Optimization of feature extraction for the prediction of macromolecular interactions : OTE-24 Approach

UNDER PEER REVIEW IN IJAR

1. Abstract

In the field of molecular biology, where every interaction between macromolecules is of crucial importance, analyzing the structural features of biological macromolecules remains a major challenge. Traditional feature extraction techniques from protein sequences often prove to be inefficient. The reliability of the extracted information is sometimes questionable due to the complexity and volume of the data involved. The volume and complexity of this biological data compel researchers in the field to turn to computational feature extraction techniques. Over the years, several computational methods have been proposed to accurately extract relevant and representative information from macromolecule sequences within these large datasets. However, these extraction techniques are sometimes impractical, and the relevance of the extracted information may be limited. In this study, we propose a large-scale feature extraction method based on the correlation analysis of two physicochemical properties of amino acids: hydrophobicity and hydrophilicity, as well as the correlation between amino acids. The results of this research, evaluated using databases commonly utilized in previous studies, show an accuracy improvement of over **2.58%** compared to existing methods.

Keywords : Molecular biology, Feature extraction, Physicochemical properties of amino acids, Hydrophobicity and hydrophilicity, Macromolecular interaction prediction

2. Introduction

In the drug development process, the study of interactions between biological macromolecules is crucial. This step is of paramount importance in the fields of biology, bioinformatics, and medical research. Biological macromolecules, such as proteins, nucleic acids, lipids, and polysaccharides, are the fundamental components of living organisms. Their interactions, whether at the cellular or macromolecular level, are responsible for regulating various biological processes, transmitting genetic information, and modulating immune responses, among other key functions [1]. Several high-throughput chemometric techniques, such as protein microarrays [2], Nuclear Magnetic Resonance (NMR) [3],[4], Biacore (Surface Plasmon Resonance) SPR [5], [6], and Isothermal Titration Calorimetry (ITC) [7], have been developed to detect these interactions. While these techniques have revealed numerous unknown interactions, they are often time-consuming and expensive. These constraints, combined with the volume and complexity of experimental data, have driven the development of computational models to predict large-scale macromolecular interactions.

Since the 1970s and 1980s, when computational techniques were introduced for detecting interactions between biological macromolecules, various approaches have been proposed to predict macromolecule-macromolecule interactions (MMI) using datasets available in biological databases. Several techniques, such as gene fusion [8],[9],[10], Archer FusionPlex panels, QIAseq RNAscan, and Oncomine Focus [11], 3D structural information [12], and gene ontology and annotation [13] ,[14], have contributed to this goal.

However, these approaches are not universal due to their high computational complexity. Their precision and reliability heavily depend on the information previously collected from the datasets used during implementation. The practical implementation of these approaches, as well as the practical information on gene annotation and ontology, is often incomplete for several reasons. First, although the Gene Ontology database is widely used, it is not exhaustive, and many annotations are incorrect or missing. This limits a comprehensive understanding of gene functions and gene products in different biological contexts [15]. Furthermore, the 3D structure of many proteins remains unknown. A significant portion of proteins has yet to be resolved using techniques such as X-ray crystallography or cryo-electron microscopy, despite considerable efforts to determine these structures [16]. Finally, macromolecule-macromolecule interactions (MMI) in many species are often rare and poorly documented. This is partly due to the limitations of current experimental techniques, which are costly and time-consuming, thereby restricting the amount of available data on MMIs [17].

Unlike amino acid data, which are widely available in biological databases, most of the proposed approaches in the past use data extracted from sequences to study and predict macromolecule interactions.

Several sequence-based approaches for macromolecule analysis have been proposed. For example, the Biological Jaccard Index [16] measures the similarity between macromolecule sequences. This method identifies k-mers (subsequences of length k) in each macromolecule and calculates the Jaccard similarity between these sets. However, this method is sensitive to variations (it performs less well for low sequence similarity) and does not account for structural information, which may limit its accuracy for certain complex interactions. Another approach is the ISLAND method, which uses various feature representations of macromolecular sequences, including amino acid composition (AAC), the average features of the BLOSUM-62 substitution matrix, Position Specific Scoring Matrix (PSSM) features, and descriptors derived from the biophysical properties of amino acids to model evolutionary relationships and physicochemical properties of macromolecules [18]. However, the diversity of features used in this method increases computational complexity, and its accuracy depends heavily on the quality of the data.

Another approach, the Stacked Autoencoder method, transforms macromolecular sequences into numerical features using methods such as autocovariance and conjoint triad, then trains an autoencoder to learn compact and informative representations of the sequences [19]. In addition to sharing the same data dependency limitation as the ISLAND method, this approach may suffer from overfitting.

N-gram-based approaches are also used for the analysis and prediction of interactions. These approaches focus on analyzing macromolecule sequences as fixed-length (n-gram) or variable-length segments [20]. The approach proposed by Kopoin et al. for predicting protein-protein interactions uses bigrams, where n = 2. It examines consecutive pairs of amino acids in the sequences. The physicochemical properties of hydrophobicity and hydrophilicity of

UNDER PEER REVIEW IN IJAR

amino acids are used to create these bigrams. This method is also combined with the Position Specific Scoring Matrix (PSSM), which provides information on the probability of amino acid substitutions according to their position. This allows for the generation of an enriched matrix that captures both the relationships between amino acids and contextual information. These features are then used to train an artificial neural network, which improves the accuracy of protein-protein interaction prediction. Although n-gram-based models effectively capture local patterns, they often experience contextual information loss. These approaches are also sensitive to the choice of n, as their performance varies depending on the size of the selected n-grams.

In this study, we propose an approach that combines the amino acid correlation calculation method proposed by Chou [21] with the bigram method proposed by Kopoin et al. in 2020 [20] to extract features from macromolecular sequences. In our research on macromolecule-macromolecule interactions, we employ the Random Forest algorithm [22], [23] to effectively learn the representations of macromolecule pairs. To evaluate the effectiveness of our model, we applied it to a large dataset of macromolecule-macromolecule interactions from the work of Vazquez et al. [24].

UNDER PEER REVIEW IN IJAR

3. Materials and Methods

3.1 General Overview

This study relies on a dataset of macromolecular interactions from the Human Protein Reference Database (HPRD), as described in the study by [25]. This reference database, widely used by many researchers for predicting interactions between macromolecules, is publicly accessible.

