Statistical Modeling of Chlorinated Chemical Compounds Bioactivity

VLADIMIR MUKHOMOROV
Physical Department
State Polytechnic University
St. Petersburg
RUSSIA

Abstract: - A general regression equation based on physical concepts of the behavior of a polar molecule in a condensed medium is derived. The regression equation makes it possible, from a unified standpoint, to statistically significantly explain the toxicity of both chlorine-substituted benzenes and saturated and unsaturated chlorinated hydrocarbons. Statistically significant explanatory molecular features that determine the bioactivity of drugs have been identified.

Key-Words: - Regression, significance, information, quality criteria, collinearity, toxicity, intermolecular, electronic, pseudopotential, chlorinated chemical compounds

Received: May 26, 2021. Revised: April 15, 2022. Accepted: May 12, 2022. Published: June 6, 2022.

1 Introduction

Much attention has recently been paid to finding various quantitative relationships that link variations in the molecular structure of chemical compounds to their biological activity. For these purposes, either abstract statistical models are used (leaving the mechanism of biological activity undisclosed) or assumed physico-chemical ideas about the possible behaviour of chemical compounds in the biosystem. This article analyzes the toxic effects of chlorobenzene derivatives, as well as saturated and unsaturated chlorine-containing compounds. Using the methods of the theory of intermolecular interactions, the corresponding regression relationships will be derived here for the purposes of statistical analysis of the relationship between the structure of a molecule and its biological activity.

2 Problem Formulation

In accordance with modern concepts, the biological activity of chemical compounds is determined by their physicochemical properties at the macroscopic level (solubility, distribution, permeability), as well as at the microscopic level (electronic characteristics of molecules). In this regard, it can be assumed that the biological effect is equally determined by two circumstances: the transport of the molecule to the site of action and the physicochemical interaction of the molecule with the receptor. Attempts to obtain appropriate regression equations that take into account different factors have been repeatedly discussed in the literature [1,2]. The authors of the papers [3,4] point out the difficulties in the physicochemical interpretation of the observations used in these studies.

3 Problem Solution

3.1 Chlorinated benzene derivatives

The Hansch model [5,6] has been the most widely used in recent years. This model relates the bioactivity of chemical compounds to their lipophilic characteristics. In many practical cases, this model has proved useful. Therefore, let us check whether the bioactivity (average lethal doses of benzene chlorine derivatives for white rats upon oral administration [7]) is really related to the partition coefficient P of the substance in the octanol–water system. We use the well-known Hansch equation

$$A = B_0 + B_1 \lg P + B_2 (\lg P)^2, \qquad (1)$$

here $A = 1000/LD_{50}$ is bioactivity, B_0, B_1, and B_2 are some unknown parameters that are defined by minimizing the squared deviation of function values (1) from known experimental values. Toxicity and $\lg P$ values for a number of substituted benzenes are given in Table 1.

Using equation (1), the coefficient of determination $R^2 = 0.237$ was determined. This coefficient characterizes the magnitude of the statistical

relationship between activity A and distribution parameters for immiscible solvents. That is, the "explaining" ability of the model is only 23.7%. The statistical significance of the multiple correlation coefficient can be tested by using the following inequality [8]: $t = |R|\cdot(N - m - 1)^{0.5}/(1 - R^2)^{0.5} = 1.67 < t_{0.05}^{cr}(f = 9) = 2.26$; where m is the number of explanatory variables, N is the sample size. Consequently, it must be recognised that there is no reliable relationship between regression (1) and the observed toxicity at the 95% confidence level. Comparison of the multiple correlation coefficient $R = 0.49$ with the critical value $R_{0.05}^{cr}(N - m - 1 = 9; m = 2) = 0.697$ [9] also indicates that the correlation coefficient R is insignificant at the significance level $\alpha = 0.05$. Moreover, as the analysis showed, the use of various modifications of equation (1) (see, for example, [10]), including the use of the Hammett constants σ, also does not allow one to establish a relationship between changes in the structure of the molecule and the variation of the biological response. Apparently, the relationship between the molecular structure and the bioactivity of a chemical compound must be sought based on other properties of this series of compounds. In molecular pharmacology, it is known that the biological action of a chemical compound depends on its ability to accumulate in certain areas of the body through interaction with sensitive local biostructures of the body.

Analysis of the relationship between various toxicity indicators of chemical compounds and their physicochemical properties showed the following [11]. The greatest number of correlations is found with the properties of low-molecular weight compounds, which are determined at the electronic level and are related to the energy of intermolecular interaction of molecules [12,13].The presence of long-range components in the energy of the intermolecular interaction must lead to a concentration gradient of chemical compounds. This contributes to the emergence of a diffusion flow of low molecular weight chemical compounds directed towards the active center.

The process of binding exogenous molecules in the body can approximately determine the additive components of pairwise intermolecular interactions. Difficulties in consistently taking into account the contributions to the intermolecular interaction are due to the variety of types of pair interactions of molecules, as well as the possibility of correctly taking into account the influence of the condensed phase on these interactions. Intermolecular interactions are usually divided into two groups according to their radius of action and relative strength: specific and non-specific (universal). The first group includes anisotropic pairwise quasi-chemical bonds (donor-acceptor complexes, hydrogen bonds) that arise when the electron shells of interacting molecules overlap markedly. Nonspecific interactions include various electrostatic interactions, as well as short-range dispersion forces. These interactions are determined not only by the individual properties of individual chemical compounds, but also by the properties of the biosubstrate, that is, the condensed phase in which exogenous chemical compounds are distributed. The proposed mathematical model should highlight the additive components of the interaction, that is, orientation interactions, polarization and short-range contributions (on a molecular scale). These contributions are related to the dipole moment, electronic polarizability, ionization potential, and also to the position of one-electron energies MO on the energy scale of an isolated molecule. This approach makes it possible to establish the existence of possible causal relationships between molecular features and bioresponse.

The analyzed series of chlorobenzene derivatives (Table 1) is interesting from the point of view that in this case it is possible to move away from the problems associated with the conformational transitions of molecules. In addition, these molecules have similar sizes and are not expected to be involved in metabolic transformations.

It is known [16,17] that chlorinated compounds have good acceptance properties. This allows them to participate in the formation of donor-acceptor molecular complexes due to electron transfer. The change in total energy (ΔE) when a bond is formed between atom s of the donor molecule and atom t of the acceptor molecule can be written as follows

$$\Delta E = - q_s q_t/(\kappa_s R_{st}) + 2\Sigma_m\Sigma_n(C_s^m C_t^n \Delta\beta_{st})^2/(\varepsilon_m - \varepsilon_n). \quad (2)$$

Here the summation occurs over the m occupied molecular orbitals (MO) of the donor and over the n vacant MO of the acceptor; ε_m and ε_n are the energies of single-particle MO of the donor and acceptor, respectively; $\Delta\beta_{st}$ is the change in the resonance integral of the interacting atoms s and t at the distance R_{st} between the atoms; C_s^m and C_t^n are the expansion coefficients of MO in atomic orbitals; κ_s is the static permittivity of the condensed medium.

Table 1
Molecular parameters of chlorobenzene derivatives and their mean lethal doses (LD_{50}) for white rats after oral administration of drugs.

№	Chemical compounds	lg*P*	ε_{nb}^{0}, eV	I_1 [*)], eV	μ_1, *D*, [14]	α_1 [**)], 10^{24}cm^3	*Z*, arb. units	*H*, *bits*	A_{exp}, 1000/LD_{50} [14]	A_{mod}, (15)
1	Chlorobenzene	2.84	-0.971	9.15	1.69	13.2	3.00	1.33	0.303	0.248
2	*p*-Dichlorobenzene	3.39	-1.441	9.11	0	15.0	3.50	1.46	0.398	0.476
3	*o*-Dichlorobenzene	3.39	-1.356	9.23	2.51	15.9	3.50	1.46	0.468	0.788
4	1,2,4,5-Tetrachlorobenzene	4.89	-2.097	9.31	0	19.7	4.50	1.46	0.667	1.036
5	2,4,6-Trichlorophenol	3.06	-1.751	9.02	1.62	19.0	4.00	1.78	1.299	0.900
6	1,2,4-Trichlorobenzene	4.13	-1.764	9.26	1.25	18.3	4.00	1.50	1.323	0.847
7	3,4-Dichloraniline	2.69	-1.292	8.24	4.16	17.8	3.43	1.73	1.429	1.405
8	*p*-Nitrochlorobenzene	2.39	-3.112	10.21	2.52	16.5	3.71	1.99	1.802	2.291
9	*m*-Nitrochlorobenzene	2.46	-3.077	10.00	3.38	17.9	3.71	1.99	2.326	2.572
10	*o*-Nitrochlorobenzene	2.53	-2.987	9.95	4.25	16.7	3.71	1.99	2.949	2.900
11	2,4-Dinitrochlorobenzene	2.45	-3.657	10.86	3.29	19.1	4.25	2.11	3.571	3.030
12	2,3,5,6-Tetrachloronitrobenzene	4.55	-3.576	9.86	5.34	24.1	5.00	1.99	4.000	4.041

[*)] Dipole moments of molecules and their ionization potentials were calculated using the MINDO/3 quantum mechanical method. [**)] The polarizabilities of the molecules were determined using the Lefevre additive scheme [15].

The surrounding dielectric medium is considered as structureless and continuous, which is characterized by the dielectric constant κ_s. The first term in equation (2) determines the electrostatic interaction between atoms, which have electronic charges q_s and q_t. Electrostatic forces favor the interaction of donor and acceptor atoms, but usually are not decisive for the stabilization of the complex. In highly polar solvents, electrostatic interactions are significantly weakened. The second term defines quasi-chemical binding, i.e. it characterizes the partial electron transfer from the donor MO to the acceptor, thereby stabilizing the molecular complex. The donor-acceptor mechanism arises when the free orbital of the acceptor overlaps with the filled orbital of the donor or donor group of atoms. The donor is assumed to have a lone pair of electrons.

For example, a nitrogen atom has a lone pair of electrons in the state $2s^2$. When two molecules approach each other, the lone pair of electrons is shared between the two molecules.

This generalization of electrons is accompanied by the formation of bonds between molecules. The interaction between the donor and acceptor leads to a decrease in the energy of the ground state of the entire system below the initial levels of the donor and acceptor. The measure of the acceptor activity of a chemical compound having a closed electron shell is the position of the lowest free MO (ε_{nb}^{0}). Moreover, the acceptor properties are stronger, the lower the level ε_{nb}^{0}. Index zero indicates that one-electron energy corresponds to an isolated molecule in vacuum.

The quantum-chemical method was used to determine the numerical values of one-electron MO energies CNDO/S' [18]. The experimental values of the lengths of interatomic bonds, bond angles of molecules, for all chemical compounds analyzed here, were taken from the reference book [19].

In a condensed polar medium, under the influence of the electrostatic field of the dipole molecules of the environment, the electronic levels are shifted relative to their position in an isolated molecule. The macroscopic electric field $\boldsymbol{E}_{eff}$ acting on a molecule in a condensed medium differs from the average macroscopic field. This is due to the effect of polarization of the dielectric in an external field, as well as due to the action of the reactive field of the polar molecules of the dielectric medium. The induced (reactive) electric field acts on the field source, changing the electronic distribution of the molecule (i.e., self-action of the polar molecule occurs). The presence of a polarizable medium between interacting molecules can significantly change their total potential energy.

As is known [20], the reactive field of L.Onsager [21] acting on a molecule is proportional to its dipole moment. To determine the effect of a reactive field on the electronic states in a molecule, we use the concepts developed in molecular spectroscopy of the condensed state [22]. Intermolecular interactions can significantly affect the optical

spectra of molecules, shifting the maxima of the absorption and emission bands, changing the intensity of the bands, and new spectral frequencies may appear.

Within the framework of the continuum theory of the reaction field, the one-electron energy ε_{nb}, under the condition of thermodynamic equilibrium of the molecule with the environment, will be determined as follows:

$$\varepsilon_{nb} = \varepsilon_{nb}^0 - f_R\mu_1^2/(2a^3) + \alpha_1 f_R^2\mu_1^2/(2a^6) - 3I_1I_3\alpha_1\alpha_3(n_3^2-1)/[2(I_1 + I_3)a^6(n_3^2+2)]. \quad (3)$$

Here, the reaction field and dispersion interactions are taken into account. $f_R = 2(\kappa_s-1)/(2\kappa_s + 1)$ is the electric field factor of the reaction of a point dipole for a polar medium with a static permittivity κ_s; a, μ_1, α_1 and I_1 are the effective size of a molecule of a low molecular weight exogenous chemical compound commensurate with the average radius of the molecule, its dipole moment, as well as the average static electronic polarizability and the first ionization potential of the molecule, respectively.

