BioSystems 256 (2025) 105554

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Full Length Article

Simulation peptide toxicity using the fragments of local symmetry in amino acid sequences

Andrey A. Toropov [a,*], Alla P. Toropova [a], Valentin O. Kudyshkin [b], Emilio Benfenati [a], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156, Milan, Italy
[b] Institute of Polymer Chemistry and Physics, Academy of Sciences of the Republic of Uzbekistan, Kodyri street 7b, 100128, Tashkent, Uzbekistan
[c] Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, MS, 39217-0510, USA
[d] Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS, 39217, USA

ARTICLE INFO

Keywords:
Peptide cytotoxicity
Semi-correlations
Fragments of local symmetry
Monte Carlo method

ABSTRACT

An efficient scheme for modeling peptide toxicity is proposed and applied. The peptide cytotoxicity is calculated in the form of a mathematical function of its constituent amino acids, represented by single-symbol abbreviations. It was found that considering the so-called fragments of local symmetry can significantly increase the predictive potential of the proposed models. The best model gives for validation set value of Matthews correlation coefficient 0.7 on the set of 2784 peptides. The prospects for applying the ideas of infodynamics to research activities related to modeling the biochemical behavior of peptides are discussed.

1. Introduction

Peptides are short chains of amino acids that are the building blocks of proteins (proteins are longer chains of amino acids). The many versatile biological roles of peptides make them ideal candidates for therapeutic use (Fosgerau and Hoffmann, 2015; Craik et al., 2013). However, in addition to their therapeutic features, they could also exhibit toxicity.

Such characteristics have to be investigated and evaluated in order to assure safe medical applications of peptides.

Numerous methods have been developed in the past for predicting peptide cytotoxicity. However, there is a heuristic need to develop new, efficient methods that would assist in the accomplishment of such an important objective (Randic et al., 2000; Gozalbes and De Julián-Ortiz, 2018; Rathore et al., 2024). To accomplish this, one has to consider novel approaches. Construction of peptide toxicity models as a mathematical function of the list of amino acids composing the peptide provides a possible way to create such models (Toropov et al., 2012).

Information theory was used in defining measures of the topological properties of molecules. On this basis, general expressions for information content of molecules can be a useful tool to establish the correlation between molecular structure and physicochemical and biochemical behavior of corresponding substances (Bonchev and Trinajstic, 1977). However, now, in addition to the traditional concept of information, the concept of infodynamics is also used (Vopson and Lepadatu, 2022). Perhaps infodynamics will find applications for modeling physical-chemical and biochemical phenomena. Comparison of some provisions of infodynamics (Vopson, 2023, 2025) (e.g., the first and second laws of infodynamics) and facts regarding the stability of amino acids as components of peptides and biopolymers indicates the possibility of some contact between infodynamics and the theory/practice of peptides and biopolymers in the future.

2. Method

2.1. Data

Balanced data on cytotoxicity 11036 peptides (5518 toxic and 5518 non-toxic) were taken from the reference (Rathore et al., 2024). The available experimental data were randomly divided into a training set (≈75 %) and a validation set (V). The training set was structured into an active training set (A), a passive training set (P), and a calibration set

* Corresponding author. Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

https://doi.org/10.1016/j.biosystems.2025.105554
Received 8 May 2025; Received in revised form 1 August 2025; Accepted 7 August 2025
Available online 8 August 2025

(C). The logic of using such data partitioning in model construction is described in the literature (Toropova et al., 2018).

2.2. Semi correlation

The idea of semi-correlations is an attempt to extend the algorithms developed for regression models for the case when the modeled quantity takes two values: 1 for those classified as toxic and −1 for non-toxic. Here, semi-correlations are used to develop a categorical model of peptide toxicity. The CORAL software, 2024 (http://www.insilico.eu/coral), used in a current study, is an efficient tool for constructing semi-correlations. The descriptor (y) is the sum of the correlation weights of different amino acids.

$$y = C_0 + C_1 \times DCW(T,N) \quad (1)$$

$$CLASS(\text{Amino Acid List}) = \begin{cases} 1 \ \ (\text{active}), \text{if } y \geq 0.5 \\ -1 \ \ (\text{inactive}), \text{if } y < 0.5 \end{cases} \quad (2)$$