The developed approach focuses on extracting features from macromolecular sequences, enabling the extraction of the physicochemical properties of amino acids. Random forests, known for their robustness and efficiency in classification and pattern recognition, were used to predict macromolecular interactions [26]. The effectiveness of this classification method guided our choice of model, allowing us to achieve promising results [26], demonstrating a significant improvement in the predictive accuracy of macromolecular interactions. A detailed illustration of the process is presented in Figure 1.

3.2 **Dataset**

In this study, we focus on implementing a model based on macromolecular sequences to predict macromolecule-macromolecule interactions (MMI). The dataset of macromolecular interactions was derived from the Human Protein Reference Database (HPRD) [27]. To ensure data quality, duplicates were removed from the carefully selected positive data. For the construction of negative pairs, which represent non-interacting macromolecule pairs, the authors [15] paired macromolecules located in distinct subcellular localizations, using the observable macromolecular localization information available in version 57.3 of the Swiss-Prot database (uniprot.org). In their approach, they excluded shorter sequences (fewer than 50 amino acids) as well as those with multiple localizations, ensuring a high level of representativeness.

Our dataset includes a total of 36,630 positive interactions involving 9,630 different human macromolecules. To balance the dataset, we also selected 36,480 negative interaction pairs derived from 1,773 macromolecules [28], [29]. We also use datasets from Swiss-Prot [30], the Protein Data Bank (PDB) [31], BioGrid, and STRING [32] to compare the effectiveness of our method with other recent approaches in the field of MMI.

3.3 Random forest

Random Forest is a supervised learning algorithm that works by creating a collection of decision trees, where each tree is built from a random sample of the training data. This technique, known as bagging (Bootstrap Aggregating), allows for the creation of subsets of data from the original dataset $\boldsymbol{X}$, where each subset $\boldsymbol{S_i}$ is a randomly drawn sample with replacement of size N.

$$\boldsymbol{S_i = Sample(X, N)} \qquad [\mathbf{1}]$$

Each tree is then trained on a distinct sample, which enhances the robustness and accuracy of the model . In classification, each tree produces a prediction, and the final result is determined by a majority vote from the trees, as represented by:

$$\boldsymbol{\mathcal{H} = mode(h_1, h_2, h_3, \ldots\ldots, h_T)} \qquad [\mathbf{2}]$$

where $\boldsymbol{h_t}$ is the prediction of the t-th tree, and T is the total number of trees. In regression problems, the final prediction is the average of the tree predictions:

$$\boldsymbol{\mathcal{H} = \frac{1}{T}\sum_{t=1}^{T} h_t} \qquad [\mathbf{3}]$$

One of the key concepts in Random Forest is the use of measures such as Gini impurity or entropy to determine the most appropriate splits in the trees. **Gini Impurity:** In binary classification, Gini impurity $\boldsymbol{G}$ measures the probability that an observation will be misclassified if it were randomly assigned according to the class distribution in the node. It is calculated using the formula :

$$\boldsymbol{G = 1 - \sum_{i=1}^{C} P_i^2} \qquad [\mathbf{4}]$$

where $\boldsymbol{P_i}$ is the proportion of instances belonging to class ii, and $\boldsymbol{C}$ is the number of possible different classes that the target variable can take. Entropy (Alternative to Gini Impurity): Entropy is another measure of node homogeneity, often used with information gain. It is defined as :

$$\text{Å}(\boldsymbol{S}) = -\sum_{i=1}^{\boldsymbol{C}} \boldsymbol{P_i\, log_2(P_i)} \qquad [\mathbf{5}]$$

These measures allow each tree to choose the features that provide the best splits by minimizing impurity or maximizing information.

UNDER PEER REVIEW IN IJAR

Resistance to Overfitting and Feature Selection

Random Forest is also resistant to overfitting, particularly when the number of trees is sufficiently large. It combines multiple weak models to create a more powerful one. Additionally, Random Forest efficiently handles datasets with a large number of features. Each tree in the forest uses a subset of these features, randomly selected at each node, which increases diversity between the trees. The importance of features can be measured by the average reduction in impurity (Gini or Entropy) for each feature $\boldsymbol{X_j}$across all trees, according to the following formula:

$$\boldsymbol{I}(\boldsymbol{X_j}) = \frac{\boldsymbol{1}}{\boldsymbol{T}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{T}} \Delta\boldsymbol{G_t}\ (\boldsymbol{X_j}) \qquad [\boldsymbol{6}]$$

where $\Delta\boldsymbol{G_t}\ (\boldsymbol{X_j})$ is the impurity reduction for tree tt when the feature $\boldsymbol{X_j}$ is used.

Key Hyperparameters of Random Forest: Random Forest has several hyperparameters that directly influence its performance. These parameters include:

- **The number of trees (n_estimators) :** This is the total number of trees in the forest. A higher number of trees tends to improve overall accuracy, although it also increases computation time. The relationship between the number of trees and model accuracy can be approximated by:

$$\boldsymbol{Accuracy_{RF}} \approx \boldsymbol{f}(\boldsymbol{n_{estimators}}) \qquad [\boldsymbol{7}]$$

- **Maximum depth (max_depth):** This controls the depth of each tree. A greater depth allows for capturing complex relationships in the data but may lead to overfitting.
- **Number of features selected at each split (max_features):** This parameter determines how many features are available for each tree when making decisions. A restricted selection promotes diversity among the trees, thus reducing the risk of overfitting.

4. Proposed Feature Extraction Approach

This section explains our feature extraction approach, named OTE-24. This approach is inspired by the Bi-gram method proposed by Kopoin et al. [20] and the method for calculating amino acid correlation features by Chou in the APAAC method [21]. The computational models proposed in the literature require learning relevant and representative features of sequence pairs from the training dataset in order to perform prediction tasks on the test dataset.

Kopoin et al.'s approach extracts protein features using physicochemical properties in the form of bigrams. It involves calculating the physicochemical distance values for each amino acid sequence in the dataset, forming an $L \times 20$ matrix represented by C, where L is the length of the amino acid sequence, and creating a bigram feature vector from the data matrix for training. It uses the ANN classifier to predict protein interactions. Chou's Amphiphilic

UNDER PEER REVIEW IN IJAR

Pseudo Amino Acid Composition (APAAC) method is an improvement over the Pseudo Amino Acid Composition (PseAAC) method, designed to capture both hydrophobic and hydrophilic features of amino acids in protein sequences [33]. This method accounts for both the order of amino acids and the physicochemical properties of proteins, which is crucial for applications such as protein function prediction or their interaction with other macromolecules. APAAC is calculated using two key properties of amino acids: hydrophobicity and hydrophilicity. These values are integrated into a correlation function, which measures the similarity between two amino acids, and are then used to generate additional descriptors related to the amino acid sequence order [34].

4.1 Description of Our Approach

Our approach uses the bigram method and the pseudo-amino acid composition with autocorrelation (APAAC) method to generate feature vectors from macromolecular sequences, thereby facilitating the prediction of interactions between biological macromolecules. First, bigrams are calculated to extract local interactions between consecutive residues based on their physicochemical properties, such as hydrophobicity and hydrophilicity, resulting in a 400-value vector. Then, the APAAC amino acid correlation calculation method is applied to integrate global correlations on a larger scale, adding *2×2×λ* additional values to capture long-distance interactions. λ represents the interaction length, defining the range of interactions between amino acid residues. This process results in a final vector of *800 + 2×2×λ* values for each sequence, providing a rich and detailed representation of the structural and functional features of macromolecules.

Our general formula is as follows:

$$A_t = \begin{cases} \sum_{i=1}^{L-1} C_{k,i}.C_{k+1;j}\ , & with\ 1 \le i \le 20\ and\ 1 \le j \le 20 \\ \dfrac{N(t)}{1+\ \varphi\sum_{t=1}^{L} f_t}\ , & with\ 1 \le t \le L \end{cases} \qquad [8]$$

Where A_α is the numerical value or feature of the amino acid A at position α in the macromolecular sequence, $N(t)$ is the number of occurrences of amino acid t in the sequence. L is the length of the macromolecule sequence. φ is the weight parameter that adjusts the influence of the physicochemical properties of the amino acids relative to their base frequency, thus balancing the contribution of residue interactions and the simple composition of the macromolecular sequence. f_l is the correlation function based on the physicochemical properties of the amino acids, calculated as follows:

$$f_l = \frac{1}{N-l}\sum_{j=1}^{N-l} H_1(j).H_1(j+l) + H_2(j).H_2(j+l) \qquad [9]$$

N is the length of the macromolecule sequence, i.e., the total number of amino acid residues. H_1 and H_2 are the values of hydrophobicity and hydrophilicity properties, respectively. They are used to represent the similarity or difference between amino acids at two positions j and $j + l$. l is the offset between two indices in the macromolecule sequence. If $l = 1$, we study interactions between adjacent amino acid residues, and for $l > 1$, we consider interactions between residues separated by a specific number of positions in the sequence, allowing us to capture long-range interactions.

UNDER PEER REVIEW IN IJAR

Here, $C_{i,j} = \frac{1}{i} D\left(R_i, R_j\right)\ with\ i = 1 \dots \dots 20\ and\ j = 1 \dots \dots 20$ [10] .

The $C_{i,j}$ represent the physicochemical distance values between amino acids in the sequence. Specifically, $C_{k,i}$ is the physicochemical distance value at position k for amino acid i. $C_{k+1,j}$ is the physicochemical distance value at position *k+1* for amino acid j. These values are used to calculate the transition frequency between amino acids i and j in the sequence. *L* is the length of the macromolecule sequence. $\frac{1}{\mathrm{i}}$ is the weighting function for rank *i*.

$$D\left(R_i, R_j\right) = \frac{1}{2} \left\{\left[h_1\left(R_j\right) - h_1(R_i)\right]^2 + \left[h_2\left(R_j\right) - h_2(R_i)\right]^2\right\} \qquad [11]$$

$R_i$and $R_j$are the amino acid residues of rank *i* and *j*, respectively. Then, $h_1\left(R_j\right)$and $h_1(R_i)$are the respective numerical values of the hydrophobicity of residues R_i and R_j, and $h_2\left(R_j\right)$ and $h_2(R_i)$are the values of hydrophilicity for $R_i$and R_j. These values are calculated using the following formulas:

$$\begin{cases} h_1(R_i) = \dfrac{H_1^0(R_i) - \sum_{k=1}^{20} H_1^0(\mathbb{R}_k)/20}{\sqrt{\sum_{t=1}^{20}[H_1^0(R_i) - \sum_{k=1}^{20} H_1^0(\mathbb{R}_k)/20]^2/20}} \\ h_2(R_i) = \dfrac{H_2^0(R_i) - \sum_{k=1}^{20} H_2^0(\mathbb{R}_k)/20}{\sqrt{\sum_{t=1}^{20}[H_2^0(R_i) - \sum_{k=1}^{20} H_2^0(\mathbb{R}_k)/20]^2/20}} \end{cases} \qquad [12]$$

The $\mathbb{R}_{\boldsymbol{k}}$ values range from 1 to 20 and represent the 20 natural amino acids according to the alphabetical order of their one-letter codes: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y.

