Journal of Information Systems Engineering and Management
2025, 10(9s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

NLP Based Protein Sequence Classification using CNN

Prativesh Pawar*[1], Dr. Pinaki Ghosh[2]
[1]*PhD scholar*
SAGE University Bhopal, India
prativesh.sait@gmail.com
[2]*Dept. Of CSE*
SAGE university Bhopal, India
pinaki.g@sageuniversity.edu.in

ARTICLE INFO

Received: 02 Nov 2024

Revised: 22 Dec 2024

Accepted: 05 Jan 2025

ABSTRACT

The capacity to modify proteins to have better or novel functions has led to a significant increase in interest in protein redesign in the pharmaceutical industry. Natural selection, amplification, and mutation processes may now be simulated in the lab because to recent technological developments. However, a significant barrier remains since protein sequences are complex structures with a large number of potential mutations. Not all possible variations of a protein can be synthesised or evaluated. Protein prediction algorithms have shown very little success in predicting protein structures, despite advances in machine learning. Furthermore, most current approaches concentrate on a narrow set of traits associated with protein sequences.

This study offers a novel approach to categorise protein sequences using artificial intelligence (AI) and convolutional neural networks (CNNs). Our research aims to evaluate three distinct prediction models for efficacy. A combination of single-amino-acid and three-dimensional protein structure-based descriptors are used to train each model. To evaluate the accuracy of our forecasts, we employed a variety of evaluation metrics, encompassing both publicly accessible and proprietary datasets. The results demonstrate the remarkable effectiveness of Convolutional Neural Network (CNN) models trained using amino acid property descriptors in addressing the challenges related to protein structure estimation. For applications in the pharmaceutical industry, this renders them highly advantageous.

Keywords: Protein sequence, NLP (Natural language processing); Deep learning

I.INTRODUCTION

The multidisciplinary area of bioinformatics combines concepts from computer technology, genetics, statistics, mathematics, and molecular biology. From the standpoint of computer science, it tackles several important and data-rich biological problems. Finding meaningful patterns in massive datasets and comprehending biological processes at the molecular level are two of this domain's biggest problems. Significant progress in genetics over the last ten years has produced enormous volumes of biological data, requiring the use of sophisticated computer methods for inference and analysis. These methods are essential for simplifying sequential data, making biological sequences able to be classified and predicted, and condensing research results from a variety of life science fields. The use of data mining and machine learning techniques has become crucial for these goals due to the exponential rise in data creation [1].

Bioinformatics has advanced significantly as a result of the growing number of biological data that is now available [2-4]. Specifically, data mining techniques have emerged as indispensable instruments for drawing conclusions from large-scale biological information. Sequential pattern mining is a subfield of data mining that looks for patterns in sequences. Sequences are common in many domains, such as commerce, research, security, and medical. Protein and DNA sequences are two prominent examples of sequences, which are basically ordered lists and are fundamental to many biological processes. Protein sequence categorization is the process of putting proteins into groups or labels according to the amino acid sequences they contain; previously identified sequences

are stored in databases. Large proteins can include up to a thousand amino acids, whereas smaller proteins only contain about 100. Proteins are the chains of amino acids which are arranged in certain and specific patterns.

Directed evolution processes are extensively employed in the pharmaceutical industry for the customization of proteins whose activities display enhanced properties. For example, for the enzymes that are utilized as catalysts in the synthetic processes for the preparation of drug materials and pre-drug materials at the commercial scale, it is vital that they perform most effectively in what might otherwise be considered almost adverse conditions. By mimicking the evolutionary process, you systematically search through a best-parent-sequence-led process in each cycle to promote a pool of variant solutions or sub-solutions which is, in the doings, screened with some in vitro assays to identify the target performances. The best performing variants under the particular circumstances remain for further optimization. Directed evolution is in essence an optimization problem out of all the possibilities among protein sequences. Therefore, identifying beneficial mutations may still be headache, even with very sensitive tests [6]. Prediction models can provide a way to accelerate beneficial mutations. Thus one would be able to predict-feeding models that will further augment the value of any sequence, and allow these Polynter-based algorithms to learn the association between a sequence and its function by learning from complementary data on extant sequences. The application of this computational approach would only regard the next rounds of experimental work in synthesizing only the most promising sequences. Therefore this problem now attractions involved with confronting the challenge through machine-learning-models and tunics [7].

This study plans to use an NLP model for contextual content extraction and constructing CNN-based protein sequence classification framework by working on an accessible dataset. The paper detailed implications of techniques such as data preprocessing, data visualization, feature generation, model training, and model evaluation to perform efficient protein classification into various groups. Contrary to the models that only focus on variant site, the CNN models deal with ordering throughout the entire protein sequence, a proven reason why they perform well. The remaining parts of this paper are organized as follows: Section II reviews current relevant literature. CNN architectures are overviewed in Section III. In Section IV, we provide a CNN-based approach for protein sequence categorization algorithms over the dataset in question. Section V discusses and ponders upon the performance results of the proposed approach in experimental evaluation to officially end the article at section VI.

II.RELATED WORK

One of the most crucial aspects of the pharmaceutical industry involves the creation of new drugs, and computational modeling has considerably reduced both the time and expense associated with such a process. Drug screening and design, on the other hand, present a number of challenges that require perhaps only a few strategic methods to come up with a kind of solution. This section mainly concerns some very deep studies and applications of machine learning that managed to go beyond previous research boundaries.

Wei-Li et al. [18] introduced an MOEA+STAAR framework, where STAAR stands for stochastic assignment of atom coordinates in real space. A modified version of Rama torsion angle sampling is incorporated here into the republication process proposed previously [16]. The main aid the structural similarity criterion provides is in reducing the error on the energy function. Using reinforcement learning, Zhou et al. [19] constructed CDNN models for the prediction of protein secondary structures. This setup takes advantage of the CDNN's ability to combine the abstraction capabilities of CNNs with sequence data processing capabilities of LSTM networks, hence providing a more robust classification performance. Training the model directly involves minimizing cross-entropy error between labels for protein secondary structures and dense layer outputs. Execution of the CDNN architecture across two diverse datasets shows the method is very efficacious in practice, with small outliers.

Moreover, You et al.'s work [20] produced a Deep ResNet model for predicting template-free protein folding and protein contact/distance. Recent developments in Deep ResNet have been noteworthy in two significant domains: tertiary structure prediction and protein-protein interaction. Although less difficult and employing less sophisticated inter-residue orientation data, the proposed three-dimensional modelling technique has correctly folded most proteins, even without using evolutionary data to estimate their natural folds.

This paper is motivated by the huge progress made by co-evolution-based classifiers in predicting protein contact maps for unrelated protein sequences. Such co-evolution-based classifiers trained from multiple alignments have

been found in many applications to be outperformed by deep-learning-based, single-model predictions on single-domain proteins. The backbone of any deep-structure problem relies on the availability of precise templates. The acquisition of this template is, however, essentially not dependent on the actual presence of sequence similarity.

Proteins having a very high sequence similarity do not gain authors' interest. In contrast, authors are excited to address proteins with no significant sequence similarity, those that could defuse the analysis of even simple amino acids in the manner of orphan proteins. Predicting contact maps from only sequence can be further boosted by new concepts that eliminate the necessity of using co-evolution-based methods throughout the total sequence examined. One may overlap co-evolution with deep structure continuously in varying proportions. The method by Du et al. [22] proposes an RGN as another viable implementer to make contact maps with the sequence intention of solving the contact map optimization problem for sequence alignments.

Guo et al. [23] created a technique based on multi-advanced deep belief networks to improve the prediction of protein secondary structures. This method, which used hidden Markov model profiles created using emission and transition probabilities, produced results with an accuracy of over 80%. The network showed uneven features despite the effectiveness of this approach.

