Ingénierie des Systèmes d'Information
Vol. 31, No. 4, April, 2026, pp. 1231-1238
Journal homepage: http://iieta.org/journals/isi

Alignment-Free SARS-CoV-2 Protein Sequence Analysis Using K-mer Representation and Latent Topic Modeling

Sura Z. Alrashid*, Hawraa S. Hamza, Sarah Hayder Hashim, Huda Kadhum Ayoob

College of Information Technology, University of Babylon, Babylon 51002, Iraq

Corresponding Author Email: sura.alrashid@uobabylon.edu.iq

https://doi.org/10.18280/isi.310418

Received: 15 January 2026
Revised: 18 March 2026
Accepted: 5 April 2026
Available online: 30 April 2026

Keywords:
SARS-CoV-2, protein sequence analysis, alignment-free methods, k-mer representation, Latent Dirichlet Allocation, topic modeling, bioinformatics

ABSTRACT

Protein sequence analysis is essential for understanding viral structure, functional motifs, and potential host–virus interaction mechanisms. This study proposes an alignment-free machine learning framework for SARS-CoV-2 protein sequence analysis using k-mer representation and latent topic modeling. Protein sequences were obtained from the Genome Warehouse (GWH) Virus Database and organized into a dataset of 1,000 sequences, with 60% used for model training and 40% for testing. After preprocessing FASTA records, ambiguous amino acid residues were removed, and each sequence was represented through physicochemical descriptors, Bag-of-Words features, one-hot encoding, and overlapping 3-mer tokens. Latent Dirichlet Allocation (LDA) was then used to discover recurring motif-like topic structures from the k-mer feature space. The learned topic distributions were further examined through dominant topic assignment, motif-weight analysis, t-SNE visualization, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)-based clustering. The results show that several latent topics captured repetitive amino acid patterns, including motifs enriched with lysine and glutamic acid residues, which may be associated with structural stability, protein-binding interfaces, or functional domains. The t-SNE visualization indicated separable groups of protein sequences according to dominant topic distributions, while motif-weight analysis provided interpretable evidence for the contribution of specific sequence fragments. Overall, the proposed framework offers a computationally efficient and interpretable route for exploratory SARS-CoV-2 protein sequence analysis without relying on pairwise sequence alignment. Future work should validate the biological relevance of the extracted motifs using larger datasets and experimentally annotated protein interaction records.

1. INTRODUCTION

A global health issue, the COVID-19 disease can cause critical respirational failure or occur asymptomatically or with symptoms of virus-related pneumonia. It spreads effortlessly from person to person by contact and droplets. The SARs-CoV2 virus has caused a considerable number of fatalities worldwide and has spread far more quickly and extensively than previous viruses [1-3].

Disease-causing pathogens, such as viruses, introduce their proteins into host cells, where they interact with the host's proteins to facilitate virus-related replication. Understanding these protein-protein interactions (PPIs) is fundamental to deciphering the mechanisms of infectious diseases, Viruses like SARS-CoV-2 interact with host proteins by altering their natural pathways to facilitate viral replication. These interactions often include hijacking the host's protein machinery, disrupting cellular processes such as signal transduction, immune responses, and protein synthesis. The biological proteins target specific host proteins through sequence-based compatibility, influencing their structure and function, which may lead to pathological outcomes [4].

All across the tree of life, proteins are the final machines of nature. Even while our understanding of protein sequences is growing at an exponential rate, one of the biggest scientific problems of our day is still figuring out how they work, which has many health implications. Protein sequences can be thought of as letters made up of amino acids. Predictive protein tasks are therefore a perfect fit for machine-learning techniques created for natural language and other sequences [5]. Protein sequence analysis is an essential field of study in virology and bioinformatics since it has become a crucial tool for comprehending the molecular causes of diseases [6].

Machine learning (ML) plays a vital role in bioinformatics by qualifying both classification and clustering of genes, DNA, and proteins. Classification, a supervised learning approach, is used to assign labels to biological data (for example, identifying gene functions, detecting disease-related DNA sequences, or predicting protein types) by means of algorithms such as support vector machines (SVM), random forests, and neural networks. Inversely, clustering is an unsupervised approach that groups similar sequences to uncover hidden patterns, (such as, preserved DNA motifs, protein families or gene co-expression clusters). Techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), hierarchical clustering, and k-means are

generally employed for this objective. Machine Learning approaches adopt disease prediction, functional annotation, and a deeper consideration of complex biological systems [7-10].

In this approach, the application of the LDA2vec algorithm is explored to analyze and cluster SARs-CoV2 protein sequences and their interactions with host proteins. Using the dataset existing at GWH Virus Database, the aim of this study is to detect meaningful patterns and dependencies within the interactions where the contributions are as follows:

(1) Utilizing SARS-CoV-2 protein sequence, applied Latent Dirichlet Allocation (LDA) topic modeling to discover recurring motif-like patterns.
(2) Deployment LDA2-vec for an alignment-free and ML-based approach for protein sequence analysis.
(3) Identified main motif features and influential topics that adopt functional annotation, biomarker discovery, and deeper realizing of protein interactions.

2. RELATED WORK

With the development of machine learning techniques that improve our knowledge of protein structures and functions, protein sequence Analysis has emerged as a critical area of study in Bioinformatics and computational biology. Recent research has demonstrated effectiveness of many machine learning methodologies in identifying and classifying proteins, which is essential for applications in drug discovery and disease treatment.

Borsari et al. [11] employed common information for gene selection to find nonlinear correlations and utilized the K-Means approach for clustering. The proposed hybrid approach efficiently refined gene expression data, progresses cluster quality, controls the noise, and minimizes computing space.

Le and Kha [12] found that vesicular transport proteins can be identified by sophisticated computational methods. By combining position-specific scoring features with amino acid compositions, an ensemble model produced maximum accuracy and sensitivity; they had a noteworthy classification performance by using a Gated Recurrent Unit model.

In addition, Draizen et al. [13] highlighted how vital high-quality datasets are for training ML modeled in structural bioinformatics. They unveiled that "Prop3D," is a tool for building and spreading protein domain libraries enhanced with evolutionary and biophysical characteristics. In addition to addressing the issues of data accessibility, their work offered a complete dataset, Prop3D-20sf, which can critically enable the creation of reproducible ML algorithms.

In order to demonstrate the versatility of machine learning techniques in medical diagnostics, Nimmagadda et al. [14], concentrated on applying various ML algorithms to identify chronic kidney disease. Despite their primary effort on chronic kidney disease (CKD), the approaches they discussed - such as classifiers like K nearest neighbor (KNN) and SVM - can be useful in protein sequence analysis.

Brierley and Fowler [15] examined the biases in the genomic composition of spike proteins and entire genome sequences from various coronaviruses using random forest models. Their results showed that these biases were around 73% accurate in predicting animal hosts.

Kshirsagar et al. [4] identified the function of protein sequences in the binding interactions between human proteins and SARS-CoV-2. Despite class imbalances, their supervised machine learning models showed a respectable prediction performance with an area under the precision-recall curve (AUC-PR) of 0.65. The significance of particular amino acid sequences in mediating viral contacts is highlighted in this work, which may help guide therapeutic approaches meant to break these relationships. There has also been an increase in the use of (AI) in COVID-19 diagnostic procedures.

Mano et al. [16], and Bhosale and Patnaik [17], presented a systematic analysis of AI methods for COVID-19 diagnosis, highlighting the use of imaging methods like chest X-rays and CT. They are illustrious that the majority of image classification tasks were handled by convolutional neural networks, demonstrating how deep learning can improve diagnostic precision. The results show how important it is to use AI to optimize diagnostic processes, which is especially important when it comes to dealing with patient care during pandemics.

Kockelbergh et al. [18] stated that the use of bulk Latent T-cell receptor (TCR) repertoire sequencing to evaluate SARs-CoV2 immune responses. They pointed out that by identifying particular clonotypes linked to efficient immune responses, this method offers insights into the dynamics of T-cell activation during infection. Monitoring TCR diversity and its relationship to disease severity may help guide the development of vaccines and treatment plans.

Kosar et al. [19] innovated knowledge of COVID-19 diagnoses by methodically investigative a range of methods, including RT-PCR, antigen testing, and imaging modalities. According to their analysis, AI is vital for increasing the precision and effectiveness of diagnosis, particularly when it comes to chest imaging by grouping and conflicting these methods.

In order to identify practical vaccine candidates based on how closely they matched authorized vaccinations, Gharaibeh and Doncker [20] used natural language processing and unsupervised learning algorithms. This Enhanced approach establishes how machine learning can speed up the creation of vaccines, which is important for battling the pandemic.

To sum up, integrating ML methods with protein sequence analysis offers a practical way to advance our understanding of SARs-CoV2 and its interactions. In this study, the application of the LDA2vec algorithm explored to cluster and analyze SARs-CoV2 protein sequences and their interactions with host proteins. Using the dataset existing at GWH Virus Database: Coronaviridae. The aim of this approach is to discover meaningful patterns and dependencies within the interactome.

Recent protein sequence analysis methods based on Transformer architectures, such as ProteinBERT and ProtBERT, generally outperform traditional topic-modeling approaches like LDA2vec by capturing long-range contextual dependencies and complex biochemical patterns through self-attention mechanisms. But, LDA2vec remains computationally efficient and more interpretable, making it suitable for smaller datasets and resource-limited environments. Therefore, the proposed LDA2vec approach can be considered an efficient and interpretable alternative within modern protein representation learning research [5, 21].

3. GENOME WAREHOUSE VIRUS DATASET

The GWH Virus Database was demonstrated as the main

source of protein sequences in this approach. To ensure reliable training and evaluation of models, the dataset was parted into training and testing subsets. The training set was utilized to construct the LDA-based topic model and extract latent sequence motifs; testing set was managed to evaluate the generalization capability of the proposed approach. Table 1 describes the dataset structure and splitting ratios employed in this approach.

Table 1. Genome Warehouse (GWH) virus dataset

Dataset	Number of Samples	Percentage (%)	Description
Total Dataset	1000	100%	Total protein sequences obtained from the GWH Virus Database
Training Set	600	60%	Protein sequences managed to train the Latent Dirichlet Allocation (LDA) model and obtain latent topics
Testing Set	400	40%	Protein sequences managed to evaluate the model performance and topic distribution

4. FREE ALIGNMENT METHODS

The commonly used algorithms in the field of bioinformatics and Computation Biology systems are alignment methods, which measure the degree of functional similarity between sequences whether they are protein sites or DNA or RNA, including their structure, function, and evolutionary relationships [22]. Global alignment and local alignment are the two groups of alignment techniques. The most advanced and refined technique that operates along the entire length of all sequences is global alignment.