In a condensed medium, each molecule is under the influence of a combination of surrounding molecules. This is partly taken into account by the factor f_R, which depends on the macroscopic properties of the condensed medium. Let us agree that index 1 refers to the molecule of an exogenous chemical compound, index 2 refers to the biosubstrate molecule with which the molecule of the exogenous substance interacts. Index 3 refers to a polar dielectric medium, the optical refractive index of which is equal to n_3. The second and third terms in equation (3) describe the interaction of the dipole moment of the molecule with the field of the electrostatic reaction of the polar dielectric medium [22] and, thus, the effect of polarization of the molecule by this field is taken into account. The third term in (3) characterizes the effect of short-range dispersion interactions (in the London approximation) on the MO levels of an exogenous chemical compound in a condensed medium.

On a macroscopic scale, the role of the reaction field manifests itself in the fact that the liquid is compressed in the region of the surrounding molecule, thereby increasing the potential energy of the polar liquid medium. This is one of the reasons for the increase in the boiling point (T_b) and decrease in the melting point (T_m) of polar liquids. Therefore, it is natural that the correlation coefficients between the bioresponse and each of the parameters T_b, T_m and μ_1 are close in magnitude to each other [11]. However, in this case, the regression equations are not informative enough, since it is not clear which approximately additive contributions of intermolecular interactions should be taken into account in each particular case.

One can make some assumptions about the molecular properties of the local region with which the exogenous molecule is associated, if the main contributions are known. For example, if dispersion forces make a dominant contribution, then a local, biological object must have high polarization properties.

When a low molecular weight chemical compound approaches a receptor, the molecule enters into a pair interaction with it. In the dipole approximation, the efficiency of this interaction is determined by the following additive physical components:

1) dipole - dipole interaction, which after averaging over all orientations of molecules at temperature T has the following form

$$E_{dip} = - 2\mu_1^2\mu_2^2/(3\kappa_s R_0^6 k_B T), \quad (4)$$

that is, the attraction between the dipoles depends on the temperature;

2) inductive interaction

$$E_{ind} = - (\alpha_2\mu_1^2 + \alpha_1\mu_2^2)/(\kappa_s^2 R_0^6); \quad (5)$$

3) dispersion interaction, which, in the London approximation, can be written in the following form:

$$E_{disp} = - 3\alpha_1\alpha_2 I_1 I_2/[2R_0^6(I_1 + I_2)]. \quad (6)$$

Approximate formulas (3) - (6) make it possible to write the pair interaction of molecules in terms of the properties of individual molecules. Here R_0 is the effective distance between the interacting molecules; α_i is the isotropic electronic polarizability of the i-th molecule; μ_i is the dipole moment of the i-th molecule; k_B is the Boltzmann constant; I_i is the first ionization potential of the i-th molecule. It should be borne in mind that the assumption of the additivity of intermolecular interactions is not sufficiently rigorous [23]. However, for solving practical problems, the approximation of interaction additivity is generally accepted. The possibility of using simplified formulas makes it possible to significantly expand the scope of the theory of intermolecular interactions.

Studies have shown [20,24] that formulas (4) - (6) are applicable to real systems. These approximations make it possible to correctly indicate the changes in the potential energy of the interaction of molecules as a function of the distance between them and the individual properties of the molecules. Before writing the final equation that determines the change in the energy of the system, taking into account specific and nonspecific

interactions, let's simplify equation (2). For homologous series of chemical compounds or related compounds, which interact with the same donor, the electron-accepting properties of the molecule mainly depend on the position on the energy scale of the lower free molecular orbital ε_{nb} of the acceptor: $\Delta E_{d\text{-}a} = f(\varepsilon_{nb})$. For the homological series of chemical compounds, the change in interaction energy is approximately proportional to the energy ε_{nb}: $\Delta E_{d\text{-}a} \approx \varepsilon_{nb}$. In general, the equation that determines the value of the stabilization energy of the complex of a low-molecular chemical compound plus a receptor will have the following form:

$$\Delta E_{d\text{-}a} = k_0 + k\varepsilon_{nb}{}^0 + \mu_1{}^2\{-kf_R/(2a^3) + k\alpha_1 f_R{}^2/(2a^6) - 2\mu_2{}^2/(3\kappa_s R_0{}^6 k_B T) - \alpha_2/(\kappa_s{}^2 R_0{}^6)\} + \alpha_1\{-\mu_2{}^2/(\kappa_s{}^2 R_0{}^6) - 3\alpha_2 I_1 I_2/(2R_0{}^6(I_1 + I_2))\} - [3\alpha_1 I_1 I_3/(2a^3(I_1 + I_2))]\cdot[(n_3{}^2 - 1)/(n_3{}^2 + 2)]. \quad (7)$$

Here k and k_0 are some numerical coefficients. For the purposes of regression analysis, this equation can be simplified. It is known that the inductive interaction $E_{ind} << E_{disp}$ and therefore in the fourth term of the series (7) the first term in the curly bracket can be ignored in comparison with the second term. In what follows, we will assume that the molecular parameters that relate to the biosubstrate and the polar medium are constant for the entire range of chemical compounds.

In this study, the active center of the biophase with which the exogenous molecules interacts is not specified. Therefore, for certainty we will assume $I_2 \approx I_3 \approx 10$ eV. This value of the ionization potential corresponds to most organic molecules [25,26]. Since the main goal is to obtain a regression equation, this approximation is quite satisfactory. Thus, equation (7) can be reduced to the following multifactorial regression equation with three explanatory variables that enter the equation additively and have a joint simultaneous effect on the resulting trait:

$$A \equiv 1000/LD_{50} = B_0 + B_1\varepsilon_{nb}{}^0 + B_2\mu_1{}^2 + B_3\alpha_1 I_1/(I_1 + 10). \quad (8)$$

For the convenience of presenting the statistical material, in what follows we introduce the following notation: $x_1 = \varepsilon_{nb}{}^0$, $x_2 = \mu_1{}^2$ и $x_3 = \alpha_1 I_1/(\alpha_1 + 10)$. Next, the problem is reduced to estimating the multiple regression coefficients B_i from the known results of sample observations.

The statistics of the populations A, x_1, x_2 and x_3 will be as follows:

$A = 1000/LD_{50}$: $N = 12$, $A^{av} = 1.71 \pm 0.36$; 95% confidence interval: 0.91 - 2.51; $A^{min} = 0.303$, $A^{max} = 4.00$, $S_A = 1.266$, $\tau^{min} = 1.12 < \tau^{max} = 1.82 < \tau_{0.05}{}^{cr,2}(N) = 2.387 < \tau_{0.05}{}^{cr,1}(N) = 2.523$; Wilk-Shapiro normality test: $W = 0.910 > W_{0.05}{}^{cr}(N) = 0.859$, David-Hartley-Pearson normality test: $U1_{0.05}{}^{cr}(N) = 2.800 < U = [(A^{max} - A^{min})/S_A] = 2.92 < U2_{0.05}{}^{cr}(N) = 3.910$, representativeness of the sample size: $N_{repr} = 10$;

x_1: $N = 12$, $x_1{}^{av} = -2.26 \pm 0.28$; 95% confidence interval: (-2.87, -1.64); $\varepsilon_{nb}{}^{0,min} = -3.67$, $\varepsilon_{nb}{}^{0,max} = -0.97$, $S_{x1} = 0.964$, $\tau^{max} = 1.34 < \tau^{min} = 1.46 < \tau_{0.05}{}^{cr,2}(N) = 2.387 < \tau_{0.05}{}^{cr,1}(N) = 2.523$; Wilk-Shapiro normality test: $W = 0.895 > W_{0.05}{}^{cr}(N) = 0.859$, David-Hartley-Pearson normality test: $U1_{0.05}{}^{cr}(N) = 2.800 \approx U = [(\varepsilon_{nb}{}^{0,max} - \varepsilon_{nb}{}^{0,min})/S_{x1}] = 2.791 < U2_{0.05}{}^{cr}(N) = 3.910$; $N_{repr} = 10$;

x_2: $N = 12$, $x_2{}^{av} = 8.82 \pm 2.54$; 95% confidence interval: 3.23 - 14.41; $\mu_1{}^{2,min} = 0$, $\mu_1{}^{2,max} = 28.52$, $S_{x2} = 8.80$, $\tau^{min} = 1.00 < \tau^{max} = 2.24 < \tau_{0.05}{}^{cr,2}(N) = 2.387 < \tau_{0.05}{}^{cr,1}(N) = 2.523$; Wilk-Shapiro normality test: $W = 0.885 > W_{0.05}{}^{cr}(N) = 0.859$, David-Hartley-Pearson normality test: $U1_{0.05}{}^{cr}(N) = 2.800 < U = [(\mu_1{}^{2,max} - \mu_1{}^{2,min})/S_{x2}] = 3.25 < U2_{0.05}{}^{cr}(N) = 3.910$; $N_{repr} = 10$;

x_3: $N = 12$, $x_3{}^{av} = 8.43 \pm 0.42$; 95% confidence interval: 7.75 - 9.58; $x_3{}^{min} = 5.276$, $x_3{}^{max} = 11.965$, $S_{x3} = 1.744$, $\tau^{min} = 1.81 < \tau^{max} = 2.03 < \tau_{0.05}{}^{cr,2}(N) = 2.387 < \tau_{0.05}{}^{cr,1}(N) = 2.523$; Wilk-Shapiro normality test: $W = 0.975 > W_{0.05}{}^{cr}(N) = 0.859$, David-Hartley-Pearson normality test: $U1_{0.05}{}^{cr}(N) = 2.800 < U = [(x_3{}^{max} - x_3{}^{min})/S_{x3}] = 3.83 = U2_{0.05}{}^{cr}(N) = 3.910$; $N_{repr} = 10$. (9)

Since the sample size is limited, before analyzing the regression (8), we perform the following procedure. Let's analyze a regression that uses only one explanatory variable, such as the variable x_2:

$$A_1(x_2) = a_0 + a_1 x_2.$$

For this regression, we get the following statistics:

$N = 12$, $m_1 = 1$; $R = 0.80 \pm 0.11$, $|R^*| = 0.82 > R_{0.05}{}^{cr}(N - 2) = 0.576$; correlation coefficient significance test based on the Fisher normalizing z-transform (with Hotelling corrections taken into account): $u_H = 1.07 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.591$; $RMSE = 0.794$; the minimum sample size sufficient for the reliability of the correlation coefficient: $N_{0.05}{}^{min} = 6$; $a_0 = 0.17 \pm 0.09$, $a_1 = -1.01 \pm 0.26$, $|t(a_1)| = 3.8 > t_{0.05}{}^{cr}(N - 2) = 2.228 > t(a_0) = 1.93$; unexplained regression residuals (perturbing variable) are normally distributed: Wilk-Shapiro test: $W = 0.938 > W_{0.05}{}^{cr}(N) = 0.859$; $F = 17.57 > F_{0.05}{}^{cr}(f_1 = 1; f_2 = 10) = 4.96$.

Here $R^* = R\cdot[1 + 0.5\cdot(1 - R^2)/(N - 3)]$ is the adjusted correlation coefficient. Thus, there is a significant relationship between the explanatory variable x_2 and bioresponse. Checking the relationship of regression residuals δA_1 with the explanatory variable x_1 also indicates the presence of a significant correlation between them:

$$\delta A_1(x_1) = a_{01} + a_{11}\,x_1,$$

$N = 12$, $m_1 = 1$; $R = -0.74 \pm 0.14$, $|R^*| = 0.76 > R_{0.05}^{cr}(N - 2) = 0.576$; correlation coefficient significance test based on the Fisher normalizing z-transform (with Hotelling corrections taken into account): $u_H = 0.92 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.591$; $RMSE = 0.533$; the minimum sample size sufficient for the reliability of the correlation coefficient: $N_{0.05}^{min} = 7$; $a_{01} = -1.31 \pm 0.41$, $a_{11} = -0.58 \pm 0.17$, $|t(a_{11})| = 3.48 > t_{0.05}^{cr}(N - 2) = 2.228$; unexplained regression residuals (perturbing variable) are normally distributed: Wilk-Shapiro test: $W = 0.944 > W_{0.05}^{cr}(N) = 0.859$; $F = 12.13 > F_{0.05}^{cr}(f_1 = 1; f_2 = 10) = 4.96$.