Wang et al. [24] trained a deep neural network (DNN) with a protein feature vector and the proposed MOS descriptor with amino acid (AA) classification in order to predict protein-protein interactions (PPIs) with accuracy. Unlike previous protein representations such as autocorrelation (AC), composition-transition-distribution (CT), and local descriptor (LD), the MOS descriptor accurately depicts the order connection of the whole AA sequence. The network's task-specific parameters, such as the ADAM optimizer, ReLU activation function, and cross-entropy cost function, were carefully selected. We were also able to determine the optimal settings for network depth, breadth, and learning rate by targeted computations. To facilitate comparisons with the recommended approach, the DNN model was trained independently using AC, CT, and LD.

Another paper, published by Jha and Saha [25], employed an LSTM-based classifier with both protein modality features, such as sequence-based and structure-based information. In their approach, three types of companion representations of proteins were initialized by structural representation of proteins. Further, feature extraction was performed on these three kinds of companions at the beginning; their best experiments implemented ResNet50 exclusively for computational usage by several other best-performing networks.

As far as sequence-based PPI prediction is concerned, a setup utilizing deep neural networks (DNNs) was introduced by Li et al. [26], wherein the network was programmed to learn from numerical inputs without having to learn from features manually. The segmentation of the protein sequence into natural numbers was randomly performed.

Essentially, a study by Gonzalez-Lopez et al. [27] explored PPI prediction without feature engineering using embedding systems and RNNs. Tokenization is putting a number to every triplet in the sequence, and this brought numerical language to expression for the sequences. The network worked on for each protein pair representation by two symmetrical branches. FC, embedding, and RNN layers followed; each layer had its clear and separate goal. Batch normalization and dropout were part of the entire network's adjustments to prevent overfitting network training.

This body of work demonstrates the astounding advancements in protein structure prediction and drug discovery made possible by the use of deep learning and machine learning techniques. These strategies have a great deal of promise for solving the issues in this crucial field by utilizing innovative techniques and intricate designs.

III.ARCHITECTURAL FRAMEWORK OF CNN

One of the most important aspects of the pharmaceutical business is that the time and cost required to create new drugs have decreased thanks to computational methodologies. Multi-approach solutions are required for drug screening and design because of various obstacles. The main focus of this section is on techniques for applying machine learning and deep learning that surpass the constraints of previous studies.

A multi-objective evolutionary strategy was developed by Wei-Li et al. [18] that includes techniques including Rama torsion angle sampling, loop-based crossover, near-native sampling, and loop-based resampling. An extra structural similarity criteria is used in this method to assist reduce energy function mistakes. Zhou et al. [19] also used convolutional deep neural networks (CDNNs) trained with reinforcement learning to predict the secondary structures of proteins. By merging the sequence data processing skills of LSTM networks with the abstraction capabilities of CNNs, the CDNN architecture produces excellent classification results. The model is trained using the cross-entropy error between dense layer outputs and protein secondary structure labels. Overall, on both examined datasets, the CDNN technique performed well, with a few outliers.

You et al.'s [20] Deep ResNet model, which predicts template-free protein folding and protein contact/distance, is a further noteworthy breakthrough. The domains of protein-protein interaction and tertiary structure prediction have benefited greatly from recent advancements in Deep ResNet. The majority of proteins have spontaneously folded using the suggested three-dimensional modelling technique, despite its simplicity and utilisation of less complex inter-residue orientation data.

The "DSA" approach from Xu et al. [21] combines template-based structural modelling and deep learning algorithms for residue-residue interaction prediction conclusions, allowing a 3-D structure prediction of over 1200 single-domain proteins for the first time with high accuracy. Even though it is deficient on using spatial coupling scores, obtained from raw frequency distributions of numerous sequence alignments, the implementation of the CCP method ensures an attractive substantial improvement compared to previous ones.