2.3. Descriptor

Two descriptors are considered in our study. The first is equal to the sum of the correlation weights of the amino acids that make up the peptide. The second is calculated using a more complex approach. It is equal to the sum of the correlation weights of the amino acids, supplemented by the sum of the correlation weights of the fragments of local symmetry:

$$DCW1(T,N) = \sum CW(A_k) \quad (3)$$

$$DCW2(T,N) = \sum CW(A_k) + \sum CW(FLS_j) \quad (4)$$

T is an integer that separates codes of amino acids into two classes: active (non-rare) and blocked (rare). N is the number of epochs of Monte Carlo optimization. A_k is the amino acid one symbol abbreviation. FLS is a so-called fragment of local symmetry.

It should be noted that the division into the four subsets affects the results of the optimization process, as different distributions influence the lists of active amino acids that directly impact the models, and the lists of rare amino acids that are ignored. The latter do not contribute to the resulting model. Choosing too large a value of T leads to an overly simple (primitive) model, while choosing too small a T allows rare amino acids to influence model construction, resulting in overfitting. A good compromise appears to be T = 3. The value N = 7 was chosen empirically, as no significant improvement was observed with larger N.

2.4. Fragments of local symmetry

Three classes of local symmetry are considered: XYX, XYYX, and XYZYX. Fig. 1 contains examples of the extraction of codes of fragments of local symmetry (FLS).

2.5. Monte Carlo method

The Monte Carlo method is used to calculate correlation weights by means of the optimization of the target function, which is calculated as:

$$TF = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 + IIC_C \times 0.3 \quad (5)$$

r_{AT} and r_{PT} are determination coefficients between the experimental and calculated endpoint values for the active and passive training sets, respectively. The IIC is the index of ideality of correlation (Toropova et al., 2018).

Table 1 contains the correlation weights for amino acids and FLS used for calculations of DCW2(3,7) (Split 1). Table 2 contains an example of DCW2(3,7) calculations.

ANALYSIS	FLS CODE
Search for XYX **GCCSDPRCRYRCGAAG** .....RCR..... .....RYR.....	XYX2
Search for XYYX **GCCSDPRCRYRCGAAG** .....GAAG	XYYX1
Search for XYZYX **GCCSDPRCRYRCGAAG** .....CRYRC....	XYZYX1

Fig. 1. An example of an extraction of FLS codes from a sequence of amino acids.

2.6. Domain of applicability

The domain of applicability for the described model, calculated with Eq. (15) defines the so-called statistical defects of SMILES attributes. These defects can be calculated as:

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N'(A_k)} + \frac{|P(A_k) - P''(A_k)|}{N(A_k) + N''(A_k)} + \frac{|P'(A_k) - P''(A_k)|}{N'(A_k) + N''(A_k)} \quad (6)$$

where $P(A_k)$, $P'(A_k)$ $P''(A_k)$ are the probability of A_k in the active training set, passive training set, and calibration set, respectively; $N(A_k)$, $N'(A_k)$, and $N''(A_k)$ are frequencies of A_k in the active training set, passive training set, and calibration set, respectively. The statistical SMILES-defects (D_j) are calculated as:

$$D_j = \sum_{k=1}^{NA} d_k \quad (7)$$

where NA is the number of non-blocked SMILES attributes in the SMILES.

A SMILES falls in the domain of applicability if

$$Dj < 2^*\overline{D} \quad (8)$$

2.7. Mechanistic interpretation

Among the obtained results, one acquires numerical data on the correlation weights of codes applied in quasi-SMILES. This was obtained as the result of several runs of the Monte Carlo optimization. From the obtained data, one can extract three categories of these codes:

(i) Codes that have a positive value of the correlation weight in all runs. These are promoters of endpoint increase.
(ii) Codes that have a negative value of the correlation weight in all runs. These are promoters of endpoint decrease.
(iii) Codes that have both negative and positive values of the correlation weight in different runs of the optimization. These are codes with unclear roles (one cannot classify these features as promoters of increase or decrease for the endpoint).

