Clustering-Based Negative Sampling Approaches for Protein-Protein Interaction Prediction

Zehra Kesemen[1(✉)], İlknur Karadeniz[1,2], and Reyhan Aydoğan[1,2,3]

[1] Computer Science, Özyeğin University, Istanbul, Turkey
zehra.kesemen@ozu.edu.tr,
{ilknur.karadeniz,reyhan.aydogan}@ozyegin.edu.tr
[2] Artificial Intelligence and Data Engineering, Özyeğin University, Istanbul, Turkey
[3] Interactive Intelligence Group, Delft University of Technology, Delft, The Netherlands

Abstract. The lack of confirmed negative interactions poses a major challenge to the prediction of protein-protein interactions. The reliable selection of these negative samples within a dataset is crucial for a better understanding of the underlying patterns and dynamics. The random sampling method is the most widely used negative sampling method, where negative pairs are randomly selected from unlabelled samples (i.e., samples not experimentally confirmed as positive interactions). However, they tend to introduce inaccurately labelled negative samples, resulting in less reliable predictions, which may affect the efficiency of the learning process. Our study aims to assess the reliability of clustering-based negative sampling methods and highlight their fundamental differences from the widely used random sampling method. To achieve this goal, we propose a hierarchical clustering-based algorithm that uses different mechanisms to select negative instances from unlabelled instances. We investigated the effectiveness of our proposed approach compared to existing clustering-based negative sampling methods and random sampling on four different datasets. The results indicate that clustering-based methods surpass the commonly used random sampling method.

Keywords: Host-pathogen interactions · Negative sampling strategy · Machine learning methods · Binary classification

1 Introduction

Viral infections result from the interaction between pathogen proteins and host proteins. The extraction of these protein interactions is crucial for understanding the mechanisms of infection. However, the experimental extraction of protein interactions faces several challenges including the complicated nature of the domain that requires expert knowledge, high experimental costs, and time constraints. Therefore, automated extraction of pathogen-host protein interactions (PHI) using computational methods has become an increasingly important

L. Cerulo et al. (Eds.): CIBB 2024, LNBI 15276, pp. 3–14, 2025.
https://doi.org/10.1007/978-3-031-89704-7_1

research topic in recent years [9]. One of the major challenges in the automated extraction of PHI interactions is the lack of experimentally verified negative samples for non-interacting protein pairs, although positive samples with experimentally confirmed interacting protein pairs are available. Therefore, the selection of reliable negative samples is crucial for building satisfactorily generalisable prediction models to gain a better understanding of pathogen-host protein interactions. In the literature, most of the currently available studies use a random sampling method [7,8,10,13,14,18], where negative pairs are randomly selected from the unlabelled samples. However, this approach tends to introduce inaccurately labelled negative samples, resulting in less reliable predictions, which may affect the efficiency of the learning process and reduce the performance of predictive models [6]. In addition, the reproduction of the results might been affected negatively due to the randomization in the selection process.

In this paper, we present a novel cluster-based approach for negative sampling. Then, we explore existing cluster-based negative sampling methods in the literature to compare them with the widely used random sampling method. The intuition of our method is to avoid selecting an unknown positive sample as a negative sample as much as possible. We aim to select unknown samples with different characteristics by forming clusters of positive samples and selecting a few examples from each cluster. To see how the similarity of the negative samples to the positive samples is affected, we investigate four different alternative selection mechanisms in which the samples are selected from the clusters depending on their distance from the centroid of the positive sample clusters. These mechanisms include selecting the closest samples, the farthest samples, samples selected uniformly from cluster centroids, and selecting both the closest and the farthest samples. We empirically compare our approach with the existing cluster-based selection approaches and the random sampling method. The results show that clustering-based methods perform better than the random sampling method.

The rest of the paper is organized as follows. Section 2 introduces the problem and highlights a recent study that serves as the basis for this paper. Section 3 describes the data sets used in this study. The proposed new clustering-based method (CNS) is explained in detail in Sect. 4. Section 5 presents an overview of the benchmark methods, experimental results, and performance comparisons. Finally, Sect. 6 concludes the paper and points to future research directions.