A unique recurrent geometric network (RGN), developed by Du et al. [22] under the empathy of structure prediction from sequence alone without any priors, is very well-geared towards addressing the issue at hand, albeit with an inaccurate amino-acid sequence alignment to orphaned today, a straightforward task in theory yet computationally intensive. This every angle is dealt with in RGN2, in which the sequence is considered to convert the backbone structure into a geometry of the Cα carbon backbone through consequent local interactions, e.g. between secondary structure elements (curvature and torsion angles of Cα carbons).

Guo et al. [23] constructed and coupled a secure framework with protein sequence in order to study improved protein secondary structure predictions via deep belief networks consisting of several progressive layers. The accuracy of the hidden Markov models extracted approximately 80% by emission and transition probabilities. Such a technique has brought success, but the network is unstable in some of its properties.

Wang et al. [24] used an MOS descriptor recommended with AA classification and protein feature vector trained to a deep neural network (DNN) to predict PPIs with due success. Unlike earlier representations for proteins like local descriptor (LD), composition-transition-distribution (CT), and autocorrelation (AC), the MOS descriptor perfectly and accurately shows the entire order relationship on a sequence of AAs. The ADAM optimizer, the ReLU activation function, and cross-entropy cost function are the big characterizing features of the network that we chose carefully. Using slight computations, we could get the best configurations of some of the parameters, like network size, depth, and learning rate. The DNN model, as trained independently while using AC, CT, and LD, was carried out to make a certain degree of comparison with our proposed method.

The LSTM-based classifier the next authors, Jha and Saha [25], implemented brings two different protein representations together-sequence-based features and structural information. Using the structural representations of the proteins, these authors have made three different representations of the proteins based upon three different properties first. The feature sets relevant to protein for prediction are identified by utilizing a ResNet50 model.

Li et al. [26], in their pioneering work, applied automatic feature engineering and deep neural networks (DNNs) to predict sequence-based PPIs without a requirement of human feature engineers: Natural numbers were inserted randomly for amino acids in the protein sequence, and the architecture could be learned directly from the numeric input.

Similar to this, Gonzalez-Lopez et al. [27] did not use feature engineering in their RNN and embedding system-based PPI prediction. Tokenisation allowed for the numerical expression of the sequences by giving each triplet of sequences an integer token. The bipartite neural network examined each protein pair's representation. The fully connected (FC), embedding, and recurrent layers of the design each performed a different function. To maintain input uniformity and prevent overfitting, batch normalisation and dropout were utilised.

The most part of this document then, underscores the incredible advancement achieved through the channels of deep learning and machine learning in drug discovery as well as protein structure prediction. The methods themselves seem rather promising in terms of the application towards addressing problems in this important area, as they involve complex models and novel techniques.

IV.PROPOSED DEEP CNN BASED PROTEIN SEQUENCE CLASSIFICATION

In order to effectively categorize diverse proteins into different kinds, this study describes e process how the data is from pre-processing to visualization and feature engineering, modelling, training, and evaluation. The whole work is majorly classified in four steps as shown in figure 1 and explained as follows: Dataset definition: in this particular work, the dataset known as Structural Protein Sequences is utilized from kaggle.com. The first section of the collection includes protein Meta data pertaining to protein classification and extraction methods. The second portion of the collection consists of protein structural sequences. Both data banks follow the "structureID" property of having the proteins as the guiding principle for the databases. Distance between these databases is that the first contains 141000 rows and 14 columns whereas the second contains 467000 rows but just 5 columns in it. The proteins have been adopted from the Protein Data Bank (PDB) at RCSB's Research Collaboratory for Structural Bioinformatics for this study.

a) Data Preprocessing: Dataset merging. In this, the "structureID" property does the work of unifying the two datasets into a single dataset. Remove rows with improper labels or sequences after merging. Then we filter out all but proteins because macromoleculeType_x comprises varied types of macromolecules. Various biologically important macromolecules are represented in the data. Most of these files contain information about proteins. Proteins are biomolecules that are directly involved in various biological pathways and cycles because of DNA being the precursor of RNA and due to the conversion of RNA into proteins. On the basis of family, it is known to have a couple of roles. For example, a Hydrolase type protein is focused on the hydrolysis of bioactive foreign elements such as external DNAs and RNAs types.

