Table 1 Correlation coefficient matrix of the selected descriptors

	Xt	MATS1m	PJI3	Mor23u	nCs	H-046
Xt	1	0.183	−0.489	0.323	−0.401	−0.226
MATS1m		1	−0.528	0.396	−0.374	−0.613
PJI3			1	−0.449	0.364	0.206
Mor23u				1	−0.648	−0.287
nCs					1	0.421
H-046						1

value of the mean effect (MF) was calculated. This calculation was performed by use of Eq. 2.

$$\mathrm{MF}_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \quad (2)$$

MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule, and m is the descriptor's number in the model. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign shows the direction of variation in the toxicity values as a result of the increase (or reduction) of the descriptor values. The mean effect values are −0.043, 1.071, −0.081, 0.035, −0.004, and 0.023 for Xt, MATS1m, PJI3, Mor23u, nCs, and H-046. By interpreting the descriptors contained in the model, it is possible to gain useful chemical insights into the toxicity of phenols. For this reason, an acceptable interpretation of the QSAR results is provided below.

The first descriptor which has appeared in the model is Xt (total structure connectivity index). Connectivity indices are among the most popular topological indices and are calculated from the vertex degree of the atoms in the H-depleted molecular graph. Xt is a connectivity index contemporarily accounting for all the atoms in the graph. Also the total structure connectivity index is the square root of the simple topological index that is proposed for measuring molecular branching [28]. The mean effect of Xt has a negative sign, which indicates that an increase in the molecular branch leads to a decrease in its pIG_{50} value.

The second descriptor is MATS1m (Moran autocorrelation—lag 1/weighted by atomic masses), which is a 2D autocorrelation descriptor. In this descriptor the Moran coefficient is a distance-type function, and is any physicochemical property calculated for each atom of the molecule, for example atomic mass, polarizability, etc. The Moran coefficient usually takes a value in the interval [−1, +1]. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values. Therefore, the molecule atoms represent a set of discrete points in space and the atomic property is the function evaluated at those points. The physicochemical property in this case is the atomic mass. MATS1m has a positive sign, illustrating a greater mean effect value than that of the other descriptors, which indicates that this descriptor had a significant effect on the toxicity and that the pIG_{50} value is directly related to this descriptor. Hence, it was concluded that by increasing the molecular mass the value of this descriptor increased, causing an increase in its pIG_{50} value.

The third descriptor is PJI3 (3D Petijean shape index), which is a geometrical descriptor. The Petitjean shape index is a topological anisometry descriptor also called a graph-theoretical shape coefficient that is calculated from the topological radius and the topological diameter obtained from the distance matrix representing the considered molecular graph. PJI3 has a negative sign, which indicates that the pIG_{50} is inversely related to this descriptor.

Mor23u is the fourth descriptor appearing in the model. It is a 3D-MoRSE descriptor. 3D MoRSE descriptors (3D molecule representation of structures based on electron diffraction) are derived from infrared spectra simulation using a generalized scattering function [28]. This descriptor was proposed as signal 23/unweighted. Mor23u has a positive sign, which indicates that the pIG_{50} is directly related to this descriptor.

The fifth descriptor is nCs which is one of the functional groups. nCs represents the number of total secondary $C(sp^3)$. The mean effect of nCs has a negative sign, which indicates that an increase in the number of secondary $C(sp^3)$ of the molecule leads to a decrease in its pIG_{50} value.

The final descriptor of the model was the H-046 (H attached to C0 (sp^3)). It is one of the atom-centered fragment descriptors that describe each atom by its own atom type and the bond types and atom types of its first neighbors. This descriptor represents the first neighbor (hydrogen) of carbon atoms. This descriptor has a positive sign, which indicates that the pIG_{50} is directly related to this descriptor.

In summary, it is concluded that the molecular branching, the molecular mass, the molecular shape, the number of secondary $C(sp^3)$ of molecules, and the first neighbor (hydrogen) of carbon atoms are of major importance in the toxicity of the compounds studied.

Genetic algorithm: multiparameter linear regression

We used a GA for selection of the most relevant descriptors. Multiparameter linear correlation of pIG_{50} values for 150 different phenolic compounds in the training set was achieved by the GA by use of the six descriptors selected, and the following equation was obtained: