ISSN 1561-8331 (Print)
ISSN 2524-2342 (Online)
УДК 577.112, 577.112.854
https://doi.org/10.29235/1561-8331-2026-62-4-333-343

Поступила в редакцию 11.08.2026
Received 11.08.2026

M. A. Shapira[1], A. D. Miguckaya[2], V. P. Bakhir[1], D. O. Dormeshkin[1]

[1]Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Minsk, Belarus
[2]Belarusian State University, Minsk, Belarus

CORRELATION ANALYSIS OF PHYSICOCHEMICAL PROPERTIES REVEALS HIDDEN STRUCTURAL RELATIONSHIPS IN NANOBODIES FRAMEWORK REGIONS

Abstract. Nanobodies (VHH) – single-domain antibody fragments derived from camelid heavy-chain-only immunoglobulins – represent a rapidly expanding class of therapeutic and research tools whose unique structural properties depend on a compact immunoglobulin fold maintained in the absence of a light chain [1]. Despite the broad interest in nanobody engineering, a systematic large-scale characterization of the physicochemical principles governing their region architecture has not been performed. In the present study, a non-redundant dataset of 1 119 experimentally determined VHH structures was compiled from three open-access structural repositories and subjected to a comprehensive statistical and correlation analysis. A comparison of the framework regions and complementarity-determining regions revealed that each structural domain possesses a distinct physicochemical signature consistent with its biological role: FR1 is the most hydrophobic and aliphatic region, FR2 and FR3 exhibit the highest solubility and protein-binding potential, FR4 provides stabilization of FR1, and the CDR regions display a high variability, with CDR3 carrying a characteristic negative charge. Position-level Pearson correlation analysis identified specific residues whose amino acid properties are predictive of global region characteristics, including position 52, whose conformational propensity correlates significantly with the CDR3 length. Interpositional distance correlation analysis further revealed long-range co-variation networks that persist after removal of direct structural contacts, forming compact spatial clusters within the VHH beta-sandwich. These findings demonstrate that the VHH structural integrity is maintained through coordinated physicochemical adaptation across multiple framework positions and provide a data-driven basis for further rational nanobody scaffold engineering and synthetic library design.

Keywords: VHH, physicochemical analysis, structural bioinformatics

For citation. Shapira M. A., Miguckaya A. D., Bakhir V. P., Dormeshkin D. O. Correlation analysis of physicochemical properties reveals hidden structural relationships in nanobodies framework regions. *Vestsi Natsyyanal'nai akademii navuk Belarusi. Seryya khimichnykh navuk = Proceedings of the National Academy of Sciences of Belarus. Chemical series*, 2026, vol. 62, no. 4, pp. 333–343 (in Russian). https://doi.org/10.29235/1561-8331-2026-62-4-333-343

М.А. Шапіра[1], А.Д. Мігуцкая[2], В.П. Бахір[1], Д.А. Дармешкін[1]

[1]Інстытут біярганічнай хіміі Нацыянальнай акадэміі навук Беларусі, Мінск, Беларусь
[2]Беларускі дзяржаўны ўніверсітэт, Мінск, Беларусь

КАРЭЛЯЦЫЙНЫ АНАЛІЗ ФІЗІКА-ХІМІЧНЫХ УЛАСЦІВАСЦЕЙ, ЯКІ ВЫЯЎЛЯЕ СХАВАНЫЯ СТРУКТУРНЫЯ СУВЯЗІ Ў КАРКАСНЫХ РЭГІЁНАХ НАНАЦЕЛ

Анатацыя. Нанацелы (VHH) – аднадаменавыя фрагменты антыцелаў, атрыманыя з цяжкаланцуговых імуноглабулінаў вярблюдавых, – прадстаўляюць хутка пашыраемы клас тэрапеўтычных і даследчых інструментаў, унікальныя структурныя ўласцівасці якіх абумоўлены кампактным фолдам імуноглабуліну, які захоўваецца ў адсутнасць лёгкага ланцуга [1]. У дадзенай рабоце з трох агульнадаступных структурных рэпазіторыяў быў сфарміраваны набор з 1 119 эксперыментальна вызначаных структур VHH, які прайшоў комплексны статыстычны і карэляцыйны аналіз. Параўнанне каркасных абласцей (FRs) і абласцей, якія вызначаюць камплементарнасць (CDRs), паказала, што кожны структурны дамен валодае ўласцівым яму фізіка-хімічным профілем, які адпавядае яго біялагічнай функцыі: FR1 з'яўляецца найбольш гідрафобнай і аліфатычнай вобласцю, FR2 і FR3 вызначаюцца найвышэйшай растваральнасцю і патэнцыялам звязвання бялкоў, FR4 забяспечвае стабілізацыю FR1, тады як CDR-вобласці характарызуюцца высокай зменлівасцю, прычым CDR3 нясе характэрны адмоўны зарад. Карэляцыйны аналіз Пірсана на ўзроўні асобных пазіцый выявіў канкрэтныя рэшткі, уласцівасці амінакіслот якіх прадказваюць глабальныя характарыстыкі даменаў; у прыватнасці, пазіцыя 52 дэманструе значную карэляцыю з даўжынёй CDR3. Аналіз міжпазіцыйных дыстанцыйных карэляцый дадаткова выявіў залежнасці, якія захоўваюцца пасля выняцця прамых структурных кантактаў і ўтвараюць кампактныя прасторавыя кластары ў межах β-структуры VHH. Атрыманыя вынікі сведчаць аб тым, што структурная суцэльнасць VHH падтрымліваецца дзякуючы скаардынаванай фізіка-хімічнай адаптацыі шматлікіх каркасных пазіцый, і фарміруюць грунт для рацыянальнай інжынерыі сінтэтычных бібліятэк.

Ключавыя словы: VHH, фізіка-хімічны аналіз, структурная біяінфарматыка

Для цытавання. Карэляцыйны аналіз фізіка-хімічных уласцівасцей, які выяўляе схаваныя структурныя сувязі ў каркасных рэгіёнах нанацел / М. А. Шапіра, А. Д. Мігуцкая, В. П. Бахір, Д. А. Дармешкін // Весці Нацыянальнай акадэміі навук Беларусі. Серыя хімічных навук. – 2026. – Т. 62, № 4. – С. 333–343 https://doi.org/10.29235/1561-8331-2026-62-4-333-343

Introduction. Nanobodies, also referred to as VHH domains, are the antigen-binding fragments of the heavy-chain-only antibodies naturally produced by members of the family *Camelidae* and certain cartilaginous fishes [2]. Unlike conventional immunoglobulins, camelid heavy-chain antibodies lack an associated light chain, and their single variable region is sufficient for high-affinity, high-specificity antigen recognition. This structural autonomy confers several practically important properties: nanobodies are approximately ten times smaller than conventional IgG molecules, exhibit high thermal and chemical stability, are readily expressed in microbial systems, and retain their function under conditions that would denature conventional antibody fragments. These characteristics have driven their adoption across therapeutic development, molecular imaging, structural biology, and biosensor design, and a growing number of nanobody-based therapeutics have entered clinical evaluation [3].