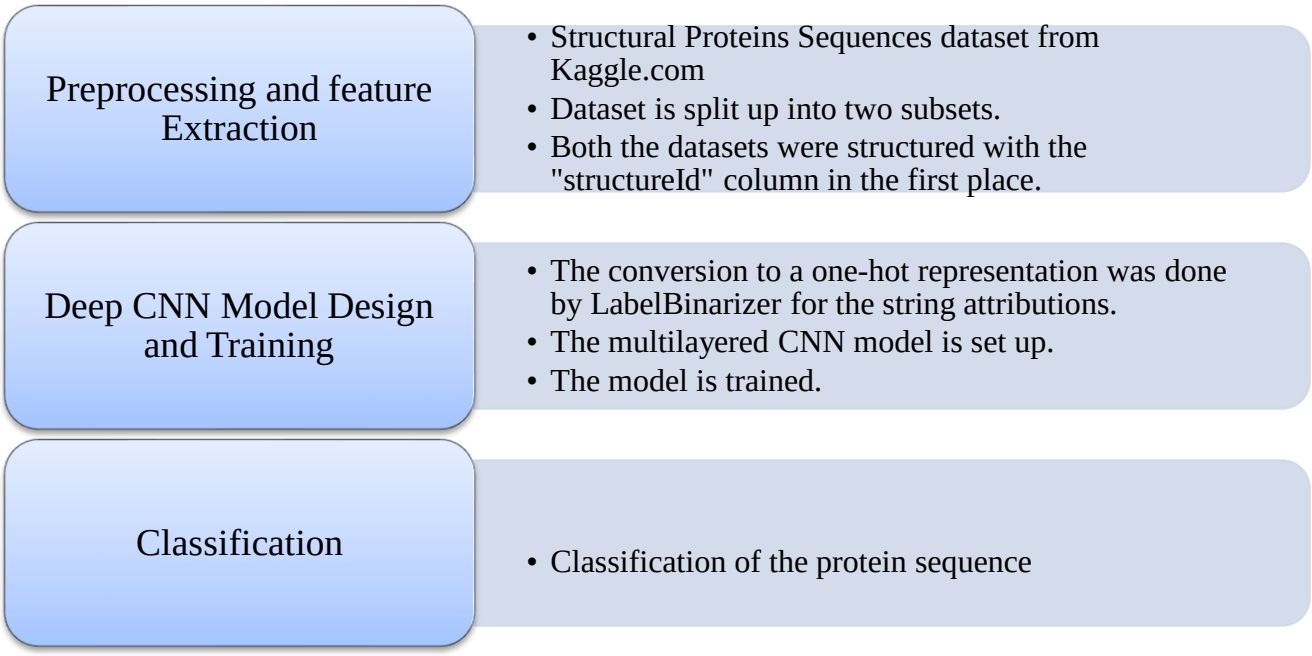


Fig.1 Proposed Framework

a) Feature Extraction

The pharmaceutical industry depends on drug development as a critical part of their workflow. Computational techniques have greatly sped up and simplified the process of developing new medications. The complexity of the problems we are confronting directly relates to the drug screening and design techniques that are required. The essay's primary subjects are deep learning and sophisticated machine learning techniques. They created a multi-objective evolutionary strategy by combining loop-based resampling, near-native sampling, stochastic rank-based selection, and Rama torsion angle sampling (Wei-Li et al., 2018). The secondary structure similarity criterion may help in remedying an inaccurate energy function. Zhou et al. [19] use reinforcement learning to train their deep neural network (CDNN) for protein prediction. CDNN, as an alternative, has a broader utility by combining identification with classifying tissues like CNNs and this pragmatic capacity is validated by the appropriate processing of the LSTM sequence data. The minimization of cross-entropy between protein structure secondary labels and dense-layer outputs is required before CNN training. Efficacy of the CDNN paradigm over two test datasets was empirically proved despite their wide divergence. The prediction goes wrong only when there are more substantial changes in these data, given that it is a weaker future projection.