Due to its ability to find similar regions within lengthy sequences that are typically varied in a wide range, local alignment is recommended over global alignment in this instance. The computational complexity involved in identifying regions of similarity between sequences is the limitation of local alignment [23].

Alignment methods are not able to interpret biological data due to its heterogeneity and substantial amount, which is growing every year. Thus, in order to decrease the high dimensionality and fast implementation time, new analysis systems, such as Free Alignment approaches, must be developed [24].

4.1 One hot encoding

The process of transforming category variables into a format that machine learning algorithms are able to use is known as one-hot encoding [25]. One type of encoding technique is one-hot encoding. It includes extending the original feature vector to a multidimensional matrix, where each dimension corresponds to a distinct state and the matrix's dimension is the number of states in the feature. As a result of this processing, only one feature matrix dimension (typically 1 is stated for a particular state, while all other state dimensions are zero) [26].

	T	T	T	G	A	C	T	C	G	T
A	0	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	0	1	0	0
G	0	0	0	1	0	0	0	0	1	0
T	1	1	1	0	0	0	1	0	0	1

Figure 1. Example of one-hot encoding of the DNA sequence "TTTGACTCGT" [27]

One-hot encoding should be used to convert a single DNA sequence into a two-dimensional matrix. A character from the original DNA sequence is represented by each column in the generated matrix. While the other locations in the same column are filled with 0, the number 1 appears where it represents the appropriate character. Figure 1 shows an example of one-hot encoding. A, C, G, and T are represented, respectively, by the first through last positions in the column. Each character in the original DNA sequence is represented by four channels when one-hot encoding is used. In Figure 1, the channels are displayed in the same column beneath one another. Since the lengths of the DNA sequences vary, they are all padded to the same length using columns that contain only zeros [27].

4.2 K-mer method

K-mers are enhanced by fragments of sequences, known as words, which could be RNA or DNA sequences. The K-mer serves as a source for DNA sequence assembly, which enhances the manufacturing of metagenomics vaccines and the identification of viruses and disorders based on diverse gene expression [28]. Since nucleotides (such as A, T, G, and C) are made up of k-mers, the K-mer has generally been used in genome computation and sequence analysis. Here, the sequences are transformed into words using the ordering process, and the words are then saved in a bag of words inside every sample. By sliding fixed windows, it is possible to extract all overlapping k-mers, represented by words, from the gene sequence, for instance, when k = 8. k-mer as determined by the formula:

$$n = (l - k) + 1 \quad (1)$$

where, l is the length of the sequence, k is the width of the word and n is times of k-mer, Figure 2 shows the processing of sequences by k-mer analysis [24].

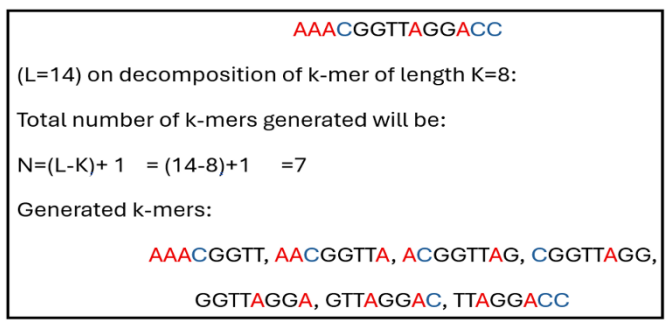

Figure 2. The processing of sequences by k-mer analysis

4.3 Bag of words

Variable selection, variable subset selection, or attribute selection, are other names for the Bag of Words (BoW) or Bag of Features (BoF). It is a versatile model that may be applied to document and image classification as well as feature selection algorithms. Creating bags for each image attribute can be used to classify images, and a BoW, also known as the document's histogram, is a vector of word occurrences used in document classification. Images are handled similarly to papers in the BoW approach, and their features match those of the words in the documents. The histogram of the features in the images can be obtained in three basic steps: "feature detection," "feature description," and "codebook production [29-31].

5. PROPOSED SYSTEM

Our proposed system as shown in Figure 3 consists of the following key stages:

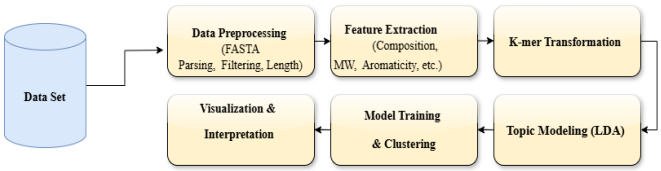

Figure 3. The proposed system

5.1 Data preprocessing

Protein sequences are extracted from FASTA files. These sequences are transformed into structured data, where each protein is represented by its sequence ID, raw sequence, and length. After that, letter filtering is used to remove proteins containing unknown residues (B, X, J, Z).

5.2 Feature extraction

Protein sequences were individually extracted from FASTA files using Biopython and filtered to remove ambiguous amino acid symbols (B, Z, X, and J).