The structural basis of VHH stability and antigen-binding diversity has been the subject of considerable investigation. The immunoglobulin fold of the VHH region is organized into four framework regions (FR1-FR4) that form the conserved beta-sandwich scaffold, and three complementarity-determining regions (CDR1-CDR3) that project from the scaffold surface to form the paratope [4]. Several hallmark positions within the framework – most notably the substitutions at positions 37, 44, 45, and 47 (the IMGT numbering scheme was applied throughout the study) that replace the canonical VH-VL interface residues – are well characterized as necessary adaptations to the absence of a light-chain partner [5]. CDR3 in VHH regions is typically longer than its conventional VH counterpart and frequently folds back over the framework scaffold, where it interacts with FR2 residues and contributes additional contacts that compensate for the missing light chain [1].

Such a description is of direct practical value. Rational engineering of VHH scaffolds for improved stability, altered biophysical properties, or expanded library diversity requires knowledge not only of which positions are conserved and why, but also of which non-conserved positions are functionally coupled and must therefore be treated as coordinated units during mutagenesis. Existing analyses have largely focused on sequence conservation, hallmark position identity, and structural classification of CDR3 conformations. A systematic statistical examination of physicochemical descriptors across structural regions, and particularly the analysis of long-range interpositional co-variation that persists after the removal of direct physical contacts, has not been reported for a large non-redundant VHH dataset.

In the present study, we address this gap by constructing a curated non-redundant database of 1 119 VHH structures drawn from three complementary structural repositories and performing a layered statistical analysis. We characterize the physicochemical profile of each structural region using six regional descriptors, identify specific positions whose amino acid properties drive regional characteristics through Pearson correlation analysis, and apply a permutation-based distance correlation protocol to detect long-range interpositional dependencies that reflect coordinated evolutionary adaptation. The results provide a comprehensive statistical map of the VHH domain and establish a quantitative foundation for structure-guided nanobody engineering.

Materials and methods. *Database construction*. Structural data were obtained from three open-access repositories: SAbDab (Structural Antibody Database [6]), NAStructuralDB (Natural Antibody Structural Database [7]), and SNAC-DB (Structural Nanobody Antigen Complex Database [8]). A total of 21 686 structural entries were retrieved from SAbDab, including full-length antibodies, VHH, scFv, Fab, and isolated V_H and V_L regions. All available entries from NAStructuralDB were retrieved, comprising 2 678 nanobody–antigen complexes and 4 160 single-chain VHH structures. In addition, all single-chain VHH and nanobody–antigen complex structures available in SNAC-DB were included.

For each entry, the atomic coordinates and associated metadata were extracted from the PDB files and corresponding summary tables. The collected data included atomic coordinates, amino acid sequences, IMGT-based sequence annotations, structural descriptions, and source information.

Chain identities and sequence integrity were validated using ANARCI (Antibody Numbering and Antigen Receptor Classify) *in silico* [9].

Sequence redundancy was removed using CD-HIT [10] with a 95 % sequence identity threshold, yielding a non-redundant dataset of 5 885 structures. Subsequent filtering for the VHH domains resulted in a final dataset of 1 119 unique structures.

Region annotation. Nanobody structures were divided into framework regions (FRs) and complementarity-determining regions (CDRs). The standardized IMGT numbering system was applied throughout the study for sequence annotation, structural analysis, and comparison of nanobody regions.

Primary structure analysis. Aromaticity was defined as the fraction of aromatic residues (phenylalanine, tryptophan, and tyrosine) within the analyzed region.

The aliphatic index (AI) was calculated according to Equation 1:

$$AI = (A + 2.900 \times V + 3.900 \times (I + L)) \times 100, \quad (1)$$

where *A*, *V*, *I*, and *L* represent the fractions of alanine, valine, isoleucine, and leucine, respectively. Higher *AI* values (> 80–100) are associated with increased protein thermostability and hydrophobicity, whereas values below 70 indicate greater structural flexibility and lower thermal stability [11]. For short peptide fragments, the AI was interpreted primarily as a measure of local hydrophobicity and aggregation propensity.

Net charge was calculated at pH 5.0, 7.4, and 9.0 using the ProteinAnalysis class implemented in Biopython (hereinafter – v. 1.76). The total charge was determined by summing the contributions of ionizable amino acid residues according to their predicted protonation states.

The Grand Average of Hydropathy (GRAVY) index was calculated as the mean Kyte–Doolittle hydropathy score across all residues in the analyzed region.

The instability index was determined from dipeptide composition using the Biopython implementation; values below 40 were considered indicative of protein stability [12].

The Boman index was computed using the Peptides Python library (v. 0.5.0). This parameter reflects the average solvation free energy of amino acid side chains, with values > 2.480 indicating a high binding potential and hydrophilicity, whereas negative values indicate predominantly hydrophobic regions [13].

Applied descriptors. The set of descriptors applied to the VHH domains is listed above. For individual residues, a list of 669 available descriptors was used during the position analysis [14]. This constituted a table in which each of the 20 possible amino acids was assigned a specific value for each property. These properties were pre-grouped into clusters covering hydrophobicity, structural propensity, volume, polarity, solvation, flexibility, conformational preferences, thermodynamic properties, contact energies, mutability, linker propensity, membrane association, and substitution matrices.

Statistical analysis. Statistical analyses were performed to identify significant differences between the structural regions (FR1–FR4 and CDR1–CDR3) and to evaluate correlations between the calculated physicochemical descriptors. All data processing and statistical testing were conducted in Python using the scipy.stats (v. 1.18.0) and statsmodels (v. 0.14.6) libraries.

The normality of each numerical parameter (aromaticity, aliphatic index, net charge, GRAVY index, instability index, and Boman index) across the seven structural regions was assessed using the Shapiro–Wilk test. Since most distributions significantly deviated from normality ($p < 0.050$), non-parametric statistical methods were applied.

