Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(8), 2308-2318
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

OTE-24LD: An Extended Descriptor Integrating Long-Distance Correlations for the Prediction of Macromolecular Interactions

Obonan Etienne Traore[1], Ndiffon Charlemagne Kopoin[2], Dagou Dangui Augustin Sylvain Legrand Koffi[2], Gbame Gbede Sylvain[3], Souleymane Oumtanaga[1]

[1]Laboratoire des Sciences de Données et Intelligence Artificielle (LASDIA), École Doctorale Polytechnique des Sciences et Technologies de l'Ingénieur (EDP-STI), Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire
[2]École Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC), Abidjan, Côte d'Ivoire
[3]Université Félix Houphouët-Boigny (UFHB), UFR Mathématiques-Informatique, Département Informatique, Abidjan, Côte d'Ivoire
Email: obonan.traore22@inphb.ci, charlemagnekopoin@gmail.com, dagousylvain@gmail.com, gbamegbedesylvain@gmail.com, souleymane.oumtanaga@inphb.ci

How to cite this paper: Traore, O.E., Kopoin, N.C., Koffi, D.D.A.S.L., Sylvain, G.G. and Oumtanaga, S. (2025) OTE-24LD: An Extended Descriptor Integrating Long-Distance Correlations for the Prediction of Macromolecular Interactions. *Open Journal of Applied Sciences*, **15**, 2308-2318.
https://doi.org/10.4236/ojapps.2025.158153

Received: July 25, 2025
Accepted: August 15, 2025
Published: August 18, 2025

Abstract

The prediction of interactions between biological macromolecules, particularly macromolecular interactions, remains a major challenge in structural and functional bioinformatics. Numerous feature extraction methods have been developed, relying primarily on the physicochemical properties of amino acids and their sequential relationships to address this issue. Among these approaches, descriptors such as AAC (Amino Acid Composition), DPC (Dipeptide Composition), CTD (Composition-Transition-Distribution), and PseAAC (Pseudo Amino Acid Composition) have been widely used to transform macromolecular sequences into numerical vectors suitable for machine learning models. In this context, the OTE-24 method was recently introduced to capture local correlations between residues based on two normalized physicochemical properties. Despite its performance, as with several earlier methods, this model suffers from an intrinsic limitation: it overlooks long-distance correlations, which play a crucial role in the formation of recognition sites and the stability of macromolecular complexes. To address this limitation, we propose in this study an optimized extension of OTE-24, named OTE-24LD (Long Distance), which enhances the original descriptor by integrating distant relationships between residues, using a decreasing weighting factor modulated by positional distance. This improvement makes it possible to capture func-

tional interactions often missed by descriptors based solely on immediate neighborhoods. The evaluation of this method, conducted on the HPRD dataset, demonstrates a 6.73% improvement in precision compared to earlier approaches.

Keywords

Macromolecular Interactions, Feature Extraction, Long-Distance Correlations, OTE-24LD Descriptor, Protein-Protein Interaction Prediction

1. Introduction

The prediction of macromolecular interactions is a key component of structural bioinformatics, with major applications in systems biology, drug discovery, and protein engineering. To effectively model these interactions, it is crucial to represent protein sequences as numerical vectors that can be processed by machine learning algorithms [1] [2]. Descriptors based on the physicochemical properties of amino acids have thus been widely employed to capture relevant sequence features [3]. Among the classical descriptors are Amino Acid Composition (AAC), Dipeptide Composition (DPC), Composition-Transition-Distribution (CTD), and Pseudo Amino Acid Composition (PseAAC) [4] [5]. These methods primarily focus on the local or global properties of sequences but present limitations, particularly regarding the consideration of long-distance interactions between residues. Recent studies have highlighted the importance of integrating structural information and long-range correlations to improve the accuracy of protein function and interaction predictions [6] [7].

In this context, the OTE-24 method was proposed to enhance the representation of macromolecular sequences by integrating two normalized physicochemical properties and capturing local correlations between adjacent residues [8]. Although this approach has demonstrated improved performance compared to traditional descriptors, it remains limited by its focus on immediate local interactions, thereby neglecting long-distance correlations, which are often critical to protein function and structure [9]-[11].

To overcome this limitation, we introduce in this study an extension of the OTE-24 method, named OTE-24LD (Long Distance). This new approach aims to integrate long-distance interactions between residues using a decreasing weighting factor based on the positional distance within the sequence. By capturing both local and distant correlations, OTE-24LD provides a more comprehensive and informative representation of protein sequences, potentially enhancing the prediction of interactions and biological functions.