Three main feature representation strategies were employed: physicochemical descriptors, BoW, k-mer and one-hot encoding representations. For the BoW approach, overlapping 3-mer subsequences were created from each protein sequence, producing a vocabulary-based sparse feature matrix. Likewise, physicochemical properties containing amino acid composition, molecular weight, uncertainty index, flexibility, isoelectric point, and secondary structure fractions were extracted using analysis functions of protein sequence.

Combining all features into a unified feature matrix where rows match protein sequences and columns relate to numerical descriptors the counts words in each protein. To reduce scale variation between heterogeneous features, numerical attributes were normalized before clustering and topic modeling.

5.3 K-mer representation and topic modeling

Protein sequences are converted into k-mer representations (subsequences of length k). These sequences are tokenized into words using k-mers (size = 3); A k-mer size of k = 3 was chosen because it provides an effective balance between biological relevance and computational efficiency. Tripeptides are capable of capturing short, conserved motifs and local amino acid relationships that may reflect structural or functional properties of proteins, while maintaining a manageable feature space for topic modeling. Smaller k values may lose important contextual information, whereas larger values increase sparsity and computational complexity. Therefore, 3-mers offer a suitable representation for extracting meaningful latent patterns from protein sequences using LDA.

5.4 Model training and evaluation

The dataset is split into training (60%) and testing (40%) subsets. The LDA model is applied to the training set, and for evaluating the model quality a coherence and perplexity scores were computed. Then the trained model is used to conclude topic distributions for unseen protein sequences.

5.5 Visualization and interpretation

The learned topic distributions are visualized. Additionally, for grouping similar proteins based on feature embeddings a clustering analysis is performed by using DBSCAN.

6. RESULTS AND DISCUSSION

The proposed model determines the potential of combining ML techniques with alignment-free methods for protein sequence analysis. By using the LDA2vec algorithm, it has become possible to cluster SARS-CoV-2 protein sequences and identify meaningful patterns in their interactions with host proteins. Also, the use of k-mer representations and topic modeling provides a new approach to understanding the interactome, offering understandings to the unique and shared features of SARS-CoV-2. Overall, the workflow of this model extracts motifs from protein sequences, models them as probabilistic topics, and clusters sequences based on topic distributions.

6.1 Topic discovery in protein sequences

LDA identifies latent topics representing recurring motif-like n-grams in the protein sequence dataset. Each topic corresponds to a distribution over motifs that frequently co-occur in the dataset. Figure 4 presents the top motif-like 3-mer patterns associated with each of the ten latent LDA topics extracted from the SARS-CoV-2 protein sequence dataset. These words represent conserved amino acid patterns contributing to each topic's identity. Many of the motifs - such as "ELEKE", "LKELE", and "KLEKL" - appear biochemically repetitive and structured, suggesting possible links to secondary structure or functional domains in proteins.

Precisely, the LDA model was trained using 10 topics with iterative optimization until convergence. The topics number was nominated experimentally based on perplexity evaluation and coherence to achieve a balance between model quality and interpretability. Additionally, main hyperparameters, such as alpha, were regulated empirically to optimize performance of sequence representation and topic separation.

6.2 Dominant topic assignment per sequence

After training the LDA model, each sequence is appointed

to a prevalent topic based on the highest topic probability. The distribution of topics among sequences is summarized in Figure 5 illustrates the dominant topic assignment for descriptive protein sequences, containing the associated motif patterns and topic probability scores. Each sequence is labeled with its highest contributing topic, associated motifs, and proportion score. This analysis confirms that certain topics (e.g., Topic 3 and Topic 7) dominate specific subsets of sequences, highlighting meaningful structure.

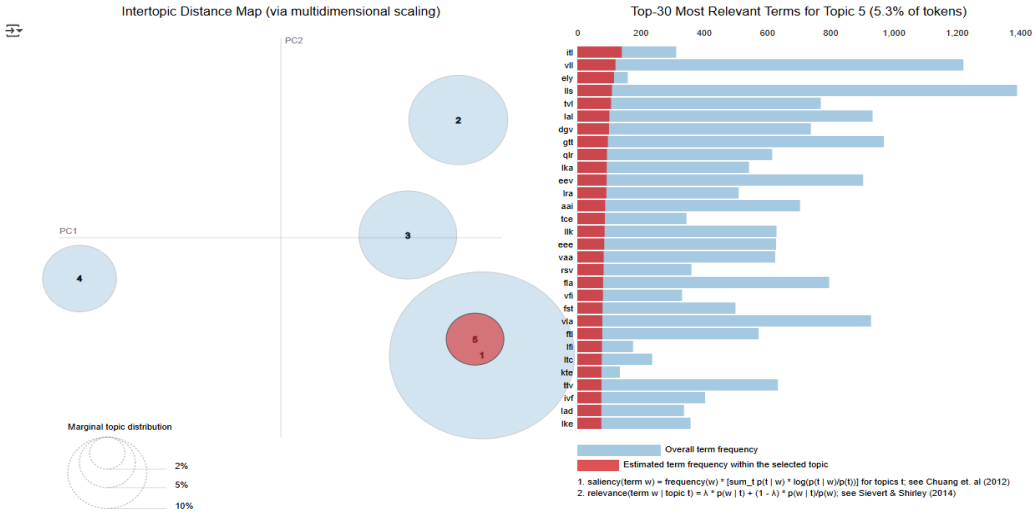

Figure 4. Top motif-like words identified by Latent Dirichlet Allocation (LDA) for each of the ten topics