It follows that the regression $\delta A_1(x_1)$ is also statistically significant. That is, the unexplained residuals of regression $A_1(x_2)$ correlate with another explanatory variable, namely variable x_1. Further verification of the relationship between the regression residuals $\delta A_1(x_1)$ and the explanatory variable x_3 showed that the residuals are not associated with the molecular index characterizing the dispersion interaction: $R = 0.31 < R_{0.05}^{cr}(N - 2) = 0.576$, $F = 1.07 < F_{0.05}^{cr}(f_1 = 1; f_2 = 10) = 4.96$. Obviously, this contribution to the interaction of molecules is insignificant for the analyzed sample.

It follows from inequalities (9) that at a significance level of 5%, the populations A, x_1, x_2, and x_3 are homogeneous and normally distributed. The homogeneity of the analyzed data was checked using the τ-criterion [27,28]. The following statistics were obtained for regression (8):

$N = 12$, $m_1 = 3$ – number of explanatory variables; multiple correlation coefficient: $R_1 = 0.966 > R_{0.05}^{cr}(m_1; N - m_1 - 1) = 0.777$, multiple determination coefficient: $R_1^2 = 0.933$, $R_1^{*2} = 0.91$; standard error of the regression estimate: $S_A = 0.381$; $B_0 = -1.29 \pm 0.82$, $B_1 = -0.78 \pm 0.18$, $B_2 = 0.06 \pm 0.02$, $B_3 = 0.09 \pm 0.12$; $|t(B_1)| = 4.27 > t(B_2) = 3.51 > t_{0.05}^{cr}(f = N - m - 1) = 2.306 > t(B_3) = 0.70$; $F = 37.29 > F_{0.05}^{cr}(f_1 = 3; f_2 = 8) = 4.07$; $\Sigma_1 = 1.1586$; $AIC_1 = -1.837$, $SC_1 = -1.5093$, $SS_1 = 0.1196$. (10)

Here Σ_1 is the sum of squared residuals; $R_{0.05}^{cr}(m_1; N - m_1 - 1)$ is the critical value of the multiple correlation coefficient [9], which determines the lower acceptable limit of the degree of association between the variations of the resulting attribute and all explanatory variables.

Statistics (10) contains information quality criteria for the linear regression equations of Akaike [29] and Schwartz [30], as well as the alternative ratio $SS = \Sigma^{0.5}/(N - m)$. For regression residuals, the Wilk-Shapiro test would be as follows: $W = 0.975 > W_{0.05}^{cr}(N) = 0.859$. The information quality criteria for the regression equation are defined as follows:

$$AIC = (2m/N) + \ln(\Sigma/N),$$

$$SC = [(m + 1)\ln N]/N + \ln(\Sigma/N)\,. \quad (11)$$

The assessment of the significance of the multiple determination coefficient (10) is carried out using the F-statistics: $F = R^2\cdot(N - m - 1)/m/(1 - R^2)$. Since $F > F_{0.05}^{cr}$ (see Eq.(10)), it can be assumed that the multiple coefficient of determination is reliably different from zero with probability $1 - \alpha = 0.95$, and the explanatory variables reliably explain variations in bioactivity. The B_i regression coefficients are significantly greater than zero if $t(B_i) > t^{cr}$ at significance level α and number of degrees of freedom $f = N - m_1 - 1$ at the two-sided critical region. Therefore, the coefficient B_3 in Eq.(10) is not statistically reliable. Regression (8) explains 93.3% of the variability in bioactivity. Only 6.7% of unexplained variations can be attributed to unaccounted for factors or random variations in the original data.

The importance of the participation of each of the independent explanatory variables in assessing the variability of the resulting sign is characterized by standardized regression coefficients. The standardized regression coefficients B_i^* are related to the normal regression coefficients (8) by the following relationships:

$$B_1^* = B_1\cdot S_{x1}/S_A = -0.596,\ B_2^* = B_2\cdot S_{x2}/S_A = 0.405,$$

$$B_3^* = B_3\cdot S_{x3}/S_A = 0.098. \quad (12)$$

On a natural scale, the regression coefficients are dimensional quantities. However, the standardized coefficients are dimensionless and this makes it possible to perform their quantitative comparison. Knowledge of standardized coefficients makes it possible to determine the proportion of explanatory variables involved in explaining the variability of the resulting sign. An approximate ratio for the multiple coefficient of determination can be used to obtain information about the comparative influence of individual variables [31]:

$$R_{appr}^2 = B_1^*\cdot r_{x1,A} + B_2^*\cdot r_{x2,A} + B_1^*\cdot r_{x3,A} =$$

$$0.535 + 0.321 + 0.075 = 0.931. \quad (13)$$

Here $r_{x1,A}$ = -0.897, $r_{x2,A}$ = 0.798 and $r_{x3,A}$ = 0.770 are the pairwise correlation coefficients between each explanatory variable and the observed bioactivity. From relation (13) it follows that the greatest contribution to the explanation of the variability in the toxicity of chemical compounds comes from the variables x_1 (53.5%) and x_2 (32.1%). The equity participation of variable x_3 is very insignificant and amounts to only 7.5%. The approximate value of the coefficient of determination (13) is very close to the value of R_1^2 = 0.933 (see Eq.(10)).

The adjusted (unbiased) coefficient of determination is determined from the following:

$$R^{*2} = 1 - (1 - R^2)\cdot(N - 1)/(N - m - 1). \quad (14)$$

An adjusted coefficient of determination, R^{*2} is applied so that models with different numbers of explanatory factors can be compared.

Thus, at the 95% confidence level, a very strong relationship can be assumed between toxicity and the explanatory variables x_1 and x_2. The coefficient of determination R_1^2 = 0.933 (10) indicates what portion (in this case, 93.3%) of the total variance of the bioresponse function is explained by the factors x_1, x_2, and x_3. Only 6.7% of the total variance cannot be explained by the model (the uncertainty factor is 0.067) and appears to be due to unaccounted for variables or random deviations in the original sample. The values of coefficients B_1^* and B_2^* reflect the significant dependence, at the 95% confidence level, of the resultant variable on the explanatory factors x_1 and x_2. At the same time, the intermolecular dispersion interaction (~ x_3) is of minor importance (not significant at the chosen significance level α), since $t(B_3) < t^{cr}$ (two-sided hypothesis evaluation). The difference of the coefficient B_3 from zero can be attributed to random fluctuations in the original sample. Therefore, nothing definite can be said about the influence of the dispersion interaction on the resultant attribute. Features that have low information content (i.e., "weight") can be excluded from further analysis.

The choice of independent explanatory variables is a process of successive refinement of the initial hypothesis. The following steps can be distinguished in this process: formation of a primary hypothesis (Eq. (8)) about the set of independent variables; analysis of structural relationships; narrowing of features and selection of significant variables for modeling.

Since the influence of dispersion interactions is insignificant, equation (8) can be replaced by the following reduced two-factor regression equation:

$$A \equiv 1000/LD_{50} = B_0 + B_1x_1 + B_2x_2, \quad (15)$$

N = 12, m_2 = 2, B_0 = -0.76 ± 0.29, B_1 = -0.86 ± 0.14, B_2 = 0.06 ± 0.02; $|t(B_1)|$ = 6.10 > $t(B_2)$ = 3.98 > $|t(B_0)|$ = 2.65 > $t_{0.05}^{cr}$ ($f = N - m_2 - 1$) = 2.26; R_2 = 0.968 > $R_{0.05}^{cr}$ (2;9) = 0.697, R_2^2 = 0.937, R_2^{*2} = 0.933; standard error of the regression estimate: S_A = 0.370; F = 59.04 > $F_{0.05}^{cr}(f_1 = m_2; f_2 = N - m_2 - 1)$ = 4.26; Σ_2 = 1.2294; AIC_2 = -1.9450, SC_2= -1.6572, SS_2 = 0.1109. (16)

Regression residuals are normally distributed. Wilk-Shapiro normality test: W = 0.940 > $W_{0.05}^{cr}(N)$ = 0.859.

If the ratio F many times (for example, not less than four times) exceeds the tabular value, then such a regression, according to [32], has predictive properties. Additional information about the significance of variable x_3 can be obtained by analyzing the relationship between the regression residuals (15) and variable x_3. As the analysis showed, the correlation coefficient is insignificant: R = 0.16 < $R_{0.05}^{cr}(N - 2)$ = 0.576, that is, the explanatory variable x_3 is not related in any way to the unexplained regression residuals (15). Therefore, this variable does not really need to be included in the regression equation. A comparison of the information tests AIC_1 and AIC_2 for regressions (8) and (15) shows that the quality of the reduced regression (15) is higher than for regression (8), although the number of explanatory variables has decreased. Similar relations hold for the tests SC_1 = -1.5093 and SC_2 = -1.6572. The values of the Akaike and Schwarz tests are associated with the ratios SS: $SS_1 = \Sigma_1^{0.5}/(N - m_1)$ = 0.1196 and $SS_2 = \Sigma_2^{0.5}/(N - m_2)$ = 0.1109. The use of AIC, SC and SS tests is justified because the R^2 criterion may not be informative. The fact is that for a model with a large number of explanatory variables, the criterion R^2 will always be no less than for a model with a smaller number of explanatory variables. The presence of collinearity between explanatory variables can significantly distort the relationship between the information tests for full regression and reduced regression.

If we analyse the residuals δA of a simple linear regression that depends on only one explanatory variable x_1: $A \equiv 1000/LD_{50} = B_0 + B_1x_1$ ($|R|$ = 0.90 > $R_{0.05}^{cr}(N - 2)$ = 0.576; F = 41.2 > $F_{0.05}^{cr}(f_1 = 1; f_2 = 10)$ = 4.96), then appears that the residuals δA of the regression are significantly correlated with the set of variables x_2: r = 0.66 > $R_{0.05}^{cr}(N - 2)$ = 0.576, F =

$7.76 > F_{0.05}^{cr}(f_1 = 1;\ f_2 = 10) = 4.96$. Thus, for the regression equation (15), both explanatory variables x_1 and x_2 are significant. However, it is necessary to check whether these variables are not significantly correlated. Collinearity between explanatory variables can lead to misjudgment of the impact of variables on the outcome variable because the explanatory variables are related. The presence of collinearity between explanatory variables can significantly distort the test relationships for full regression and reduced regression. Checking the interconnectedness of the explanatory variables x_1 and x_2 is carried out as follows:

$N = 12;\ \ x_1(x_2) = b_0 + b_1x_2,\ \ r_{1,2} = -0.564 \pm 0.216,\ |r_{1,2}| = 0.564 < R_{0.05}^{cr}(N-2) = 0.576;\ |t(b_1)| = 2.16 < t_{0.05}^{cr}(N-2) = 2.228;\ b_0 = -2.80 \pm 0.81,\ b_1 = -6.15 \pm 2.38$; the minimum sample size sufficient for the reliability of the correlation coefficient: $N_{0.05}^{min} = 11$; $F = 4.67 < F_{0.05}^{cr}(f_1 = 1;\ f_2 = 10) = 4.96.$ (17)

Since $F < F^{cr}$, the relationship between x_1 and x_2 is not significant at the 95% confidence level. For small sample sizes ($N \le 15$), the best estimate of the correlation coefficient is the adjusted correlation coefficient [33]:

$$r^* = r\cdot[1 + 0.5\cdot(1 - r^2)/(N-3)]. \quad (18)$$

For regression (15), the residuals are normally distributed (16). In this case, a more accurate quantification of collinearity can be used, as suggested by Farrar and Glauber [34]. Farrar-Glauber test, has a chi-square distribution with $f = m(m-1)/2$ degrees of freedom:

$$\chi^2 = -(N - 1 - (2m_2 + 5)/6)\cdot\ln(r_{1,1}\cdot r_{2,2} - r_{2,1}\cdot r_{1,2})$$
$$= 3.64 < \chi_{0.05}^{2,cr}(f = 1) = 3.841. \quad (19)$$

Here $m_2 = 2$ is the number of explanatory variables. Since $\chi^2 < \chi^{2,cr}$, then the hypothesis of the independence of the explanatory variables does not contradict the original data. The absence of a significant relationship between the features ε_{nb}^0 and μ_1^2 is also indicated by the inequality:

$$t = |r_{1,2}|\cdot(N-2)^{0.5}/(1 - r_{1,2}^2)^{0.5} =$$
$$2.16 < t_{0.05}^{2,cr}(N-2) = 2.23. \quad (20)$$

That is, the correlation coefficient $|r_{1,2}|$ statistically insignificant. It is also possible to obtain an estimate of the collinearity of the variables x_1 and x_2 by using the following relation [8]:

$$F = (N - m_2)r_{1,2}^2/[(m_2 - 1)(1 - r_{1,2}^2)] =$$
$$4.67 < F_{0.05}^{cr}(f_1 = 1;\ f_2 = 10) = 4.90. \quad (21)$$

Inequality (21) also indicates that the explanatory variables ε_{nb}^0 and μ_1^2 can be recognized as independent. Consequently, the explanatory variables x_1 and x_2, have a simultaneous effect on the resultant variable, in a significant, multidirectional and independent manner.

