a conditional probability of activity of kind A_k if the descriptor D_i is present in a set of molecule's descriptors; $P(A_k)$ is a priori probability to find a compound with activity of kind A_k. For any kind of activity A_k, if $P(A_k|D_i)$ is equal to 1 for all descriptors of a molecule, then $B_k = 1$; if $P(A_k|D_i)$ is equal to 0 for all descriptors of a molecule, then $B_k = -1$; if there is no relationship between the molecule's descriptors and activity of kind A_k, and, so, $P(A_k|D_i) \approx P(A_k)$, then $B_k \approx 0$.

Up to the PASS version 1.703 the algorithm of prediction was based on the following data:
n is the total number of compounds in the SAR Base;
n_i is the number of compounds containing descriptor D_i in the structure description;
n_k is the number of compounds containing the kind of activity A_k in the activity spectrum;
n_{ik} is the number of compounds containing both the kind of activity A_k and the descriptor D_i.
And the estimations of probabilities $P(A_k)$, $P(A_k|D_i)$ are given by:

$$P(A_k) = n_k/n, \qquad P(A_k|D_i) = n_{ik}/n_i.$$

In PASS version 1.703 and later instead of integers n_i and n_{ik} the sums g_i and g_{ik} of descriptors weights w are used, where $w = 1/m$, and m is the number of MNA descriptors of individual molecule. This modification increases the accuracy of prediction significantly. So, right now the estimations of probabilities $P(A_k|D_i)$ are given by:

$$P(A_k|D_i) = g_{ik}/g_i.$$

The calculations are done by using $n–1$, $g_i–w$, and, when the kind of activity A_k is contained in its activity spectrum in the SAR Base, by using $n_k–1$ and $g_{ik}–w$. Here $w = 1/m$, and m is a number of MNA descriptors in molecule under prediction and its equivalent in the SAR Base. The B_k values are calculated using MNA descriptors, which are found in SAR Base, i.e., for descriptors of a molecule under prediction with $g_i > 0$ or $g_i–w > 0$, in the case of structure «exclusion».

To take the «yes/no» qualitative prediction it is necessary to determine B-statistics threshold values for each kind of activity A_k. Using theory of statistical decision this can be done on the basis of risk function's minimization. But nobody can a priori specify the risk functions for all activity kinds and all possible practical tasks. Therefore, the predicted activity spectrum in PASS is presented by the rank-order list of activities with probabilities «to be active» *Pa* and «to be inactive» *Pi*, which are the functions of *B*-statistics for a molecule under prediction. The *B*-statistics functions *Pa* and *Pi* are the results of the training procedure described below. The list is arranged in descending order of *Pa–Pi*; thus, the more probable activity kinds are at the top of the list. The list can be shortened at any desirable cutoff value, but *Pa*>*Pi* is used by default. If the user chooses rather higher value of *Pa* as a cutoff for selection of probable activities, the chance to confirm the predicted activities by the experiment is also high, but many existing activities will be lost. For instance, if *Pa*>80% is used as a threshold, about 80% of real activities will be lost; for *Pa*>70%, the portion of lost activities is 70%, etc.

For each compound from the training set MNA descriptors are generated and its known activity spectrum and set of descriptors are stored in the SAR Base. If this compound has the equivalent structure in SAR Base, only new activities are added to activity spectrum. After inclusion of all information from the training set(s) into SAR Base the values n, g_i, n_k, g_{ik} are calculated. For each compound in the SAR Base and for each activity kind A_k, values B_k of B-statistics are calculated. Calculations are done taking into account the described above «exclusion» of processed compound. For each activity kind A_k, the calculated values B_k are subdivided into two samples: for active and inactive compounds. These obtained samples are used for calculation the smooth estimations of B-statistics distribution functions on the following basis.