Differences among the seven regions were evaluated using the Friedman test, which is a non-parametric alternative to repeated-measures ANOVA. Significant results ($p < 0.050$) were followed by pairwise comparisons using two-tailed Wilcoxon signed-rank tests across all 21 possible region pairs. Multiple testing correction was performed using the Bonferroni method, resulting in an adjusted significance threshold of $\alpha = 0.050/21.00 \approx 0.00238$. Differences were considered statistically significant at $p < 0.00238$.

Associations between the physicochemical descriptors (e. g., hydrophobicity and aliphatic index) were assessed using Kendall's rank correlation coefficient (τ). Kendall's τ was selected as a robust non-parametric measure suitable for the outlier-resistant analysis of monotonic relationships. The correlation significance was evaluated at $\alpha = 0.050$.

Pearson correlation analysis. To identify residue-level physicochemical determinants underlying the structural region properties, Pearson correlation coefficients (r) were calculated between position-specific amino acid descriptors and aggregated regional characteristics.

A vectorized parallel processing pipeline was implemented in Python to enable a large-scale analysis across all structures, amino acid properties, and regional metrics. The sequence-position matrices (669 properties per position) and regional property tables were indexed by unique antibody object identifiers and pre-grouped according to the IMGT positions and structural regions. Correlation matrices were calculated using vectorized column-wise operations (pandas. corrwith).

The statistical significance of the Pearson correlation coefficients was assessed using a two-tailed Student's *t*-test (equation 2) under the null hypothesis of no correlation ($H_0 : \rho = 0$):

$$t = (r\sqrt{(n-2)}) / \sqrt{(1-r^2)}\,, \tag{2}$$

where r is the Pearson correlation coefficient, and n is the number of paired observations with non-null values. The resulting t-statistic was used to calculate two-tailed p-values from the cumulative t-distribution.

Entries lacking defined sequence positions were excluded from the analysis. To focus on biologically relevant associations, only strong correlations ($|r| \geq 0.500$) were initially retained. To correct for multiple testing, a group Bonferroni correction (equation 3) was applied within each amino acid property group:

$$p_adj = \min(p_pearson \times M_g, 1.0)\,, \tag{3}$$

where M_g represents the total number of tested positions for a given property group. Correlations with adjusted significance values of $p_adj < 0.05$ were considered statistically significant.

To minimize potential bias from low sample sizes, candidate correlations were stratified into high-confidence ($n > 30$) and exploratory ($n < 30$) groups before ranking.

Distance correlation analysis. To rigorously validate non-random co-variation and spatial-functional dependencies between sequence alignment positions (pos_i and pos_j), a high-performance permutation protocol was implemented. Position-specific physicochemical profiles were standard-scaled across valid antibody structures ($mean = 0$, $\sigma = 1$), and pairwise feature divergence was quantified via Euclidean distance. First, to construct the empirical null distribution (H_0) for all objects, their indices of position j were randomly permuted across 10 iterations while keeping position i fixed. For each object, position j was permuted within the fixed set of amino acids to speed up the calculations.

Empirical p-values (equation 4) and standardized effect sizes (Z-scores) were derived as follows:

$$p = \frac{\sum(\left|d_{shuffled} - \mu_{shuffled}\right| \geq \left|d_{real} - \mu_{shuffled}\right|) + 1}{N_{perm} + 1}, \tag{4}$$

$$Z = \frac{d_{real} - \mu_{shuffled}}{\sigma_{shuffled}}. \tag{5}$$

Subsequently, the results of this rapid, approximate test were filtered based on a Z-score > 3.0 and an adjusted p-value > 0.050. The remaining pairs were then subjected to further permutation testing – specifically, 1 000 permutations per pair. In this step, position j was permuted, thereby preserving position-specific amino acid frequencies while systematically disrupting inter-positional associations.

Multiple hypothesis testing across all evaluated position-descriptor pairs was controlled using the Benjamini–Hochberg False Discovery Rate (FDR) procedure ($p_{fdr} < 0.050$). To reduce feature redundancy inherently present in biological descriptor databases, individual amino acid properties were mapped to higher-level functional clusters. Pairwise links were deduplicated by retaining exclusively the maximum effect-size edge (Z-score) per unique cluster-position pair (pos_i, $cluster_i$) ↔ (pos_j, $cluster_j$).

All obtained data were filtered by Z-score > 2, correlations between positions that usually have physical interactions in structures or have one common "neighbor" were also eliminated. For each residue, the total amount of correlations (by higher-level functional clusters) and the number of correlated positions were collected. The normalized interaction index was obtained by dividing the total number of correlations by the number of correlated positions.

Results and discussion. *Statistical analysis of* the *VHH domain properties*. During the analysis of the basic chemical and physical descriptors of the antibody structural regions we have found that among the VHH framework regions, FR1 exhibited the most pronounced hydrophobic and aliphatic character, with significantly higher GRAVY index (0.273 ± 0.210) and aliphatic index (100.720 ± 9.950) values compared to other regions. At the same time, it is strikingly low in aromatic residues (aromaticity index of 0.002 ± 0.009). That is well described by the enrichment of FR1 in aliphatic residues (*A*, *V*, *I*, *L*). Such an amino acid composition also explains the lowest Boman index (0.612±0.323) and the highest instability index (62.660 ± 13.300) values. Nevertheless, the instability of FR1 is well balanced by the other frameworks. Especially by FR4 (instability index –15.570 ± 11.240), which is indirectly but clearly confirmed by the high stability of VHH molecules in solution.