The regression coefficients (15) need to be transformed again to reveal the comparative effect on bioactivity of the explanatory variables:

$$B_1^* = B_1\cdot S_{x1}/S_A = -0.656, \quad B_2^* = B_2\cdot S_{x2}/S_A = 0.428. \quad (22)$$

Thus, in contrast to the results (15), the standardized coefficients B_1^* and B_2^* do not differ very significantly in absolute value. Using the approximate relation (13), we determine the share contribution of each variable to the explanation of the variability of bioactivity:

$$R_{appr}^2 = B_1^*\cdot r_{x1,A} + B_2^*\cdot r_{x2,A} = 0.562 + 0.377 = 0.939. \quad (23)$$

The largest contribution to (23) is made by the explanatory variable ε_{nb}^0 (56.2%). The approximate coefficient of determination $R_{appr}^2 = 0.939$ is very close to the coefficient of determination (16): $R_2^2 = 0.937$.

In order to compare the coefficients of determination of the two models with different numbers of explanatory factors m, their adjusted R^{*2} values must be calculated (14). For regression (8), the adjusted coefficient of determination is $R_1^{*2} = 0.90$ for $m_3 = 3$. However, for multiple regression (16) $R_2^{*2} = 0.933$; $m_2 = 2$. The difference in the coefficients of determination $|R_1^{*2} - R_2^{*2}|$ determines the measure of additional explanation for the variation of the resultant variable, by including another explanatory variable in the regression. Next, we use the following relation, which has the Fisher F-distribution:

$$F = |R_1^{*2} - R_2^{*2}|\cdot(N - m_1 - 1)/(m_1 - m_2)/(1 - R_1^{*2}). \quad (24)$$

If $F > F_{0.05}^{cr}(f_1 = m_1 - m_2;\ f_2 = N - m_1 - 1)$, then the additional explanatory variable must be retained in the regression equation. In this case there is an inverse inequality of $F = 2.75 < F_{0.05}^{cr}(f_1 = 1;\ f_2 = 8) = 5.32$. That is, the additional explanatory variable x_3 does not improve the regression. The purpose of the regression equation is not only to describe the experiment satisfactorily. The regression equation should indicate the physical phenomena associated with the variability of bioactivity.

Thus, independent and simultaneously acting explanatory molecular variables ε_{nb}^0 and μ_1^2 are

associated with the toxicity of substituted chlorobenzenes. Model (15) makes it possible to make some assumptions about the properties of the biophase region with which an exogenous molecule can interact. In accordance with the regression equation (15), biophase molecules must have a significant dipole moment or charge. In addition, the energy of the highest occupied molecular orbital $\varepsilon_{\text{hocc}}$ of the biophase should be close to the energy $\varepsilon_{\text{nb}}^0$ of the exogenous molecule. If it is possible to briefly describe the initial information, then there is confidence that some objective regularity has been revealed that exists in the structure of the feature space, allowing this reduction to be carried out [18].
The correlation between the experimental toxicity values of substituted chlorobenzenes and theoretical values is shown in Fig. 1. The energy of stabilization of the molecular complex is the higher, the greater the energy of the donor-acceptor complex.
Table 1 also lists the Z and H molecular features for chlorine-substituted compounds. Z is the average number of electrons in the outer shell of atoms in a molecule: $Z = \Sigma_i n_i Z_i/N$ [36,37]. Here n_i is the number of atoms of the i-th sort with the number of electrons Z_i on the outer electron shell. The summation is performed on all atoms in the molecule; $\Sigma_i n_i = N$ is the total number of atoms. The electronic factor Z is related to the pseudopotential of the molecule [38].

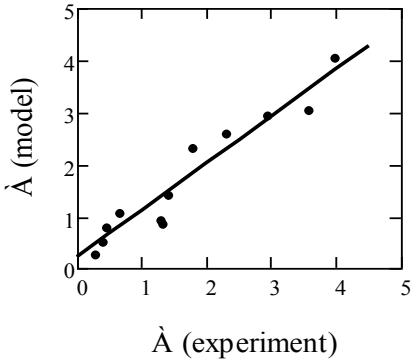

Fig. 1. Scatterplot and regression line. Calculated and experimental values of average lethal doses ($A \equiv 1000/\text{LD}_{50}$) of chlorobenzene derivatives (Table 1). The regression line is given by the equation $A_{\text{mod}} = a_0 + a_1 A_{\text{exp}}$, $N = 12$, $a_0 = 0.191 \pm 0.168$, $a_1 = 0.904 \pm 0.080$; $R = 0.96 \pm 0.03$; $R^* = 0.97 > R_{0.05}^{\text{cr}}(N-2) = 0.576$, $F = 159.2 > F_{0.05}^{\text{cr}}(f_1 = 1;\ f_2 = 10) = 4.96$; straightness index: $K = (N\cdot(1 - R^2))^{0.5} = 0.84 < K^{\text{thr}}$(threshold value) = 3.00 [35].

The information function H [39], for a discrete data set, is quantified as follows: $H = -\Sigma_j p_j \log_2 p_j$. The ratio $p_i = n_i/N$ satisfies the following conditions: $0 \le p_i \le 1$, $\Sigma_i p_i = 1$. In which connection, $p_i = 0$ means the impossibility of the occurrence of the i-th event; $\Sigma_i n_i = N$; N is the number of atoms in the molecule. The ratio n_k/N determines the share holding of the kth kind of atom in the molecule. For chemical compounds from Table 1, it was found that the molecular factor Z is closely related to the polarizability of the α_1 molecule: the linear correlation coefficient is 0.93 ± 0.12. The statistical significance of the correlation coefficient is determined by the inequality:

$$t = |R|\cdot(N-2)^{0.5}/(1-R^2)^{0.5} = 12.36 > t_{0.05}^{\text{cr}}(N-2) = 2.23. \quad (25)$$

Inequality (25) uses a two-sided critical region for the t-quantile. The Chaddock scale defines the pairwise correlation coefficient as corresponding to a "very close relationship" [40].
The relationship between the values of one-electron energies $\varepsilon_{\text{nb}}^0$ and the values of the molecular factor Z was also verified. A very close linear relationship was found (Fig. 2A) between these factors for chlorine-substituted benzenes (chemical compound numbers No = 1 – 6 in Table 1):

$\varepsilon_{\text{nb}}^0(Z)_1 = a_{01} + a_{11}\cdot Z$, $N_1 = 6$, $R_1 = -0.996 \pm 0.004$, $|R_1^*| = 0.998 > R_{0.05}^{\text{cr}}(N_1 - 2) = 0.811$; criterion of significance of the correlation coefficient based on the normalizing Fisher z-transform (Hotelling corrections is taken into account [33]): $u_{\text{H}} = 2.98 > u_{0.05}(N) = z_{0.975}\cdot(N-1)^{-0.5} = 0.86$; $S_1 = 0.040$; sufficient sample size to ensure the validity of the correlation coefficient: $N_{0.05}^{\text{min}} < 4$; $a_{01} = 1.23 \pm 0.13$, $a_{11} = -0.75 \pm 0.03$, $|t(a_{11})| = 21.7 > t(a_{01}) = 9.46 > t_{0.05}^{\text{cr}}(N_1 - 2) = 2.776$; $F = 469.5 > F_{0.05}^{\text{cr}}(f_1 = 1; f_2 = 4) = 7.71$; straightness index: $K = 0.22 < K^{\text{thr}} = 3.0$. (26)

Because the corrected correlation coefficient $|R_1^*| = 0.998$, and the value of $S_1 = 0.040$, then this relationship of features is close to the functional dependence.

Z and ε_{nb} statistics:

$N_1 = 6$, $Z^{\text{av}} = 3.75 \pm 0.21$; 95% confidence interval (3.20-4.30), $Z^{\text{min}} = 3.00$, $Z^{\text{max}} = 4.50$, $S_{Z1} = 0.524$, $\tau^{\text{max}} = 1.43 = \tau^{\text{min}} = 1.43 < \tau_{0.05}^{\text{cr},2}(N_1) = 1.996 < \tau_{0.05}^{\text{cr},1}(N_1) = 2.184$; Wilk-Shapiro normality test: $W = 0.960 > W_{0.05}^{\text{cr}}(N_1) = 0.788$, David-Hartley-Pearson normality test: $U1_{0.05}^{\text{cr}}(N_1) = 2.200 < U = [(Z^{\text{max}} - Z^{\text{min}})/S_Z] = 2.863 < U2_{0.05}^{\text{cr}}(N_1) = 3.012$, $N_{\text{repr}} = 5$; (27)

$N_1 = 6$, $\varepsilon_{\text{nb}}^{\text{av}} = -1.56 \pm 0.16$; 95% confidence interval (-1.97,-1.15), $\varepsilon_{\text{nb}}^{\text{min}} = -2.097$, $\varepsilon_{\text{nb}}^{\text{max}} = -0.971$, $S_\varepsilon =$

0.392, $\tau^{min} = 1.37 < \tau^{max} = 1.50 < \tau_{0.05}^{cr,2}(N_1) = 1.996 < \tau_{0.05}^{cr,1}(N_1) = 2.184$; Wilk-Shapiro normality test: $W = 0.975 > W_{0.05}^{cr}(N_1) = 0.788$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_1) = 2.200 < U = [(\varepsilon_{nb}^{max} - \varepsilon_{nb}^{min})/S_\varepsilon] = 2.870 < U2_{0.05}^{cr}(N_1) = 3.012$; $N_{repr} = 5$. (28)

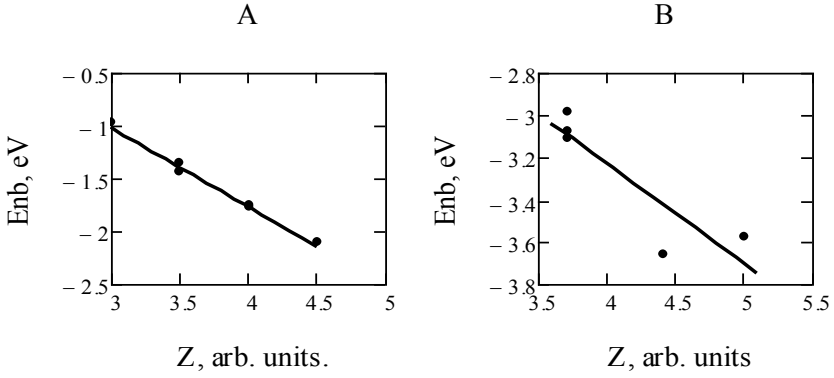

Fig. 2. Scatterplots and regression lines. Relationship between the energy of the lowest free molecular orbital Enb ≡ ε_{nb}^0 of an exogenous molecule and the molecular factor *Z*. (A) Chloro-substituted benzenes. (B) Nitrochlorobenzenes.