Suppose the sample x_1, ..., x_n of n values of random variable X, which has an unknown distribution function $F(x)$. Using of an empirical step-function for approximation of F often faults because of small n. To provide the smooth estimation of $F(x)$, the inverse function $x(F)$ is calculated as the conditional expectation of random variable X:

$$x(F) = \sum_i (n–1)!\bullet F^{i-1}/(i–1)!\bullet(1–F)^{n-i}/(n–i)!\bullet x'_i,$$

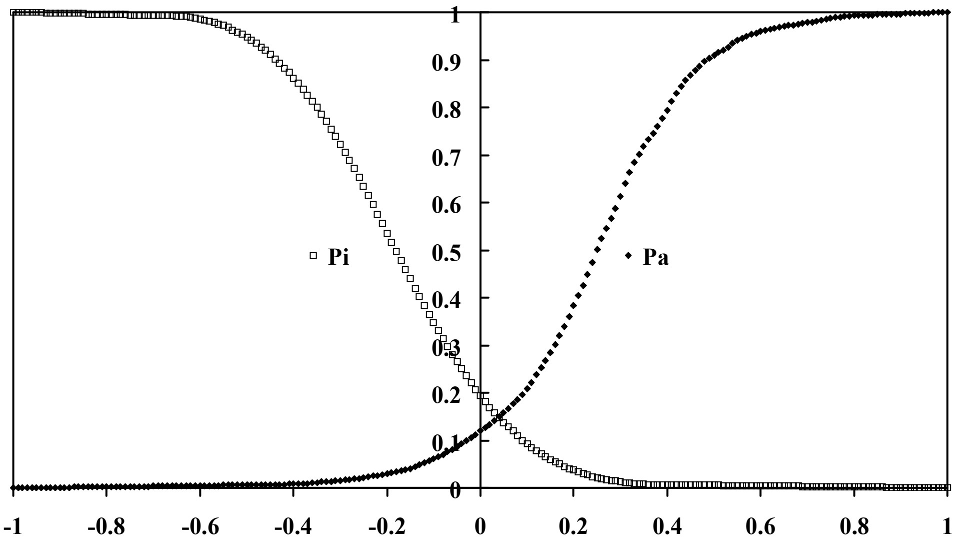
where $(n–1)!\bullet F^{i-1}/(i–1)!\bullet(1–F)^{n-i}/(n–i)!$ is the binomial distribution, and x'_1, ..., x'_n ($x'_1 < x'_2 < ... < x'_n$) is the ranked sample x_1, ..., x_n. The distribution function $F(x)$ is given reciprocal function of quantiles $x(F)$.

Each sample of B values for active compounds is arranged in the ascending order; each sample of B values for inactive compounds is arranged in descending order. The described above quantiles $b(F)$ are calculated. As a result, for each appropriate kind of activity the probabilities Pa and Pi are given by:

$$b_{active}(Pa) = B, \qquad b_{inactive}(Pi) = B.$$

By definition the probabilities Pa and Pi are also the probabilities of the 1st and 2nd kinds of prediction error at the threshold B, respectively. They can be also interpreted as the measures of belonging to fuzzy subsets of «active» and «inactive» compounds. Both interpretations of probabilities Pa and Pi are equivalent and can be used for interpreting the results of prediction. They can also be used for construction of different criteria for prediction results' analysis corresponded to specific practical problems.

The example of the probabilities $Pa(B)$ and $Pi(B)$ for activity «Alpha adrenoreceptor antagonist» as functions of B-statistics value is shown in the figure.



PASS prediction accuracy is estimated using average ***IEP*** for all predictable activity kinds.
IEP is calculated for each type of activity in PASS prediction:

$$\boldsymbol{IEP = \#(\,B_0 > B_1)/(\,N_oN_1),}$$

where $\mathbf{B_0}$ and $\mathbf{B_1}$ are values of ***B***-statistics for some pair of inactive and active compounds in the training set,
$\mathbf{N_o}$ is the number of inactive compounds in the training set,
$\mathbf{N_1}$ is the number of active compounds in the training set.
If the values of ***B***-statistics for any active compounds are higher than the values of ***B***-statistics for all inactive compounds, then **IEP=0**. It means that all «active» and «inactive» compounds for the current type of activity from the training set were divided absolutely correct during LOO CV procedure. If the values of ***B***-statistics for any active compounds are the same as the values of ***B***-statistics for all inactive compounds, then **IEP=0.5**. This means that prediction is not correct.