UNDER PEER REVIEW IN IJAR

4.2 The architecture (flowchart)

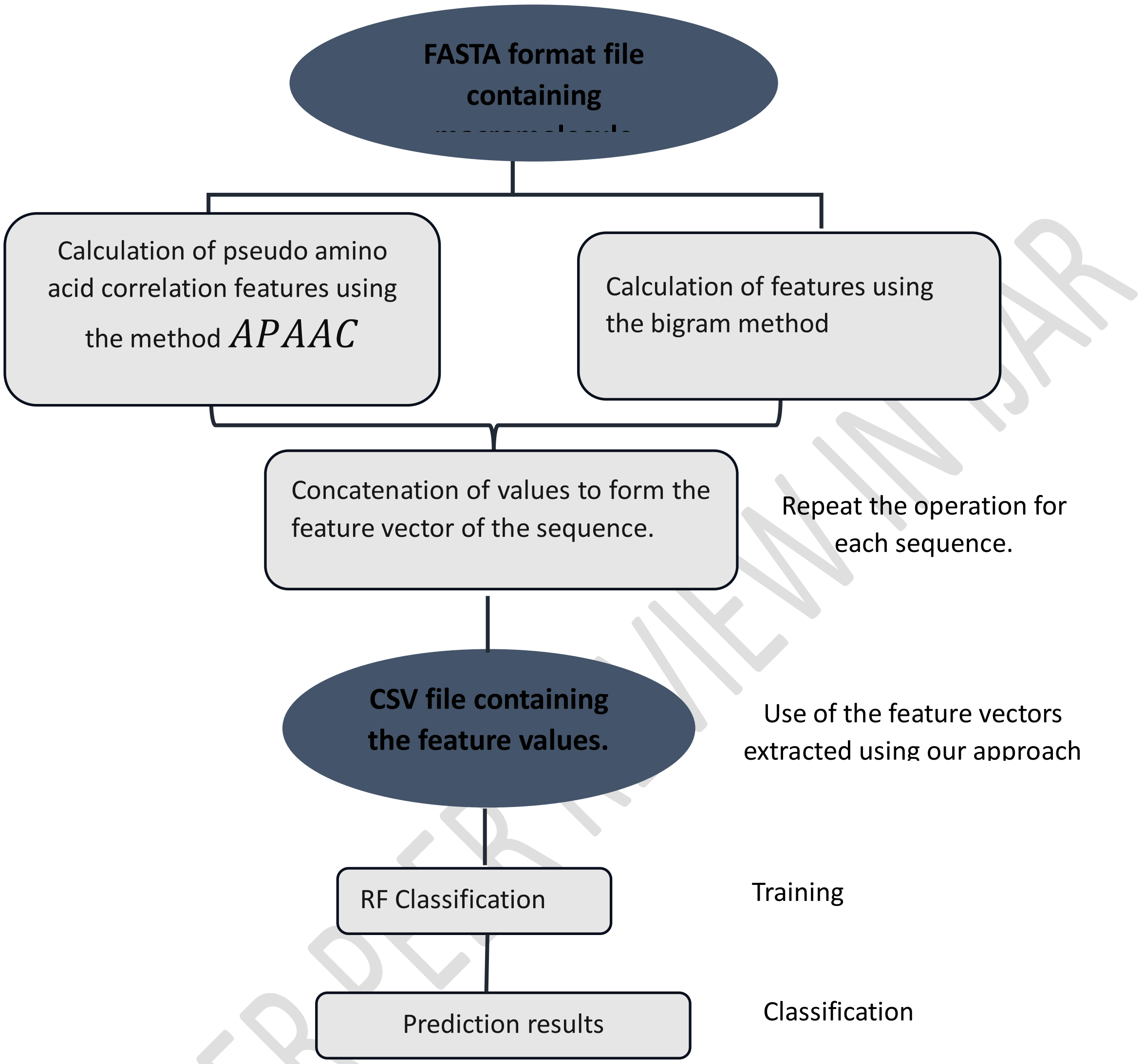
FASTA format file
containing
Calculation of pseudo amino acid correlation features using the method $APAAC$
Calculation of features using the bigram method
Concatenation of values to form the feature vector of the sequence.
Repeat the operation for each sequence.
CSV file containing the feature values.
Use of the feature vectors extracted using our approach
RF Classification
Training
Prediction results
Classification

Figure 1 : Algorithmic diagram of the OTE-24 approach

In our study, we undertook a methodical approach to extract meaningful features from amino acid sequences. First, for each sequence, we calculated the physicochemical distance values, which allowed us to construct an $L×20$ matrix, where L is the length of the sequence. Next, we determined the pseudo-amino acid components specific to each amino acid, which are essential for capturing important information about the sequence.

At the same time, we generated a bigram feature vector from the data of the matrix C. These bigram vectors capture the local relationships between amino acids, taking into account successive pairs, enriching the sequence representation. These two vectors (the APAAC component vector and the bigram vector) were then concatenated to form a global vector that captures both the physicochemical characteristics and sequential relationships.

Finally, this global vector was fed into a classifier based on the Random Forest algorithm for the learning and prediction phases.

UNDER PEER REVIEW IN IJAR

4.3 The evaluation metrics of the model.

We use widely recognized measurement criteria in the literature [35], [36] to evaluate the performance of our proposed approach and compare it with other existing models. These criteria include accuracy (Acc), precision (Pre), sensitivity (Sen), negative predictive value (NPV), F1 score (F1), and Matthews correlation coefficient (MCC). Accuracy (Acc) assesses the overall proportion of correct predictions made by the model, including both correctly predicted positives and negatives. Precision (Pre) measures the proportion of positive predictions made by the model that are actually correct, indicating its ability to limit false positives. Sensitivity (Sen), also called recall, evaluates the proportion of true positives detected by the model among all the actual true positives, which is crucial for identifying all real interactions. Negative predictive value (NPV) quantifies the proportion of true negatives among all negative predictions, ensuring that the model minimizes false negatives. The F1 score (F1) is a harmonic mean of precision and sensitivity, offering a balance between the ability to detect true positives and avoid false positives. The Matthews correlation coefficient (MCC) evaluates the correlation between the model's predictions and the actual observations, taking into account all cells of the confusion matrix to provide a global assessment of the model’s performance. These metrics allow us to assess the model's ability to effectively discriminate between interactions and non-interactions between biological macromolecules, which is crucial for the reliability and practical usefulness of the model in biomedical research [27]. The AUROC and AUPRC measures are essential for evaluating models predicting interactions between biological macromolecules. AUROC assesses the model's ability to distinguish true interactions from non-interactions by integrating the ROC curve, which represents sensitivity versus 1 - specificity across all classification thresholds. A high AUROC score near 1 indicates strong discrimination capability. In contrast, AUPRC focuses on precision and recall across different classification thresholds, with a high value indicating good precision and high recall, both of which are essential for applications requiring accurate detection of biological interactions. These metrics provide a comprehensive evaluation of the model’s performance by integrating both its discriminative ability and its precision across the full range of decision thresholds for interactive and non-interactive macromolecules. The formulas for calculating these measures are:

- **Accuracy (Acc)**

$$\boldsymbol{AAC} = \frac{TP+TN}{TP+TN+FN+FP} \qquad [13]$$

- **Precision (Pre)**

$$\boldsymbol{Pre} = \frac{TP}{TP+FP} \qquad [14]$$

- **Sensitivity (Sen) (Recall)**

$$\boldsymbol{Sen} = \frac{TP}{TP+FN} \qquad [15]$$

- **Negative Predictive Value (NPV)**

$$\boldsymbol{NPV} = \frac{TN}{TN+FN} \qquad [16]$$

- **Score F1 (F1)**

$$\boldsymbol{F1} = 2.\frac{Pre\ .*\ Sen}{Pre\ +\ Sen} \qquad [17]$$

- **Specificity (Spe)**

UNDER PEER REVIEW IN IJAR

$$Spe = \frac{TN}{\boldsymbol{TN+FP}} \quad [18]$$

- **Matthews Correlation Coefficient (MCC)**

$$\boldsymbol{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(\boldsymbol{TP+FP})*(\boldsymbol{TP+FN})*(\boldsymbol{TN+FP})*(\boldsymbol{TN+FN})}} \quad [19]$$

- **Area Under the ROC Curve (AUC-ROC)**

$$\boldsymbol{AUCROC} = \int_0^1 Sen\big(FRP^{-1}(t)\big)d(1 - Spe\big(FRP^{-1}(t)\big) \quad [20]$$

Where $FRP^{-1}(t)$ is the inverse function of the false positive rate for a decision threshold *t*.

