Statistical Modeling of Antiemetic Activity of Chemical Compounds

VLADIMIR K. MUKHOMOROV Physical Department University "Federico II" Via Cintia, I-80126, Napoli ITALY

Abstract: - A model has been built linking the molecular structure of drugs with their antiemetic activity. It is shown that the bioactivity of drugs depends on the pseudopotential of the molecule, the information function of the molecule, and the geometric size of the substituents. The critical conditions for the emergence of effective antiemetic activity of molecules have been established. The optimal sizes of the substituents were determined, which correspond to the maximum bioactivity of the drugs. The mechanism of the threshold action of substituted benzamides was revealed.

Key-Words: - Antiemetic, statistical criteria, threshold, information function, pseudopotential, linear dimension, benzamides.

Received: February 22, 2021. Revised: July 10, 2021. Accepted: July 19, 2021. Published: July 30, 2021.

1 Introduction

The search for effective antiemetics is currently receiving much attention, since nausea and vomiting are among the most frequent and severe adverse reactions, for example, during radiation and chemotherapy of cancer patients [1]. The reaction of the body can be so significant that it forces you to interrupt or completely stop treatment [2]. Analysis of literature materials indicates that at present the most promising agents with antiemetic activity are compounds of a number of substituted benzamides, which do not give side effects and are well tolerated by patients [3,4]. In the article, a statistical study was undertaken of the relationship of such molecular factor-signs that make it possible to establish quantitative relationships between the structure of the molecule and its biological activity in the series of 4-substituted 5-nitro-2-methoxy-N-(2-diethylaminoethyl) benzamides (Fig. 1) [5]. In the investigated series of benzamides, effective modifications of biological activity were obtained by varying the substituent in the fourth position of the benzene ring (Table 1). The antiemetic activity (A) of the most effective drug dimetpromide used in clinical practice was conventionally assessed by ten points [6].

2 Problem Formulation

It is assumed [7] that the bioactivity of chemical compounds of a number of benzamides is essentially due to the distance between the substituent in the fourth position of the benzene ring (R^4) and the cationic head N⁺(C₂H₅)₂ (Fig. 1). The authors of this article also suggested that the value of this distance should have some optimal value.

Studies of the electronic structure of molecules from Table 1 were carried out by the semi-empirical quantum chemical method MINDO/3 [8]. This method also makes it possible to obtain the optimal spatial geometry of the molecules. Calculations have shown that the conformation of molecules is slightly different from each other. As a result, the distance between the substituent R⁴ and the nitrogen atom for all compounds in Table 1 remains practically unchanged and fluctuates around the value of 9.20 Å. The slight difference that exists is mainly due to the difference in bond lengths C(4)–H μ C(4)–R⁴.

To interpret the changes in the bioactivity of the compounds upon variation of the R^4 substituent, various electronic parameters of the molecules were tested. The following parameters were calculated: charges on atoms, energies of molecular orbitals, as well as the electronic sign *Z* associated with the pseudopotential of the molecule and the information function of the molecule *H* [9,10]. In addition, the steric parameters of the Verloop substituents were

used [11].



Figure 1. Structural formula of the molecule of 4-substituted 5-nitro-2-methoxy-N- (2-diethylamino-ethyl) benzamide.

A.Verloop together with his colleagues, proposed a multiparametric approach to determining the spatial dimensions of substituents. The spatial dimensions of the substituent are characterized by five steric dimensions: L, B_1 , B_2 , B_3 , and B_4 [11]. The L parameter turned out to be the most informative parameter. The value of the parameter L determines the linear size of the substituent along the chemical bond line. Parameters B_1 and B_2 determine the size of the substituent along the perpendicular to the Laxis up to the assumed maxima of the substituent boundaries (left and right). Similarly, estimates are made for dimensions B_3 and B_4 (left and right) perpendicular to line L and line B_1 - B_4 . Table 1 also shows the melting points of benzamides. The verification did not give positive results on the relationship between the variability of the bioactivity of chemical compounds and the melting temperature T_{melt} .

It is known, the physiological effect of biologically active molecules is often determined by their spatial geometry, that is, the location and size of the atoms and groups of atoms included in the molecule. In this case, the interaction between the molecule of the biological object (receiver) and the drug (signal source) is very specific. As a result of "recognition" the biological object develops a specific reaction in the process of interaction.

3 Problem Solution

Let us check the presence of a trend for chemical compounds from Table 1 to determine the significance of the relationship between bioresponse and the explanatory variable L. To do this, we use the Abbe-Linnik test [12]:

$$q = 0.5 \cdot \sum_{i=1}^{N-1} (L_{i+1} - L_i)^2 / \sum_{i=1}^{N} (L_i - L^{av})^2$$

= 0.159 < $q_{0.05}^{cr}(15) = 0.602$,

$$Q^* = -(1-q) \cdot [(2N+1)/(2-(1-q)^2)]^{0.5} = -28.66$$

$$< u_{0.05} = -1.645.$$
 (1)

Here N = 15. The arithmetic mean value: $L^{av} = 4.69$ (in arbitrary units); $u_p/2$ is the quantile of the normal distribution at p = 0.10. Since $q < q^{cr}$ the null hypothesis of series homogeneity should be rejected at a confidence level of 95%. Consequently, a number of L_i values cannot be recognized as random. Additional statistical information on the linearity of the relationship between the bioactivity A_{exp} and the value of the molecular parameter L for chemical compounds can be obtained by compiling a variation series from grouped subsets of the sample. Let us check whether the relationship is linear between the experimentally determined activity of molecules A_{exp} and the value of the molecular parameter L. The ranked series of chemical compounds was divided into three equal groups in accordance with the Sturges ratio $n \approx 1 + 1$ $3.32 \cdot \lg(N)$ [13]. As a result, the following sequence of parameter values $b_{LA,i-j}$ was obtained (here i and *i* are group numbers):

$$b_{LA,1-2} = \frac{A_{(1)}^{av} - A_{(2)}^{av}}{L_{(1)}^{av} - L_{(2)}^{av}} = \frac{8.83 - 8.67}{2.81 - 3.76} = -0.17,$$

$$b_{LA,1-3} = \frac{A_{(1)}^{av} - A_{(3)}^{av}}{L_{(1)}^{av} - L_{(3)}^{av}} = \frac{8.83 - 6.00}{2.81 - 4.80} = -1.42,$$

$$b_{LA,2-3} = \frac{A_{(2)}^{av} - A_{(3)}^{av}}{L_{(2)}^{av} - L_{(3)}^{av}} = \frac{8.67 - 6.00}{3.76 - 4.80} = -2.57.$$
 (2)

The values of the parameters $b_{LA,i\cdot j}$ should not differ significantly for a linear relationship. However, it is obvious from the sequence (2) that the parameters $b_{LA,i\cdot j}$ differ significantly. Therefore, the relationship for a group of chemical compounds (Nos. 1–9) is most likely to be non-linear. Let us also compare the mean values for two samples Nos. 1–9 (first group) and Nos. 10–15 (second group). Let's first compare the sample variances for f_1 and f_2 degrees of freedom. The Fisher distribution is used as a significance criterion:

$$F = S_{A1}^2 / S_{A2}^2 = 26.9 >$$

$$F_{0.05}^{\rm cr}(f_1 = N_1 - 1; f_2 = N_2 - 1) = 4.82.$$
(3)

That is, the sample variances for bioactivities differ significantly at the 95% confidence level.

			*)	*)		A _{exp.}	$A_{\text{mod.1}}$	$A_{\rm mod.2}$
N⁰	Substituent R ⁴	$T_{\text{melt.}}$,	$L^{*)}$	$Z^{(i)}$	Н,	[6]	Eq.(20)	Eq.(32)
		°C [6]			bits		points	
1	$N(CH_3)_2$	175-176	3.53	2.68	1.60	10.0	8.8	-
2	NH ₂	178-179	2.93	2.77	1.66	9.0	9.2	-
3	NO ₂	176-177	3.44	3.00	1.75	9.0	8.9	-
4	NHCH ₃	206-207	3.53	2.72	1.63	8.5	8.8	-
5	Н	206-207	2.06	2.76	1.62	8.5	8.6	-
6	Ι	191-192	4.23	2.91	1.76	7.5	7.7	-
7	$N(C_2H_5)_2$	148-149	4.50	2.61	1.55	6.0	7.1	-
8	NHC ₂ H ₅	208-209	4.96	2.68	1.60	6.0	5.7	-
9	NH(CH ₂) ₂ C	172-173	4.96	2.80	1.72	6.0	5.7	-
10	NHCH ₂ CH=CH ₂	188-189	5.11	2.71	1.60	1.0	-	0.8
11	NH(CH ₂) ₇ CH ₃	202-203	6.50	2.50	1.48	0.5	-	0.2
12	N	167-168	5.50	2.67	1.58	0.5	-	0.6
13	N	185-186	6.02	2.73	1.61	0.5	-	0.4
14	N	170-171	6.06	2.59	1.51	0.2	-	0.4
15	N N-CH3	208-209	6.15	2.64	1.58	0.2	-	0.35

Table 1.

^{*)} In arbitrary units.

Consequently, the samples belong to different general populations. In this case, for a quantitative comparison of the average values of bioactivities, one can use the inequality [14]:

$$|A_1^{av} - A_2^{av}| = 7.35 > T =$$

$$\frac{v_1 t_{0.05}^{cr} (f_1 = 8) + v_2 t_{0.05}^{cr} (f_2 = 5)}{(v_1 + v_2)^{0.5}} = 1.21.$$
(4)

Here $N_1 = 9$, $N_2 = 6$; $v_1 = S_{A1}^2 / N_1 = 0.257$, $v_2 = S_{A2}^2 / N_2 = 0.014$. Inequality (4) rejects the null hypothesis of equality of means. Consequently, the difference between the bioactivities of the first and second groups of drugs is significant (not a coincidence).

The data of Table 1, which presents a series of

compounds ranked by the L (sorted from lowest value to highest value), indicate the following. First, there is a trend in the activity of chemical compounds (1). Second, there is a threshold for the activity of chemical compounds, which depends on the linear size L of the substituent. We will use Cochran's criterion to statistically evaluate the significance of the existence of a threshold [12, 15]. Consider the following series of bioactivities (n = 7), ranked in order from lowest value to highest value:

$$A_i: \quad 0.2; \ 0.2; \ 0.5; \ 0.5; \ 0.5; \ 1.0; \ 6.0 \text{ (points)}, \\ i = 1, 2, \dots, 7. \tag{5}$$

The following average values of groups of consecutive m and n - m elements were calculated, respectively:

$$y_{3}^{(1)} = (A_{1} + A_{2} + A_{3})/3 = 0.3, \qquad m = 3;$$

$$y_{3}^{(2)} = (A_{4} + A_{5} + A_{6} + A_{7})/4 = 2.00,$$

 $y_4^{(1)} = (A_1 + A_2 + A_3 + A_4)/4 = 0.35, \qquad m = 4;$ $y_4^{(2)} = (A_5 + A_6 + A_7)/3 = 2.50,$

 $y_5^{(1)} = (A_1 + A_2 + A_3 + A_4 + A_5)/5 = 0.38, \quad m = 5;$ $y_5^{(2)} = (A_6 + A_7)/2 = 3.50.$ (6)

Using the mean values (6), the values were calculated χ_m^2 :

$$\chi_m^2 = m(n-m)(y_m^{(1)} - y_m^{(2)})^2 / n / A^{\text{av}},$$

$$A^{\text{av}} = \sum_{i=1}^n A_i / n = 1.27,$$

$$\chi_2^2 = 2.53, \quad \chi_3^2 = 3.90, \quad \chi_4^2 = 6.24, \quad \chi_5^2 = 10.95.$$
(8)

If the value of χ_m^2 is significantly greater than the tabular value of $\chi_{0.05}^{2,cr}(f=1)=3.841$, then the change in the average value by a jump is considered significant at the 95% confidence level. From Eq.(8) it follows that after i = 6 Cochran's criterion indicates the presence of a jump in the average value of bioactivity (drug 10 in Table 1). This result does not contradict inequality (4). The region of the jump in bioactivity corresponds to the linear size of the substituent $L^{thr} \approx 5$ in arbitrary units. You can also additionally use the following test [15]:

$$\chi^{2} = (N_{1} \cdot N_{2} / N) \cdot (A_{1}^{av} - A_{2}^{av})^{2} / A^{av} = 39.74 >$$

$$\chi^{2, cr}_{0.05} (f = 1) = 3.841.$$
(9)

Here, the average values of the observed antiemetic activity are used for bioactive $(N_1 = 9)$ compounds $A_1^{av} = 7.833$ (points), for weakly active $(N_2 = 6)$ compounds $A_2^{av} = 0.4833$ (points), respectively, as well as the total (N = 15) average $A^{av} = 4.893$ (in points). Since $\chi^2 >> \chi^{2,cr}$, it is necessary to accept the hypothesis of the presence of a jump (or threshold) in the biological activity of chemical compounds.

The stability (uniformity) of the tendency of the ranked series can also be checked using the Chow test [16]:

$$F = \frac{(\Sigma_0 - \Sigma_1 - \Sigma_2)(N - 2m - 2)}{(\Sigma_1 + \Sigma_2)(m + 1)} = 39.95 >$$

$$F_{0.05}^{\rm cr}(f_1 = m + 1, f_2 = N - 2m - 2) = 3.98.$$
(10)

Here $N_1 = 9$ and $N_2 = 6$ are the sizes of the sample populations; $\Sigma_1 = 17.21$ and $\Sigma_2 = 0.36$ are the

residual sums of the squares of the regressions; m = 1 is the number of explanatory variables; $N = N_1 + N_2$ is the total sample; $\Sigma_0 = 143.28$. It follows from inequality (10) that the structures of the two subsamples differ significantly and the null hypothesis of the homogeneity of the two subsamples is rejected. Therefore, it is not recommended to combine them into one sample. That is, each subsample should be considered separately, since there is a structural jump or break in the trend.

We also performed a statistical comparison of partial samples Nos. 1–9 and Nos. 10–15 for the bioactivity A_{exp} and the molecular parameter *L* of the substituent. Sample statistics (Nos. 1–9) will be as follows:

 $N_1 = 9, A_1^{av} = 7.83 \pm 0.51; 95\%$ confidence interval is (0.134 - 0.165); $A_1^{min} = 6.0, A_1^{max} = 10, S_{A1} =$ 1.52; Grubbs-Romanovsky homogeneity criterion: $\tau^{min} = 1.20 < \tau^{max} = 1.43 < \tau^{cr,2}_{0.05}(f = 9) = 2.237$ $< \tau^{cr,1}_{0.05}(f = 9) = 2.392;$ Wilk-Shapiro normality test: $W = 0.872 > W_{0.05}^{cr}(9) = 0.829,$ David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_1) = 2.59 < U =$ $[(A_1^{max} - A_1^{min})/S_{A1}] = 2.63 < U2_{0.05}^{cr}(N_1) = 3.552;$

set statistics of L_1 :

 $N_1 = 9, \ L_1^{av} = 3.79 \pm 0.32; 95\%$ confidence interval is: 3.05 - 4.53; $L_1^{min} = 2.06, \ L_1^{max} = 4.96, \ S_{L1} = 0.96, \ \tau^{max} = 1.22 < \tau^{min} = 1.80 < \tau^{cr,2}_{0.05}(f = 9) = 2.237 < \tau^{cr,1}_{0.05}(f = 9) = 2.392;$ Wilk-Shapiro normality test: $W = 0.940 > W_{0.05}^{cr}(9) = 0.829,$ David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_1) = 2.59 < U = [(L_1^{max} - L_1^{min})/S_{L1}] = 3.01 < U2_{0.05}^{cr}(N_1) = 3.552;$ (11)

set statistics of $A_{3\kappa cn2}$ (Nos. 10–15):

 $N_2 = 6, \ A_2^{av} = 0.48 \pm 0.12; \ 95\% \text{ confidence interval}$ is: 0.134 - 0.165; $A_2^{min} = 0.20, \ A_2^{max} = 1.00, \ S_{A2} = 0.293, \ \tau^{min} = 0.96 < \tau^{max}_{0.05} = 1.77 < \tau^{cr,2}_{0.05}(f = 6)$ = 1.996 < $\tau^{cr,1}_{0.05}(f = 6) = 2.184;$ Wilk-Shapiro normality test: $W = 0.836 > W^{cr}_{0.05}(6) = 0.788,$ David-Hartley-Pearson normality test: $U1_{0.05}^{cr}(N_2) =$ 2.200 < $U = [(A_2^{max} - A_2^{min})/S_{A2}] = 2.73 <$ $U2_{0.05}^{cr}(N_2) = 3.012;$

set statistics of L_2 (Nos.10 – 15):

 $N_{2} = 6, \ L_{2}^{av} = 5.89 \pm 0.20; \ 95\% \text{ confidence interval}$ is: 5.37 - 6.41; $L_{2}^{\min} = 5.11, \ L_{2}^{\max} = 6.5, \ S_{L2} = 0.5, \ \tau^{\max} = 1.22 < \tau^{\min} = 1.56 < \tau^{cr,2}_{0.05}(f = 6) = 1.996 < \tau^{cr,1}_{0.05}(f = 6) = 2.184; \text{Wilk-Shapiro normality test:}$ $W = 0.937 > W_{0.05}^{cr}(6) = 0.778, \ David-Hartley Pearson normality test: <math>U1_{0.05}^{cr}(N_{2}) = 2.200 < U$ = $[(L_{2}^{\max} - L_{2}^{\min})/S_{L2}] = 2.78 < U2_{0.05}^{cr}(N_{2}) = 3.012.$ (12)

Thus, samples (11) and (12) are homogeneous and normally distributed. Now let us compare the variances of samples for the molecular parameter L:

$$F = S_{L1}^2 / S_{L2}^2 = 3.69 <$$

$$F_{0.05}^{\rm cr}(f_1 = N_1 - 1; f_2 = N_2 - 1) = 4.82.$$
(13)

It follows from inequality (13) that the difference between the variances of the samples turns out to be insignificant. Therefore, to assess the significance of the difference in mean values, one can use the following inequality [17]:

$$|L_{1}^{av} - L_{2}^{av}| = 2.10 >$$

$$t_{0.05}^{cr}(f) \left\{ \frac{N[(N_{1} - 1)S_{L1}^{2} + (N_{2} - 1)S_{L2}^{2}]}{N_{1}N_{2}(N_{1} + N_{2} - 2)} \right\}^{0.5} = 0.93.$$
(14)

Here the number of degrees of freedom is $f = N_1 + N_2 - 2$. Inequality (14) indicates that the mean values of the samples differ significantly. Hence, the discrepancy between the samples is not accidental. Therefore, the idea of the threshold of action of drugs does not deviate.

Let us find out what is the possible non-linear relationship between the resulting indicator and the explanatory variable. For example, nonlinear dependence can be approximated by a polynomial. Let's find out what degree the polynomial is preferable to use. The essence of the method for determining the degree of a polynomial is the sequential calculation of the differences. For a ranked series, the first, second, third, etc. level differences are calculated [18, 19].

$$\Delta_N^{(1)} = (A_{\exp})_N - (A_{\exp})_{N-1}, \quad \Delta_N^{(2)} = \Delta_N^{(1)} - \Delta_{N-1}^{(1)},$$
$$\Delta_N^{(3)} = \Delta_N^{(2)} - \Delta_{N-1}^{(2)}, \dots$$
(15)

The calculation of these differences continues until the differences become approximately close in magnitude. Table 2 shows the calculated level differences for the first group of molecules Nos. 1-9. If the assumed dependence is linear, then the differences for equally spaced neighboring values should be the same. The order of the polynomial is determined by comparing the residual differences. For the second level, the increments have approximately close values (Table 2). It is generally accepted that in this case the order determines the degree of the polynomial. Thus, the given empirical series can be approximated by a polynomial of the second degree. For empirical series, a strict coincidence of the values of all increments in any order is guite rare. Additional information on the possible analytical form of the relationship between bioactivity and an explanatory attribute can be obtained if the regression residuals are checked for linear regression (first group: Nos. 1-9). The residuals of linear regression were analyzed:

$$A_{\rm mod1}(L) = b_{0,+} + b_1 \cdot L, \tag{16}$$

N = 9, correlation coefficient: $R = -0.78 \pm 0.24$, |R|> $R_{0.05}^{cr}(f = 7) = 0.666$, corrected correlation coefficient: $|R^*| = |R| \cdot [1 + 0.5 \cdot (1 - R^2)/(N - 3)] =$ 0.805 (for N < 20); sample size sufficient for the validity of the correlation coefficient: $N_{0.05}^{min} = 6$; RMSE = 1.1355; $b_0 = 12.94 \pm 1.63$, $b_1 = -1.37 \pm$ 0.42, $|t(b_0)| = 7.96 > |t(b_1)| = 3.30 > t_{0.05}^{cr}(f = 7) =$ 2.365; significance of the coefficient of determination: $F = 10.86 > F_{0.05}^{cr}(f_1 = 1; f_2 = 7) =$ 5.59; $\Sigma_1 = 9.0263$; AIC₁ = 0.3756; SC₁ = 0.4912, SS₁ = 0.3755.