2 Problem Statement

As widely used random sampling method generates negative samples for host-pathogen protein-protein interactions, which randomly selects from the unlabelled pairs (i.e. the pairs we cannot ensure about their interactions as they were not experimentally verified). However, this method may lead to a higher number of false negatives, potentially affecting the learning process and reducing the sensitivity in predicting protein-protein interactions. Regarding the negative sampling strategy, the biggest risk is to select unlabelled samples as negative, although there is an interaction between these protein pairs that has not

yet been discovered by scientists. This can mislead the prediction model and lead to potential interactions being overlooked. To address this issue, the main task of this study is to identify unlabelled protein pairs through a novel cluster-based negative sampling approach and evaluate their performance against existing methods.

This study is based on a recently published article Koca *et al.* [7], which obtained state-of-the-art results on the prediction of protein-protein interactions between human and virus proteins. Their study incorporates the topological properties into the amino acid embeddings as part of the graph convolution process using the GraphSAGE model [4]. For encoding of protein sequences, they used Doc2Vec [11] along with the Byte Pair Encoding (BPE) method [2] to convert variable-length text segments into vector representations as the amino acid sequences are considered as documents.

The overall workflow, as illustrated in Fig. 1, shows how these topological properties and sequence embeddings are integrated into the prediction model. Our study adopted this workflow for the protein-protein interaction prediction problem in order to examine the effect of negative sampling strategies.

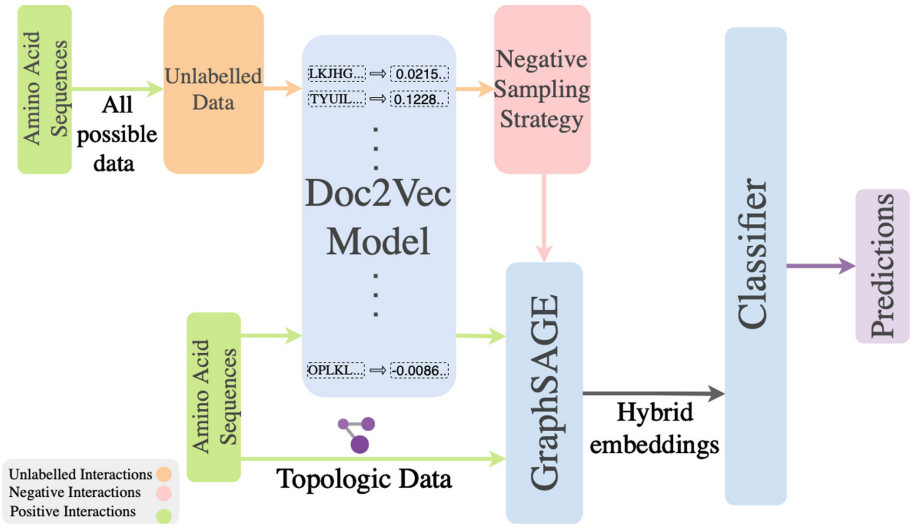


Fig. 1. Prediction framework for protein-protein interaction between human and virus proteins, as proposed by Koca *et al.* [7].

3 Datasets

The main dataset of this study, as used by Koca et al. [7], is the PHISTO dataset [1], which contains 39, 544 interactions, including 6, 571 viral proteins and

1,715 human proteins. Although this dataset provides a comprehensive basis for the analysis, additional datasets were utilized to ensure the reliability and generalizability of the methods, as detailed below:

- **Gordon's SARS-CoV-2 dataset** is based on the SARS-CoV-2 and human protein-protein interaction, as detailed in the study by [3]. This study identified 332 high-confidence protein-protein interactions, involving 332 SARS-CoV-2 proteins and 27 human proteins.
- **Human-Virus dataset** which includes 8,929 interactions from the DeepTrio study [5] is used as a benchmark dataset. It originally compiled DeepViral study by Liu-Wei et al. [10].
- **Tsukiyama et al. dataset** from the study [15] includes 22,383 interactions, comprising 5,882 viral proteins and 996 human proteins.