- **Area under the precision-recall curve (AUPRC)**

$$\boldsymbol{AUPRC} = \int_0^1 Pre\big((t)\big)d(Recall) \quad [21]$$

True Positives (TP): Represent the interactions between macromolecules that we correctly predicted. For example, when we predict that an interaction occurs between two proteins, and this prediction is confirmed by experimental data, it constitutes a true positive.

True Negatives (TN): Correspond to pairs of macromolecules for which we correctly predicted that no interaction occurs. For example, if we predict that a specific enzyme does not interact with a particular substrate, and this is confirmed by the data, it is counted as a true negative.

False Positives (FP): Are situations where we incorrectly predicted that an interaction occurred between two macromolecules, while in reality, it does not occur. For example, if our model suggests that protein A interacts with protein B, but this interaction is not observed experimentally, it constitutes a false positive.

False Negatives (FN): Occur when our model fails to detect an interaction that actually exists between two macromolecules. For example, if two macromolecules do interact but our model does not predict this interaction, it is counted as a false negative.

By analyzing these categories **(TP, TN, FP, FN)**, we evaluate the overall performance of our prediction models. This evaluation is crucial for refining our approaches and improving the accuracy of our results in predicting interactions between biological macromolecules.

5. Results

In this section, we present the results obtained and compare them with those reported by other researchers using different methods. For this project, we developed a method based on sequence analysis to predict interactions between biological macromolecules. Unlike some previous studies, we used Python version 3.11.6 with JupyterLab in the Anaconda environment version 2.5.4, which allowed us to benefit from improved dependency management and a powerful interactive development environment.

5.1 Predictive performance of the proposed approach

In the predictive part, we used the same principle of splitting our datasets to train the chosen model with our extracted data. These features, in the form of numerical vectors, are used as input for our OTE-24 model. We performed 5-fold cross-validation on our reference dataset, which allowed us to train 3 different models. The results obtained are presented in Figure 2.

UNDER PEER REVIEW IN IJAR

Model 2 showed the best performance with a precision of 99.29%, an accuracy of 99.52%, a recall of 99.75%, an F1-score of 99.52%, and an area under the ROC curve (ROC AUC) of 99.99%. On average, the performances are 98.83% for precision (PRE), 99.4721% for accuracy (ACC), 98.83% for recall (SEN), 98.67% for the F1-score, and 83.67% for the ROC AUC. The high values of these different metrics, all above 98% except for the ROC AUC, indicate excellent predictive performance.

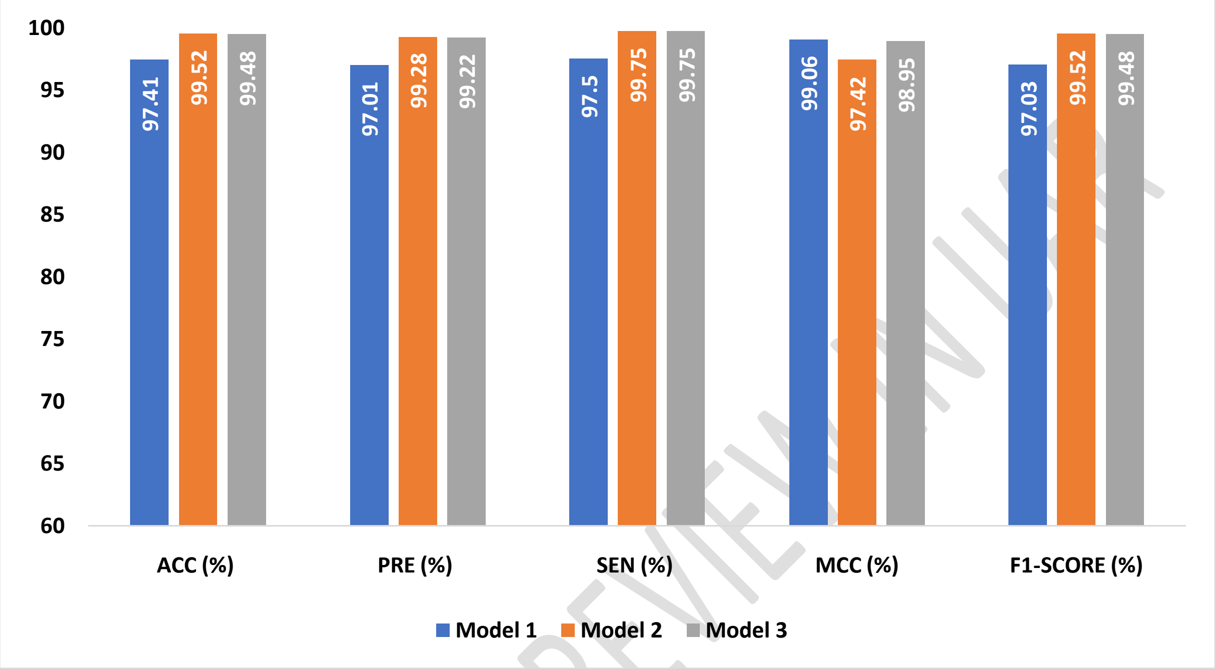

Figure 2 : Performance comparison of the three models trained on the HPRD database.

5.2 Comparison of our approach with other techniques

We compared our method with several other commonly used feature extraction techniques from the literature, applied to the same human dataset. These techniques include the bigram method [20], DWKNN (Ensemble) [37], BOW-GBDT [38], and DTI-BERT [39]. The comparison is based on various evaluation metrics. Figure 3 highlights these different comparison metrics between our approach and the approaches from the literature.