Regression residuals are normally distributed: $W = 0.978 > W_{0.05}^{cr}(N=9) = 0.829$. AIC and SC are informational criteria of quality for linear regression of Akaike (AIC) and Schwarz (SC) [20,21]:

AIC =
$$2m/N + \ln(\Sigma/N)$$
,
SC = $(m + 1)\ln(N)/N + \ln(\Sigma/N)$,
SS = $\Sigma^{0.5}/(N - m)$. (17)

Here *m* is the number of explanatory variables; Σ is the sum of the squares of the regression residuals; *N* is number of observations. It also suggests using the

SS relation, which is conjugated with the AIC and SC tests.

N	R^4	L,	$A_{\rm exp}$,	$\Delta^{(1)}$	$\Delta^{(2)}$	$\Delta^{(3)}$	$\Delta^{(4)}$	$\Delta^{(5)}$
		arb. unit.	points					
1	Н	2.06	8.5	-	-	-	-	-
2	NH ₂	2.93	9.0	0.5	-	-	-	-
3	NO_2	3.44	9.0	0	-0.5	-	-	-
4	$N(CH_3)_2$	3.53	10.0	1.0	1.0	1.5	-	-
5	NHCH ₃	3.53	8.5	-1.5	-2.5	-3.5	-5.0	-
6	Ι	4.23	7.5	-1.0	0.5	3.0	6.5	11.5
7	$N(C_2H_5)_2$	4.50	6.0	-0.5	0.5	0	-3.0	-9.5
8	NHC ₂ H ₅	4.96	6.0	0	0.5	0	0	3.0
9	NH(CH ₂) ₂ Cl	4.96	6.0	0	0	-0.5	-0.5	-0.5

Table 2. Calculation of residual differences $\Delta_N^{(i)}$.

For small sample sizes (N < 50), the significance of the correlation coefficient is also estimated using the Fisher normalizing *z*-transform (taking into account Hotelling's corrections) [15]):

$$u_{\rm H} = z - (3z + R)/4/N = 0.985 >$$

$$z_{0.975}(N - 1)^{-0.5} = 0.693,$$

$$z = 0.5 \cdot \ln((1 + R)/(1 - R)). \tag{18}$$

Here $z_{\alpha} = 1.960$ is the quantile of the normal distribution at the significance level $\alpha = 0.05$. At the beginning of the studied series of compounds, the regression residuals (16) are positive, with a decreasing trend, and at the end of the series, they are negative with an increasing trend. This alternation of residues indicates that the relationship between bioactivity and factor *L* has the character of a non-linear relationship. We will use a second order polynomial to reveal the quantitative relationship (the first group of chemical compounds) of the structure of the molecule and the variation of its bioactivity:

$$A_{\text{mod}1}(L) = b_0 + b_1 \cdot L + b_2 \cdot L^2.$$
(19)

Function (19) is linear with respect to the regression parameters and non-linear with respect to the explanatory variable *L*. Equation (19) can be linearized. To do this, it is enough to make the following replacement $LL \equiv L^2$.

The sets of the explanatory variables L and the resulting indicator A are homogeneous and satisfy the normal distribution (11). Thus, instead of

nonlinear regression in the explanatory variable, we obtain a two-factor linear regression:

$$A_{\text{MOJ.1}}(L) = b_0 + b_1 \cdot L + b_2 \cdot LL.$$
(20)

The statistics of regression (20) are as follow:

N = 9; multiple correlation coefficient: $R = 0.92 > R_{0.05}^{cr}(v = 6; m = 2) = 0.795$, $R^2 = 0.846$, $R^{*2} = 0.795$; RMSE = 0.693; $b_0 = 2.22 \pm 3.45$, $b_1 = 4.80 \pm 1.96$, $b_2 = -0.83 \pm 0.27$; $|t(b_2)| = 3.10 > t(b_1) = 2.45 > t_{0.05}^{cr}(f = 6) = 2.447$; significance of the multiple determination coefficient: $F = 16.26 > F_{0.05}^{cr}(f_1 = m; f_2 = N - m - 1) = 5.14$; $\Sigma_2 = 2.8811$; AIC₂ = - 0.6946; SC₂ = - 0.4066, SS₂ = 0.2425.

Here m = 2 is the number of explanatory variables. The regression residuals are normally distributed: $W = 0.949 > W_{0.05}^{cr}(N = 9) = 0.829$. Information tests by Akaike and Schwarz, as well as the SS ratio, indicate an improvement in the quality of regression (20) compared to regression (16). Regression coefficients b_1 and b_2 are statistically significantly different from zero. The following standardized values of the regression coefficients were obtained:

$$b_1^* = S_L b_1 / S_{A1} = 3.32$$
, $b_2^* = S_{LL} b_2 / S_{A1} = -4.12$. (21)

Here, the standard deviations of the bioresponse and the explanatory variables *L* and *LL* are used: $S_{A1} = 1.52$, $S_L = 0.96$, $S_{LL} = 6.97$.

Statistics for a set of *LL* parameters:

 $N_{1} = 9; \ LL_{1}^{av} = 15.21 \pm 2.36; \ 95\% \text{ confidence}$ interval: 9.78 - 20.65; $LL_{1}^{\min} = 4.243, \ LL_{1}^{\max} = 24.60, \ S_{LL} = 7.07, \ \tau^{\max} = 1.33 < \tau^{\min} = 1.55 < \tau_{0.05}^{cr,2}(f = 9) = 2.237 < \tau_{0.05}^{cr,1}(f = 9) = 2.392; \text{ Wilk-Shapiro normality test: } W = 0.938 > W_{0.05}^{cr}(9) = 0.829, \ David-Hartley-Pearson normality test:$ $<math>U1_{0.05}^{cr}(N_{1}) = 2.59 < U = [(LL_{1}^{\max} - LL_{1}^{\min})/S_{LL}] = 2.88 < U2_{0.05}^{cr}(N_{1}) = 3.552.$ (22)

Regression coefficients are statistically significantly different from zero at the 95% confidence level. The standardized regression coefficients are dimensionally independent and therefore allow comparison of the contributions that explanatory variables make to the regression. In statistics (20), the corrected coefficient of determination is used, which is valid for small samples $N \le 15$:

$$R^{*2} = 1 - (1 - R^2) \cdot (N - 1)/(N - m),$$
(23)

as well as the critical value of the multiple correlation coefficient $R_{0.95}^{cr}(\nu = 6; m = 2)$ [22]. The null hypothesis that the multiple correlation coefficient is equal to zero is rejected at the significance level $\alpha = 0.05$.