4 Proposed Approach: Clustering-Based Negative Sampling Strategy (CNS)

In this section, we propose a novel clustering-based sampling method to select reliable negative samples from the unlabelled samples. Our intuition for this method is to avoid selecting an unknown positive sample as a negative sample as much as possible. Therefore, we aim to select unknown samples with different characteristics by forming clusters of positive samples and selecting samples from each cluster. To minimize the randomness in sample selection, we propose a clustering-based negative sampling strategy employing Agglomerative Clustering [12]. In contrast to the K-means method, Agglomerative Clustering is deterministic, consistently producing the same clustering structure when applied to the same dataset. We explain our proposed method step by step, as illustrated in Fig. 2. We apply the Agglomerative Clustering Algorithm to the positive sample set P (i.e., experimentally confirmed pairs). Agglomerative Clustering is an unsupervised data mining technique used to create a hierarchy of clusters. It operates in a bottom-up manner, where each sample starts as its cluster, and clusters are progressively merged based on similarity, resulting in a hierarchy. Our algorithm organizes the positive sample set P into different k clusters. The positive sample set P is represented in Eq. 1, where each cluster is denoted as C_n. The centroid of each cluster, $C_{n\text{-center}}$, serves as a representative of the corresponding cluster, which is calculated by averaging the positive samples within the cluster.

$$P = \bigcup_{n=1}^{k} C_n, \quad n \in \{1, 2, \ldots, k\}. \tag{1}$$

The next step is to assign the unlabelled protein sequences from the set U to the closest cluster. Since Agglomerative Clustering does not support assigning unseen data to existing clusters and is only capable of clustering the training data, we assigned those sequences in a similar way to K-means. That

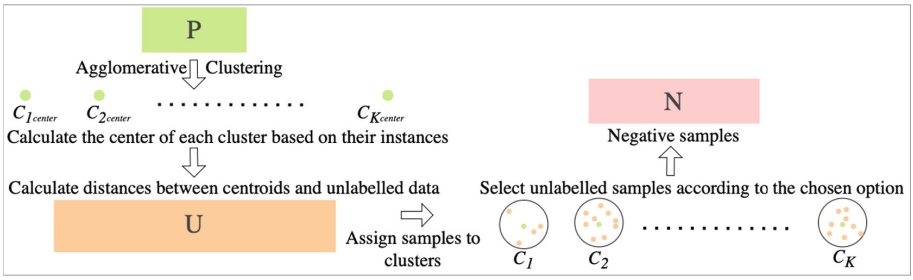


Fig. 2. The process of our clustering-based sampling approach.

is, we assigned the protein sequences to the clusters based on their similarity to the cluster centroids. To assess the similarity between positive protein sequences x and unlabelled protein sequences y, we use three distance metrics such as $d_{\text{Canberra}}(x, y)$, $d_{\text{Euclidean}}(x, y)$ and $d_{\text{Cosine}}(x, y)$, which are calculated according to the corresponding equations (See Eqs. 2–4). In all equations, $x = (x_1, x_2, \ldots, x_m)$ stands for the vector of attributes of the cluster centroid and $y = (y_1, y_2, \ldots, y_m)$ for the vector of attributes of the unlabelled sample.

- **Canberra distance** normalizes the differences between elements by their sum, making it effective for measuring relative differences.
- **Euclidean distance** measures the straight-line distance between two points in a multidimensional space and calculates the absolute geometric distance between the cluster centroid and the unlabelled sample.
- **Cosine distance** evaluates how the vectors are in terms of direction. It assesses the cosine of the angle between two vectors x and y, where $x \cdot y$ is the dot product of the vectors, and $|x|$ and $|y|$ represent Euclidean norms.

$$d_{\text{Canberra}}(x, y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{2}$$

$$d_{\text{Euclidean}}(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{3}$$

$$d_{\text{Cosine}}(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \tag{4}$$

We pose the question of which unlabelled samples should be selected as negative samples from each cluster. To address this question, we propose four methods for utilizing the cluster structure in the selection process:

- **Closest selection** selects the elements closest to the centroids to ensure that the most representative elements of each cluster are selected.
- **Farthest selection** selects the elements farthest from the centroids, therefore, the more outlier instances within each cluster are captured.

– **Closest and Farthest selection** selects both the closest and farthest elements to create a balance between representative and diverse elements.
– **Uniform selection** ensures an equitable distribution by selecting elements evenly across all clusters (See Fig. 3 for detailed information).