The FR2 and FR3 regions are notably less aliphatic compared to FR1 (46.620 ± 15.440 and 62.900 ± 6.910, respectively). At the same time, both have an average aromaticity (0.147 ± 0.034 and 0.139 ± 0.019, respectively) that is higher than that of FR1 but lower than in the CDR regions. For FR2, it is primarily explained by the hallmark positions: 42 (where F/Y account for 89.2 %), 49, 50 and 52 (where F/W/Y account for 51.7 %), that are crucial for the antibody architecture. This pattern is consistent with the tight packing of FR2 and FR3 within the VH/VL interface, where aromatic side chains frequently contribute to hydrophobic stabilization and interdomain complementarity [15]. Thus, the observed physicochemical profiles of FR2 and FR3 likely reflect structural constraints imposed by their role in forming the VH/VL interface rather than CDR-like functional diversification. Also, position 41 is strictly conserved as an aromatic residue (*W* 99.9 %, *Y* 0.01 %). In FR3, the key structural positions are 66, 67, 76, 88, 102, and 103. Despite the significant number of positions listed, their contribution is distributed across the sequence of FR3, which represents the longest region of the VHH domain (average length of 37.940 ± 0.660 AA). At the same time, positions 67, 76, and 102 display their side chains inside the protein globule and form a single spatial cluster, interacting with each other and contributing to the protein fold. Residue 103 frequently interacts with CDR3 and is highly conserved (*Y* 90.7 %). Only positions 66 and 88 are exposed to the solvent and are not involved in structure stabilization. While position 66 has wide amino acid variability, position 88 is quite conserved (*Y* 87.2 %). The reason for such conservation is still unclear. These characteristics do not contradict the high Boman index values of FR2 (2.085 ± 0.538) and FR3 (2.273 ± 0.296). On the one hand, these two regions are the key elements determining protein folding, and their protein-protein interaction potential must be high enough to ensure the binding of all other VHH domains. On the other hand, the Boman index is calculated by summing the free energies of transfer from cyclohexane to water for each amino acid side chain; therefore, it indirectly describes the solubility of the peptide. Nevertheless, it also does not contradict the physicochemical principles (expected physical behavior), as FR2 and FR3 contain many hydrophilic amino acids. Thus, FR2 and FR3 GRAVY index values are –0.713 ± 0.276 and –0.627 ± 0.152 respectively (the lowest among all other regions), which describes them as soluble protein regions. Another notable characteristic of FR2 is its average charge at pH 7.4. It is the only framework region that has a positive charge (0.52 ± 0.775) at the physiological pH, while all other framework regions have a slightly negative or close-to-zero charge. At the same time, all CDRs have a negative charge (it seems to be a characteristic property of CDRs), but CDR3 has the lowest charge (–0.820 ± 1.528). These properties facilitate the interaction between CDR3 and FR2, thereby compensating for the absence of the light chain.

FR4, which is the shortest (10.690 ± 0.830 AA) framework region, has an average aliphatic index value (57.560 ± 12.760) and a low aromaticity (0.087 ± 0.031). Along with a low Boman index value (0.917 ± 0.513), these characteristics highlight its main function – the stabilization of FR1, to which it is spatially close.

Because of the high sequence variability of the CDRs, most of the statistical analyses performed were not sufficiently informative for this region. Aside from the CDR3 charge, the only exception was the CDR2 instability index (24.860 ± 36.863), which showed a statistically significant difference (p-value < 0.05) and seems to be of interest. Such a value of the instability index is surpassed only by that of FR4, and as FR4 is spatially located next to FR1, this suggests that CDR2 All the obtained data are illustrated in Figure 1.