UNDER PEER REVIEW IN IJBR

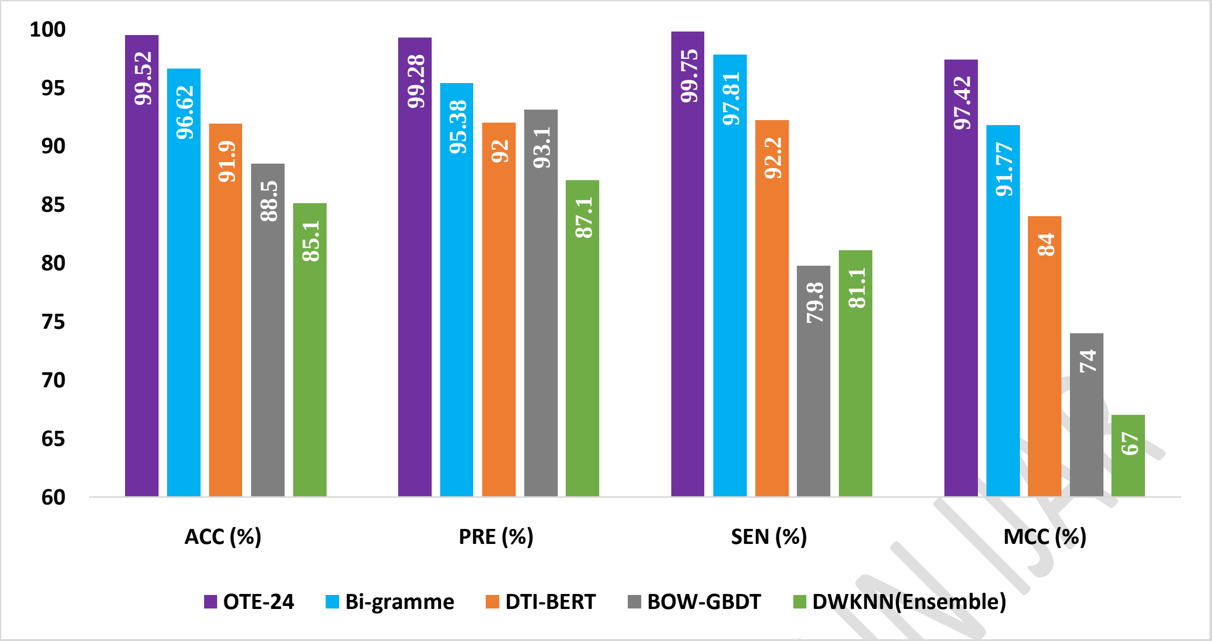

Figure 3: Comparison of the OTE-24 model with models from the literature

We compared our method with several other commonly used feature extraction techniques found in the literature, applied to the same human dataset. These techniques include the bi-gram method [20], DWKNN (Ensemble) [35], BOW-GBDT [36], and DTI-BERT [37]. The comparison is based on various performance metrics. Figure 3 highlights the differences in these metrics between our approach and those from the literature.

This comparison revealed an accuracy (ACC) of 96.62%, 91.90%, 88.50%, and 85.10% for the bi-gram, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble) methods, respectively, compared to 99.52% for our approach (OTE-24). This represents an improvement of 2.90%, 7.62%, 11.02%, and 14.42%, respectively.

Regarding precision (PRE), the rates are 95.38%, 92%, 93.10%, and 87.10% for the bi-gram, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble) methods, respectively, compared to 99.28% for our approach. This corresponds to an improvement of 3.90%, 7.28%, 6.18%, and 12.18%, respectively.

For sensitivity (SEN), we observed rates of 97.81%, 92.20%, 79.80%, and 81.10% for the bi-gram, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble) methods, respectively, compared to 99.75% for our approach. This results in an improvement of 1.94%, 7.55%, 19.95%, and 18.65%, respectively.

As for the Matthews Correlation Coefficient (MCC), the rates are 91.77%, 84%, 74%, and 67% for the bi-gram, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble) methods, respectively, compared to 97.42% for our approach. This represents an improvement of 5.65%, 13.42%, 23.42%, and 30.42%, respectively.

Our analysis shows that our technique surpasses the bi-gram method by at least 1.94%, DWKNN (Ensemble) by 12.18%, BOW-GBDT by 6.18%, and the DTI-BERT method by 7.28% on all the studied metrics.

UNDER PEER REVIEW IN IJAR

In the study of macromolecular interaction prediction, authors commonly use classification algorithms such as Support Vector Machine (SVM) [20], Random Forest (RF) [38], and K-Nearest Neighbors (KNN). In our case, we used the Random Forest algorithm and determined the hyperparameters using the grid search method. The optimal hyperparameters obtained are:

Bootstrap: True, **max_depth**: None, **min_samples_leaf**: 1, **min_samples_split**: 5, **n_estimators**: 100.

6. Discussion

The results obtained in our study reveal better performance of our method for predicting interactions between biological macromolecules. Through five-fold cross-validation, we trained three distinct models, and the performances achieved, particularly for model 2, demonstrate the efficiency of our approach. With a precision (PRE) of 99.29%, an accuracy (ACC) of 99.52%, a recall (SEN) of 99.75%, an F1-score of 99.52%, and an area under the ROC curve (ROC AUC) of 99.99%, our method significantly outperforms other techniques compared in the literature. On average, the observed performances, with values of 98.83% for precision, 99.4721% for accuracy, 98.83% for recall, 98.67% for F1-score, and 83.67% for ROC AUC, confirm the robustness and effectiveness of our approach.

A comparison with commonly used feature extraction techniques in the literature, such as the bi-gram method, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble), highlighted the superiority of our method. For instance, our approach achieves an accuracy rate of 99.52%, surpassing the bi-gram, DTI-BERT, BOW-GBDT, and DWKNN (Ensemble) methods by 2.90%, 7.62%, 11.02%, and 14.42%, respectively. Similarly, for precision, our method outperforms the other techniques by 3.90% to 12.18%. The Matthews Correlation Coefficient (MCC) also shows an improvement ranging from 5.65% to 30.42%, depending on the method compared. These results not only confirm the efficiency of our approach but also its ability to better capture the complex interactions between biological macromolecules.

One of the main strengths of our approach lies in the optimized use of Random Forest, combined with a particularly effective feature extraction method. Our feature extraction method appears to better capture the relevant information from amino acid sequences compared to other methods. Unlike models like DTI-BERT, which may require larger data volumes for effective learning, our method seems more suitable even for moderately sized datasets. The choice of Random Forest proved to be wise due to its ability to handle complex datasets with nonlinear relationships. Moreover, the optimization of hyperparameters through the grid search method allowed us to maximize the model's performance, making our method not only precise but also robust and generalizable to other datasets.

Another key advantage of our method is its flexibility. Unlike methods like DTI-BERT, which require substantial data volumes for optimal learning, our approach performs well even with smaller datasets. This feature is particularly valuable in the context of predicting interactions between biological macromolecules, where data can be limited.

Although our method shows exceptional overall performance, certain limitations deserve to be discussed. The average value of the ROC AUC, although respectable at 83.67%, is lower than the other metrics. This could suggest sensitivity to false positives or false negatives, an aspect that could be improved in future work.