The cubic regression test showed the statistical insignificance of the regression equation. All regression coefficients were found to be insignificant at the 95% confidence level. The following estimates of the quality criteria of regressions were also obtained: $\Sigma_3 = 2.9670$, AIC₃ = -0.4430, SC₃ = -0.1331, SS₃ = 0.2871. These tests indicate a decrease in the quality of cubic regression compared to quadratic regression (19). Checking the relationship of the regression residuals (19) with the explanatory variable L^3 also led to an insignificant result. Thus, the performed statistical analysis, as well as the AIC, SC and SS tests, do not contradict the difference method (Table 2). Consequently, preference is given to the analytical form of regression, which corresponds to the minimum of tests.

Checking the collinearity between the variables L and LL showed that the correlation coefficient between them is 0.994 (close relationship). However, this collinearity can be eliminated. Let's introduce new variables: $(L - L^{av})$ and $(L - L^{av})^2$ [23]. Correlation coefficient decreases significantly for new variables: $r_{1,2} = -0.31$. The regression will be as follows:

$$A_{\rm mod}(L) = a_0 + a_1 \cdot (L - L^{\rm av}) + a_2 \cdot (L - L^{\rm av})^2,$$
(24)

$$\begin{split} N_1 &= 9; \ R = 0.92 > R_{0.05}^{\rm cr}(\nu = 6; m = 2) = 0.795, \ R^2 = \\ 0.846, \ R^{*2} &= 0.80, \ RMSE = 0.693; \ a_0 = 8.52 \pm 0.27, \\ a_1 &= -1.48 \pm 0.27, \ a_2 = -0.83 \pm 0.27; \ t(a_0) = 26.09 > \\ |t(a_1)| &= 5.51, \ t(a_2)| = 3.10 > t_{0.05}^{\rm cr}(f = 6) = 2.447; \ F \\ &= 16.26 > F_{0.05}^{\rm cr}(f_1 = 2; f_2 = 6) = 5.14. \end{split}$$

The statistics of the sets $\Delta L = L - L^{av}$ and $\Delta L^2 = (L - L^{av})^2$ will be as follows:

 $N_{1} = 9; \ \Delta L^{\text{av}} = 0.00 \pm 0.32; \ 95\% \text{ confidence}$ interval: (-0.74, 0.74); $\Delta L^{\text{min}} = -1.73, \ \Delta L^{\text{max}} = 1.17, \ S_{\Delta L} = 0.96; \ \tau^{\text{max}} = 1.21 < \tau^{\text{min}} = 1.80 < \ \tau^{\text{cr},2}_{0.05}(f = 9)$ = 2.237 < $\tau^{\text{cr},1}_{0.05}(f = 9) = 2.392;$ Wilk-Shapiro normality test: $W = 0.940 > W^{\text{cr}}_{0.05}(9) = 0.829,$

 $N_1 = 9; \Delta L^{2,av} = 0.83 \pm 0.32; 95\%$ confidence interval: (0.08–1.57), $\Delta L^{2,min} = 0.0693, \Delta L^{2,max} =$ 2.004, $S_{\Delta L2} = 0.97, \tau^{min} = 0.78 < \tau^{max} = 1.21 <$ $\tau^{cr,2}_{0.05}(f = 9) = 2.237 < \tau^{cr,1}_{0.05}(f = 9) = 2.392;$ Wilk-Shapiro normality test: $W = 0.800 \approx W^{cr}_{0.05}(9) =$ 0.829. (25)

The standardized regression coefficients are:

$$a_1^* = -0.94$$
, $a_2^* = -0.53$. (26)

It is important to note that in absolute value the coefficients (21) are opposite to the coefficients (26). This indicates that the values of the standardized regression coefficients are not always associated with the values of the relative contribution of the explanatory variables to the variability of the resulted indicator. We can make an estimate of the relative contributions of the variability of an effective trait, given the magnitudes of the correlation coefficients between the explanatory variables and the observed bioactivity:

$$R_{\text{appr.}^{2}} = a_{1}^{*} \cdot r_{\Delta L,\text{Act}} + a_{2}^{*} \cdot r_{\Delta L2,\text{Act}} = 0.720 + 0.123 = 0.843.$$
(27)

Here $r_{\Delta L,Act} = -0.77$ and $r_{\Delta L2,Act} = -0.23$ are the pair correlation coefficients. From relation (27) it follows that the greatest contribution to the variability of the effective factor comes from the explanatory variable ΔL . We also note that the approximate coefficient of determination (27) is very close to the coefficient of determination (24).

Comparing the results for regressions (24) and (20), it is easy to verify their complete identity. Obviously, the regression coefficients a_i and b_j are interrelated. For example, there is the following relationship: $b_1 = a_1 - 2a_2L^{av}$. Regression residuals (24) are normally distributed (Wilk-Shapiro test: W= 0.961 > $W_{0.05}^{cr}(9) = 0.829$). Therefore, a quantitative assessment of collinearity between explanatory variables can be carried out using the Farrar-Glauber [24] relation, which has a *chi-square* distribution for f = m(m - 1)/2:

$$\chi^{2} = -[N - 1 - (2m + 5)/6] \cdot \ln(1 - r_{1,2}^{2}) = 0.66 <$$
$$\chi^{2, \, cr}_{0.05}(f = 1) = 3.841.$$
(28)

It follows from inequality (28) that the new explanatory variables $(L - L^{av})$ and $(L - L^{av})^2$ are not collinear. However, the results of the two regressions (20) and (24) completely coincide.

Consider now a regression that does not take into account the linear term *L*:

$$A_{\text{mod}}(L) = b_0 + b_1 \cdot LL.$$
⁽²⁹⁾

For this regression the following statistic was obtained:

N = 9; $R_1 = -0.83 \pm 0.21$, $m_1 = 1$ is the number of explanatory variables; corrected correlation coefficient: $|R_1^*| = 0.85 > R_{0.05}^{cr}(7) = 0.666$; the minimum sample size, sufficient for the reliability of the correlation coefficient: $N_{0.05}^{\min} = 6$; a criterion for the significance of the correlation coefficient based on the Fisher normalizing z-transform (taking into account Hotelling's corrections): $u_{\rm H} = 1.151 >$ $u_{0.05}(N) = z_{0.975} \cdot (N-1)^{-0.5} = 0.693; RMSE = 0.908;$ $b_0 = 10.54 \pm 0.75, b_1 = -0.17 \pm 0.05, t(b_0) = 13.99 >$ $|t(b_1)| = 3.93 > t_{0.05}^{cr}(f = 7) = 2.365; F = 16.46 >$ $F_{0.05}^{\rm cr}(f_1 = 1; f_2 = 7) = 5.59; \ \Sigma_4 = 6.7506, \ AIC_4 =$ -0.0653, SC₄ = 0.2007, SS₄ = 0.3248. (30)

The regression residuals are normally distributed: $W = 0.903 > W_{0.05}^{cr}(N=9) = 0.829$. Correlation coefficients between the values of information tests: $r_{AIC-SS} = 0.997 \text{ M} r_{SC-SS} = 0.994 (R_{0.05}^{cr}(4) = 0.811)$. According to the AIC, SC and SS tests, the quality of the regression (30) is lower than the quality of the regression (20). That is, the relationship between tests is statistically significant. The sample uses data: Σ_i , AIC_i, SC_i, SS_i for i = 1,2,3,4. According to the AIC, SC and SS tests, the quality of the regression (30) is lower than the quality of the regression (30) is low regression (20).