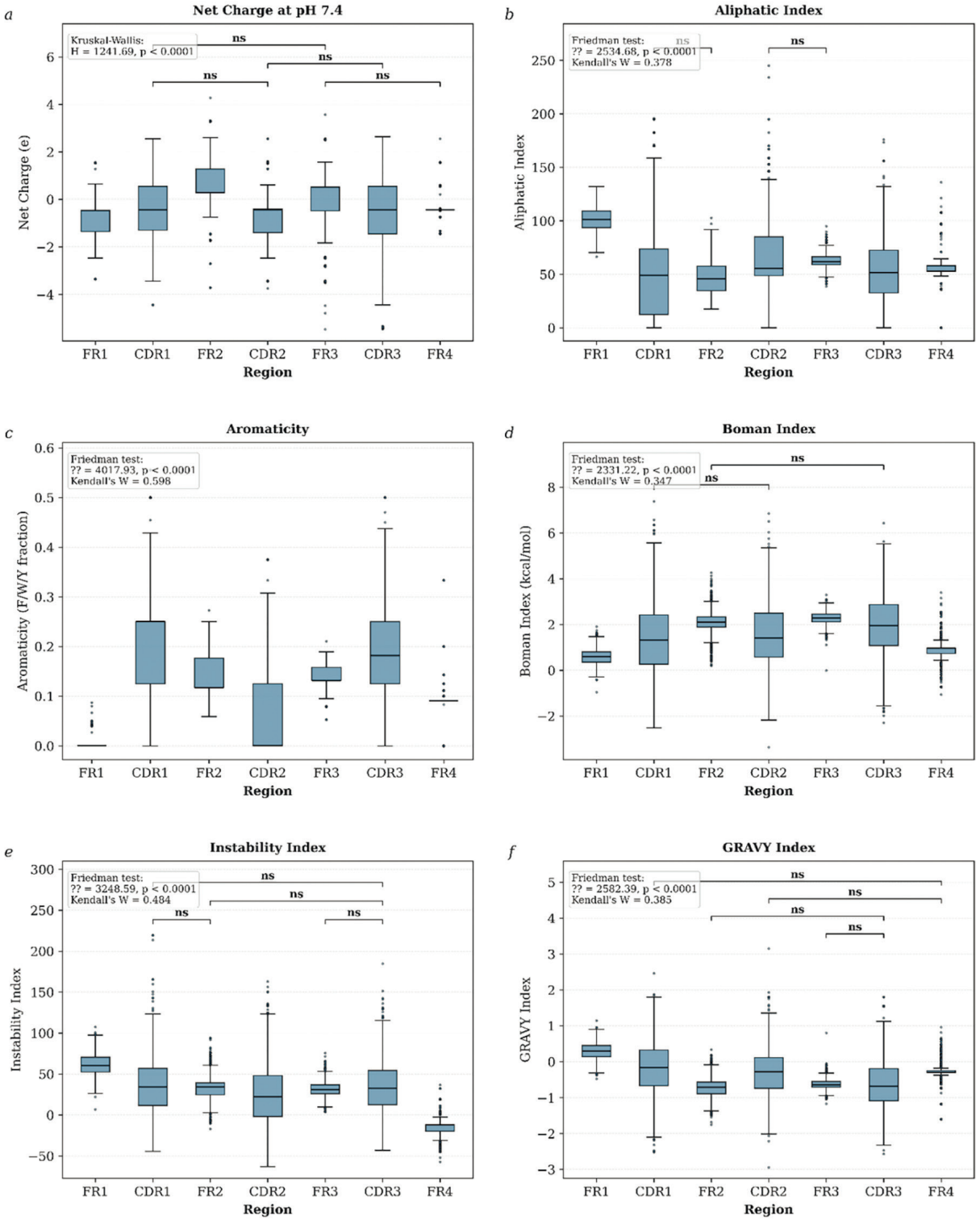


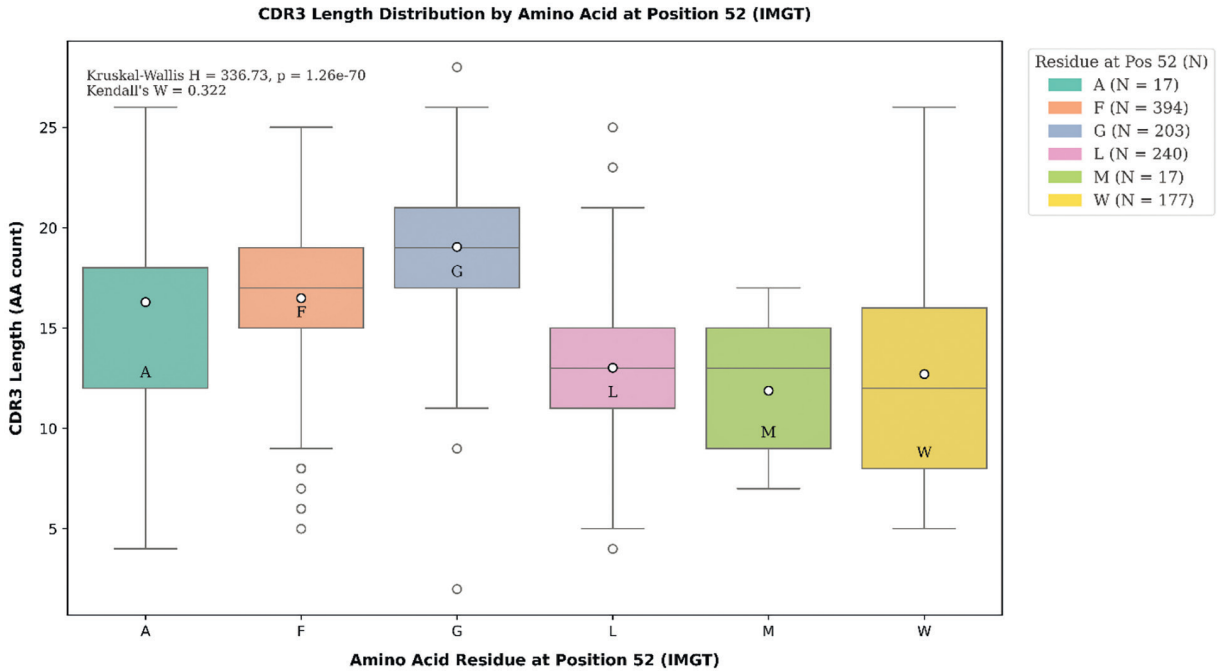
Fig. 1. Statistical evaluation of VHH domain physicochemical descriptors. Comparison across the framework regions (FR1–FR4) and CDRs of: (*a*) the net charge distribution at physiological pH 7.4, (*b*) the aliphatic index, (*c*) the aromaticity, (*d*) the Boman index, (*e*) the instability index, and (*f*) the GRAVY index. Only statistically insignificant correlations are displayed in the graphs; the absence of brackets indicates a statistically significant difference ($p < 0.050$)

All hallmark positions demonstrated moderate ($r > 0.500$) to strong ($r > 0.750$) correlations with several FR2-associated characteristics. Among these, the strongest correlation ($r = 0.540$) was observed for hallmark position 52, described by the "Normalized frequency of turn in all-alpha class (PALJ8101)"

descriptor [22] and the CDR3 length (Figure 2). This descriptor reflects the statistical propensity of amino acids to adopt turn or loop conformations. Although VHH position 52 resides within a β-strand adjacent to the CDR3 loop, its turn-propensity metric may indicate an increased likelihood of local β-sheet destabilization, thereby facilitating conformational deviation from the canonical β-architecture. Given that position 52 is highly variable and located in direct spatial proximity to CDR3, an influence on certain CDR3 structural properties is plausible. Another mechanistic interpretation is that variation at position 52 may modulate the elevation or orientation of the β-hairpin formed by positions 44–50 relative to the FR2 β-sheet plane, potentially affecting hairpin–antigen interactions.

The next notable association involves position 58, located within CDR2, which exhibits a moderate negative correlation with the aromaticity of the entire CDR2 region (r = –0.55; descriptor "AA composition of total proteins» (NAKH9001)"[23]). Although aromatic residues occupy position 58 considerably more frequently (22.6 %) than their average abundance in proteins (8.32 %) [24], the observed negative correlation indicates that the physicochemical profile of this position does not simply follow the overall aromatic character of CDR2. Instead, the relationship suggests that variation at position 58 is coordinated with changes in regional aromaticity, potentially reflecting compensatory constraints within the loop.

Analysis of the FR4 region identified residue 118 as the most prominent position. It has high r-value correlations with multiple FR4-related properties. However, the invariant presence of tryptophan at position 118 (90.8 %) at the FR4 boundary is widely recognized as essential for stabilizing the β-sheet core structure of the immunoglobulin fold. Moreover, its universal prevalence in functional repertoires largely reflects germline inheritance and evolutionary conservation rather than post-recombination selection for specific physicochemical properties.



p_adj	F	G	L	M	W
A	ns	0.159	0.050*	0.085	0.074
F	-	3.5e-16*	8.1e-31*	2.9e-05*	5.7e-20*
G	-	-	6.5e-48*	1.5e-08*	7.0e-33*
L	-	-	-	ns	ns
M	-	-	-	-	ns
W	-	-	-	-	-

Fig. 2. CDR3 length distribution by amino acids at position 52. The p-values of the pairwise comparison are presented in the table below

During *V*(*D*)*J* recombination, the RAG1/2 recombinase recognizes and cleaves the DNA exclusively at the conserved recombination signal sequences (RSSs), independently of the downstream coding sequence and its encoded amino acids [25, 26]. Subsequent processing of the coding ends by the non-homologous end joining (NHEJ) machinery, together with the terminal deoxynucleotidyl transferase (TdT), generates extensive junctional diversity through nucleotide trimming and non-templated nucleotide addition [27]. However, because these modifications are largely confined to the immediate coding ends, the highly conserved TGG codon encoding *W*118 within the germline IGHJ segment is generally preserved, maintaining this characteristic residue in the mature antibody repertoire [26].

Interposition correlation analysis. Not only can position-region correlation be useful for mining nonobvious structural features. During the interpositional correlation analysis, several positions influencing other structures were selected based on having the highest correlation coefficients (Figure 3).

Thus, Figure 4 demonstrates the strongest interpositional correlation networks identified for the selected framework residues from Cluster 3 as those with the highest interaction index. Despite the exclusion of direct physical contacts and neighboring residues during data processing, the correlated positions are not randomly distributed throughout the VHH structure. Instead, they form compact spatial clusters combined according to their spatial proximity, indicating that the detected associations reflect coordinated structural modules rather than isolated residue-residue interactions.