Route to protein structures and functions without a template-fate-wise beginning to the prediction of the proteins can be predicted by building the Deep ResNet [20]. Recent advances from Deep ResNet focus on protein-protein interaction and tertiary structure prediction. Today, inter-residue orientation data-based prediction of 3D structures is quite crude by most 3D modeling approaches that have been proposed. For this low-level quality design, the deep ResNet does not need to rely on evolutionary domain for accurate prediction of the folding of most of the designed protein.

Xu et al. developed a technique, "deep structural inference," designed to predict residue interactions in proteins. This method is a combination of deep learning and template-based structural modeling. The methodology had achieved, for the first time, an all-atom structure prediction for more than 1200 within single-dome in proteins. This indicates that from the results of the statistical evolutionary studies that the coupling scores coming from CCMPred, forming multiple sequence alignments' raw frequency distributions are inadequate. Duet al. [22] presented a specialized kind of recurrent geometric networks (RGNs) capable of doing structural predictions from sequences without any prior knowledge. This computationally efficient method is particularly beneficial for the orphan and designer proteins, for which the enough similarity could not be gathered for multiple-sequence alignment to work effectively. RGN2 endeavored to take a very simple approach for exposing the geometry of the C backbone. This method could work due to its reliance on the primitive reconstruction of the backbone. It could only handle local interactions between C atoms, such as curvature and torsion angles.

Guo et al. in [23] used a multilayer deep belief network (DBN) to predict the secondary structure of proteins. When presented, the prediction performance was recalibrated by approximately 80% co-operatively. Consequently, the profiles based on the transition and emission probabilities of the hidden structure can predict the secondary structure quite well. However, the network's inter-function will be enormously diverse. Wang et al. [24] integrated the most recommended MOS descriptor with AA classification into a DNN to give a reliable prediction of PPI from a protein feature vector.

In a further creative and fascinating study, Jha and Saha [25] combined data from two different protein modalities—sequence-based and structure-based—using a long short-term memory (LSTM) classifier. We constructed three distinct protein representations based on three distinct qualities using the structural representations of the proteins, which we then used to train a ResNet50 model to extract matching feature sets. The required methodology was presented by Li et al. [26] in order to forecast sequence-based PPIs utilising DNs and auto-feature engineering (i.e., features that were not developed manually). Only numerical input may be used to update the NN design. By assigning a random natural number to each amino acid, the protein's sequence was changed.

Gonzalez-Lopez et al. [27] were able to anticipate PPIs using RNNs and embedding systems without the need of feature engineering. Tokenization is a quantitative method of representing sequences in which each triplet in a series is given an integer token. For every protein, there appeared to be two identical branches handling the NN's pair representation. Without the other two, none of the three layers—FC, embedding, and recurrent—could have carried out their respective design tasks. Branch normalization and Dropout were used to further ensure input uniformity and avoid over-fitting.

$$idf(w) = log\frac{nd}{df(d,w)} + 1$$

After tokenizing the training text, TF-IDF (word frequency-inverse document frequency) may be computed by exploiting the statistical distribution of word frequencies within the dataset. After tokenization, each word in the list has to have its proper IDF value assigned in order to calculate TF-IDF scores. The Word2Vec model may then generate word vectors using the tokenised corpus. The Continuous Bag of phrases (CBOW) architecture was utilized to capture the context and identify phrases that have semantic similarity to a specified keyword. This approach represents the inputs with projections that resemble neural networks. This approach noticeably removes the conventional non-linear hidden layer in order to simplify the time series processing. Moreover, the context of each word serves as the input for the projection layer, sharing the same data across all words.

b) CNN Model Training and Testing

Protein engineering often uses many cycles of experimental treatments to expand the sequence space. Mutations can occur at any of infinite known or unknown sites. To facilitate real-world situations, splitting data into test and training sets according to a timestamp is the best option. For tasks with visual input, convolutional neural networks (CNNs) typically known to accelerate input from the inner layer are used. Rather than utilizing fully linked layers, condensed input from the layer above each hidden node through the convolutional neural network (CNN) model ensures consideration of location connectivity within images and identifies high-level properties. In general, the convolutional layers are better than fully linked ones for spatial dependency management in pictures and for extracting critical high-level features due to their incorporating of such local connections. Mostly, CNN models usually receive 2D images with numerous channels-such as RGB color channels, however, for some types of data, 1D CNN models come into play.