Let us check quantitatively whether the joint explanatory variables L and LL have a significant effect on the variability of bioactivity. To do this, we'll use the F statistics:

$$F = \frac{(R^{*2} - R_1^2)(N - m - 1)}{(m - m_1)(1 - R^{*2})} = 10.97 >$$

$$F_{0.05}^{\rm cr}(f_1 = m - m_1; f_2 = N - m - 1) = 5.99.$$
(31)

It follows from inequality (31) that the additional explanatory variable *L* in regression (20), together with the variable *LL*, has a significant effect on the variation of the bio-response. Therefore, in what follows we will use the regression (20). The range of applicability of equation (20) is determined by the inequality $L \le L^{\text{thr}}$. For the threshold value of the explanatory variable, we take the following value: $L^{\text{thr}} \approx 5.0$ (arbitrary units). Near the threshold, there is a sharp decrease (or increase) in the activity of 4-substituted 5-nitro-2-methoxy-N-(2-diethylamino-ethyl) benzamides (Fig. 2).

The effect of the hydrophobic properties of the substituents on bioactivity was also tested. Probably, it was not possible to obtain a statistically significant regression dependence, due to the small size of the sample. However, it was found that for chemical compounds with close values of the linear size *L* (Nos, 2-5), the effect of hydrophobic contributions of substituents on bioactivity is noted. So, for example, for chemical compound No. 4, which has the maximum bioactivity (substituent N(CH₃)₂), the hydrophobic contribution is positive and equal to $\pi = 0.18$ arb. units [25].

The hydrophobic regions of molecules are usually not polarized and are unable to participate in the formation of hydrogen bonds. Hydrophobic substituents include hydrocarbon radicals, for example, N(C₆H₅)₂, CH₂C₆H₅, OC₆H₅, Si(CH₃)₃. At the same time, chemical compounds Nos. 2, 3 and 5, the bioactivities of which are in the region of maximum, but below the maximum (No. 4) bioactivity $A_1 = 10$ points, have substituents for which the hydrophobicity is negative (i.e., they are hydrophilic) and is equal, respectively: -1.23, -0.28 μ - 0.47 arb. units.



Fig. 2. Relationship between the observed values of the bioactivity of benzamides with the linear size L of the substituent. 1 – The first group of compounds (Nos. 1–9). 2 – The second group of compounds (Nos. 10–15). The dotted line indicates a discontinuity in the regression line associated with the threshold phenomenon (cutoff value $L^{\text{thr}} \approx 5.0$ arb. units).

Strong hydrophilic substituents include polar groups of atoms, for example, COCH₃, NHOH, SO₂CH₃. The data of Table 1 suggest that the bioactivity of the drug may depend on whether the region of the R⁴ substituent is hydrophobic (hydrocarbon) or hydrophilic (polar). However, the linear size L may also be a limiting factor. In addition, it should be noted that, in the series of substituents Nos. 1-9, the N(CH₃)₂ (No.1) substituent has the lowest known electron affinity energy of the analyzed substituents: A = 1.08 eV[26]. For example, the substituents NH_2 (No. 2), NO₂ (No.3) and NHCH₃ (No. 4) adjacent to the substituent N(CH₃)₂ in Table 1 have a higher electron affinity of 1.12, 3.10 and 1.56 eV, respectively. It is known, that the affinity for an electron characterizes the oxidizing ability of an atom or a group of atoms and determines the amount of energy released when an electron is attached. The greater the electron affinity, the higher the oxidizing properties of the substituent. For a hydrogen atom, the electron affinity is lower (0.754)eV) than for the N(CH₃)₂ substituent. However, bioactivity can be limited (see equation (19)) by the linear size of the hydrogen atom. The linear size of the hydrogen atom is noticeably less than the optimal size of 2.89 arb. units (see below).

For the second group (Nos.10-15) chemical compounds (weakly bioactive drugs; range of values $L > L^{\text{thr}} \approx 5$ arb. units), the relationship between the variation in bioactivity and changes in the molecular structure can be approximated by a linear relationship (Fig. 2):

$$A_{\text{mod }2}(L) = a_0 + a_1 \cdot L, \quad N = 6, R = -0.71 \pm 0.35,$$

$$|R^*| = 0.77 \approx R_{0.05}^{cr}(4) = 0.811, RMSE = 0.23, a_0$$

= 2.94 ± 1.22, $a_1 = -0.42 \pm 0.21.$ (32)

In accordance with the Chaddock scale, the corrected correlation coefficient falls within the range of $0.7 < R^* < 0.9$. This relationship between bioactivity and the explanatory variable can be characterized as "tight". Thus, there is a structural shift here when comparing the relationship of bioactivity with an explanatory molecular factor for bioactive and inactive drugs. It turned out that the relationship between the effective factor and the explanatory factor *L* for areas 1 and 2 is either significantly different, or for area 2 there is no relationship at all.

The nonlinear relationship determined by the regression equation (19) (or (20)) suggests that a chemical compound with the maximum activity should exist in the studied series of compounds. We can determine the optimal value of the linear size $L_{\text{opt}} = -b_1/(2b_2) = 2.89$ (in arbitrary units) from the condition $dA_{\text{mod.1}}(L)/dL = 0$, at which the regression line (19) reaches its maximum (Fig. 2). From equation (24) we obtain the same value: $L_{opt} = L^{av}$ $a_1/(2a_2) = 2.89$ (in arbitrary units). These results are consistent with the assumption of the optimal spatial correspondence of the substituent in the lowmolecular-weight chemical compound and the region of interaction of the biological object. On the one hand, mathematical modeling demonstrates a statistically significant relationship between the antiemetic activity of benzamides and the linear size L of the substituent. On the other hand, modeling allows us to point out an important result, namely, that one can hardly expect to obtain a drug whose activity is significantly higher than that of the dimetpromide molecule by varying only the linear size of the substituent.

Figure 3 shows a very close agreement between the observed values and those calculated using models (19) and (32). Thus, we can accept the hypothesis that the variation in the bioactivity of drugs is due to the variability of the linear size L of the substituent.

Apparently, the necessary condition for the manifestation of the bioactivity of the benzamide molecule is the complementarity of the chemical compound to the local area of the biological object. However, for the manifestation of the bioactivity of the molecule, it is not enough only the structural correspondence of the drug to the localization area. It is also necessary that the benzamide molecule has a stable interaction with the local site of the biological object.



Model, points

Fig. 3. Correlation field and scatter diagram of indicators of the observed antiemetic activity of benzamides and calculated values of bioactivity according to equations (19) and (32). — regression line: $A_{\text{mod}} = a_0 + a_1 \cdot A_{\text{exp}}$, N = 15; $R = 0.99 \pm 0.03$, RMSE = 0.489, $a_0 = 0.03 \pm 0.21$, $a_1 = 1.00 \pm 0.03$; $F = 878.5 > F_{0.05}^{\text{cr}}(f_1 = 1; f_2 = 13) = 4.67$.