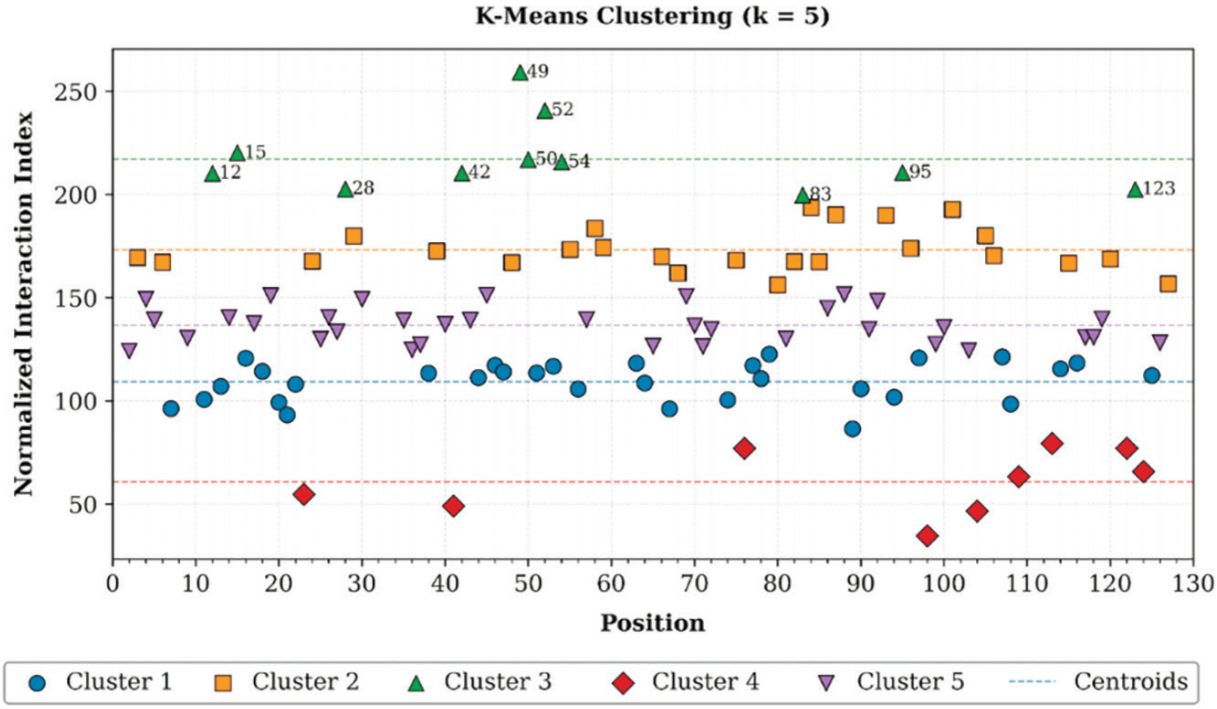


Fig. 3. Results of k-means clustering of the normalized interaction index for VHH positions. Cluster 3 was chosen for further analysis. Hallmark positions were excluded from the analysis, as modifications at these sites must be approached with caution

The most extensive network was observed for positions 13 and 123 (Figure 4, *a*).

Although these residues are located at the opposite termini of the VHH sequence, they occupy the framework scaffold and are spatially positioned beneath one face of the β-sandwich. The most strongly correlated positions (49–50, 83–84, and 115–116) are concentrated within specific structural regions rather than around the highlighted residues themselves. This pattern suggests that variability at positions 13 and 123 is associated with coordinated physicochemical adaptation across an entire framework surface instead of being limited to local structural contacts. Such behavior is consistent with their potential role as scaffold anchor residues contributing to the maintenance of the global architecture of the immunoglobulin fold.

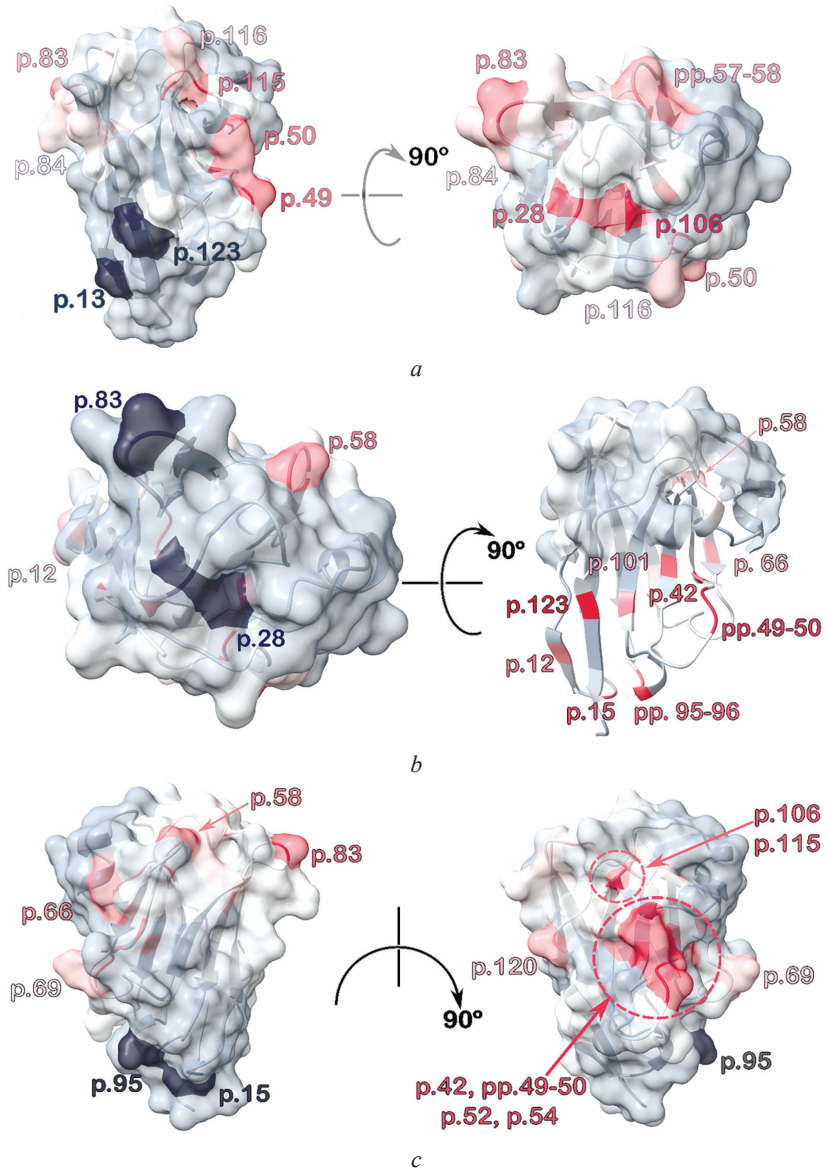


Fig. 4. The most significant correlations between positions from Cluster 3 (pointed black) with other positions. Color scale representing normalized interaction index from the lowest (gray) to intermediate (white) and highest (red): *a* – positions 13 and 123, *b* – positions 28 and 83, *c* – positions 15 and 95. The color saturation of the labels corresponds directly to the correlation level. VHH domain structure with PDB ID 1f2x was used as a template