The sets *Z* and ε_{nb} are homogeneous and have a normal distribution. A quantitative relationship between these features was also found (Fig. 2B) for chlorine-substituted nitrobenzenes (numbers of chemical compounds Nos. = 8 - 12 in Table 1):

$\varepsilon_{nb}^0(Z)_2 = a_{02} + a_{12}Z$, $N_2 = 5$, $R_2 = - 0.89 \pm 0.12$, $|R_2^*| = 0.94 > R_{0.05}^{cr}(N_2 - 2) = 0.8783$; $S_2 = 0.166$; criterion of significance of the correlation coefficient based on the normalizing Fisher *z*-transform (Hotelling corrections is taken into account): $u_H = 1.904 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.98$; sufficient sample size to ensure the validity of the correlation coefficient: $N_{0.05}^{min} < 4$; $a_{02} = -1.37 \pm 0.59$, $a_{12} = -0.47 \pm 0.14$, $|t(a_{12})| = 3.32 > t_{0.05}^{cr}(N_2 - 2) = 3.182$; $F = 11.04 > F_{0.05}^{cr}(f_1 = 1; f_2 = 3) = 10.13$; straightness index: $K = 1.02 < K^{thr} = 3.0$. (29)

Z and ε_{nb} statistics:

$N_2 = 5$, $Z^{av} = 4.11 \pm 0.26$; 95% confidence interval (3.38-4.83), $Z^{min} = 3.71$, $Z^{max} = 5.00$, $S_{Z2} = 0.584$, $\tau^{min} = 0.682 < \tau^{max} = 1.527 < \tau_{0.05}^{cr,2}(N_2) = 1.869 < \tau_{0.05}^{cr,1}(N_2) = 2.080$; Wilk-Shapiro normality test: $W = 0.773 > W_{0.05}^{cr}(N_2) = 0.762$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_2) = 2.200 < U = [(Z^{max} - Z^{min})/S_Z] = 2.21 < U2_{0.05}^{cr}(N_2) = 3.222$; $N_{repr} = 4$;

$N_2 = 5$, $\varepsilon_{nb}^{av} = -3.28 \pm 0.14$; 95% confidence interval (-3.66,-2.30), $\varepsilon_{nb}^{min} = -3.657$, $\varepsilon_{nb}^{max} = -2.987$, $S_\varepsilon = 0.310$, $\tau^{max} = 0.95 < \tau^{min} = 1.22 < \tau^{max} = 1.527 < \tau_{0.05}^{cr,2}(N_2) = 1.869 < \tau_{0.05}^{cr,1}(N_2) = 2.080$; Wilk-Shapiro normality test: $W = 0.831 > W_{0.05}^{cr}(N_2) = 0.762$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_2) = 2.200 \approx U = [(\varepsilon_{nb}^{max} - \varepsilon_{nb}^{min})/S_\varepsilon] = 2.16 < U2_{0.05}^{cr}(N_2) = 3.012$, $N_{repr} = 4$. (30)

An approximate comparative estimate of the regression coefficients a_{11} and a_{12} can be made using the following relation [41]:

$$t = |a_{11} - a_{12}|/[S_1^2/(N_1 - 1)/S_{Z1}^2 + S_2^2/(N_2 - 1)/S_{Z2}^2]^{0.5} = 1.918 < t_{0.05}^{cr}(N_1 + N_2 - 4) = 2.365. \quad (31)$$

That is, estimates of regression coefficients differ insignificantly, since $t < t^{cr}$.

Dispersion interaction does not make a statistically significant contribution to the explanation of the bioresponse of a number of compounds from Table 1. However, there is a relationship between the molecular feature *Z* and the value of the dispersion contribution:

$x_3(Z) = a_0 + a_1Z$, $N = 12$, $R = 0.96 \pm 0.03$, $R^* = 0.97 > R_{0.05}^{cr}(N - 2) = 0.576$; sufficient sample size to ensure the validity of the correlation coefficient: $N_{0.05}^{min} < 5$; criterion of significance of the correlation coefficient based on the normalizing Fisher *z*-transform (Hotelling corrections is taken into account): $u_H = 1.942 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.591$; $a_0 = -1.10 \pm 0.95$, $a_1 = 2.52 \pm 0.24$, $t(a_1) = 10.42 > t_{0.05}^{cr}(N - 2)$; standard error of the regression estimate: 0.44; $F = 108.5 > F_{0.05}^{cr}(f_1 = 1; f_2 = 10) = 4.96$. (32)

Statistics of set *Z*:

$N = 12$, $Z^{av} = 3.87 \pm 0.16$; 95% confidence interval: 3.52 - 4.22; $Z^{min} = 3.00$, $Z^{max} = 5.00$, $S_Z = 0.548$, $\tau^{min} = 1.59 < \tau^{max} = 2.06 < \tau_{0.05}^{cr,2}(N) = 2.387 < \tau_{0.05}^{cr,1}(N) = 2.523$; Wilk-Shapiro normality test: $W = 0.951 > W_{0.05}^{cr}(N) = 0.859$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.800 < U = [(Z^{max} - Z^{min})/S_Z] = 3.64 < U2_{0.05}^{cr}(N) = 3.910$; $N_{repr} = 10$.

(33)

It follows that the sets of indices Z is homogeneous and satisfies a normal distribution. The authors of [42] present the additive contributions to the energy of the pair interaction of a tetramethyluric acid molecule with aromatic hydrocarbon molecules. Including the magnitude of the dispersion contribution is reported. Our check showed that in this case, too, there is a statistically significant relationship between the molecular factor Z and the dispersion energy value of the pairwise interaction.

If there are chemical compounds capable of forming hydrogen bonds, the energy contribution due to hydrogen bonds must be taken into account in the regression equations (8) and (15). The hydrogen bond is essentially a quasi-chemical short-range interaction of molecules. Therefore, the properties of a hydrogen bond are difficult to describe using the properties of isolated molecules. However, some quantitative estimates can be made. Considering that the terms of equation (8) are proportional to the contributions to the binding energy, it can be assumed that the contributions to the bioactivity from these interactions are also approximately equal.

The application of model (15) to the calculation of the bioactivity of the 2,4-dichlorophenol molecule leads to the following value of the resultant feature $A = 0.58$ ($\varepsilon_{nb}^0 = -1.4$eV, $\mu_1 = 1.5D$), which is markedly lower than the experimental value 2.08. However, it should be kept in mind that in an isolated 2,4-dichlorophenol molecule there is an intramolecular hydrogen bond between the hydroxyl group proton and the chlorine atom in the *ortho*-position. When a molecule enters a polar condensed medium, the intramolecular hydrogen bond is broken due to the electric field of the reaction. This state of the molecule corresponds to a lower total energy. In other words, the molecular state is stabilized and the hydroxyl group has the opportunity to take part in the formation of an intermolecular hydrogen bond. In a polar dielectric medium, with an increase in the dipole moment of a molecule, the equilibrium of molecular forms with and without an intramolecular hydrogen bond shifts towards a molecular state with a large dipole moment. That is, without the formation of an intramolecular hydrogen bond. Such a situation has indeed been observed experimentally. The formation of an intermolecular hydrogen bond is accompanied by an additional contribution to the interaction energy. This in turn increases the bioresponse A by about 1.0 (contribution to the total interaction energy $\approx -0.85\varepsilon_{nb}^0$). Therefore, the total calculated value of bioactivity A will be approximately equal to 1.58, which is close to the experimental value. However, these remarks cannot be applied to the 2,4,6-trichlorophenol molecule (Table 1). This molecule contains a hydroxyl group. However, the molecule is not involved in the formation of intermolecular complexes through hydrogen bonds. The fact is that the proton of the hydroxyl group oscillates between two neighboring chlorine atoms, which are characterized by significant electronegativity. Oscillations are the result of a proton tunneling through a potential barrier. In this way, an intramolecular transition from one equilibrium position to another is carried out. Spectroscopic studies of 2,4,6-trichloro-, 2,4,5,6-tetrachloro- and pentachlorophenols confirm the existence of intramolecular proton migration.

The application of the regression equation (8) to predict the bioactivity of molecules is complicated by the fact that the researcher is required to know many molecular parameters. In particular, knowledge of the energies of single-electron molecular orbitals, which can be determined by complicated and cumbersome quantum mechanical calculations, is required. The results of these calculations require professional analysis.

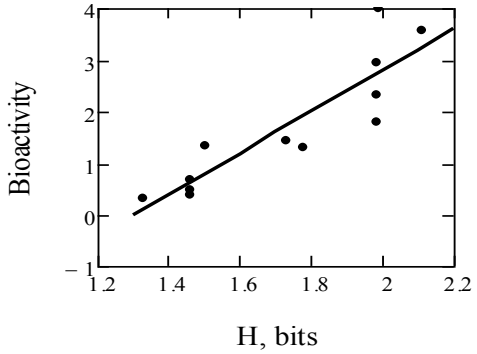

Fig.3. Scatterplot and regression line. The regression line is approximated by the linear equation: $A(H)_{mod} = b_0 + b_1 H$, $N = 12$; $b_0 = -5.25 \pm 1.15$, $b_1 = 4.02 \pm 0.07$, $N_{0.05}^{min} = 5$; $R = 0.89 \pm 0.07$; $R^* = 0.90 > R_{0.05}^{cr}(N - 2) = 0.576$, $S = 0.604$, $F = 37.6 > F_{0.05}^{cr}(f_1 = 1; f_2 = 10) = 4.96$; straightness index: $K = 1.59 < K^{thr} = 3.0$.

However, as the analysis showed, some rapid assessment of the toxicity of chlorine-substituted

benzenes can be done using the information function H (Table 1). Figure 3 shows a linear relationship between the toxicity of chlorine-substituted benzenes and the information function of molecules. Using the regression equation in Fig.3, we obtain the following estimate for the bioactivity of 2,4-dichlorophenol (A_{exp} = 2.08), which was not used in the original sample: A_{mod} = 1.74 (H = 1.738 *bits*, Z = 3.692 arb. units.).

There is also a significant relationship between the toxicity of chemical compounds and the explanatory factor Z:

$A(Z) = b_0 + b_1Z$, $N = 12$, $R = 0.62 \pm 0.20$, $R^* = 0.64 > R_{0.05}^{cr}(N - 2) = 0.576$; sufficient sample size to ensure the validity of the correlation coefficient: $N_{0.05}^{min} = 10$; $b_0 = -3.82 \pm 2.29$, $b_1 = 1.43 \pm 0.59$, $t(b_1) = 2.44 > t_{0.05}^{cr}(N - 2) = 2.228$; standard error of the regression estimate: $S_A = 1.044$; criterion of significance of the correlation coefficient based on the normalizing Fisher z-transform (Hotelling corrections is taken into account): $u_H = 0.697 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.591$; straightness index: $K = 2.7 < K^{cr} = 3.0$. (34)

For 2,4-dichlorophenol, we obtain the following toxicity estimate from equation (34): $A = 1.46$ at $Z = 3.692$ arb. units. The higher the Z and H values, the higher the toxicity of the chemical compound. The existence of such a trend for factor Z is also indicated by the Abbe-Linnick criterion [27]:

$$q = 0.5\cdot\Sigma_{i=1}^{N-1}(Z_{i+1} - Z_i)^2/\Sigma_{i=1}^{N}(Z_i - Z^{av})^2 =$$

$$0.415 < q_{0.05}^{cr}(N) = 0.5636,$$

$$Q^* = -(1 - q)\cdot[(2N + 1)/(2 - (1 - q)^2)]^{0.5} =$$

$$-2.27 < u_{0.05} = -1.645. \quad (35)$$

A similar trend takes place for the molecular information factor H:

$$q = 0.231 < q_{0.05}^{cr}(N) = 0.5636,$$

$$Q^* = -3.24 < u_{0.05} = -1.645. \quad (36)$$

Regression (15) indicates only an important initial stage of the manifestation of its biological action by a chemical compound, associated with the fixation of the molecule.

3.2 Saturated and unsaturated chlorine-containing compounds

In this part of the article, we will again use the general equation (8) to interpret the narcotic effect of a series of saturated and unsaturated chlorine-containing compounds. Isotoxic concentrations (C in units of mM/l) of vapours of compounds causing lateral positioning of 50% of white mice are used as the biological response (Table 2). We use the following regression equation, which is similar in form to regression (8):

$$A_{mod} \equiv 100/C = B_0 + B_1x_1 + B_2x_2 + B_3x_3,$$

$N = 15$, $B_0 = -11.88 \pm 1.11$, $B_1 = -0.96 \pm 0.35$, $B_2 = 0.90 \pm 0.13$, $B_3 = 2.72 \pm 0.27$, $R_3 = 0.980 > R_{0.05}^{cr}(m_3; N - m_3 - 1) = 0.703$, $R_3^2 = 0.960$, $R_3^{*2} = 0.949$; $m_3 = 3$; standard error of the regression estimate: $S_3 = 0.905$; $|t(B_0)| = 10.71 > t(B_3) = 10.24 > t(B_2) = 6.94 > t(B_1) = 2.79 > t_{0.05}^{cr}(f = N - m_3 - 1) = 2.20$; $F = 88.04 > F_{0.05}^{cr}(f_1 = 3; f_2 = 11) = 3.59$; $\Sigma_3 = 9.0072$; $AIC_3 = -0.1100$, $SC_3 = 0.2121$, $SS_3 = 0.2501$. (37)

Bioactivity statistics:

A: $N = 15$, $A^{av} = 5.27 \pm 1.04$; 95% confidence interval: 3.04 - 7.49; $A^{min} = 0.71$, $A^{max} = 13.33$, $S_A = 4.01$, $\tau^{min} = 1.14 < \tau^{max} = 2.01 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.881 = W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(A^{max} - A^{min})/S_A] = 3.15 < U2_{0.05}^{cr}(N) = 4.170$; $N_{repr} = 12$.