We evaluated the performance of OTE-24LD on several benchmark datasets for macromolecular interaction prediction derived from the work of Vazquez *et al.* [12]. To assess the effectiveness of our model, we employed the Random Forest

algorithm.

2. OTE-24LD Approach

2.1. Overview of OTE-24

The OTE-24 method is a macromolecular sequence characterization technique based on the integration of the physicochemical properties of amino acids using a bigram approach. It involves computing matrices that represent the distances between amino acids according to properties such as hydrophobicity and hydrophilicity, and then extracting feature vectors by combining these matrices through correlation and concatenation methods. This approach captures both the local order of amino acids and their immediate contextual relationships within the sequence [8].

The OTE-24 approach offers the advantage of generating normalized, fixed-dimensional feature vectors, facilitating their use in classification models. It effectively exploits local relationships between amino acids, which is crucial for identifying functional or structural motifs. Furthermore, the combination of multiple physicochemical properties enhances the richness of the extracted information, contributing to improved prediction accuracy for molecular interactions.

However, a notable limitation of this method is its limited capacity to model long-distance interactions between amino acids, which play a critical role in the three-dimensional structure and function of proteins. This restriction to the analysis of local relationships can reduce the model's precision when predicting complex structural behaviors or interactions that require a more global overview of the sequence.

2.2. Motivation and Principles of OTE-24LD

In macromolecular structures, it is common for amino acids that are distant within the sequence to interact directly within the three-dimensional structure. These interactions play a decisive role in the stability and function of biological macromolecules. However, the OTE-24 method, limited to successive amino acid pairs, is unable to capture these long-distance relationships. To address this limitation, it is necessary to introduce a mechanism capable of quantifying the correlations between the physicochemical property values of amino acids separated by a given distance. The proposed extension, named OTE-24LD, consists of enhancing the classical OTE-24 descriptor by integrating weighted correlations at various distances. The idea is to preserve the local information provided by OTE-24 while adding components that represent interactions between amino acids separated by a defined number of positions within the sequence. Taking these weighted correlations into account improves the model's ability to capture the complexity of molecular structures and interactions at multiple scales.

2.3. Mathematical Formulation of OTE-24LD

Let a protein sequence of length L be composed of amino acids denoted as A_1,

A_2, A_3, A_4, $\cdots$, A_L. Each amino acid A_k is associated with a numerical value for each considered physicochemical property, denoted as $C_{(k,p)}$, where p represents the property index (in this study, $P=2$). The long-distance correlation between amino acids separated by d positions is calculated as follows:

$$V_d = \sum_{k=1}^{L-d}\left(\omega_d \cdot \sum_{p=1}^{P} C_{A_k,p} C_{A_{k+d},p}\right) \text{ with } 1 \le k \le L-1 \tag{1}$$

where:

d represents the distance between two amino acids, varying from 1 to $D_{\max}$,

ω_d is the weight associated with the distance d, and

P is the number of physicochemical properties considered.

This formulation makes it possible to quantify the interaction between all pairs of amino acids separated by d positions for each of the selected properties.

To modulate the importance of interactions based on distance, we introduce a weighting factor ω_d, which follows an exponential decay defined as:

$$\omega_d = \mathrm{e}^{-\alpha d} \tag{2}$$

where α is an adjustable parameter that controls the rate of decrease in the influence of interactions as the distance increases. This mechanism preserves the effect of nearby correlations while progressively reducing the influence of more distant relationships.

For each macromolecular sequence, the V_d values are computed for all distances d ranging from 1 to $D_{\max}$. The results are then concatenated to form a unique feature vector:

$$V = \left[V_1 V_2 V_3 \cdots V_{D_{\max}}, V_{2,1}, V_{2,2}, V_{2,3} \cdots V_{2,20}\right] \tag{3}$$

This vector thus brings together all the weighted long-distance correlations between amino acids for each of the selected properties. In this study, we considered two properties: hydrophobicity and hydrophilicity.

2.4. Selection and Normalization of Physicochemical Properties

For this OTE-24LD extension, two essential properties were selected. Hydrophobicity, a key property in the structuring of macromolecules and the formation of internal hydrophobic regions, and hydrophilicity, which influences the exposure of amino acids to solvents and interactions with the aqueous environment. These two properties were chosen because they are widely used in studies of macromolecular structural behavior.