Table 1
Correlation weights for amino acids and FLS.

A_k or FLS	NA[a]	NP	NC	d_k	Non-blocked
A	1964	1956	1994	0.0000	TRUE
C	1295	1318	1276	0.0000	TRUE
D	1509	1589	1504	0.0000	TRUE
E	1489	1517	1452	0.0000	TRUE
F	1488	1554	1511	0.0000	TRUE
G	2181	2225	2203	0.0000	TRUE
H	892	900	911	0.0000	TRUE
I	1738	1786	1687	0.0000	TRUE
K	1765	1765	1767	0.0000	TRUE
L	2079	2070	2040	0.0000	TRUE
M	942	1015	942	0.0000	TRUE
N	1528	1525	1489	0.0000	TRUE
P	1797	1798	1833	0.0000	TRUE
Q	1124	1206	1168	0.0000	TRUE
R	1616	1605	1627	0.0000	TRUE
S	1981	1986	1946	0.0000	TRUE
T	1604	1614	1593	0.0000	TRUE
V	1819	1862	1809	0.0000	TRUE
W	713	699	696	0.0000	TRUE
Y	1196	1189	1193	0.0000	TRUE
[xyx10]	0	1	0	1.0000	FALSE
[xyx12]	0	1	1	1.0000	FALSE
[xyx13]	1	0	0	1.0000	FALSE
[xyx0]	972	973	971	0.0000	TRUE
[xyx1]	876	888	896	0.0000	TRUE
[xyx2]	537	556	504	0.0000	TRUE
[xyx3]	242	243	242	0.0000	TRUE
[xyx4]	82	85	98	0.0000	TRUE
[xyx5]	39	36	30	0.0001	TRUE
[xyx6]	10	9	9	0.0000	TRUE
[xyx7]	3	0	3	1.0000	FALSE
[xyx8]	0	1	0	1.0000	FALSE
[xyx9]	1	0	0	1.0000	FALSE
[xyyx0]	2460	2492	2438	0.0000	TRUE
[xyyx1]	280	276	291	0.0000	TRUE
[xyyx2]	21	20	22	0.0000	TRUE
[xyyx3]	0	4	3	1.0000	FALSE
[xyyx5]	1	0	0	1.0000	FALSE
[xyyx9]	1	1	0	1.0000	FALSE
[xyzyx0]	2490	2518	2496	0.0000	TRUE
[xyzyx1]	247	253	245	0.0000	TRUE
[xyzyx2]	22	19	10	0.0002	TRUE
[xyzyx3]	1	2	2	1.0000	FALSE
[xyzyx4]	1	0	0	1.0000	FALSE
[xyzyx5]	1	0	0	1.0000	FALSE
[xyzyx6]	0	0	1	1.0000	FALSE
[xyzyx7]	0	1	0	1.0000	FALSE
[xyzyx8]	1	0	0	1.0000	FALSE

[a] NA = frequency of code (amino acid or FLS) in active training set; NP = frequency in passive training set; NC = frequency in calibration set; d_k = statistical defect of a code; non-blocked, i.e., involved in the simulation process (TRUE) or blocked (FALSE).

Table 2
An example of DCW2(3,7) calculation for peptide IWKS: DCW(3,7) = 0.7166.

Code	CW(Code)
I	−0.1314
W	0.7418
K	0.4350
S	−0.4150
[xyx0]	0.8017
[xyyx0]	0.2221
[xyzyx0]	−0.9375

3. Results and discussion

Table 3 contains the data related to the statistical quality of models obtained using DCW1(3,7). Table 4 gives the statistical quality of models obtained using DCW2(3,7). The average determination coefficient value for the validation set is 0.63 ± 0.03, whereas in the case of DCW2(3,7), the characteristics are 0.68 ± 0.01 (Fig. 2).