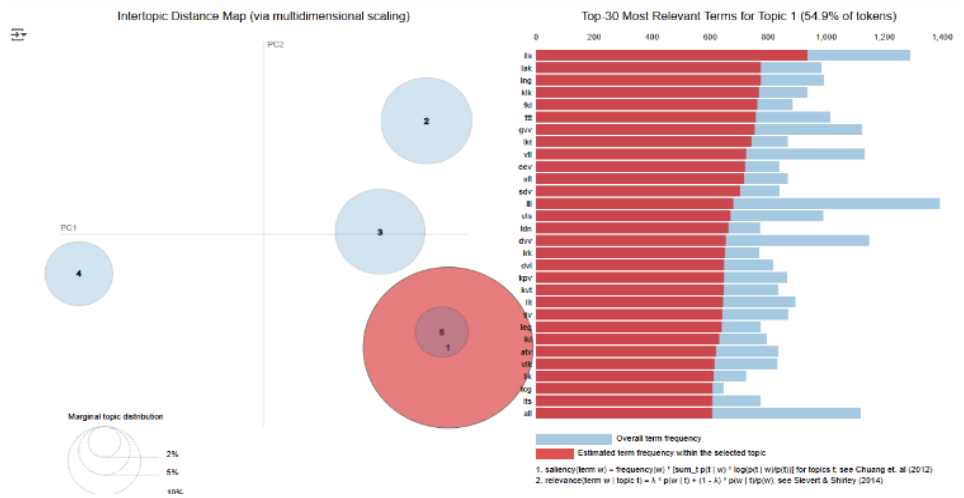

Figure 5. Latent Dirichlet Allocation (LDA)-dominant topic assignments for sample protein sequences

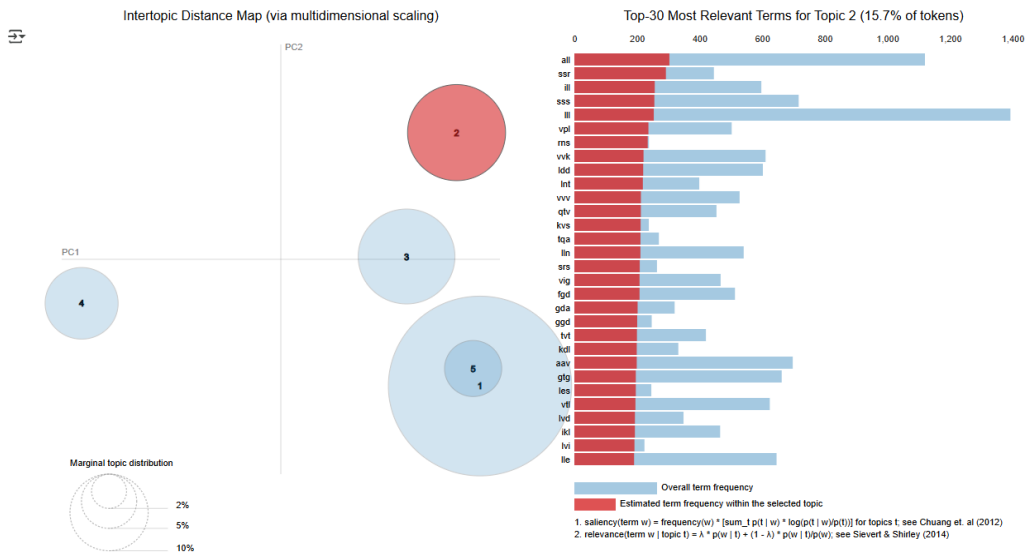

Figure 6. Topic proportions for each protein sequence

6.3 Topic composition of protein sequences

A bar chart visualization of topic proportions for several sequences shows that while some sequences are highly associated with one topic, others contain a mixture of multiple topics. Figure 6 shows the topic composition of individual protein sequences, demonstrating that some sequences contain mixtures of multiple latent topics. Multitopic compositions suggest overlapping motif features across sequences. Such soft clustering reflects the biological reality of proteins being composed of multiple functional motifs or domains.

6.4 Model training and evaluation

LDA-based topic distributions were reduced to two dimensions using t-SNE to visualize the clustering behavior of sequences in a reduced feature space. Figure 7 visualizes the clustering of protein sequences in the reduced t-SNE space, where sequences are color-coded according to their dominant LDA topic. Clear clusters emerge, confirming the separation power of topic features. Also, Figure 7 indicates that sequences grouped under the same topic tend to cluster together, supporting the effectiveness of LDA in feature extraction and unsupervised grouping.