3. Experimental Protocol

3.1. Dataset

In this section, a class of possible solutions is proposed and the insertion of these solutions in the main model is investigated to check where they will lead to. This is initiated by the statement of the following claim.

The study was conducted using data from the Human Protein Reference Data-

base (HPRD), a reference resource for the functional and structural analysis of human proteins [13]. Developed through a collaboration between the Bioinformatics Institute of Bangalore (India) and the Pandey Laboratory at Johns Hopkins University (USA), HPRD compiles manually curated scientific annotations covering the majority of human macromolecules.

The latest version of this dataset includes over 36,500 unique macromolecular interactions involving 25,000 proteins and 6360 isoforms [14] [15].

In this study, we maintained the same data proportions as the original OTE-24 method [8]. We performed preprocessing on the macromolecular sequences prior to feature extraction. First, incomplete and redundant sequences were filtered out to ensure the integrity and independence of the samples. Then, verification and homogenization of the macromolecular or protein alphab*et al.* allowed us to retain only sequences containing the 20 standard amino acids. Finally, the physicochemical properties of amino acids used in the descriptor calculations were normalized using the following transformation:

$$C_{A,p}^{\text{norm}} = \frac{C_{A,p} - C_p^{\min}}{C_p^{\max} - C_p^{\min}} \tag{4}$$

where $C_{A,p}$ is the raw value of a given property for amino acid A, and $C_p^{\min}$ and $C_p^{\max}$ are respectively the minimum and maximum values of this property across the 20 standard amino acids.

3.2. Experimental Parameters

The parameters involved in the computation of long-range correlation descriptors were optimized based on preliminary experiments conducted on the HPRD dataset. A grid search procedure was employed to identify the most relevant combinations of values for the parameters $D_{\max}$ and α, using a validation set extracted from the training data (80% of the HPRD dataset).

The parameter $D_{\max}$, which defines the maximum distance between two amino acids considered for computing correlations between physicochemical properties, was evaluated over the range 10 to 30 with a step size of 5. The highest average performance was achieved for $D_{\max} = 20$, which was therefore selected for the final model.

Similarly, the exponential weighting factor α, which controls the attenuation of interaction influence as a function of distance, was explored within the interval [0.01, 0.10], with an increment of 0.01. The optimal value identified was $\alpha = 0.05$.

With respect to the physicochemical properties considered, two fundamental attributes were retained: hydrophobicity and hydrophilicity, due to their crucial role in protein–protein interaction mechanisms.

To ensure an unbiased performance assessment and minimize training-related biases, we adopted a stratified 5-fold cross-validation procedure in conjunction with the Random Forest algorithm. The HPRD dataset was randomly partitioned into five balanced subsets based on class labels (positive/negative). In each itera-

tion, four subsets were used for training and the remaining one for testing. This process was repeated five times, with each subset serving once as the test set. Final performance metrics were computed as the average of the scores obtained across the five iterations.

3.3. Comparison Methods

To highlight the performance of the OTE-24LD method, it was compared to several classical descriptors from the literature, including AAC, DPC, CTD, PseAAC, and OTE-24. Recall that the AAC (Amino Acid Composition) model encodes the proportion of each amino acid within the sequence without order information [16] [17]. The DPC (Dipeptide Composition) model quantifies the frequency of successive amino acid pairs, thereby incorporating local order information [18] [19]. The CTD (Composition, Transition, Distribution) encodes sequences based on groupings of physicochemical properties and their distribution [20] [21]. The PseAAC (Pseudo Amino Acid Composition) enriches the AAC encoding by considering correlations between distant positions [22] [23]. Finally, the OTE-24 technique is a coding method based on weighted local interactions, considered here as the immediate reference method.

We applied these different methods on the same HPRD dataset and used the same cross-validation procedure, thus ensuring a fair and rigorous comparison of their performance.

4. Results and Discussion

4.1. Evaluation Metrics for Overall Model Performance

In this section, we present the performance of the proposed OTE-24LD model, trained using the Random Forest algorithm, for the prediction of interactions between biological macromolecules. The performance was assessed using several metrics relevant to the study context. These metrics reflect different aspects of the model's predictive behavior, particularly in the context of imbalanced classes. We used accuracy to provide a general measure of prediction correctness, while precision, recall, and specificity assess the model's ability to correctly identify positive and negative interactions. The F1-score offers a meaningful balance between precision and recall, especially useful in imbalanced class scenarios. Balanced accuracy corrects for the impact of class imbalance, and the Matthews Correlation Coefficient (MCC) provides a robust evaluation even when classes are unequal. Additionally, the ROC and PR curves, along with their respective areas under the curve (AUC and AUPR), enable graphical and quantitative analysis of the model's performance across various classification thresholds. The confusion matrix completes this analysis by offering a detailed visualization of prediction errors. These metrics were selected to ensure a rigorous and multidimensional evaluation of performance, tailored to the specific contextual characteristics of the problem studied. They are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (8)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensibilite}}{\text{Precision} + \text{Sensibilite}} \quad (9)$$

$$\text{Balanced Accuracy} = \frac{\text{Sensibilite} \times \text{Specificite}}{2} \quad (10)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