Statistics of explanatory variables:

x_1: $N = 15$, $x_1^{av} = -1.19 \pm 0.27$; 95% confidence interval: -1.77, -0.61; $x_1^{min} = -2.75$, $x_1^{max} = 0.45$, $S_{x1} = 1.04$, $\tau^{max} = 1.04 < \tau^{min} = 1.50 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.930 > W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(x_1^{max} - x_1^{min})/S_{x1}] = 3.08 < U2_{0.05}^{cr}(N) = 4.170$, $N_{repr} = 12$,

x_2: $N = 15$, $x_2^{av} = 3.59 \pm 0.61$; 95% confidence interval: 2.29 - 4.89; $x_2^{min} = 0$, $x_2^{max} = 7.95$, $S_{x2} = 2.35$, $\tau^{min} = 1.53 < \tau^{max} = 1.86 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.957 > W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(x_2^{max} - x_2^{min})/S_{x2}] = 3.39 < U2_{0.05}^{cr}(N) = 4.170$; $N_{repr} = 12$,

x_3: $N = 15$, $x_3^{av} = 4.71 \pm 0.30$; 95% confidence interval: 4.07 - 5.34; $x_3^{min} = 3.308$, $x_3^{max} = 7.353$, $S_{x3} = 1.149$, $\tau^{min} = 1.22 < \tau^{max} = 2.31 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.923 > W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(x_3^{max} - x_3^{min})/S_{x3}] = 3.52 < U2_{0.05}^{cr}(N) = 4.170$; $N_{repr} = 12$. (38)

Thus, the populations x_1, x_2, x_3 and A are homogeneous and normally distributed.

For regression (37), the t-values of all coefficients $|t(B_i)| > t_{0.05}^{cr}(N)$. Consequently, the coefficients characterize the significant effect of each of the intermolecular interaction contributions on the toxicity of chemical compounds. To determine the comparative influence of individual contributions of intermolecular interactions, we turn to standardized regression coefficients. We will use relations (22) for this purpose:

$$B_1^* = B_1 \cdot S_{x1}/S_A = -0.251,\ B_2^* = B_2 \cdot S_{x2}/S_A = 0.525,$$
$$B_3^* = B_3 \cdot S_{x3}/S_A = 0.778. \quad (39)$$

Let us determine the contribution of each variable to the variability of bioactivity. The approximate value of the multiple coefficient of determination is determined by the relation (13):

$$R_{appr}^2 = B_1^* \cdot r_{x1,A} + B_2^* \cdot r_{x2,A} + B_3^* \cdot r_{x3,A} =$$
$$0.098 + 0.179 + 0.686 = 0.963. \quad (40)$$

Here, the pair correlation coefficients are $r_{x1,A} = -0.392$, $r_{x2,A} = 0.337$ and $r_{x3,A} = 0.880$. The approximate multiple coefficient of determination (40) practically coincides with the value of the coefficient of determination $R_3^2 = 0.960$ (37). Given the values of the standardized regression coefficients (39), it can be noted that the explanatory variables do not equally affect the variability of the bioactivity. In contrast to the series of chlorobenzenes (Table 1), for which the contribution from the variable $x_1 = \varepsilon_{nb}^0$ is maximum, for a number of chemical compounds from Table 2 this contribution to regression (37) is minimal. At the same time, the maximum contribution to the variability of bioactivity is made by pair dispersion interactions $x_3 = \alpha_1 I_1/(I_1 + 10)$. For chlorobenzenes, the sequence of influence on the bioactivity of intermolecular interactions will be as follows (13): $x_1(53.5\%) > x_2(32.1\%) > x_3(7.5\%)$. Whereas for a number of compounds from Table 2 the hierarchy of influence will be the opposite: $x_1(9.8\%) < x_2(17.9\%) < x_3(68.6\%)$. The parentheses indicate the percentage contributions of the explanatory variables. Note that the signs of the coefficients of B_i^* (37) remain the same as for the regression equation (10). Consequently, the direction of influence of the explanatory variables on the bioactivity of the molecules does not change. Thus, the molecular factors taken into account by the model explain 96.0% of the variation in the bioresponse ($R_3^2 = 0.960$) and only 4.0% remain unexplained.

The hierarchy of values of the regression coefficients B_i^* makes it possible to indicate the relative importance of the intermolecular contributions taken into account in the regression. The maximum contribution to the regression equation is made by the explanatory variables that determine the pair dispersion interaction. The influence of acceptor interactions ($\sim \varepsilon_{nb}^0$) can be considered as corrections to the dispersion and dipole-dipole contributions. It seems that the molecular region of the biophase with which the molecule interacts has the highest filled molecular orbital energy level, which lies below the ε_{nb}^0 level on the energy scale. That is, such an arrangement of energy levels does not favor the transfer of an electron to a free one-electron level ε_{nb}^0.

Testing for collinearity between explanatory variables resulted in the following values for the pairwise correlation coefficients between explanatory variables: $r_{1,2} = 0.541$, $r_{1,3} = -0.545$ и $r_{2,3} = 0.064$. To quantitatively check the presence of collinearity between explanatory variables, we use the Farrar-Glauber test (19):

$$\chi^2 = -(N - 1 - (2m + 5)/6) \cdot \ln(\det|r_{i,j}|) = 12.16 >$$
$$\chi_{0.05}^{2,cr}(f = m(m-1)/2) = 7.82;\ i = 1,2,3;\ \ j = 1,2,3. \quad (41)$$

Since $\chi^2 > \chi^{2,cr}$, it is necessary to reject the null hypothesis about the absence of collinearity between the explanatory variable at a significance level of $\alpha = 0.05$. To determine which explanatory variable generates the greatest interdependence between variables, the following relationship is used [8]:

$$t_{i,k} = r_{i,k} \cdot (N - m)^{0.5}/(1 - r_{i,k}^2)^{0.5}, \quad (42)$$

which has a t-distribution with $f = N - m$ degrees of freedom. Using relation (42), we obtain the following sequence of inequalities: $|t_{1,3}| = 2.251 > t_{1,2} = 2.228 > t_{0.05}^{cr}(f = N - m_3) = 2.179 > |t_{2,3}| = 0.222$. From these inequalities it follows that the variable x_1 generates the interdependence of features.

Table 2
Physico-chemical parameters of chlorine-containing compounds and isotoxic concentrations (C, mM/l) of the vapors of these compounds, causing the lateral position of 50% of white mice.

№	Chemical compound	*Z*, arb. units	*H*, *bits*	d_{20}, g/ml [43]	n_{20} [43]	ε_{nb}^0, eV	$I_1^{*)}$, eV	$\mu_1^{*)}$, *D*,	$\alpha_1^{**)}$, 10^{24}cm^3	$A_{exp.}$, 100/C, [44]	$A_{mod.}$, Eq.(37)
1	Ethyl chloride	2.50	1.299	0.8978	1.3676	0.151	10.70	2.0 [14]	6.40	0.71	0.55
2	Propyl chloride	2.36	1.241	0.8909	1.3879	0.445	10.44	1.8 [14]	8.24	1.23	2.03
3	Vinyl chloride	3.00	1.460	0.9999	1.4046	-0.522	9.82	1.44[14]	7.83[15]	1.56	1.02
4	1,1-Dichlorovinyl	4.00	1.586	1.2180	1.4249	-1.216	9.59	1.13	8.07	2.50	1.17
5	1,2-Dichlorovinyl	4.00	1.586	1.2837	1.4490	-1.214	9.15	1.77	7.78[15]	2.50	2.20
6	1,1-Dichloroethane	3.25	1.500	1.1757	1.4164	-1.206	10.60	1.80[14]	8.38	3.08	3.90
7	Chloromethylene	2.80	1.372	1.3255	1.4242	-1.215	10.79	2.40[14]	6.48[15]	3.08	3.59
8	Trichlorovinyl	3.45	1.539	1.4642	1.4773	-2.276	9.48	1.01	10.06	4.00	4.53
9	Chloroform	4.86	1.449	1.4832	1.4459	-2.467	11.00	1.51	8.23	5.00	4.25
10	Tetrachlorovinyl	6.00	0.919	1.6227	1.5053	-2.623	9.66	0	12.03	5.00	6.70
11	1,2-Dichloroethane	3.25	1.500	1.2531	1.4476	-0.089	10.92	2.43	8.45	5.71	5.49
12	1,2-Dichloropropane	2.91	1.436	1.1559	1.4394	0.099	10.60	2.82	10.20	9.52	9.42
13	1,1,2-Trichloroethane	4.00	1.562	1.4714	1.4940	-1.264	10.89	2.72	10.28	10.00	10.5
14	1,1,2,2-Tetrachloroethane	4.75	1.500	1.5953	1.4940	-1.680	10.89	2.17	12.15	11.76	11.2
15	Pentachloroethane	5.50	1.299	1.6796	1.5025	-2.753	10.88	1.37	14.11	13.33	12.4

*) Dipole moments and ionization potentials of the molecules are calculated by the quantum-chemical method MINDO/3.
**) Polarizabilities of molecules are calculated using the Clausius-Mossotti formula: $\alpha = (n_{20}^2 - 1)3M/(n_{20}^2 + 2)/(4\pi d_{20}N_A)$; d_{20} and n_{20}^2 are the density and refractive index at 20^0C, respectively.

To eliminate or reduce the collinearity of the explanatory variables, perform a linear transformation. For example, for the variable x_2 : $x_2^* = x_1 - x_2$. We write multiple regression (37) as follows:

$$A_{mod} \equiv 100/C = B_0 + B_1x_1 + B_2x_2^* + B_3x_3,$$

$N = 15$, $B_0 = -11.88 \pm 1.11$, $B_1 = -0.064 \pm 0.287$, $B_2 = -0.897 \pm 0.129$, $B_3 = 2.716 \pm 0.265$, $R_3 = 0.980 > R_{0.05}^{cr}(m_3 = 3;\ v = N - m_3 - 1) = 0.703$, $R_3^2 = 0.960$, $R_3^{*2} = 0.954$; $m_3 = 3$; standard error of the regression estimate: $S_A = 0.905$; $|t(B_0)| = 10.71 > t(B_3) = 10.24 > t(B_2) = 6.94 > t_{0.05}^{cr}(f = N - m_3 - 1) > |t(B_1)| = 0.224$; $F = 88.06 > F_{0.05}^{cr}(f_1 = 3; f_2 = 11) = 3.59 = 3.59$; $\Sigma_3 = 9.007$, $AIC_3 = -0.1100$, $SC_3 = 0.2121$, $SS_3 = 0.2501$; $B_1^* = -0.0168$, $B_2^* = -0.4448$, $B_3^* = 0.7783$. (43)

The approximate coefficient of determination R_{appr}^2 practically coincides with the coefficient of determination $R_3^2 = 0.960$ (43):

$$R_{appr}^2 = B_1^* \cdot r_{x1,A} + B_2^* \cdot r_{x*2,A} + B_3^* \cdot r_{x3,A} =$$

$$0.007 + 0.269 + 0.685 = 0.961. \quad (44)$$

The regression coefficient B_1 (43) is statistically insignificant at the 95% confidence level. Correlation coefficients were also determined between the explanatory variables x_1, x_2^* and x_3: $r_{1,2*} = 0.112$, $r_{1,3} = -0.545$ and $r_{2*,3} = -0.207$. Between variables x_1 and x_2^* there was a significant decrease in the correlation coefficient (compare with $r_{1,2} = 0.541$). From the Farrar-Glauber relation (41) we now obtain the following inequality: $\chi^2 = 5.06 < \chi_{0.05}^{2,cr}(f = 3) = 7.82$. Thus, the null hypothesis that there is no significant multicollinearity between the explanatory variables can be accepted.