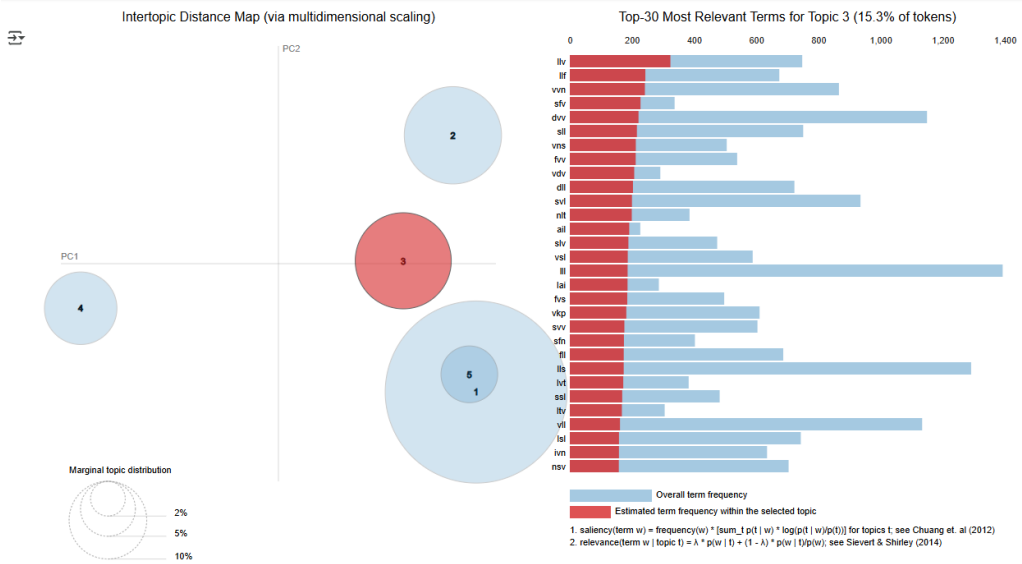

Figure 7. The t-SNE plot of protein sequences, color-coded by their dominant Latent Dirichlet Allocation (LDA) topic

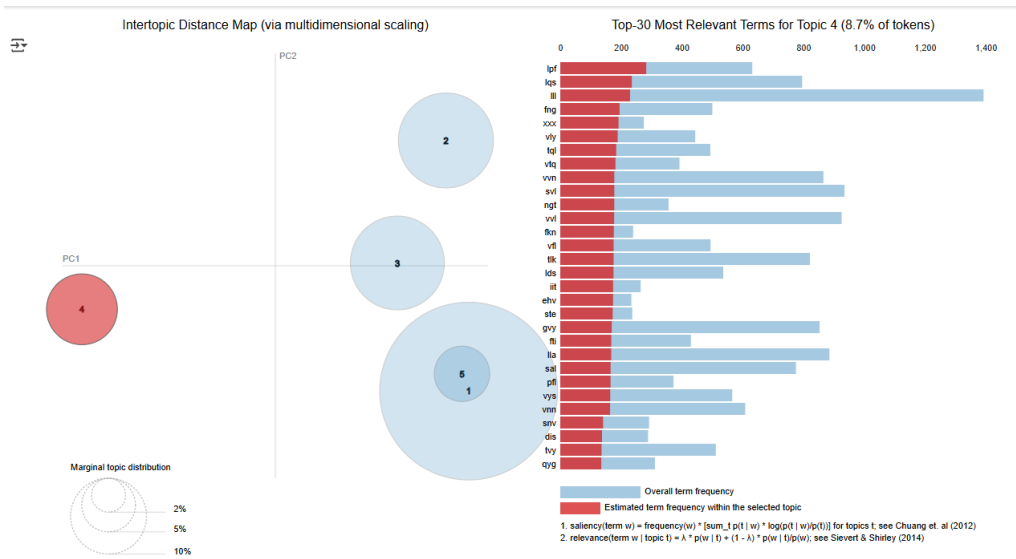

Figure 8. Top motif weight for selected Latent Dirichlet Allocation (LDA) topics

6.5 Topic-motif importance via word weights

To interpret the motifs driving each topic, the top weighted words for selected topics are visualized using bar charts. Figure 8 highlights the most influential motif weights for selected LDA topics, indicating motif patterns that contribute strongly to topic formation. These motifs are candidates for domain-level functional annotation or biomarker discovery in downstream bioinformatics tasks.

The results validate our hypothesis that protein sequences can be modeled as mixtures of latent motifs using LDA. Each topic matches a different sequence pattern, and the t-SNE-based visualization approves that these patterns cluster implicitly in low-dimensional space. This method suggestions addressing main challenges in analysis of protein sequence and using both qualitative interpretability and quantitative clustering. Unlike fixed feature extractors and traditional clustering methods, LDA proposes soft assignment and

interpretable motifs, aligning greatly with the modular and multifunctional nature of protein structures.

Precisely, several high-frequency motifs identified by the LDA model, such as LKELE, ELEKE, and related repetitive amino acid patterns, may correspond to conserved regions associated with structural stability, protein-binding interfaces, or functional domains in SARs-CoV2 proteins. These patterns are enriched in amino acids such as lysine (K) or glutamic acid (E), that are identified to contribute to static interactions, folding of protein, and host–virus binding mechanisms. In addition, the dominance of certain latent topics across subsets of sequences suggests the presence of shared functional or structural characteristics among SARs-CoV-2 proteins. The multitopic composition observed in some sequences may reflect the integrated nature of virus-related proteins, where multiple motifs contribute to processes such as viral replication, host immune evasion, and protein–protein interactions. The extracted motifs were highlighted could support downstream functional annotation and discovery of biomarkers by identifying conserved sequence regions potentially linked to virus-related pathogenicity.

The amino acid patterns and "ELEKE" are Motifs that are linked with protein-binding interfaces in Interaction of SARs-Cov2, moreover the topic of dominant reflects joint functional features between two or many virus proteins including many roles in PPIs and virus replication, these perceptions improve the connections between the discovered pattern and biological significance.