- **TP** (True Positives): number of positive interactions correctly predicted,
- **TN** (True Negatives): number of negative interactions correctly predicted,
- **FP** (False Positives): number of negative interactions incorrectly predicted as positive,
- **FN** (False Negatives): number of positive interactions incorrectly predicted as negative.

Together, these indicators allow for a comprehensive analysis of the model's performance, both in terms of overall prediction accuracy and the differentiation between positive and negative classes. The following table summarizes the performances obtained across all test sets (Table 1).

Table 1. Overall performance of the OTE-24LD model using random forest.

Metrics	Values (%)
Accuracy	93.51
Precision	96.73
Recall	90.10
Specificity	96.94
F1-Score	93.30
Balanced Accuracy	93.52
Matthews Correlation Coefficient	87.23
AUC (ROC)	97.99
AUPR	98.17

The model achieved an accuracy of 93.51%, indicating that a large proportion of the total predictions were correct. However, in contexts where class imbalance exists, accuracy alone can be misleading. For this reason, balanced accuracy was also computed, reaching 93.52%, reflecting a balanced ability to correctly predict

both classes. The recall of 90.10% demonstrates the model's ability to correctly identify positive cases, while the specificity of 96.94% highlights its capacity to avoid false positives. The F1-score, which synthesizes precision and recall, reached 93.30%, indicating a good trade-off between interaction detection and error minimization. The Matthews Correlation Coefficient (MCC), recognized for its robustness in imbalanced class scenarios, achieved 87.23%, further confirming the overall quality of the model.

4.2. Comparison with Existing Methods

To assess the effectiveness of our proposed approach, we conducted a comparative evaluation against several widely recognized models from the literature, including AAC, DPC, and PseAAC, each coupled with different classifiers. The original OTE-24 model, primarily designed to capture short-range Bigram interactions, was used as a baseline in our experiments. It achieved an accuracy of 86.61%, an F1-score of 87.33%, a ROC-AUC of 89.43%, and a Matthews Correlation Coefficient (MCC) of 86.19. In contrast, our enhanced method, OTE-24LD, which incorporates both physicochemical properties and long-range interaction features, demonstrates a 6.73% improvement in accuracy over OTE-24.

This performance gain is further supported by the results presented in the table below, which compares our approach to conventional methods that also consider long-distance interactions (Table 2).

Table 2. Performance comparison between OTE-24LD and other approaches.

Approach	Classifier	Accuracy (%)	F1-Score (%)	ROC-AUC (%)	MCC
AAC	SVM	69.20	80.60	58.00	0.10
DPC	RF	70.70	82.50	61.40	0.06
PseAAC	SVM	65.40	76.80	58.20	0.09
OTE-24LD	**RF**	**93.57**	**93.30**	**97.99**	**0.8723**

The OTE-24LD method outperforms classical approaches in terms of accuracy, F1-score, and AUC, confirming the value of integrating physicochemical properties and long-distance interactions.

4.3. Discussion

The OTE-24LD (Long Distance) method stands out from existing approaches in the literature by explicitly integrating long-range correlations between residues. While conventional descriptors such as AAC and PseAAC limited to local frequencies or immediate neighborhoods struggle to exceed 70% accuracy, OTE-24LD achieves a global accuracy of 93.51%. This significant improvement highlights the importance of capturing extended sequential dependencies in the prediction of macromolecular interactions.

This enhancement is also reflected in the F1-score, which increases from

82.50% for DPC (used as the reference classifier) to 93.30% with OTE-24LD. Such a gain demonstrates a better balance between recall and precision, made possible by our adaptive weighting scheme. This mechanism preserves the influence of distant residues without introducing redundancy or diluting relevant local information. The AUC reaches 97.99%, a value close to the theoretical optimum, indicating an excellent trade-off between sensitivity and specificity. Likewise, the Matthews Correlation Coefficient (MCC), a metric on which traditional methods often underperform in this context, reaches 0.8723, reflecting a well-calibrated and reliable model capable of limiting both false positives and false negatives.