Statistics of the population of the explanatory variable x_2^*:

$N = 15$, $x_2^{*av} = -4.78 \pm 0.51$; 95% confidence interval: (-5.88, - 3.68); $x_2^{*min} = -8.66$, $x_2^{*max} = -2.49$, $S_{x*2} = 1.989$, $\tau^{max} = 1.151 < \tau^{min} = 1.951 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.919 > W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(x_2^{*max} - x_2^{*min})/S_{x*2}] = 3.10 < U2_{0.05}^{cr}(N) = 4.170$. (45)

Therefore, the set of elements x_2^* is homogeneous and normally distributed. The influence of the explanatory factor $x_1 = \varepsilon_{nb}^0$ on the variability of bioactivity is markedly smaller compared to the

contributions from variables x_2 (or x_2^*) and x_3. Therefore, we will perform the following comparative analysis. Write the following reduced regression:

$$A \equiv 100/C = B_0 + B_2x_2 + B_3x_3. \quad (46)$$

The regression (46) statistics will be as follows:

$N = 15$, $m_4 = 2$, $R_4 = 0.965 > R_{0.05}^{cr}(m_4; N - m_4 - 1) = 0.627$, $R_4^2 = 0.932$, $R_4^{*2} = 0.921$; standard error of the regression estimate: $S_4 = 1.131$; $B_0 = -12.063 \pm 1.384$, $B_2 = 0.681 \pm 0.129$, $B_3 = 3.164 \pm 0.264$; $|t(B_3)| = 12.00 > |t(B_0)| = 8.72 > t(B_2) = 5.28 > t_{0.05}^{cr}(f = N - m_4 - 1) = 2.18$, $F = 82.04 > F_{0.05}^{cr}(f_1 = 2; f_2 = 12) = 3.88$; $\Sigma_4 = 15.354$, $AIC_4 = 0.2900$, $SC_4 = 0.5649$, $SS_4 = 0.3014$. (47)

There is no collinearity between variables x_2 and x_3 ($r_{2,3} = 0.064$). The multiple coefficient of sample correlation R_4 significantly exceeds the tabular value of the correlation coefficient $R_{0.05}^{cr}(m_4; N - m_4 - 1)$. Thus, 93.2% of the total variance is due to the variability of the molecular factors x_2 and x_3. The uncertainty factor is 6.8%. However, all three comparative information tests AIC_4, SC_4, SS_4 indicate that the quality of the regression (46) is markedly reduced compared to the regression (37).
The standardized regression coefficients (46) are:

$$B_2^* = 0.399, \qquad B_3^* = 0.907. \quad (48)$$

Adjusted coefficients of determination are $R_3^{*2} = 0.95$ for regression (37) and $R_4^{*2} = 0.921$ for reduced regression (46), respectively. Using the standardized regression coefficients (48) an approximate coefficient of determination can be determined:

$$R_{appr}^2 = B_2^* \cdot r_{x2,A} + B_3^* \cdot r_{x3,A} = 0.14 + 0.80 = 0.94. \quad (49)$$

This value of the coefficient of determination is very close to the value of $R_4^2 = 0.932$ (47).

The significance of the contribution of x_1 to the regression equation (37) can be checked with the following statistics:

$$F = (|R_3^{*2} - R_4^{*2}|)(N - m_3 - 1)/(m_3 - m_4)/(1 - R_3^{*2}) = 6.04 > F_{0.05}^{cr}(f_1 = m_3 - m_4; f_2 = N - m_3 - 1) = 4.84. \quad (50)$$

Since $F > F^{cr}$ the additional explanatory variable ε_{nb}^0 is significant at the 95% confidence level. Let us also check whether the decrease in the variance of equation (37) is the result of an increase in the number of connections compared to regression (46). For normally distributed sets, the comparison of the two variances of equations (37) and (46) is done using a statistic that has an F-distribution:

$$F = S_4^2/S_3^2 = 1.56 < F_{0.05}^{cr}(f_1 = N - m_4 - 1; f_2 = N - m_3 - 1) = 2.79. \quad (51)$$

Therefore, at the 95% confidence level, the decrease in the variance of the regression equation (37) is not due to an increase in the number of explanatory variables. Thus, variable x_1 must be retained in the regression equation. All three AIC, SC and SS information tests do not contradict inequality (50). Preference is given to the model for which the information test has the least value. The test comparison is only valid for models built for samples containing the same number of observations. A comparison of the quality criteria of regression equations (10), (16), (37) and (46) is presented in Table 3.

Table 3
Quality criteria for linear regression equations

i	m	N	Σ_i	AIC_i	SC_i	SS_i	Equation
1	3	12	1.1586	-1.8377	-1,5093	0.1196	(10)
2	2	12	1.2294	-1.9450	-1.6572	0.1109	(16)
3	3	15	9.0072	-0.1100	0.2121	0.2501	(37)
4	2	15	15.354	0.2900	0.5649	0.3014	(46)

As expected, the quality criteria of the regressions correlate with each other: $R_{AIC\text{-}SC} = 0.9997$, $R_{AIC\text{-}SS} = 0.997$ and $R_{SC\text{-}SS} = 0.995$. From the inequalities for the t-values of the regression coefficients (38), one can indicate the following sequence of intermolecular interactions: dispersion interaction (short - range) > dipole - dipole, induction and polarization interactions > interaction associated with electron transfer. Each of the contributions of this sequence has a very specific physical meaning. This, in turn, makes it possible to associate interactions with certain physical characteristics of biophase molecules. Highlighting interactions is useful for determining the presumed characteristic

molecular properties of the local centers with which a bioactive molecule interacts.

The fact that the main contribution to the regression equation comes from the dispersion interaction allows us to make some assumptions about the molecular properties of the biophase with which the bioactive molecule forms a complex. According to the definition (6) the interaction of molecules of a number of chlorinated chemical compounds (Table 2) is most likely to take place with the biophase region, which is characterized by high values of electronic polarizability (α_2) and has a high first ionization potential (I_2). It is also noteworthy that for models (15) and (46) the dipole interaction turns out to be significant. This result indicates the importance of mutual orientations between bioactive molecules and biophase molecules.

Figure 4 shows the correlation between the observed bioactivity values of molecules and the calculated bioactivities of chemical compounds using equation (37).

A relationship was also found between the dispersion contribution $E_{disp} \sim x_3$ and the molecular factor Z of chlorine-substituted hydrocarbons presented in Table 2:

$x_3(Z) = a_0 + a_1 Z$, $N = 15$, $R = 0.71 \pm 0.14$, $R^* = 0.73 > R_{0.05}^{cr}(N-2) = 0.514$; sample size sufficient for the reliability of the correlation coefficient $N_{0.05}^{min} = 7$; correlation coefficient significance test based on the Fisher normalizing z-transform (Hotelling corrections is taken into account): $u_H = 0.87 > u_{0.05}(N) = z_{0.975}\cdot(N-1)^{-0.5} = 0.523$; $a_0 = 1.91 \pm 0.81$, $a_1 = 0.74 \pm 0.21$, $t(a_1) = 3.58 > t(a_0) = 2.36 > t_{0.05}^{cr}(N-2) = 2.16$; standard error of the regression estimate: 0.846; $F = 12.85 > F_{0.05}^{cr}(f_1 = 1; f_2 = 13) = 4.70$; straightness index: $K = 2.73 < K^{thr} = 3.00$. (52)

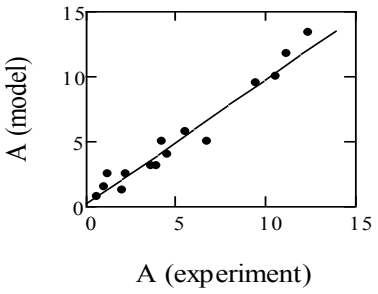

Fig.4. Scatter plot and regression line. The regression line is defined by the equation $A_{mod} = a_0 + a_1 A_{exp}$; $N = 15$, $a_0 = 0.212 \pm 0.355$, $a_1 = 0.960 \pm 0.054$; $R = 0.98 \pm 0.01$, $R^* = 0.982 > R_{0.05}^{cr}(N-2) = 0.514$; $S = 0.81$; $F = 313.6 > F_{0.05}^{cr}(f_1 = 1; f_2 = 13) = 4.7$; $K = 0.77 < K^{thr} = 3.0$.

Statistics of explanatory variable Z:

Z: $N = 15$, $Z^{av} = 3.78 \pm 0.28$; 95% confidence interval: 3.17 - 4.38; $Z^{min} = 2.36$, $Z^{max} = 6.00$, $S_Z = 1.10$, $\tau^{min} = 1.29 < \tau^{max} = 2.03 < \tau_{0.05}^{cr,2}(N) = 2.493 < \tau_{0.05}^{cr,1}(N) = 2.617$; Wilk-Shapiro normality test: $W = 0.934 > W_{0.05}^{cr}(N) = 0.881$, David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N) = 2.970 < U = [(Z^{max} - Z^{min})/S_Z] = 3.31 < U2_{0.05}^{cr}(N) = 4.170$; $N_{repr} = 12$. (53)

The sets of elements x_3 (39) and Z (53) satisfy the homogeneity and normality conditions at the 95% confidence level.

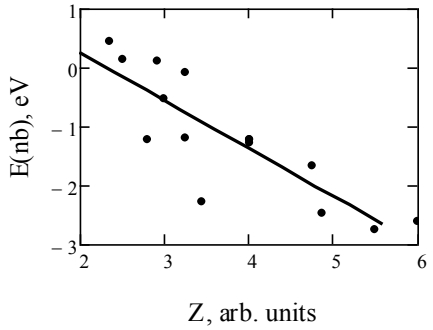

Fig.5. Scatter plot and regression line. The regression line is defined by equation (54).

For the chemical compounds from Table 2 there is also a significant relationship between the molecular factor Z and the energy ε_{nb}^0 (Fig. 5):

$\varepsilon_{nb}^0(Z) = a_0 + a_1 Z$, $N = 15$, $R = -0.85 \pm 0.08$, $|R^*| = 0.85 > R_{0.05}^{cr}(N-2) = 0.514$; sample size sufficient for the reliability of the correlation coefficient: $N_{0.05}^{min} = 5$; correlation coefficient significance test based on the Fisher normalizing z-transform (Hotelling corrections is taken into account): $u_H = 1.179 > u_{0.05}(N) = z_{0.975}\cdot(N-1)^{-0.5} = 0.523$; standard error of the regression estimate: 0.575; $a_0 = 1.87 \pm 0.55$, $a_1 = -0.81 \pm 0.14$, $|t(a_1)| = 5.79 > t(a_0) = 3.41 > t_{0.05}^{cr}(N-2) = 2.16$; $F = 33.50 > F_{0.05}^{cr}(f_1 = 1; f_2 = 13) = 4.70$; straightness index: $K = 2.04 < K^{thr} = 3.00$. (54)

The statistical significance of the correlation coefficient is characterized by the inequality (26): $t = 4.35 > t_{0.05}^{cr}(N-2) = 2.16$. According to the Chaddock scale, the correlation coefficient is in the range of 0.7 – 0.9, which is characterized as “close relationship”.For the chemical compounds presented in Table 2, a relationship was found between bioactivity and the molecular factor Z. There is a statistically significant trend between the Z value and the bioactivity of chemical compounds. Indeed,

using the Abbe-Linnick test (35) we obtain the following inequality:

$$q = 0.473 < q_{0.05}^{cr}(N = 15) = 0.6027,$$
$$Q^* = -2.24 < u_{0.05} = -1.645. \quad (55)$$

Here $Z^{av} = 3.78$ is the arithmetic mean of the set Z_i. Since inequalities (55) are satisfied, the null hypothesis about the absence of a trend is rejected. The alternative hypothesis of a trend is accepted. The following significant linear relationship of bioactivity with factor Z was also obtained:

$A(Z) = d_0 + d_1Z$, $N = 15$, $R = 0.54 \pm 0.20$, $R^* = 0.56 > R_{0.05}^{cr}(N - 2) = 0.514$; sample size sufficient for the reliability of the correlation coefficient $N_{0.05}^{min} = 13$; correlation coefficient significance test based on the Fisher normalizing z-transform (Hotelling corrections is taken into account): $u_H = 0.592 > u_{0.05}(N) = z_{0.975}\cdot(N - 1)^{-0.5} = 0.523$; $d_0 = -2.23 \pm 3.35$, $d_1 = 1.99 \pm 0.95$, $t(d_1) = 2.33 > t_{0.05}^{cr}(N - 2) = 2.160$; standard error of the regression estimate: $S_A = 3.50$; $F = 5.42 > F_{0.05}^{cr}(f_1 = 1; f_2 = 13) = 4.70$. (56)

It should be noted that regression (56) does not make it possible to distinguish between *cis*- and *trans*-isomers. The Z factor correlates with the electronic polarizability of the molecule α_1 ($R = 0.72 > R_{0.05}^{cr}(N - 2) = 0.514$; $t = 3.14 > t_{0.05}^{cr}(N - 2) = 2.160$). There is also a relationship ($R = 0.71$; $t = 3.07 > t_{0.05}^{cr}(N - 2)$) with the value of $\alpha_1 I_1/(I_1 + 10)$. There is also a relationship between the Z factor and the MO energy ε_{nb}^0 ($|R| = 0.85 > R_{0.05}^{cr}(N - 2) = 0.514$; $t = 4.35 > t_{0.05}^{cr}(N - 2) = 2.160$) (Fig. 5). For chemical compounds (Table 2) having the general formula $C_2H_xCl_y$ (sample volume $N = 11$) there is a very close linear relationship between the factor Z and the value of the one-electron MO energy ε_{nb}^0: $R = -0.95$, $|R^*| = 0.96 > R_{0.05}^{cr}(N - 2) = 0.602$; $t = 5.18 > t_{0.05}^{cr}(N - 2) = 2.262$. Note that the energy ε_{nb}^0 of the molecule actually characterizes the affinity of the molecule for the electron.