One can conclude that DCW2(3,7) gives better models. These models are the following:

$$y = -0.0092\ (\pm 0.0003) + 0.12230\ (\pm 0.00005) * DCW(3,7) \quad (9)$$

$$y = -0.0240\ (\pm 0.0003) + 0.07026\ (\pm 0.00003) * DCW(3,7) \quad (10)$$

$$y = -0.2016\ (\pm 0.0003) + 0.09660\ (\pm 0.00004) * DCW(3,7) \quad (11)$$

$$y = 0.0169\ (\pm 0.0003) + 0.12976\ (\pm 0.00006) * DCW(3,7) \quad (12)$$

$$y = -0.0339\ (\pm 0.0003) + 0.09401\ (\pm 0.00004) * DCW(3,7) \quad (13)$$

The models developed and studied in this work offer certain advantages over those previously described in the literature, particularly by Rathore et al. (2024). A key benefit of the present approach is its simplicity and practical applicability: it requires only basic input data — the amino acid sequence of each peptide and its classification as toxic or non-toxic. This makes the method more accessible and easier to implement in practice, compared to earlier approaches that likely depend on more complex or extensive datasets.

The model described in the literature, with the best predictive potential model is based on Recurrent Neural Networks. It provides MCC = 0.72 (n = 2208). Our study resulted in a similar correlation coefficient. The best predictive potential in this work, observed in the case of the model based on DCW2(3,7), provides MCC = 0.70 (n = 2784).

Table 5 contains the results of five starts of the described Monte Carlo optimization for split 1. One can see that in the list of attributes that can be considered as promoters of an increase or decrease of toxicity present both amino acids and FLS represent internal configurations of peptides as a system of amino acids. Thus, our results are in agreement with the conception about the significance of internal configuration and self-control of peptides examined in the corresponding studies (Angelova et al., 2011; Hu et al., 2020).

The prevalence of FLS involved in Table 5 is significant enough. Hence, their influence on the models considered is natural. According to Table 5, the most significant promoters of toxicity increase are leucine (L), proline (P), lysine (K), asparagine (N), and cystine (C). Significant promoters of toxicity increase are the absence of three-member FLS [xyx0], as well as [xyx1], [xyzyx1], and [xyx3]. Promoters of toxicity decrease according to Table 5 are the absence of five-member FLS [xyzux0], as well as glycine (G), serine (S), alanine (A), valine (V), isoleucine (I), arginine (R), and others (see, Table 5).

At present, the idea of information dynamics, which has some analogy with thermodynamics, is gaining more and more popularity. Mass, Energy, and Information united in a new concept of Universe architecture (Vopson and Lepadatu, 2022).

When accepting such an ideology, it is quite logical to consider the fact of the existence of stable “building blocks of life,” that is, amino acids, as one of the indicators of the non-increase in entropy (chaos) in relation to the life systems. Further, if one talks about traditional human intelligence as a system of information movement, it is quite logical to expect not an increase in the complexity of information, but an increase in the convenience of using it.

Thus, there is reason to believe that the described approach to modeling peptide toxicity is convenient and appropriate for practical applications.

4. Conclusions

The proposed approach is an attractive way to construct predictive models of peptide toxicity, since it considers not only the presence of amino acids, but also their arrangement transmitted through fragments of local symmetry (FLS). It has been established that the effect of FLS on toxicity is comparable to the effect of various amino acids. The Monte Carlo optimization scheme used aims to maximize the MCC for the

Table 3
Model peptide toxicity obtained using Monte Carlo optimization with DCW_1.