Definitely, the LDA model was trained using 10 topics with iterative optimization until convergence. The number of topics was selected experimentally based on coherence and perplexity evaluation to achieve a balance between interpretability and model quality. Moreover, key hyperparameters, including alpha and eta, were tuned empirically to improve topic separation and sequence representation performance. These performance details have now been clarified in the methodology section.

7. CONCLUSIONS

For recognizing virus-related protein interactions; this approach highlights the importance of integrating ML techniques with bioinformatics to achieve the above objective. And by using LDA2vec for clustering and analyzing SARs-CoV2 protein sequences, an interpretable and scalable approach to interactome analysis was presented. Future work could importance on improving dataset quality, exploring different feature extraction methods, and extending the analysis to other virus-related families. By progressing our recognizing of PPIs, this approach contributes to the advance of diagnostic and therapeutic strategies for emerging infectious diseases.

The proposed LDA2vec-LDA-based method is suitable for many fields such as protein family classification, motif discovery, functional annotation, and exploratory bioinformatics analysis. It provides several advantages, including interpretability, efficiency in computation, and the ability to locate latent motif consensus from protein sequences without requiring large-scale pretrained models. However, the approach also has limitations, including reduced capability in capturing long-range sequence dependencies. In addition, the quality of the extracted topics may be sensitive to the size of k-mer and topics number. Future work may attention on integrating embeddings of Transformer with topic modeling techniques, evaluating the method on larger databases, and incorporating biological domain knowledge to improve accuracy and interpretability of classification.

ACKNOWLEDGMENTS

This work is supported by College of Information Technology, University of Babylon.

REFERENCES

[1] Ersozlu, T., Gultekin, E. (2020). Tracheostomy and tracheostomy care during the covid-19 pandemic. Namık Kemal Medical Journal, 8(3): 551-557. https://doi.org/10.37696/nkmj.746867

[2] Rodriguez-Morales, A.J., Bonilla-Aldana, D.K., Tiwari, R., Sah, R., Rabaan, A.A., Dhama, K. (2020). Covid-19, an emerging coronavirus infection: Current scenario and recent developments-an overview. Journal of Pure & Applied Microbiology, 14(1): 5-12. https://doi.org/10.22207/JPAM.14.1.02

[3] Haas, L.E., Termorshuizen, F., den Uil, C.A., de Keizer, N.F., de Lange, D.W., Dutch COVID-19 Research Consortium. (2023). Increased mortality in ICU patients≥ 70 years old with COVID-19 compared to patients with other pneumonias. Journal of the American Geriatrics Society, 71(5): 1440-1451. https://doi.org/10.1111/jgs.18220

[4] Kshirsagar, M., Tasnina, N., Ward, M.D., Law, J.N., Murali, T.M., Lavista Ferres, J.M., Klein-Seetharaman, J. (2020). Protein sequence models for prediction and comparative analysis of the SARS-CoV-2—Human interactome. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, Kohala Coast, Hawaii, USA, 26: 154-165. https://doi.org/10.1142/9789811232701_0015

[5] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics, 38(8): 2102-2110. https://doi.org/10.1093/bioinformatics/btac020

[6] Aminah, S., Ardaneswari, G., Husnah, M., Deori, G., Prasetyo, H.B. (2023). Detection of COVID-19 using protein sequence data via machine learning classification approach. Journal of Applied Mathematics, 2023(1): 9991095. https://doi.org/10.1155/2023/9991095

[7] AlRefaai, N., AlRashid, S.Z. (2023). Classification of gene expression dataset for type 1 diabetes using machine learning methods. Bulletin of Electrical Engineering and Informatics, 12(5): 2986-2992. https://doi.org/10.11591/eei.v12i5.4322

[8] AL Raheim Hamza, L.A., Lafta, H.A., Al Rashid, S.Z. (2023). Classification of DNA sequence for diabetes mellitus type using machine learning methods. In International Conference on Micro-Electronics and Telecommunication Engineering, Ghaziabad, India, 894: 87-102. https://doi.org/10.1007/978-981-99-9562-2_8

[9] Dey, T.K., Mandal, S., Mukherjee, S. (2022). Gene expression data classification using topology and machine learning models. BMC Bioinformatics, 22 (Suppl 10): 627. https://doi.org/10.1186/s12859-022-04704-z

[10] García-Jaramillo, M., Luque, C., León-Vargas, F. (2024). Machine learning and deep learning techniques