In contrast, the networks centered on positions 28 and 83 (Figure 4, *b*) and positions 15 and 95 (Figure 4, *c*) exhibit a more localized organization. In both cases, the majority of highly correlated residues occupy compact structural neighborhoods within a β-sandwich, with pronounced enrichment around positions 42, 49–54, and 106–115. Although these residues are distant in the primary sequence, they converge into spatially continuous patches in the folded protein. The recurrence of these regions across independent correlation networks suggests that they represent structurally coordinated modules that accommodate physicochemical variation while preserving the integrity of the VHH framework.

Collectively, these observations indicate that the proposed statistical approach reconstructs higher-order evolutionary relationships rather than direct residue-residue interactions. The emergence of well-defined structural clusters after removing neighboring contacts implies that long-range compensatory dependencies are embedded within the VHH scaffold. Such dependencies likely reflect the coordinated adaptation of physicochemical properties that preserves β-sandwich geometry and provides structural support for the highly variable antigen-binding loops.

Conclusions. In this study, we performed a comprehensive statistical analysis of a non-redundant dataset of 1 119 experimentally determined VHH structures to identify the physicochemical principles underlying nanobody architecture. A comparison of the framework and complementarity-determining regions demonstrated that each structural domain possesses a distinct physicochemical profile closely related to its biological function. FR1 was identified as the most hydrophobic and aliphatic region, whereas FR2 and FR3 exhibited the highest protein-binding potential and solubility, consistent with their central role in maintaining the immunoglobulin scaffold. In contrast, the CDR regions displayed substantially greater variability, with CDR3 distinguished by its characteristic negative charge, suggesting a previously underappreciated contribution of electrostatic interactions to the stabilization of the heavy-chain-only antibody architecture.

Residue-level correlation analysis further demonstrated that global regional properties are governed by a relatively small number of structurally important positions. Several hallmark framework residues were found to correlate with regional physicochemical characteristics, while position 52 exhibited significant associations with the CDR3 length and other structural features. These findings indicate that local amino acid composition contributes not only to immediate structural stability but also to the modulation of distant functional regions.

The interpositional correlation analysis revealed an additional level of structural organization that cannot be explained by direct residue contacts alone. After removing neighboring residues and local interaction effects, the correlated positions remained organized into compact spatial clusters within the VHH domain fold. Rather than forming isolated residue pairs, the identified framework positions acted as hubs linking multiple distant regions of the β-sandwich. This observation suggests the existence of long-range compensatory relationships that preserve the physicochemical balance and mechanical integrity of the immunoglobulin scaffold during sequence variation.

Collectively, these findings demonstrate that the statistical analysis of large structural datasets can recover a biologically meaningful organization beyond conventional conservation or contact analyses. The presented approach identifies framework residues that participate in coordinated evolutionary adaptation and highlights structural modules that are likely to contribute to nanobody stability while accommodating extensive diversity within antigen-binding loops. These results provide a data-driven foundation for the rational engineering of VHH scaffolds and may facilitate the design of synthetic nanobody libraries with improved stability and functional diversity.

Acknowledgements. The authors thank the staff of the Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus for their assistance in performing the experiments. This study was financially supported by the State Committee on Science and Technology of the Republic of Belarus (grants B26BF-012 and X25U-006).

Благодарности. Авторы выражают благодарность сотрудникам Института биоорганической химии НАН Беларуси за помощь в проведении экспериментов. Работа была финансово поддержана Государственным комитетом по науке и технологиям Республики Беларусь (гранты B26BF-012 и X25U-006).

References

1. Mullin M., McClory J., Haynes W., Grace J., Robertson N., van Heeke G. Applications and challenges in designing VHH-based bispecific antibodies: leveraging machine learning solutions. *mAbs*, 2024, vol. 16, no. 1, art. 2341443. https://doi.org/10.1080/19420862.2024.2341443

2. Ye Z., Li X., Zhan M. Nanobodies as next-generation targeting platforms: From discovery technologies to translational biomedicine. *VIEW*, 2026, art. 20260027. https://doi.org/10.1002/VIW.20260027

3. Steeland S., Vandenbroucke R. E., Libert C. Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discovery Today*, 2016, vol. 21, no. 7, pp. 1076–1113. https://doi.org/10.1016/j.drudis.2016.04.003

4. Evers A., Guarnera E., Pekar L., Zielonka S. From discovery to the clinic: structural insights, engineering options, clinical, and "next wave" applications of camelid-derived single-domain antibodies. *mAbs*, 2025, vol. 17, no. 1, art. 2583210. https://doi.org/10.1080/19420862.2025.2583210

5. Wang X., Zhang L., Zhang Y., Li J., Xu W., Zhu W. Distinct types of VHHs in Alpaca. *Frontiers in Immunology*, 2024, vol. 15, art. 1447212. https://doi.org/10.3389/fimmu.2024.1447212

6. Capel H. L., Vavourakis O., Williams B. H., Taylor C. R., Deane C. M. SAbDab2: The structural antibody database in the age of machine learning. SAbDab2. *bioRxiv* [preprint], 2026. https://doi.org/10.64898/2026.06.16.732554

7. Chomicz D., Dudzic P., Wrobel S., Gawlowski T., Demharter S., Spreafico R., Minoux H., Phillips A., Krawczyk K. NAStructuralDB : structural database to facilitate computational studies of molecular modeling and recognition of proteins with special focus on antibody–antigen interactions. *mAbs*, 2026, vol. 18, no. 1, art. 2630438. https://doi.org/10.1080/1942086.2026.2630438