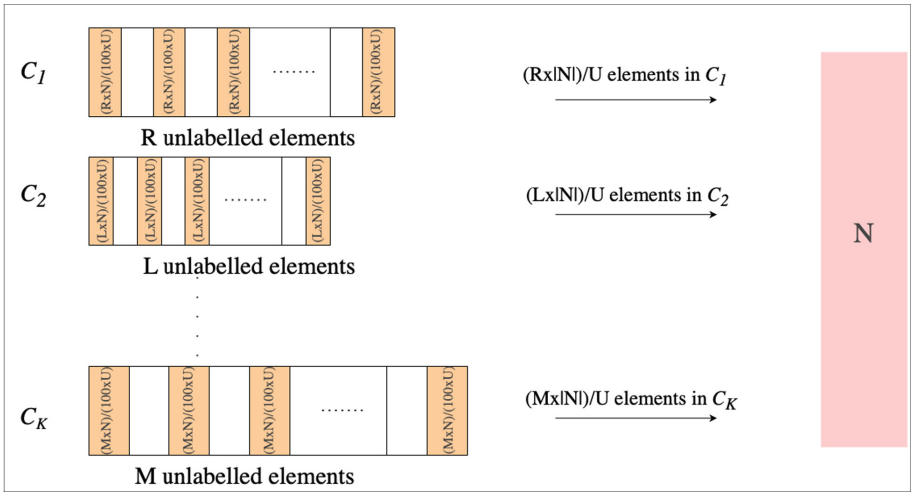


Fig. 3. The sets $R, L, \ldots, M$ denote unlabelled samples, with a total sum of U. The equation $\frac{R\times|N|}{U} + \frac{L\times|N|}{U} + \ldots + \frac{M\times|N|}{U} = |N|$ ensures proportional selection from each set, where $|N|$ is the total number of negative samples required.

We investigate the degree of similarity between the unlabelled samples and their positive centroids according to the selection mechanism to understand how these unlabelled samples are positioned within their assigned clusters. In each selection method, the number of elements chosen from each cluster is adjusted to be proportionate to its size. Here, $|N|$ denotes the total number of negative samples we seek to select, while $|C_n|$ signifies the number of elements within a given cluster where $n \in (1, k)$. The number of elements chosen from this cluster is determined by the $|N|$ to $|C_n|$ ratio, guaranteeing proportional representation relative to cluster size.

5 Evaluation

There are two notable studies in the literature [16,17] that use clustering-based methods for negative sample selection, which we compare with our proposed model. We compare our approach with these clustering-based strategies for negative sampling. *Wang et al.* [16] propose a clustering-based negative sampling strategy, in which all unlabelled data U are clustered using the K-means clustering algorithm, while our approach clusters the positive samples P using

hierarchical clustering. In Wang's study, the number of unlabelled samples to be selected from each cluster is determined based on the proportion of samples within the cluster relative to the total unlabelled data. This ensures that the negative samples are proportionally distributed across clusters. Finally, the samples closest to the center of each cluster are selected based on the estimated number of instances. On the other hand, *Wei et al.* [17] suggest applying the MiniBatchKMeans clustering algorithm to the entire dataset consisting of both positive interactions P and unlabelled interaction sets U. The ratio of unlabelled examples within each cluster for the entire cluster is calculated and the clusters are sorted in descending order according to this ratio. The unlabelled instances are extracted from the clusters whose unlabelled instance density is the highest regarding the aforementioned sort operation.

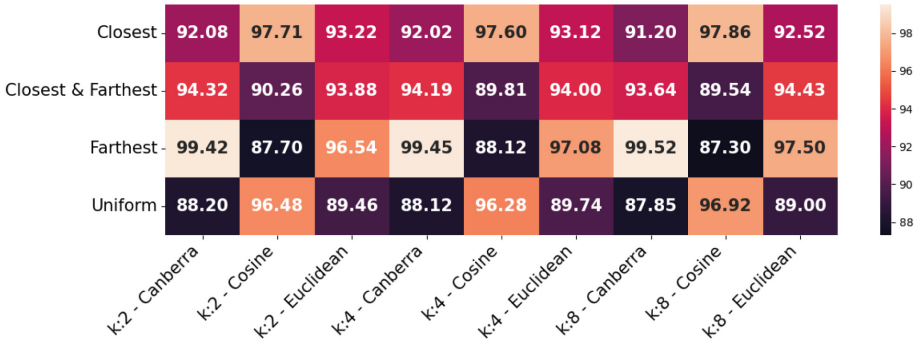
To assess the effectiveness of each clustering-based approach for negative sample selection, a binary classifier is trained to predict protein interaction (i.e., positive and negative samples are labelled 1 and 0 respectively). To increase the reliability of the performance evaluation, we applied 5-fold cross-validation technique in which the dataset is divided into five subsets and each subset in turn serves as a test set, to obtain a comprehensive evaluation of the overall performance of the model. We chose a ratio of 1 : 10 for positive and negative samples, as recommended in previous research in this field reflecting the typical imbalance in real-world data. For the Agglomerative Clustering Algorithm, we used the average linkage method, which tends to produce more balanced clusters.