Beyond these quantitative results, the ability to capture spatially distant but functionally coupled residues enables more accurate prediction of interaction interfaces, better modeling of complex stability, and identification of key regions involved in molecular recognition. This level of precision contributes to a deeper understanding of interaction mechanisms and supports the rational design of therapeutic biomolecules by reducing dependence on costly and time-consuming in vitro experiments.

Nevertheless, the OTE-24LD approach presents certain limitations that should be acknowledged. It relies solely on information derived from the primary sequence and does not incorporate structural or evolutionary features, which could further refine predictive performance. Additionally, computing long-distance correlation descriptors incurs a non-negligible computational cost. The simulations conducted in this study were performed on a high-performance server (25.1 GHz CPU, 128 GB RAM, 830 GB SSD storage). Although resource consumption remained within reasonable limits, this requirement may constitute a constraint for large-scale or real-time applications without further optimization.

By reducing reliance on experimental validation, OTE-24LD has the potential to accelerate the development and refinement of therapeutic candidates. This approach offers a robust, relevant, and generalizable solution for leveraging long-range sequence information in the prediction of macromolecular interactions. Through its performance, interpretability, and practical scope, it represents a promising advancement for research in structural and functional bioinformatics.

5. Conclusions and Suggestions

This study enabled us to develop a predictive model for macromolecular interactions based on the OTE-24LD vector, enhanced by the integration of long-distance correlations and normalized physicochemical properties of amino acids. The simulations conducted on the HPRD database confirm the superiority of our approach, notably achieving an accuracy of 93.51%, an F1-score of 93.30%, an AUC of 97.99%, and an MCC of 0.8723. These metric values reflect the model's reinforced ability to accurately distinguish true interactions from false signals, well beyond the performance levels observed with traditional methods such as AAC, DPC, or PseAAC.

The introduction of a decreasing weighting factor for distant residues within

the sequence allowed the model to capture long-range functional patterns, often underestimated by conventional descriptors. This strategy highlights the critical importance of considering distance effects in the prediction of macromolecular interfaces, offering a comparative precision gain of over 20% and achieving near-ideal discrimination according to ROC metric values.

Beyond validation on the HPRD dataset, this work lays the foundation for promising future extensions. On the one hand, incorporating 3D structural data could further enrich the OTE-24LD method; on the other hand, applying deep learning techniques or hybrid ensemble approaches to the generated feature vectors opens the way for even more robust predictive models.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Char, S., Corley, N., Alamdari, S., Yang, K.K. and Amini, A.P. (2025) ProtNote: A Multimodal Method for Protein-Function Annotation. *Bioinformatics*, **41**, btaf170. https://doi.org/10.1093/bioinformatics/btaf170

[2] Taha, K. (2025) Protein-Protein Interaction Detection Using Deep Learning: A Survey, Comparative Analysis, and Experimental Evaluation. *Computers in Biology and Medicine*, **185**, Article ID: 109449. https://doi.org/10.1016/j.compbiomed.2024.109449

[3] Kiouri, D.P., Batsis, G.C. and Chasapis, C.T. (2025) Structure-Based Approaches for Protein-Protein Interaction Prediction Using Machine Learning and Deep Learning. *Biomolecules*, **15**, Article 141. https://doi.org/10.3390/biom15010141

[4] Shen, H. and Chou, K. (2008) Pseaac: A Flexible Web Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition. *Analytical Biochemistry*, **373**, 386-388. https://doi.org/10.1016/j.ab.2007.10.012

[5] Chou, K. and Cai, Y. (2003) Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **53**, 282-289. https://doi.org/10.1002/prot.10500

[6] Emonts, J. and Buyel, J.F. (2023) An Overview of Descriptors to Capture Protein Properties—Tools and Perspectives in the Context of QSAR Modeling. *Computational and Structural Biotechnology Journal*, **21**, 3234-3247. https://pmc.ncbi.nlm.nih.gov/articles/PMC10781719/?utm_source=chatgpt.com

[7] Tang, T., Li, T., Li, W., Cao, X., Liu, Y. and Zeng, X. (2024) Anti-Symmetric Framework for Balanced Learning of Protein-Protein Interactions. *Bioinformatics*, **40**, btae603. https://doi.org/10.1093/bioinformatics/btae603