Split	Set[a]	Sens	Spec	Acc	MCC	TN	TP	FP	FN	All
1	A	0.6118	0.8580	0.7372	0.4858	829	1208	200	526	2763
	P	0.6035	0.8762	0.7329	0.4946	886	1161	164	582	2793
	C	0.6901	0.9295	0.8123	0.6402	931	1306	99	418	2754
	V	0.6984	0.9174	0.8092	0.6322	940	1266	114	406	2726
2	A	0.6227	0.7806	0.7025	0.4087	835	1071	301	506	2713
	P	0.6364	0.7785	0.7074	0.4192	884	1079	307	505	2775
	C	0.7050	0.8395	0.7716	0.5490	994	1161	222	416	2793
	V	0.7083	0.8482	0.7782	0.5620	976	1168	209	402	2755
3	A	0.6376	0.8339	0.7366	0.4813	864	1150	229	491	2734
	P	0.6667	0.8230	0.7456	0.4961	926	1167	251	463	2807
	C	0.7552	0.8869	0.8218	0.6484	1012	1216	155	328	2711
	V	0.7573	0.8933	0.8233	0.6548	1086	1206	144	348	2784
4	A	0.6358	0.8492	0.7459	0.4980	859	1222	217	492	2790
	P	0.6390	0.8389	0.7378	0.4872	878	1125	216	496	2715
	C	0.7214	0.9086	0.8149	0.6413	989	1243	125	382	2739
	V	0.7314	0.9007	0.8145	0.6400	1040	1234	136	382	2792
5	A	0.6418	0.8381	0.7394	0.4892	878	1134	219	490	2721
	P	0.6512	0.8504	0.7562	0.5143	844	1228	216	452	2740
	C	0.7038	0.9107	0.8059	0.6270	986	1244	122	415	2767
	V	0.7254	0.9085	0.8137	0.6419	1054	1231	124	399	2808

[a] A = active training set; P = passive training set; C = calibration set; V = validation set.

Table 4
Model peptide toxicity obtained using Monte Carlo optimization with DCW_2.

Split	Set	Sens	Spec	Acc	MCC	TN	TP	FP	FN	All
1	A	0.6568	0.8608	0.7608	0.5298	890	1212	196	465	2763
	P	0.6424	0.8815	0.7558	0.5354	943	1168	157	525	2793
	C	0.7472	0.9317	0.8413	0.6925	1008	1309	96	341	2754
	V	0.7459	0.9072	0.8276	0.6627	1004	1252	128	342	2726
2	A	0.6764	0.8353	0.7567	0.5187	907	1146	226	434	2713
	P	0.6796	0.8600	0.7697	0.5486	944	1192	194	445	2775
	C	0.7532	0.9053	0.8285	0.6655	1062	1252	131	348	2793
	V	0.7620	0.9165	0.8392	0.6867	1050	1262	115	328	2755
3	A	0.6568	0.8448	0.7516	0.5112	890	1165	214	465	2734
	P	0.6710	0.8378	0.7553	0.5165	932	1188	230	457	2807
	C	0.7739	0.9023	0.8388	0.6824	1037	1237	134	303	2711
	V	0.7852	0.9119	0.8466	0.7007	1126	1231	119	308	2784
4	A	0.6610	0.8603	0.7638	0.5336	893	1238	201	458	2790
	P	0.6681	0.8546	0.7602	0.5314	918	1146	195	456	2715
	C	0.7535	0.9189	0.8361	0.6816	1033	1257	111	338	2739
	V	0.7496	0.9182	0.8324	0.6761	1066	1258	112	356	2792
5	A	0.6798	0.8426	0.7607	0.5292	930	1140	213	438	2721
	P	0.6798	0.8414	0.7650	0.5301	881	1215	229	415	2740
	C	0.7566	0.9165	0.8356	0.6809	1060	1252	114	341	2767
	V	0.7619	0.9144	0.8355	0.6815	1107	1239	116	346	2808

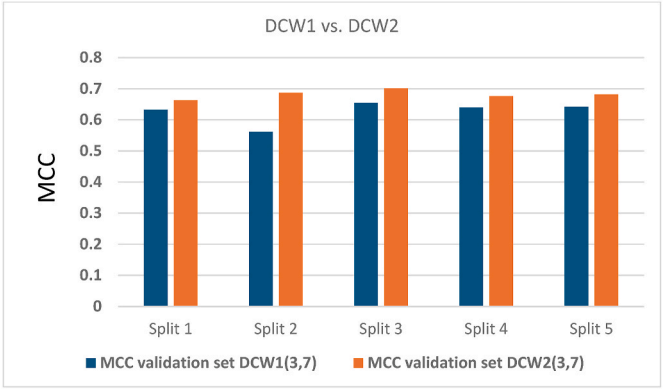

Fig. 2. The comparison of the predictive potential observed in the case of DCW1(3,7), i.e., without FLS, and in the case of DCW2(3,7), i.e., using FLS.