UNDER PEER REVIEW IN IAR

Furthermore, the complexity of the Random Forest model, although beneficial for precision, can pose challenges in terms of computation time, especially during hyperparameter optimization. Future research could explore alternative approaches to reduce this complexity without sacrificing precision, such as integrating lighter ensemble learning techniques or using more efficient feature selection methods.

7. Conclusion

In this study, we presented a new feature extraction method to predict interactions between biological macromolecules. By generating feature vectors from macromolecule sequences using a combination of bigram methods and pseudo-amino acid descriptors, our approach demonstrated its effectiveness. The results obtained, with precision and accuracy rates exceeding 99%, attest to the robustness and reliability of our method.

The superiority of our approach compared to traditional techniques lies in its ability to extract relevant and representative information from macromolecule sequences, even from moderately sized datasets. This flexibility, combined with the use of an optimized Random Forest model, allowed us to maximize predictive performance while ensuring a high generalization of results.

We can therefore conclude that our proposed extraction approach constitutes a significant advancement in the field of molecular biology. It offers a practical and effective solution for the analysis of macromolecular interactions, thereby contributing to the understanding of fundamental biological processes and the development of new therapeutic applications.

UNDER PEER REVIEW IN IJAR

8. References

[1] A. W. Senior et al., « Improved protein structure prediction using potentials from deep learning », Nature, vol. 577, no 7792, Art. no 7792, janv. 2020, doi: 10.1038/s41586-019-1923-7.

[2] A. R. Jalalvand, « Chemometrics in investigation of small molecule-biomacromolecule interactions: A review », Int. J. Biol. Macromol., vol. 181, p. 478-493, juin 2021, doi: 10.1016/j.ijbiomac.2021.03.184.

[3] I. P. Gerothanassis, « Ligand-observed in-tube NMR in natural products research: A review on enzymatic biotransformations, protein–ligand interactions, and in-cell NMR spectroscopy », Arab. J. Chem., vol. 16, no 3, p. 104536, mars 2023, doi: 10.1016/j.arabjc.2022.104536.

[4] U. Salar, Atia-tul-Wahab, et M. Iqbal Choudhary, « Biochemical evaluation and ligand binding studies on glycerophosphodiester phosphodiesterase from Staphylococcus aureus using STD-NMR spectroscopy and molecular docking analysis », Bioorganic Chem., vol. 144, p. 107153, mars 2024, doi: 10.1016/j.bioorg.2024.107153.

[5] W. Wang, S. Thiemann, et Q. Chen, « Utility of SPR technology in biotherapeutic development: Qualification for intended use », Anal. Biochem., vol. 654, p. 114804, oct. 2022, doi: 10.1016/j.ab.2022.114804.

[6] E. A. FitzGerald et al., « Multiplexed experimental strategies for fragment library screening against challenging drug targets using SPR biosensors », SLAS Discov., vol. 29, no 1, p. 40-51, janv. 2024, doi: 10.1016/j.slasd.2023.09.001.

[7] V. M. Patil, S. P. Gupta, N. Masand, et K. Balasubramanian, « Experimental and computational models to understand protein-ligand, metal-ligand and metal-DNA interactions pertinent to targeted cancer and other therapies », Eur. J. Med. Chem. Rep., vol. 10, p. 100133, avr. 2024, doi: 10.1016/j.ejmcr.2024.100133.

[8] L. Zhang, D. Lu, X. Bi, K. Zhao, G. Yu, et N. Quan, « Predicting disease genes based on multi-head attention fusion », BMC Bioinformatics, vol. 24, no 1, p. 162, avr. 2023, doi: 10.1186/s12859-023-05285-1.

[9] M. Romero, O. Ramírez, J. Finke, et C. Rocha, « Feature extraction with spectral clustering for gene function prediction using hierarchical multi-label classification », Appl. Netw. Sci., vol. 7, no 1, p. 28, déc. 2022, doi: 10.1007/s41109-022-00468-w.

[10] J. Costa-Silva, D. S. Domingues, D. Menotti, M. Hungria, et F. M. Lopes, « Computational methods for differentially expressed gene analysis from RNA-Seq: an overview », 8 septembre 2021, arXiv: arXiv:2109.03625. doi: 10.48550/arXiv.2109.03625.

[11] C. Heydt et al., « Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation », BMC Med. Genomics, vol. 14, no 1, p. 62, févr. 2021, doi: 10.1186/s12920-021-00909-y.

[12] J. Wang et al., « MIFNN: Molecular Information Feature Extraction and Fusion Deep Neural Network for Screening Potential Drugs », Curr. Issues Mol. Biol., vol. 44, no 11, Art. no 11, nov. 2022, doi: 10.3390/cimb44110382.

[13] X. Zhao et al., « PermuteDDS: a permutable feature fusion network for drug-drug synergy prediction », J. Cheminformatics, vol. 16, no 1, p. 41, avr. 2024, doi: 10.1186/s13321-024-00839-8.

UNDER PEER REVIEW IN IJAR

[14] W. Yang, Q. Zhou, M. Yuan, Y. Li, Y. Wang, et L. Zhang, « Dual-band polarimetric HRRP recognition via a brain-inspired multi-channel fusion feature extraction network », Front. Neurosci., vol. 17, août 2023, doi: 10.3389/fnins.2023.1252179.

[15] B. M. Good et al., « Reactome and the Gene Ontology: digital convergence of data resources », Bioinformatics, vol. 37, no 19, p. 3343-3348, oct. 2021, doi: 10.1093/bioinformatics/btab325.

[16] Q. Chen et al., « Network-based methods for gene function prediction », Brief. Funct. Genomics, vol. 20, no 4, p. 249-257, juill. 2021, doi: 10.1093/bfgp/elab006.

[17] S. Rapposelli, E. Gaudio, F. Bertozzi, et S. Gul, « Editorial: Protein–Protein Interactions: Drug Discovery for the Future », Front. Chem., vol. 9, nov. 2021, doi: 10.3389/fchem.2021.811190.

[18] W. A. Abbasi, A. Yaseen, F. U. Hassan, S. Andleeb, et F. U. A. A. Minhas, « ISLAND: in-silico proteins binding affinity prediction using sequence information », BioData Min., vol. 13, no 1, p. 20, nov. 2020, doi: 10.1186/s13040-020-00231-w.

[19] G. Czibula, A.-I. Albu, M. I. Bocicor, et C. Chira, « AutoPPI: An Ensemble of Deep Autoencoders for Protein–Protein Interaction Prediction », Entropy, vol. 23, no 6, Art. no 6, juin 2021, doi: 10.3390/e23060643.