These researchers aim to suggest that a one-dimensional convolutional neural network (CNN) model should be used on protein sequence data. The protein feature input is displayed as a matrix in Figure 2. With the 1D convolution filter, we first insert the knowledge from neighboring regions to extract high-level features by moving along the amino-acid sequence (columns). The columns could represent many various amino acid properties or different input channels. Figure 2 presents an example of a common one-dimensional convolutional neural network (CNN) setup.

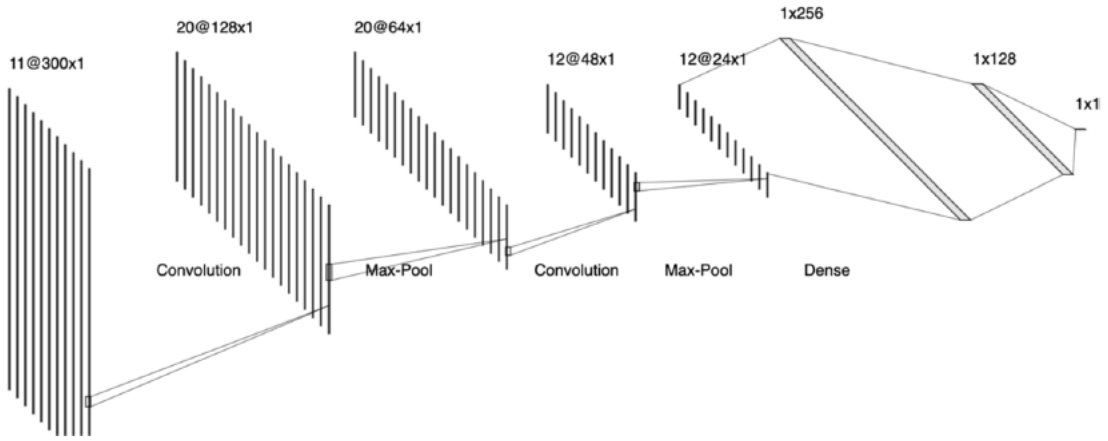


Figure 2 Common layout for a single-dimensional CNN [33]

In this scenario, we use the 11-dimensional PCscores descriptor as a distinguishing amino acid characteristic for input protein sequences longer than 300 amino acids. The first block of Figure 2 displays an 11x300 matrix, with each vertical line denoting an input characteristic for a specific protein. The max-pooling layer that follows each convolutional layer lowers the quantity of features obtained by down sampling along the sequence dimension after it has received input from neighboring sites. Each convolutional layer has filters that function similarly to sliding windows to exclude particular input kinds from the bottom layer's data channels. Figure 2 shows twenty vertical lines representing the output properties of a convolution layer with twenty filters applied to produce twenty channels. From the output of the convolutional layers, a "fingerprint vector" is produced, which is used to finish the regression process and understand the high-level features. We'll use a lot more related layers after that. CNN architectures need to be described, much like MLP architectures. The CNN's Hyperparameters are adjustable

in a number of ways, enabling the model to be fine-tuned architecturally. Several hyperparameters variables, such as the learning rate and the minibatch size, were used in a grid search.

Classification

Identification of protein sequences with a higher degree of ligand exchange is the main thrust of commercial protein engineering initiatives. This means that machine learning models are required to demarcate between protein sequences to be highly likely to exchange premier substrates from those least likely to. As a result, a two-tier classification label of affirmative and negative was generated and delineated with actual measured exchange in order to assess the predication model's classification performance.