calibration set in the hope that this will be accompanied by high values for the validation set. As a result, all statistical characteristics turned out to be good for the calibration and validation sets, and slightly higher than for the active and passive training sets. The most influential promoters of toxicity increase according to the described computer experiments were the presence of amino acids Leucine (L), Proline (P), Lysine (K), and the following FLS: xyx0, xyx1, xyxyx1 (Table 5). The most influential promoters of toxicity decrease according to the described computer experiments were the absence of FLS xyzyx (i.e., xyzyx0) and the presence of the following amino acids: Glycine (G), Serine (S), and Alanine (A) (Table 5). The data from Table 5 can be used for a preliminary assessment of toxicity for new non-considered earlier peptides. It should be noted that the distribution in the four described sets has an influence on the predictive potential of the models considered. The comparison of the predictive potential of models constructed here with predictive potential from the literature (Rathore et al., 2024) confirms the perspectives of the suggested approach.

CRediT authorship contribution statement

Andrey A. Toropov: Writing – review & editing, Writing – original

Table 5
Observed promoters of an increase or decrease in toxicity (split 1).

Attribute	Run 1	Run 2	Run 3	Run 4	Run 5	NA[a]	NP	NC	d_k
Increase									
L	0.1493	0.2533	0.1980	0.2070	0.2639	2079	2070	2040	0.0000
P	0.0178	0.0106	0.0744	0.0542	0.0487	1797	1798	1833	0.0000
K	0.3542	0.6585	0.4165	0.5230	0.7257	1765	1765	1767	0.0000
N	0.0702	0.0830	0.1007	0.1059	0.1196	1528	1525	1489	0.0000
C	1.4568	2.6188	1.5241	1.9614	2.7244	1295	1318	1276	0.0000
[xyx0]	1.4266	2.0819	1.5432	1.8135	2.2687	972	973	971	0.0000
H	0.2667	0.4461	0.2559	0.3823	0.4847	892	900	911	0.0000
[xyx1]	0.8405	1.0252	0.9176	1.0239	1.1501	876	888	896	0.0000
W	0.6497	1.1420	0.6800	0.9005	1.2232	713	699	696	0.0000
[xyzyx1]	0.6328	1.0542	0.9183	0.4676	1.5898	247	253	245	0.0000
[xyx3]	0.4026	0.2996	0.4996	0.4190	0.3493	242	243	242	0.0000
[xyzyx2]	1.5096	2.5436	1.7217	1.6470	3.1751	22	19	10	0.0002
Decrease									
[xyzyx0]	−0.2855	−0.5467	−0.0708	−0.7937	−0.1027	2490	2518	2496	0.0000
G	−0.1330	−0.1717	−0.1325	−0.1441	−0.1886	2181	2225	2203	0.0000
S	−0.3273	−0.6774	−0.3571	−0.4778	−0.6820	1981	1986	1946	0.0000
A	−0.0784	−0.1417	−0.1013	−0.1090	−0.1388	1964	1956	1994	0.0000
V	−0.4449	−0.7930	−0.4505	−0.6209	−0.8790	1819	1862	1809	0.0000
I	−0.1352	−0.1600	−0.1581	−0.1326	−0.2326	1738	1786	1687	0.0000
R	−0.1019	−0.1719	−0.1308	−0.1539	−0.1831	1616	1605	1627	0.0000
T	−0.5105	−0.8950	−0.5665	−0.6662	−0.9097	1604	1614	1593	0.0000
D	−0.2143	−0.3933	−0.2111	−0.3414	−0.4198	1509	1589	1504	0.0000
E	−0.2830	−0.4955	−0.3035	−0.3396	−0.4771	1489	1517	1452	0.0000
F	−0.0588	−0.1573	−0.1343	−0.1068	−0.1717	1488	1554	1511	0.0000
Y	−0.4961	−0.8498	−0.5156	−0.6661	−0.9148	1196	1189	1193	0.0000
Q	−0.3652	−0.6180	−0.4020	−0.4850	−0.6600	1124	1206	1168	0.0000
M	−0.5460	−1.0153	−0.5908	−0.7665	−1.0428	942	1015	942	0.0000
[xyx5]	−1.6419	−3.3742	−1.6442	−2.3317	−3.4707	39	36	30	0.0001