That is, paired intermolecular interactions should play an important role in the manifestation of antiemetic activity. For a quantitative assessment of possible intermolecular interactions, we will use the electronic factor-attribute Z (or information factorattribute H, which correlates with the value of Z [27, 28]). Molecular information function is:

$$H = -\sum_{i} p_i \log_2 p_i , \qquad (33)$$

For relations $p_i = n_i / N$ the following conditions are fulfilled: $0 \le p_i \le 1$, $\sum p_i = 1$, where $p_i = 0$ means the impossibility of occurrence of the *i*-th event; $\sum n_i = N$, n_i is the number of atoms of type *i* in a molecule, N is the total number of atoms in a molecule. The ratio n_i / N determines the fractional participation of the *i*-th sort of atom in the molecule. In fact, here, instead of the probabilities used by Shannon, we use in formula (33) Kolmogorov's combinatorial representation of the fractional participation of the elements of a set in the general set, as applied to molecules (the elements of the set are atoms) [28]. Factor Z is associated with the pseudopotential of the molecule. Information factor H and electronic factor Z turned out to be useful in quantitative determination the of paired intermolecular interactions [9.27].

Let us check whether the mean values of the information function differ significantly for bioactive and inactive chemical compounds. The following statistics were obtained for the information functions H_1 and H_2 :

 $N_1 = 9$; $H_1^{\text{av}} = 1.65 \pm 0.03$; 95% confidence interval: (1.60-1.71), $H_1^{\text{min}} = 1.55$, $H_1^{\text{max}} = 1.76$, $S_{H1} = 0.074$, $\tau^{\min} = 1.415 < \tau^{\max} = 1.442 < \tau^{cr,2}_{0.05}(f=9) = 2.237 < \tau^{cr,1}_{0.05}(f=9) = 2.392; \text{ Wilk-Shapiro normality test:}$ $W = 0.924 > W^{cr}_{0.05}(9) = 0.829, \text{ David-Hartley-Pearson normality test:} U1^{cr}_{0.05}(N_1) = 2.59 < U = [(H_1^{\max} - H_1^{\min})/S_{H1}] = 2.84 < U2^{cr}_{0.05}(N_1) = 3.552;$ (34)

$$\begin{split} N_2 &= 6; \, H_2^{\rm av} = 1.56 \pm 0.02; \, 95\% \text{ confidence interval:} \\ 1.51-1.62, \, H_2^{\rm min} = 1.48, \, H_2^{\rm max} = 1.61, \, S_{H2} = 0.053, \\ \tau^{\rm max} &= 0.94 < \tau^{\rm min} = 1.51 < \tau^{\rm cr,2}_{0.05} (f=6) = 1.906 < \\ \tau^{\rm cr,1}_{0.05} (f=6) &= 2.184; \, \text{Wilk-Shapiro normality test:} \\ W &= 0.859 > W^{\rm cr}_{0.05} (6) = 0.788, \, \text{David-Hartley-} \\ \text{Pearson normality test:} \, U1^{\rm cr}_{0.05} (N_2) = 2.200 < U = \\ [(H_2^{\rm max} - H_2^{\rm min})/S_{H2}] = 2.45 < U2^{\rm cr}_{0.05} (N_2) = 3.222. \\ (35) \end{split}$$

The variances of populations do not differ significantly:

$$F = S_{H1}^2 / S_{H2}^2 = 1.92 <$$

$$F_{0.05}^{\rm cr}(f_1 = N_1 - 1 = 8; f_2 = N_2 - 1 = 5) = 4.82,$$
(36)

Therefore, the average values of H_1^{av} and H_2^{av} can be compared using relation (14):

$$t = |H_1^{av} - H_2^{av}| = 0.094 > t^{av} =$$

$$t_{0.05}^{cr} (f = N_1 + N_2 - 2) \times$$

$$\frac{N \cdot [(N_1 - 1) \cdot S_1^2 + (N_2 - 1) \cdot S_2^2]}{N_1 \cdot N_2 \cdot (N_1 + N_2 - 2)} \bigg\}^{0.5} = 0.062.$$
(37)

The one-sided *t*-test is applied: $t_{0.05}^{cr}(f=13) = 1.77$. Inequality (37) indicates that the mean values for bioactive chemical compounds and low-active drugs differ significantly. Inequality (37) is also preserved for the two-sided *t* - test. It can be assumed that bioactive chemical compounds have such values of the molecular trait *H*, which are preferably localized around the average value $H_1^{av} = 1.65$ bits. Whereas for inactive (or weakly active) drugs, most likely, the values of factor *H* are localized around the value $H_2^{av} = 1.56$ bits (Fig. 4). Figure 4 demonstrates the threshold of the dependence of the bioactivity of molecules on the value of the parameter *H*.

Figure (4) demonstrates the threshold of the dependence of the bioactivity of molecules on the value of the information function *H*. To the right of the approximate threshold value ($H^{\text{thr}} \approx 1.61$ bits),

there is a rapid (jump) increase in the bioactivity of chemical compounds. It is important to note here that the molecular parameter L refers only to the substituent, while the information function H characterizes the molecule as a whole.



Fig. 4. Scatter plot. Δ – bioactive drugs (first group). • – inactive or weakly active drugs (second group). The solid line is the regression: $A_{\text{mod.3}}(H) = b_0 + b_1 \cdot H$, N = 15; $R^* = 0.62 > R_{0.05}^{\text{cr}}(f = 13) = 0.514$; $b_0 = -42.71 \pm 17.54$, $b_1 = 29.45 \pm 10.84$, $t(b_1) = 2.72 > |t(b_0)| = 2.44 > t_{0.05}^{\text{cr}}(f = 13) = 2.160$; $F = 7.38 > F_{0.05}^{\text{cr}}(f_1 = 1; f_2 = 13) = 4.67$. The dotted line indicates the approximate threshold value of the molecular information function.

In a similar way, the following inequalities were obtained for the molecular factor *Z*:

$$F = S_{Z1}^2 / S_{Z2}^2 = 2.04 <$$

$$F_{0.05}^{cr}(f_1 = N_1 - 1 = 8; f_2 = N_2 - 1 = 5) = 4.82,$$

$$t = |Z_1^{av} - Z_2^{av}| = 0.13 > t^{av} = 0.10.$$
(38)

Here $S_{Z1} = 0.1214$, $S_{Z2} = 0.0850$, $Z_1^{av} = 2.77$, $Z_2^{av} = 2.64$, $N_1 = 9$, $N_2 = 6$. Hence, it follows that bioactive compounds are grouped around Z_1^{av} , and weakly active ones around Z_2^{av} . The relationship between factor Z and antiemetic activity has a threshold similar to Fig. 4. At the 95% confidence level, the mean values of factor Z are statistically significantly different for active and inactive drugs: $Z_1^{av} > Z_2^{av}$.