[8] Obonan Etienne, T., Ndiffon Charlemagne, K., Tchimou Guepie Euloge, N. and Souleymane, O. (2025) Optimization of Feature Extraction for the Prediction of Macromolecular Interactions: OTE-24 Approach. *International Journal of Advanced Research*, **13**, 577-589. https://doi.org/10.21474/ijar01/20598

[9] Hosseini, S., Golding, G.B. and Ilie, L. (2024) Seq-Insite: Sequence Supersedes Structure for Protein Interaction Site Prediction. *Bioinformatics*, **40**, btad738. https://doi.org/10.1093/bioinformatics/btad738

[10] Detlefsen, N.S., Hauberg, S. and Boomsma, W. (2022) Learning Meaningful Repre-

sentations of Protein Sequences. *Nature Communications*, **13**, Article No. 1914. https://doi.org/10.1038/s41467-022-29443-w

[11] Cao, M., Zainudin, S. and Daud, K.M. (2025) Feature Fusion with Attributed Deepwalk for Protein-Protein Interaction Prediction. *Scientific Reports*, **15**, Article No. 12255. https://doi.org/10.1038/s41598-025-96510-9

[12] Wang, G., Liu, X., Wang, K., Gao, Y., Li, G., Baptista-Hon, D.T., *et al.* (2023) Deep-Learning-Enabled Protein-Protein Interaction Analysis for Prediction of SARS-CoV-2 Infectivity and Variant Evolution. *Nature Medicine*, **29**, 2007-2018. https://doi.org/10.1038/s41591-023-02483-5

[13] Kandasamy, R.K., *et al.* (2025) Human Proteinpedia: A Unified Discovery Resource for Proteomics Research. *Nucleic Acids Research*, **37**, D773-D781. https://www.researchgate.net/publication/23410255_Human_Proteinpedia_A_unified_discovery_resource_for_proteomics_research

[14] Goel, R., Harsha, H.C., Pandey, A. and Prasad, T.S.K. (2012) Human Protein Reference Database and Human Proteinpedia as Resources for Phosphoproteome Analysis. *Molecular BioSystems*, **8**, 453-463. https://doi.org/10.1039/c1mb05340j

[15] Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., *et al.* (2009) Human Protein Reference Database—2009 Update. *Nucleic Acids Research*, **37**, D767-D772. https://doi.org/10.1093/nar/gkn892

[16] Breimann, S. and Frishman, D. (2024) AAclust: *k*-Optimized Clustering for Selecting Redundancy-Reduced Sets of Amino Acid Scales. *Bioinformatics Advances*, **4**, vbae165. https://doi.org/10.1093/bioadv/vbae165

[17] Jiao, S., Ye, X., Sakurai, T., Zou, Q. and Liu, R. (2024) Integrated Convolution and Self-Attention for Improving Peptide Toxicity Prediction. *Bioinformatics*, **40**, btae297. https://doi.org/10.1093/bioinformatics/btae297

[18] Ullah, F., Salam, A., Nadeem, M., Amin, F., AlSalman, H., Abrar, M., *et al.* (2024) Extended Dipeptide Composition Framework for Accurate Identification of Anticancer Peptides. *Scientific Reports*, **14**, Article No. 17381. https://doi.org/10.1038/s41598-024-68475-8

[19] Yan, C., Geng, A., Pan, Z., Zhang, Z. and Cui, F. (2024) MultiFeatVotPIP: A Voting-Based Ensemble Learning Framework for Predicting Proinflammatory Peptides. *Briefings in Bioinformatics*, **25**, bbae505. https://doi.org/10.1093/bib/bbae505

[20] Ghafoor, H., Abbasi, A.F., Asim, M.N. and Dengel, A. (2024) CTD-Global (CTD-G): A Novel Composition, Transition, and Distribution Based Peptide Sequence Encoder for Hormone Peptide Prediction. *Informatics in Medicine Unlocked*, **50**, Article ID: 101578. https://doi.org/10.1016/j.imu.2024.101578

[21] Akbar, S., Raza, A. and Zou, Q. (2024) Deepstacked-AVPs: Predicting Antiviral Peptides Using Tri-Segment Evolutionary Profile and Word Embedding Based Multi-Perspective Features with Deep Stacking Model. *BMC Bioinformatics*, **25**, Article No. 102. https://doi.org/10.1186/s12859-024-05726-5

[22] Qiu, W., Xiao, X., Lin, W. and Chou, K. (2014) iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *BioMed Research International*, **2014**, Article ID: 947416. https://doi.org/10.1155/2014/947416

[23] Chou, K. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024