5.1 Effect of Distance Metrics on Method's Performance

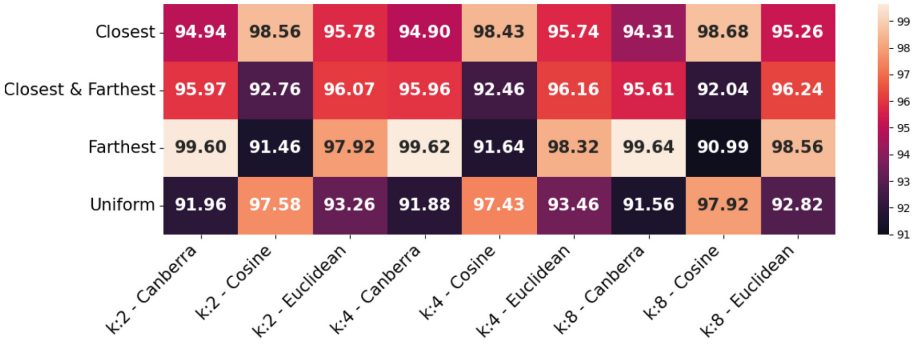
The performance of the proposed sampling methods may vary depending on parameters such as the number of clusters, distance metrics, and selection mechanisms. We conducted further analyses to understand the impact of these parameters. First, the impact of distance metrics on the performance of our method was evaluated using the PHISTO dataset. In this phase, the GA^{2}M classifier was selected aligning with the work of Koca *et al.* [7]. Figure 4(a) shows the recall metric (i.e., the percentage of true positive predictions compared to the confirmed positive samples) where the y-axis represents the selection criterion, the x-axis shows the number of clusters (k) and the type of distance metric used. As can be seen in Fig. 4, the choice of distance metric has significant effects on recall and F1 scores. Regardless of the number of clusters, it can be seen that the Canberra distance performs better than other methods with the farthest selection. Based on these results, the Canberra distance was used as the distance metric for our method in the further experimental studies below.

5.2 Effect of Cluster Numbers

We analyzed the effects of different cluster numbers, including 2, 4, and 8, on the PHISTO dataset. As can be seen in Table 1, the optimal cluster number varied depending on the method. For example, the best performance for the Closest, Closest and Farthest, and Uniform methods was achieved with 2 clusters, while



(a) Recall scores.



(b) F1 scores.

Fig. 4. Performance metrics of GA^{2}M on the PHISTO dataset.

the Farthest method performed best with 8 clusters. Similarly, Wang's method achieved the highest F1 and recall values with 4 clusters, while Wei's method showed the best results with 8 clusters. Consequently, the optimal number of clusters for each method was fixed based on their performance in the following analyses.

5.3 Analysis of Classifier Performance

In this section, we assess the performance of different classifiers with the sampling strategies. Table 2 shows the F1 and recall values of each classifier. We used the classifiers aligned with Koca's study [7], including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Generalized Additive 2 Model (GA^{2}M). The results show that the proposed approach with the farthest selection slightly outperforms the other strategies. Of the four classifiers, the Random Forest classifier outperformed the others, showing slightly better results than GA^{2}M. Therefore, we advocated using the Random Forest.

Table 1. The F1 and recall values in percent for the PHISTO dataset with different cluster numbers.