It should be noted that the electronic factor Z of a chemical compound and the information function H are determined based on different concepts; nevertheless, they turned out to be highly correlated. For chemical compounds (Table 1), molecular factors Z and H are "very closely" (on the Chaddock scale) interconnected with each other (Fig. 5).



Fig. 5. Relationship between molecular factors Z and H. $\underline{H}(Z) = b_0 + b_1 \cdot Z$, N = 15; $R = 0.96 \pm 0.08$, $R^* = 0.963 > R_{0,05}^{cr}(f = N - m - 1) = 0.514$; sample size sufficient for the validity of the correlation coefficient: $N_{0.05}^{min} < 5$; RMSE = 0.037, $b_0 = -0.06 \pm$ 0.14, $b_1 = 0.62 \pm 0.05$, $t(b_1) = 12.15 > t_{0.05}^{cr}(13) =$ 2.160; $F = 147.6 > F_{0.05}^{cr}(f_1 = 1; f_2 = 13) = 4.67$.

Thus, the level of antiemetic activity can be influenced not only by the linear dimensions of the substituent, but also by the value of the molecular parameter Z (or H), which correlates with the energy of pair intermolecular interaction [9]. Statistical analysis also shows that the bioactivity of benzamides will be the greater, the greater the value of the factor Z (or H).

That is, it can be assumed that the pairwise intermolecular interaction for bioactive molecules is on average more intense and, therefore, the activity index (in points) may be the higher, the greater the value of the explanatory factor.

4 Conclusion

The statistical approach proposed in this article establishes rules for separating bioactive chemicals from inactive chemicals. These rules can be practically useful in preliminary prediction of the antiemetic activity of new chemical compounds. It is important to emphasize that simple calculations of molecular traits require only knowledge of the structural formula of the molecule. When predicting possible new chemical compounds in the series of benzamides, the substituent at position R⁴ should have not only L close to $L_{opt} \approx 2.89$ arb. units, but the molecule must have a high value of the factor attribute $Z \approx Z_1^{av} = 2.77$ (in arb. units) > Z_2^{av} . The complementarity of a substituent is a necessary, but, apparently, not a sufficient condition for the manifestation of a chemical compound of its biological activity. That is, the intensity of the effect

of the pseudopotential of the molecule on the region of the biophase, which is characterized by the value of the molecular characteristic Z, is also important.

References:

- [1] M.L. Gershanovich, M.D. Paykin, *Symptomatic Treatment for Malignant Neoplasms*, Moscow, 1986 (in Russian).
- [2] B.M.Rakhmilevich, L.P. Simbirtseva, Yu. L. Varshamov, *Medical Radiology and Radiation Safety*, No. 1, 1983, pp.16-18 (in Russian).
- [3] Copyright Certificate 1022360 USSR. Discovery, 1983, No. 48 (in Russian).
- [4] K. Schulze-Debrein, *Gastroenterology*, Vol.77, 1979, pp.768-779.
- [5] V.I.Legeza, M.G.Shagoyan, K.S.Martirosov. *Pharmacology and Toxicology*, No.5, 1984, pp.55-59 (in Russian).
- [6] V.K.Mukhomorov, G.K.Semenova, M.G. Shagoyan, *Chemical and Pharmaceutical Journal*, No.9, 1988, pp.1108-1113 (in Russian).
- [7] G. Paunatier, L. Anker, A Theoretical Conformational Study of Substituted o-Anisamides as Models of a Class of Dopamine Antagonists, J. Pharm. Pharmacol., Vol.33, No.1, 1981, pp.145-149.
- [8] G.A. Schembelov, V.M. Ustynyuk, V.M. Mamaev, *Quantum-Chemical Methods for Calculating Molecules*, Moscow State University, 1980.
- [9] V.K.Mukhomorov, An Alternative Model of the Relationship Between the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons and their Molecular Structure, *Journal of Biological & Scientific Opinion*, Vol.6, No. 3, 2018, pp. 58-65.
- [10] V.K. Mukhomorov, Interrelation of Carcinogenic Properties of Metabolites with their Information and Electronic Descriptors, *SCIREA Journal of Biology*, Vol.3, 2018, pp.1-14.
- [11] A.Verloop, W.Hoogenstraaten, J.Tipker, Development and Application of New Steric Substituent Parameters in Drug Design. In: *Drug Design*. Ed. F.J. Ariens, Acad. Press, New-York, 1976, Vol.7, pp.165-207.
- [12] A.I. Kobzar, *Applied Mathematical Statistics*. *For Engineers and Scientists*, M., Fizmatlit, 2016. (in Russian).
- [13] H.A.Sturges, J. Amer. Statist. Assoc., Vol. 21, 1926, pp.65-66.
- [14] E.I.Pustyl'nik, *Statistical Methods of Analysis* and Processing of Observations, M., Nauka, 1968. (in Russian).

- [15] L.Sachs, Statistische Auswertungsmetoden. Springer-Verlag, Berlin, Heidelberg, New-York, 1972.
- [16] G.C.Chow, Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, Vol. 28, 1960, pp.591-605.
- [17] E.Förster, B.Rönz, Metohden der Korrelations- und Regressionsanalyse, Verlag Die Wirtschaft, Berlin, 1979.
- [18] G.M.Gambarov, N.M.Zhuravel, Yu.G. Korolev, *Statistical Modeling and Forecasting*. M., Finance and Statistics, 1990 (in Russian).
- [19] V.Yu.Urbakh, Statistical Analysis in Biological and Medical Research, M, Medicine, 1975 (in Russian).
- [20] Akaike Hirotogu, A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, Vol.19, No.6, 1974, pp.716-723.
- [21] G.Schwarz, Estimating the Dimension of Model, Annals of Statistics, Vol.6, 1978, pp.461-464.
- [22] Jiřî Likeš, Josef Laga, Základni Statisticke Tabulky, Praha, SNTL-Nakladatelstvi Technické Literatury, 1978.
- [23] F.Mosteller, J.W. Tukey, *Data Analysis and Regression. A Second Course in Statistics*, Addison-Wesley Publ. Comp., 1980.
- [24] D.E.Farrar, R.R.Glauber, Multicollinearity in Regression Analysis: The Problem Revisited, *The Review of Economics and Statistics*, Vol.49, No.1, 2018, pp.92-107.
- [25] C.Hansch, A.Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, John Wiley & Sons, New York, Chichester, Bribane, Toronto, 1979.
- [26] L.V.Gurevich, G.V.Karachevtsev, V.N. Kondrat'ev, Yu.A. .Lebedev, V.A. Medvedev, V.K. Potapov, Yu. S. Khodeev, *The Energies of Breaking Chemical Bonds. Ionization Potentials and Electron Affinity*, M., Science, 1974 (in Russian).
- [27] V.K.Mukhomorov, Statistical Aspects of the Interrelation Between the Biological Activity of Chemical Compounds and Their Molecular Structure, *Biomedical Statistics and Information*, Vol.1, No.3, 2016, pp.24-34.
- [28] A.N.Kolmogorov, *Information Theory and Theory of Algorithms*, M., Science, 1987. (in Russian).

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US