8. Gupta A., Rivero B. M., Li R., Roel-Touris J., Nanfack Y. F., Wendt M., Qiu Y., Furtmann N. SNAC-DB: An ML-Ready Database for Antibody and NANO-BODY® VHH–Antigen Complexes with Expanded Structural Diversity and Real-World Benchmarking. SNAC-DB. *Protein Science*, 2026, vol. 35, no. 7, e70655. https://doi.org/10.1002/pro.70655

9. Dunbar J., Deane C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 2016, vol. 32, no. 2, pp. 298–300. https://doi.org/10.1093/bioinformatics/btv552

10. Li W., Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 2006, vol. 22, no. 13, pp. 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

11. Ikai A. Thermostability and Aliphatic Index of Globular Proteins. *The Journal of Biochemistry*, 1980, vol. 88, no. 6, pp. 1895–1898, https://doi.org/10.1093/oxfordjournals.jbchem.a133168

12. Kyte J., Doolittle R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 1982, vol. 157, no. 1, pp. 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

13. Boman H. G. Antibacterial peptides: basic facts and emerging concepts. *Journal of Internal Medicine*, 2003, vol. 254, no. 3, pp. 197–215. https://doi.org/10.1046/j.1365-2796.2003.01228.x

14. Kawashima S., Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Research*, 2000, vol. 28, no. 1, art. 374. https://doi.org/10.1093/nar/28.1.374

15. Seidler C. A., Spanke V. A., Gamper J., Bujotzek A., Georges G., Liedl K. R. Data-driven analyses of human antibody variable domain germlines: pairings, sequences and structural features. *mAbs*, 2025, vol. 17, no. 1, art. 2507950. https://doi.org/10.1080/19420862.2025.2507950

16. Rackovsky S., Scheraga H. A. Differential Geometry and Polymer Conformation. 2. Development of a Conformational Distance Function. *Macromolecules*, 1980, vol. 13, no. 6, pp. 1440–1453. https://doi.org/10.1021/ma60078a017

17. Rackovsky S., Scheraga H. A. Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids. *Macromolecules*, 1982, vol. 15, no. 5, pp. 1340–1346. https://doi.org/10.1021/ma00233a025

18. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*, 1974, vol. 185, no. 4154, pp. 862–864. https://doi.org/10.1126/science.185.4154

19. Saerens D., Ghassabeh G., Muyldermans S. Single-domain antibodies as building blocks for novel therapeutics. *Current Opinion in Pharmacology*, 2008, vol. 8, no. 5, pp. 600–608. https://doi.org/10.1016/j.coph.2008.07.006

20. Peng X., Zhu B. Different features identified by machine learning associated with the HIV compartmentalization in semen. *Infection, Genetics and Evolution*, 2022, vol. 98, art. 105224. https://doi.org/10.1016/j.meegid.2022.105224

21. Koshi J., Goldstein M. R. A. Context-dependent optimal substitution matrices. *Protein Engineering Design and Selection*, 1995, vol. 8, no. 7, pp. 641–645. https://doi.org/10.1093/protein/8.7.641

22. Malvar S., Bhagavathula A., de Luis Balaguer M. A., Sharma S., Chandra R. Machine learning can guide experimental approaches for protein digestibility estimations. *arXiv*, 2022. https://doi.org/10.48550/arXiv.2211.00625

23. NakashimaH., Nishikawa K., Ooi T. Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins. *Proteins*, 1990, vol. 8, no. 2, pp. 173–178. https://doi.org/10.1002/prot.340080207

24. Schatz D. G., Swanson P. C. V(D)J Recombination: Mechanisms of Initiation. *Annual Review of Genetics*, 2011, vol. 45, no. 1, pp. 167–202. https://doi.org/10.1146/annurev-genet-110410-132552

25. Liu C., Zhang Y., Liu C. C., Schatz D. G. Structural insights into the evolution of the RAG recombinase. *Nature Reviews Immunology*, 2022, vol. 22, no. 6, pp. 353–370. https://doi.org/10.1038/s41577-021-00628-6

26. Gellert M. V(D)J Recombination: RAG Proteins, Repair Factors, and Regulation. *Annual Review of Biochemistry*, 2002, vol. 71, no. 1, pp. 101–132. https://doi.org/10.1146/annurev.biochem.71.090501.150203

Information about the authors

S*hapira Michail A.* – Researcher. Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus (5/2, Academician Kuprevich Str., 220084, Minsk, Republic of Belarus). E-mail: mshapira2016@gmail.com, https://doi.org/0000-0003-3284-5860

Miguckaya Arianna A. – Student. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: bio.miguckayAD@bsu.by, https://doi.org/0009-0002-3595-8748

Bakhir Veranika P. – Ph.D. Student Junior Researcher., Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus (5/2, Academician Kuprevich Str., 220084, Minsk, Republic of Belarus). E-mail: bakhirveranika@gmail.com, https://doi.org/0009-0001-5480-2178

Dormeshkin Dmitry O. – Ph. D. (Chemistry), Associate Professor, Leading Researcher. Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus (5/2, Academician Kuprevich Str., 220084, Minsk, Republic of Belarus). E-mail: Dormeshkin@gmail.com, https://doi.org/0000-0003-2505-3005

Інфармацыя аб аўтарах

Шапіра Міхаіл Анатольевіч – навуковы супрацоўнік. Інстытут біаарганічнай хіміі НАН Беларусі (вул. Акадэміка Купрэвіча, 5/2, 220084, Мінск, Рэспубліка Беларусь). E-mail: mshapira2016@gmail.com, https://doi.org/0000-0003-3284-5860

Мігуцкая Арыяна Дзмітрыеўна – студэнт. Беларускі дзяржаўны ўніверсітэт (4, пр-т Незалежнасці, 220030, Мінск, Рэспубліка Беларусь). E-mail: bio.miguckayAD@bsu.by, https://doi.org/0009-0002-3595-8748

Бахір Вераніка Паўлаўна – аспірант, малодшы навуковы супрацоўнік. Інстытут біаарганічнай хіміі НАН Беларусі (вул. Акадэміка Купрэвіча, 5/2, 220084, Мінск, Рэспубліка Беларусь). E-mail: bakhirveranika@gmail.com, https://doi.org/0009-0001-5480-2178

Дармешкін Дзмітрый Алегавіч – кандыдат хімічных навук, дацэнт, вядучы навуковы супрацоўнік. Інстытут біаарганічнай хіміі НАН Беларусі (вул. Акадэміка Купрэвіча, 5/2, 220084, Мінск, Рэспубліка Беларусь), E-mail: Dormeshkin@gmail.com, https://doi.org/0000-0003-2505-3005