[20] C. N. Kopoin, N. T. Tchimou, B. K. Saha, et M. Babri, « A Feature Extraction Method in Large Scale Prediction of Human Protein-Protein Interactions using Physicochemical Properties into Bi-gram », in 2020 IEEE International Conf on Natural and Engineering Sciences for Sahel’s Sustainable Development - Impact of Big Data Application on Society and Environment (IBASE-BF), Ouagadougou, Burkina Faso: IEEE, févr. 2020, p. 1-7. doi: 10.1109/IBASE-BF48578.2020.9069594.

[21] K.-C. Chou, « Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes », Bioinformatics, vol. 21, no 1, p. 10-19, janv. 2005, doi: 10.1093/bioinformatics/bth466.

[22] R. Veevers et D. MacLean, « Improved K-mer Based Prediction of Protein-Protein Interactions With Chaos Game Representation, Deep Learning and Reduced Representation Bias », 23 octobre 2023, arXiv: arXiv:2310.14764. Consulté le: 20 juin 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2310.14764

[23] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, et D. Beaini, « Recipe for a General, Powerful, Scalable Graph Transformer », 15 janvier 2023, arXiv: arXiv:2205.12454. Consulté le: 20 juin 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2205.12454

[24] G. Wang et al., « Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution », Nat. Med., vol. 29, no 8, p. 2007-2018, août 2023, doi: 10.1038/s41591-023-02483-5.

[25] L. Xian et Y. Wang, « Advances in Computational Methods for Protein–Protein Interaction Prediction », Electronics, vol. 13, no 6, Art. no 6, janv. 2024, doi: 10.3390/electronics13061059.

[26] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski, et D. Spinello, « Protein–protein interaction prediction with deep learning: A comprehensive review », Comput. Struct. Biotechnol. J., vol. 20, p. 5316-5341, sept. 2022, doi: 10.1016/j.csbj.2022.08.070.

[27] H.-N. Tran, P.-X.-Q. Nguyen, F. Guo, et J. Wang, « Prediction of Protein–Protein Interactions Based on Integrating Deep Learning and Feature Fusion », Int. J. Mol. Sci., vol. 25, no 11, Art. no 11, janv. 2024, doi: 10.3390/ijms25115820.

[28] L. Hu, X. Wang, Y.-A. Huang, P. Hu, et Z.-H. You, « A survey on computational models for predicting protein–protein interactions », Brief. Bioinform., vol. 22, no 5, p. bbab036, sept. 2021, doi: 10.1093/bib/bbab036.

UNDER PEER REVIEW IN IJAR

[29] X. Hu, C. Feng, T. Ling, et M. Chen, « Deep learning frameworks for protein–protein interaction prediction », Comput. Struct. Biotechnol. J., vol. 20, p. 3223-3233, janv. 2022, doi: 10.1016/j.csbj.2022.06.025.

[30] F. Soleymani, E. Paquet, H. L. Viktor, W. Michalowski, et D. Spinello, « ProtInteract: A deep learning framework for predicting protein–protein interactions », Comput. Struct. Biotechnol. J., vol. 21, p. 1324-1348, janv. 2023, doi: 10.1016/j.csbj.2023.01.028.

[31] C. Chen et al., « Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier », Comput. Biol. Med., vol. 123, p. 103899, août 2020, doi: 10.1016/j.compbiomed.2020.103899.

[32] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, et Y. Zhang, « DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks », J. Chem. Inf. Model., vol. 57, no 6, p. 1499-1510, juin 2017, doi: 10.1021/acs.jcim.7b00028.

[33] C. Malbranke, W. Rostain, F. Depardieu, S. Cocco, R. Monasson, et D. Bikard, « Computational design of novel Cas9 PAM-interacting domains using evolution-based modelling and structural quality assessment », PLOS Comput. Biol., vol. 19, no 11, p. e1011621, nov. 2023, doi: 10.1371/journal.pcbi.1011621.

[34] W.-R. Qiu, A. Xu, Z.-C. Xu, C.-H. Zhang, et X. Xiao, « Identifying Acetylation Protein by Fusing Its PseAAC and Functional Domain Annotation », Front. Bioeng. Biotechnol., vol. 7, déc. 2019, doi: 10.3389/fbioe.2019.00311.

[35] Åbo Akademi University, Faculty of Social Sciences, Business and Economics, Turku, Finland, L. Davoodi, J. Mezei, et Åbo Akademi University, Faculty of Social Sciences, Business and Economics, Turku, Finland, « A Comparative Study of Machine Learning Models for Sentiment Analysis: Customer Reviews of E-Commerce Platforms », in 35 th Bled eConference Digital Restructuring and Human (Re)action, University of Maribor Press, 2022, p. 217-231. doi: 10.18690/um.fov.4.2022.13.

[36] A. Hernandez-Guedes, I. Santana-Perez, N. Arteaga-Marrero, H. Fabelo, G. M. Callico, et J. Ruiz-Alzola, « Performance Evaluation of Deep Learning Models for Image Classification Over Small Datasets: Diabetic Foot Case Study », IEEE Access, vol. 10, p. 124373-124386, 2022, doi: 10.1109/ACCESS.2022.3225107.

[37] P. Wang, X. Huang, W. Qiu, et X. Xiao, « Identifying GPCR-drug interaction based on wordbook learning from sequences », BMC Bioinformatics, vol. 21, no 1, p. 150, avr. 2020, doi: 10.1186/s12859-020-3488-8.

[38] W. Qiu, Z. Lv, Y. Hong, J. Jia, et X. Xiao, « BOW-GBDT: A GBDT Classifier Combining With Artificial Neural Network for Identifying GPCR–Drug Interaction Based on Wordbook Learning From Sequences », Front. Cell Dev. Biol., vol. 8, févr. 2021, doi: 10.3389/fcell.2020.623858.

[39] J. Zheng, X. Xiao, et W.-R. Qiu, « DTI-BERT: Identifying Drug-Target Interactions in Cellular Networking Based on BERT and Deep Learning Method », Front. Genet., vol. 13, juin 2022, doi: 10.3389/fgene.2022.859188.

[40] A. W. Senior et al., « Improved protein structure prediction using potentials from deep learning », Nature, vol. 577, no 7792, Art. no 7792, janv. 2020, doi: 10.1038/s41586-019-1923-7.

UNDER PEER REVIEW IN IJAR