[a] NA, NP, and NC are frequencies of attribute (amino acid or FLS) in active training set, passive training set, and calibration set, respectively.

draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Alla P. Toropova:** Writing – review & editing, Writing – original draft, Resources, Formal analysis, Data curation, Conceptualization. **Valentin O. Kudyshkin:** Writing – review & editing, Writing – original draft, Resources, Data curation. **Emilio Benfenati:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration. **Danuta Leszczynska:** Writing – review & editing, Writing – original draft, Visualization, Validation. **Jerzy Leszczynski:** Writing – review & editing, Writing – original draft, Visualization, Validation, Funding acquisition.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by National Science Foundation (NSF) award number OIA-2414444 (J.L.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biosystems.2025.105554.

Data availability

Data will be made available on request.

References

Angelova, A., Angelov, B., Mutafchieva, R., Lesieur, S., Convreur, P., 2011. Self-assembled multicompartment liquid crystalline lipid carriers for protein, peptide, and nucleic acid drug delivery. Acc. Chem. Res. 4 (2), 147–156. https://doi.org/10.1021/ar100120v.

Bonchev, D., Trinajstić, N., 1977. Information theory, distance matrix, and molecular branching. J. Chem. Phys. 67, 4517–4533. https://doi.org/10.1063/1.434593.

Craik, D.J., Fairlie, D.P., Liras, S., Price, D., 2013. The future of peptide-based drugs. Chem. Biol. Drug Des. 81, 136–147. https://doi.org/10.1111/cbdd.12055.

Fosgerau, K., Hoffmann, T., 2015. Peptide therapeutics: current status and future directions. Drug Discov. Today 20, 122–128. https://doi.org/10.1016/j.drudis.2014.10.003.

Gozalbes, R., De Julián-Ortiz, J.V., 2018. Applications of chemoinformatics in predictive toxicology for regulatory purposes, especially in the context of the EU REACH legislation. Int. J. Quant. Struct. Property Relationships 3, 1–24. https://doi.org/10.4018/IJQSPR.2018010101.

Hu, F., Angelov, B., Li, S., Li, N., Lin, X., Zou, A., 2020. Single-molecule study of peptides with the same amino acid composition but different sequences by using an aerolysin nanopore. Chembiochem 21, 2467–2473. https://doi.org/10.1002/cbic.202000119.

Randić, M., Vračko, M., Nandy, A., Basak, S.C., 2000. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci. 40 (5), 1235–1244. https://doi.org/10.1021/ci000034q.

Rathore, A.S., Choudhury, S., Arora, A., Tijare, P., Raghava, G.P.S., 2024. ToxinPred 3.0: an improved method for predicting the toxicity of peptides. Comput. Biol. Med. 179, 108926. https://doi.org/10.1016/j.compbiomed.2024.108926.

Toropov, A.A., Toropova, A.P., Raska Jr., I., Benfenati, E., Gini, G., 2012. QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. Struct. Chem. 23 (6), 1891–1904. https://doi.org/10.1007/s11224-012-9995-0.

Toropova, A.P., Toropov, A.A., Benfenati, E., Leszczynska, D., Leszczynski, J., 2018. Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES. Biosystems 169–170, 5–12. https://doi.org/10.1016/j.biosystems.2018.05.003.

Vopson, M.M., 2023. The second law of infodynamics and its implications for the simulated universe hypothesis. AIP Adv. 13 (10), 105308. https://doi.org/10.1063/5.0173278.

Vopson, M.M., 2025. Is gravity evidence of a computational universe? AIP Adv. 15 (4), 045035. https://doi.org/10.1063/5.0264945.

Vopson, M.M., Lepadatu, S., 2022. Second law of information dynamics. AIP Adv. 12 (7), 075310. https://doi.org/10.1063/5.0100358.