applied to diabetes research: A bibliometric analysis. Journal of Diabetes Science and Technology, 18(2): 287-301. https://doi.org/10.1177/19322968231215350
[11] Borsari, B., Frank, M., Wattenberg, E.S., Xu, K., Liu, S.X., Yu, X., Gerstein, M. (2025). The chronODE framework for modelling multi-omic time series with ordinary differential equations and machine learning. Nature Communications, 16(1): 7021. https://doi.org/10.1038/s41467-025-61921-9
[12] Le, N.Q.K., Kha, Q.H. (2023). A sequence-based prediction model of vesicular transport proteins using ensemble deep learning. In Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Houston, USA, pp. 1-6. https://doi.org/10.1145/3584371.3612950
[13] Draizen, E.J., Readey, J., Mura, C., Bourne, P.E. (2024). Prop3D: A flexible, Python-based platform for machine learning with protein structural properties and biophysical data. BMC Bioinformatics, 25(1): 11. https://doi.org/10.1186/s12859-023-05586-5
[14] Nimmagadda, S.M., Agasthi, S.S., Shai, A., Khandavalli, D.K.R., Vatti, J.R. (2023). Kidney failure detection and predictive analytics for ckd using machine learning procedures. Archives of Computational Methods in Engineering, 30(4): 2341-2354. https://doi.org/10.1007/s11831-022-09866-w
[15] Brierley, L., Fowler, A. (2021). Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLoS Pathogens, 17(4): e1009149. https://doi.org/10.1371/journal.ppat.1009149
[16] Mano, L.Y., Torres, A.M., Morales, A.G., Cruz, C.C.P., Cardoso, F.H., Alves, S.H., Werneck, V.M.B. (2023). Machine learning applied to COVID-19: A review of the initial pandemic period. International Journal of Computational Intelligence Systems, 16(1): 73. https://doi.org/10.1007/s44196-023-00236-3
[17] Bhosale, Y.H., Patnaik, K.S. (2023). Application of deep learning techniques in diagnosis of covid-19 (coronavirus): A systematic review. Neural Processing Letters, 55(3): 3551-3603. https://doi.org/10.1007/s11063-022-11023-0
[18] Kockelbergh, H., Evans, S., Deng, T., Clyne, E., Kyriakidou, A., Economou, A., Soilleux, E.J. (2022). Utility of bulk T-cell receptor repertoire sequencing analysis in understanding immune responses to COVID-19. Diagnostics, 12(5): 1222. https://doi.org/10.3390/diagnostics12051222
[19] Kosar, A., Asif, M., Ahmad, M.B., Akram, W., Mahmood, K., Kumari, S. (2024). Towards classification and comprehensive analysis of AI-based COVID-19 diagnostic techniques: A survey. Artificial Intelligence in Medicine, 151: 102858. https://doi.org/10.1016/j.artmed.2024.102858
[20] Gharaibeh, T., de Doncker, E. (2021). Unsupervised learning model to uncover. In Computational Science and Its Applications – ICCSA 2021, pp. 544-559. https://doi.org/10.1007/978-3-030-86960-1_38
[21] Li, Z., Lu, J., Cui, J., Cao, G., et al. (2026). Functional customization of peptide linkers in fusion proteins through multimodal deep learning approach. Synthetic and Systems Biotechnology, 13: 362-375. https://doi.org/10.1016/j.synbio.2026.02.003
[22] Barissi, S., Sala, A., Wieczór, M., Battistini, F., Orozco, M. (2022). DNAffinity: A machine-learning approach to predict DNA binding affinities of transcription factors. Nucleic Acids Research, 50(16): 9105-9114. https://doi.org/10.1093/nar/gkac708
[23] Polyanovsky, V.O., Roytberg, M.A., Tumanyan, V.G. (2011). Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms for Molecular Biology, 6(1): 25. https://doi.org/10.1186/1748-7188-6-25
[24] Bonham-Carter, O., Steele, J., Bastola, D. (2014). Alignment-free genetic sequence comparisons: A review of recent approaches by word analysis. Briefings in Bioinformatics, 15(6): 890-905. https://doi.org/10.1093/bib/bbt052
[25] Qiao, Y., Yang, X., Wu, E. (2019). The research of BP neural network based on one-hot encoding and principle component analysis in determining the therapeutic effect of diabetes mellitus. In IOP Conference Series: Earth and Environmental Science, Guangzhou, China, 267(4): 042178. https://doi.org/10.1088/1755-1315/267/4/042178
[26] Yu, L., Zhou, R., Chen, R., Lai, K.K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation? Emerging Markets Finance and Trade, 58(2): 472-482. https://doi.org/10.1080/1540496X.2020.1825935
[27] Zhang, X., Beinke, B., Kindhi, B.A., Wiering, M. (2020). Comparing machine learning algorithms with or without feature extraction for DNA classification. arXiv preprint arXiv:2011.00485. https://doi.org/10.48550/arXiv.2011.00485
[28] Zizelski Valenci, G., Rubinstein, M., Afriat, R., Rosencwaig, S., Dveyrin, Z., Rorman, E., Nissan, I. (2020). Draft genome sequences of Cronobacter muytjensii Cr150, Cronobacter turicensis Cr170, and Cronobacter sakazakii Cr611. Microbiology Resource Announcements, 9(44). https://doi.org/10.1128/mra.00660-20
[29] Qader, W.A., Ameen, M.M., Ahmed, B.I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. In 2019 International Engineering Conference (IEC), Erbil, Iraq, pp. 200-204. https://doi.org/10.1109/IEC47844.2019.8950616
[30] Sunagar, P., Kanavalli, A., Shetty, N.D. (2020). Feature extraction and selection techniques for text classification: A survey. International Journal of Advanced Research in Engineering and Technology, 11(12): 2871-2881. https://doi.org/10.34218/IJARET.11.12.2020.268
[31] Madasu, A., Elango, S. (2020). Efficient feature selection techniques for sentiment analysis. Multimedia Tools and Applications, 79(9): 6313-6335. https://doi.org/10.1007/s11042-019-08409-z

NOMENCLATURE

l	the length of the sequence
k	the width of the word
n	times of k-mer