Cluster Numbers	$k = 2$		$k = 4$		$k = 8$	
	F1	Recall	F1	Recall	F1	Recall
Closest	**94.94**	**92.08**	94.90	92.02	94.31	91.20
Closest & Farthest	**95.97**	**94.32**	95.96	94.19	95.61	93.64
Farthest	99.60	99.42	99.62	99.45	**99.64**	**99.52**
Uniform	**91.96**	**88.20**	91.88	88.12	91.56	87.85
Wang's method	91.56	87.85	**96.18**	**93.95**	94.28	91.58
Wei's method	97.73	95.81	97.84	96.08	**99.22**	**98.90**

Table 2. The F1 and recall values in percent for the PHISTO dataset with different classifiers and the optimal cluster number for each method.

Classifiers	GA^{2}M		LR		SVM		RF	
	F1	Recall	F1	Recall	F1	Recall	F1	Recall
Closest	**94.94**	**92.08**	92.76	89.30	94.33	91.28	94.90	91.46
Closest & Farthest	95.97	**94.32**	93.42	91.18	95.97	94.20	**96.34**	94.05
Farthest	**99.64**	**99.52**	99.49	99.24	99.58	99.40	**99.64**	99.44
Uniform	91.96	**88.20**	88.60	83.67	91.30	87.05	**92.36**	87.72
Wang's method	96.18	**93.95**	94.00	90.76	95.45	92.96	**96.44**	93.91
Wei's method	99.22	98.90	98.88	98.37	98.99	98.36	**99.36**	**99.10**
Random sampling	73.32	67.94	62.34	54.44	70.97	63.06	**79.39**	**73.52**

5.4 Performance Across Other Datasets

To evaluate the effectiveness of the sampling strategies across different pathogen-host protein-protein interaction datasets, we further analyze the performance of the Random Forest classifier on three additional datasets, as described in Sect. 3. Table 3 presents the results for cluster numbers 2, 4, and 8 across all datasets. Based on the results in Table 3, Wei's method consistently achieved the highest performance across most datasets and cluster numbers for Random Forest. In the PHISTO dataset, our method with the farthest selection outperformed other methods and reached an F1 score of 99.64% with Random Forest at $k = 8$.

It is worth noting that there is no superior sampling strategy resulting in terms of prediction accuracy in all datasets. On the other hand, the results show that all cluster-based sampling approaches outperformed random sampling. Therefore, we suggest that the clustering-based method is preferable to the random sampling method in this domain.

Table 3. Performance of the Random Forest classifier across all datasets with different cluster numbers.

Datasets	Cluster Numbers	Farthest		Wang's Method		Wei's Method		Random Sampling	
		F1	Recall	F1	Recall	F1	Recall	F1	Recall
PHISTO	$k = 2$	**99.62**	**99.30**	95.96	93.01	97.15	94.71		
	$k = 4$	**99.62**	**99.40**	96.44	93.91	97.39	95.36	79.39	73.52
	$k = 8$	**96.64**	**99.44**	94.53	90.94	99.36	99.10		
Tsukiyama et al.	$k = 2$	96.02	92.68	96.17	93.06	**96.62**	**93.78**		
	$k = 4$	95.85	92.62	96.42	93.44	**96.50**	**93.39**	68.84	59.80
	$k = 8$	96.06	92.62	93.78	89.24	**97.60**	**95.54**		
Human-Virus	$k = 2$	92.34	87.04	92.12	86.52	**95.10**	**91.36**		
	$k = 4$	92.06	86.88	89.76	83.53	**96.54**	**93.68**	72.32	64.16
	$k = 8$	92.31	87.12	89.00	82.48	**97.52**	**95.74**		
SARS-CoV-2	$k = 2$	69.26	55.92	70.96	55.70	**87.32**	**79.84**		
	$k = 4$	67.58	55.48	68.80	53.96	**80.89**	**69.30**	67.40	55.51
	$k = 8$	84.08	75.79	76.74	64.20	**91.99**	**85.64**		

6 Conclusion

In this study, we address the challenge of selecting negative samples from unlabelled samples for pathogen-host protein-protein interactions. We propose a novel clustering-based approach to enhance the reliability of negative sample selection. We highlight the fundamental differences between our method and other clustering-based approaches, as well as the widely used random sampling method. Our experiments, conducted on four different datasets of virus-human protein interactions, employ four selection methods and four classifiers. The results demonstrate that clustering-based approaches outperform the widely used random sampling method. We believe that providing a more precise definition of protein feature representation will result in more distinguishable outcomes in the selection of negative samples. As future work, it would be interesting to analyze the characteristics of the datasets and their correlation with the obtained results on that dataset regarding the effectiveness of the negative sampling methods.