When analyzing the data in Table 2, of the two possible structures of the 1,2-dichlorvinyl molecule, the *cis*-isomer structure corresponding to the lower total energy state of the molecule in a polar medium was chosen (assuming $\kappa_s = 80$ for the static dielectric permittivity of the polar medium and $n_3^2 = 1.777$ for the refractive index). The quantitative determination of the difference in the total energies of a molecule in *cis*- and *trans*-configurations is associated with the calculation of the difference between two large quantities, the accuracy of which depends significantly on the accuracy of the quantum-chemical method used. Nevertheless, some approximate quantitative estimates can be made if the same method for calculating the electronic structure of molecules is used in both cases. It is assumed that possible inaccuracies of the quantum-chemical method can be compensated for. The total electronic energies of the *trans*-isomer and *cis*-isomer calculated using the CNDO/2 method are $E_{trans} = -1295.97$ eV and $E_{cis} = -1295.95$ eV, respectively. However, in a polar dielectric medium, the energy of the dipole *cis*-isomer, compared to the nondipole *trans*-isomer, decreases on the energy scale by

$$E_R = -(\kappa_s - 1)(n_3^2 + 2)\mu_1^2/[3a^3(2\kappa_s + n_3^2)] = -08 \text{ eV}. \quad (57)$$

Here it is accepted: $\mu_1 = 1.77D$; $a = 2.5$ Å is the effective size of the molecule. This energy value is noticeably higher in absolute value of the thermal energy of the translational motion of a molecule at room temperature (thermal energy ≈ 0.02 eV). The difference in the total electronic energies of the molecules is practically compensated. Then, according to Boltzmann's statistics, one can estimate the number of molecules in *cis*- and *trans*-configurations in a condensed dielectric medium as follows:

$$N_{cis} = N_{trans}\exp[(E_{trans} - E_{cis} - E_R)/k_BT] = 13N_{trans}. \quad (58)$$

Consequently, at room temperature, there are more than ten molecules in the *cis*-configuration per molecule in the *trans*-configuration. That is, most of the molecules in a polar dielectric medium will preferably be in the *cis*-configuration.

Comparing the regressions (15) and (46), we can make some assumptions about the selective nature of the biological action at the molecular level of the analyzed chlorine-containing chemical compounds. In the first case (Table 1), chemical chlorinated compounds are most likely to interact with those areas in the body characterized by strong donor properties. In the second case (Table 2), the active sites of the biophase are more prone to the formation of molecular associates, which are stabilized due to dispersion and dipole interactions. This is typical for the active centers of the biophase, which have high polarization properties (large values of polarizability α_2). Since short-range interactions are important for equations (15) and (46), then for the manifestation of the biological activity of chlorine-containing chemical compounds, the molecules must approach the active centers of the biosubstrate at close distances less than 5Å.

4 Conclusion

Thus, the studied variability in the toxicity of chlorinated hydrocarbons (Tables 1 and 2) is probably due to the accumulation (at least in the initial stages of action) of exogenous chemical compounds in the body's active centers. The change in toxicity can be associated with the electronic structure of exogenous molecules, which determines their ability to form bound molecular complexes due to various types of intermolecular interactions. In the case when the donor electron level ε_{don} lies higher on the energy scale of the acceptor level ε_{acc}, a real electron transfer from the impurity molecule to the biophase molecule is possible. In this case, in the local region of the biophase, through relaxation mechanisms, energy is radiationless released, approximately equal to the difference $\varepsilon_{don} - \varepsilon_{acc}$. This energy can be used to destroy the equilibrium structure of the biophase.

Multiple regression (8), (15), (37) reflect the existence of objective relationships between the bioresponse and a variety of molecular factors that characterize the interaction of molecules in a condensed medium. The general structure of the response function (8) with a clear physical meaning of the explanatory variables is able to cover various aspects of the manifestation of the toxicity properties of the analyzed chlorine-containing chemical compounds of two different classes. Additive molecular features x_1, x_2 and x_3 are intensive indicators that determine the cause-and-effect relationships of bioactivity - the molecular structure of chemical compounds. At the same time, the models statistically significantly reflect the relationship between individual molecular factors and the biological response of the body. It is important to note that the relatively simple mathematical formulas derived from rigorous theoretical concepts made it possible to obtain a statistically significant relationship between bioresponse and the molecular parameters of chemical compounds. Such a relationship is difficult to establish by the usual direct deduction from the general to the particular.

References:

[1] Zahradnik R., *Arch. Int. Pharmacodyn*., Vol.135, 1962, p.311.

[2] Zahradnik R., *Experientia*, Vol.18, 1962, p.534.

[3] Boček K., Kopecky J., Krivucova M., *Experientia*, Vol.20, 1964, p.657.

[4] Kopecky J., Boček K., *Experientia*, Vol.23, 1967, p.125.

[5] Hansch C., On the Use of Quantitative Structure-Activity Relationships in Drug Design (Review). *Chem.-Pharm. Journal*, Vol.14, No.10, 1980, pp.15-30. (in Russian).

[6] Hansch C., Leo A., *Substituent Constant for Correlation Analysis in Chemistry and Biology*. John & Sons, New York, Chichester, Brisbane, Toronto, 1979.

[7] Aibinder N. E., Bezdvorny V.N., Krasovitskaya M. L., In: *Study of Biological Action of New Products of Organic Synthesis and Natural Compounds*. Perm. 91 p., 1981, (in Russian).

[8] Förster E., Rönz B., *Metohden der Korrelations- und Regressionsanalyse*, Verlag Die Wirtschaft Berlin, 290 p., 1979.

[9] Likeś J., Laga J., *Základnŷ. Statistice Tabulky*, Praha, 356 p., 1978.

[10] Kubini H., In: *Biological Activity and Chemical Structure*, Amsterdam, 1977.

[11] Golubev A.A., Lublina E.I., Tolokontsev N.A., *Quantitative Toxicology*, Leningrad, 1973 (in Russian).

[12] Mukhomorov V.K., Frumin G.T., Quantitative Ratios of Bioactivity - Electronic Characteristics of Halocarbons of the Aliphatic Series, *Chem.-Pharm. Journal*, Vol.16, No.10, 1982, pp.70-74 (in Russian).

[13] Mukhomorov V.K., Structure-Activity Relationships. Intermolecular Interactions and Toxicity of Compounds. *Toxicology Letters*, Vol.88, Issue S1, 1996, p.87.

[14] Osipov V.A., Minkin V.M., *Reference Book on Dipole Moments*, Moscow, 1965 (in Russian).

[15] Vereshchagin A.N., *Polarizability of Molecules,* Moscow, 1980, (in Russian).

[16] Frumin G.T., Mukhomorov V.K., Comparative Analysis of Statistical Distributions as the Basis of Bioactivity-Structure Models, *Chem.-Pharm. Journal.*, Vol.16, No.10, 1982, pp,70-74, (in Russian).

[17] Klopman M., Simonetta M., Fujimoto H., *Reactivity of Molecules and Reaction Pathways*, Moscow, 1977 (in Russian).

[18] Shchembelov G.A., Ustinyuk V.M., Mamaev V.M., Ischenko V.M., Gloriozov I.P., Luzhkov V.B., Orlov V.V., Simkin V.Y., Pupyshev V.I., Burmistrov V.N., (1980). *Quantum-Chemical Methods of Molecule Calculation*. Edited by V.M. Ustynyuk, Moscow, Publishing House Chemistry, 1980 (in Russian).

[19] Sutton L.E., *Tables of Interatomic Distances, Configuration in Molecules*, London, 1958, 1965.

[20] Frölich H., *Theory of Dielectrics*, Oxford, 1958.

[21] Onsager L., Electric Moments of Molecules in Liquids, *J. Am. Chem. Phys*., Vol.58, No. 8, 1936, p.1486-1493.
[22] Bakhshiev N.G., *Spectroscopy of Intermolecular Interactions,* Leningrad, 1972 (in Russian).
[23] Willcams D., Schaad L., Murrel J., Deviations from Pairwise Additivity in Intermolecular Potentials, *J. Chem. Phys.*, Vol.47, No 12, 1967, pp.4916-4923.
[24] Kestner N., Sinanoglu O., Effective Intermolecular Pair Potentials in Nonpolar Media, *J. Chem. Phys*., Vol.38, No.7, 1963, pp.1730-1740.
[25] Gurevich L.V., Karachevtsev G.V., Kondratyev V.N., Lebedev Yu. A., Medvedev V.A., Potapov V.K., Khodeev Yu. S., *Energies of Breaking Chemical Bonds. Ionization Potentials and Electron Affinity*, Moscow, Nauka, 1974 (in Russian).
[26] Bazilevsky M.V., *Method of Molecular Orbits. Reactivity of Organic Molecules*, Moscow, Publishing House Chemistry, 1969 (in Russian).
[27] Kobzar A.I., *Applied Mathematical Statistics. For Engineers and Scientists,* FizMatLit., Moscow, 2006 (in Russian).
[28] Bolshev L.N., Smirnov N.V., *Tables of Mathematical Statistics,* Moscow, Nauka, 1983 (in Russian).
[29] Akaike Hirotogu, A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, Vol.19, No.6, 1974, pp. 716-723.
[30] Schwarz G., (1978). Estimating the Dimension of Model, *The Annals of Statistics*, Vol.6, No.2, 1978, pp. 461-464.
[31] Dmitriev E.A., *Mathematical Statistics in Soil Science,* Moscow State University Press, 1995.
[32] Draper N.R., Smith H., *Applied Regression Analysis*, John Wiley & Sons, New York, 1996.
[33] Sachs L., *Statistische Auswertungsmetoden*, Springer Verlag, Berlin, New York, 1972.
[34] Farrar D.E., Glauber R.R., Multicollinearity in Regression Analysis: The Problem Revisited, In: *The Review of Economics and Statistics*. Vol.49, No.1, 2018, pp. 92-107.
[35] Zaitsev G.N., *Mathematics in Experimental Botany*, Moscow, 1990, (in Russian).
[36] Mukhomorov V.K., Statistical Asptcts of the Interrelation, *Biomedical Statistics and Informatics*, Vol.1, No.1, 2016, pp. 24-34.
[37] Mukhomorov V.K., Entropy Approach to the Study Biological Activity Chemical Compounds, *Advances in Biological Chemistry*, Vol.1, No.1, 2011, pp. 1-5.
[38] Veljkovič V., Lalovič D., General Model Pseudopotential for Positive Ions, *Phys. Lett. A*, Vol.45, No1, 1973, pp. 59-62.
[39] Quastler H., Information Theory in Biology, Moscow, IL, 1960, (in Russian).
[40] Chaddock R.E., *Principles and Methods of Statistics*, Boston, New York, 1925.
[41] Johnson N.L., Leone F.C., *Statistics and Experimental Design,* Vol.1, Second Edition, John Wiley & Sons, New York, London, Toronto, 1977.
[42] Caillet J., Pullman B., In: *Molecular Associations in Biology*, Ed. B. Pullman, Academic Press, 1968.
[43] *Handbook of Chemistry and Physics*, 52d. Ed. R.C.Weast, Ohio, 1971.
[44] Golubev A.A., Lublina E.I., Tolokontsev N.A., *Quantitative Toxicology*, Leningrad, 1973.