Acknowledgments. The authors would like to thank Mehmet Burak Koca for his support in this work. Special thanks to Eymen Küçükçakır for his contribution and to Amin Deldari Alamdari for his technical support. Zehra Kesemen was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) through the 2210-C National M.Sc. Scholarship Program in the Priority Fields and Science, and the project grant 120N680.

References

1. Durmuş Tekir, S., et al.: PHISTO: pathogen-host interaction search tool. Bioinformatics **29**(10), 1357–1358 (2013). https://doi.org/10.1093/bioinformatics/btt137
2. Gage, P.: A new algorithm for data compression. C Users J. Arch. **12**, 23–38 (1994). https://api.semanticscholar.org/CorpusID:59804030
3. Gordon, D.E., et al.: A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature **583**(7816), 459–468 (2020). https://doi.org/10.1038/s41586-020-2286-9
4. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. arXiv:1706.02216 (2018)
5. Hu, X., Feng, C., Zhou, Y., Harrison, A., Chen, M.: Deeptrio: a ternary prediction system for protein-protein interaction using mask multiple parallel convolutional neural networks (2021). https://doi.org/10.1093/bioinformatics/btab737
6. Iuchi, H., et al.: Bioinformatics approaches for unveiling virus-host interactions. Comput. Struct. Biotechnol. J. **21**, 1774–1784 (2023). https://doi.org/10.1016/j.csbj.2023.02.044
7. Koca, M.B., Nourani, E., Abbasoğlu, F., Karadeniz, İ., Sevilgen, F.E.: Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses. Comput. Biol. Chem. **101**, 107755 (2022). https://doi.org/10.1016/j.compbiolchem.2022.107755, https://www.sciencedirect.com/science/article/pii/S1476927122001359
8. Kshirsagar, M., Carbonell, J., Klein-Seetharaman, J.: Multitask learning for host-pathogen protein interactions. Bioinformatics **29**(13), i217–i226 (2013). https://doi.org/10.1093/bioinformatics/btt245
9. Lian, X., Yang, X., Yang, S., Zhang, Z.: Current status and future perspectives of computational studies on human–virus protein-protein interactions. Briefings Bioinform. **22**(5), bbab029 (2021). https://doi.org/10.1093/bib/bbab029
10. Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N.J., Tegnér, J., Hoehndorf, R.: DeepViral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. Bioinformatics **37**(17), 2722–2729 (2021). https://doi.org/10.1093/bioinformatics/btab147
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)
12. Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J. Classif. **274–295**(3) (2014). https://doi.org/10.1007/s00357-014-9161-z
13. Sledzieski, S., Singh, R., Cowen, L., Berger, B.: D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. Cell Syst. **12**(10), 969-982.e6 (2021). https://doi.org/10.1016/j.cels.2021.08.010
14. Song, B., Luo, X., Luo, X., Liu, Y., Niu, Z., Zeng, X.: Learning spatial structures of proteins improves protein–protein interaction prediction. Briefings Bioinform. **23**(2), bbab558 (2022). https://doi.org/10.1093/bib/bbab558
15. Tsukiyama, S., Hasan, M.M., Fujii, S., Kurata, H.: LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. Briefings in Bioinformatics **22**(6) (2021). https://doi.org/10.1093/bib/bbab228
16. Wang, B., et al.: Imbalance data processing strategy for protein interaction sites prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. **18**(3), 985–994 (2021). https://doi.org/10.1109/tcbb.2019.2953908

17. Wei, Z., Yao, D., Zhan, X., Zhang, S.: A clustering-based sampling method for mirna-disease association prediction. Front. Genet. **13** (2022). https://doi.org/10.3389/fgene.2022.995535
18. Yang, X., Yang, S., Li, Q., Wuchty, S., Zhang, Z.: Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput. Struct. Biotechnol. J. **18**, 153–161 (2020). https://doi.org/10.1016/j.csbj.2019.12.005