V.IMPLEMENTATION AND RESULTS

To evaluate the efficacy of the proposed technique for predicting protein properties, intelligent prediction and classification models are trained over the existing dataset. It has utilized all of the available descriptor combinations to enhance the diversity in the training phase. Due to CNN's need for 2D matrix input, 44 distinct technique and descriptor pairs are allowed. The hyperparameters are adjusted as given in table 1:

Table 1. Hyperparameters [33]

Parameters	**Description/Values**		**Parameters**	**Description/Values**
Learning rates	0.001, 0.005, 0.01		Momentum	0.9
Optimization technique	Nesterov_momentum		Objective	mse
Kernel	GeLU		Maximum training epochs	500
Number of convolution layers	4		Initialization	HeNormal
Number of dense layers	2		Size of evaluation batch	16
Batch size	20, 40, 60		Size of embedding	128

Together with private and public data, success has been achieved by using the above-said approach. Publicly available collection is expected to cover a large portion of predictable empirically observed properties in relation to a wide range of protein classes, such as globular and membrane.

This may be useful in testing how universally applicable the principles suggested in this article are. Hence, the nickname by which the enzymes are named ranges from "Enzyme A" to "Enzyme D" to indicate to what important roles they belong in the proprietary data sets. Each of the proprietary enzymes may have a different chemical composition involved, but the identical procedure is what gives rise to one type of patent. This is simply because the ultimate outcome for all patented enzymes is substrate conversion, or the rate at which an enzyme turns a substrate into a product. As a result, this has resulted in such phenomena. The raw experimental data are converted and normalized for quantitative modeling.

Each of the 300 amino acids constituting the input protein sequence was represented as a separate amino acid feature by an 11-dimensional PC score descriptor. This means that each protein is represented by an input feature as an 11x300 matrix, with 11 lines standing up as in the declining block of Figure 2. From there, each convolutional layer samples the next set of locations in its immediate locality and max-pools down some sequence directions thus eliminating varied features. Turns out you wind up with fewer features being fed where the kernel is pointing. So, what happens is that every convolutional layer has a set of filters-that is, those filters that extract specific characteristics from the data channels of the layer below them, similar to the window sliding across it. In other words, each loaded filter pulls property features from data channels of the layer below it, much like it would as in a sliding window.

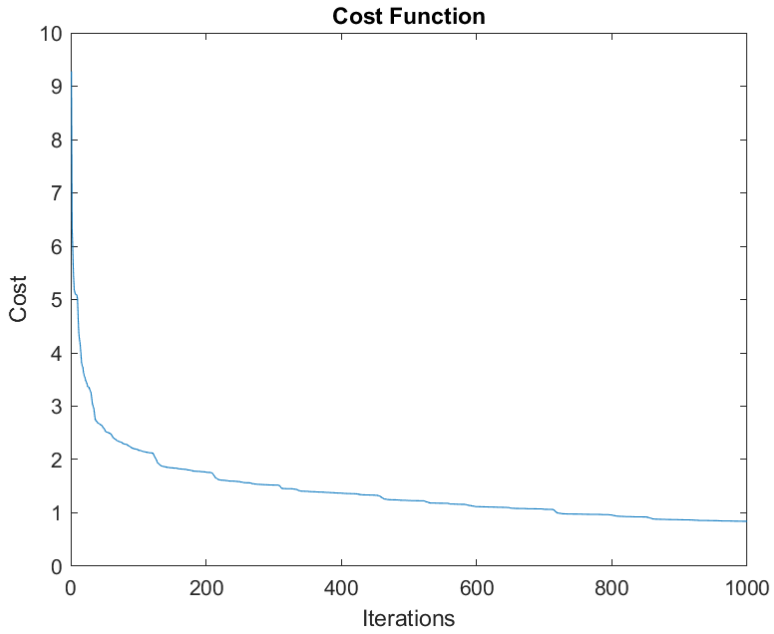


Figure 3 Cost function v/s time

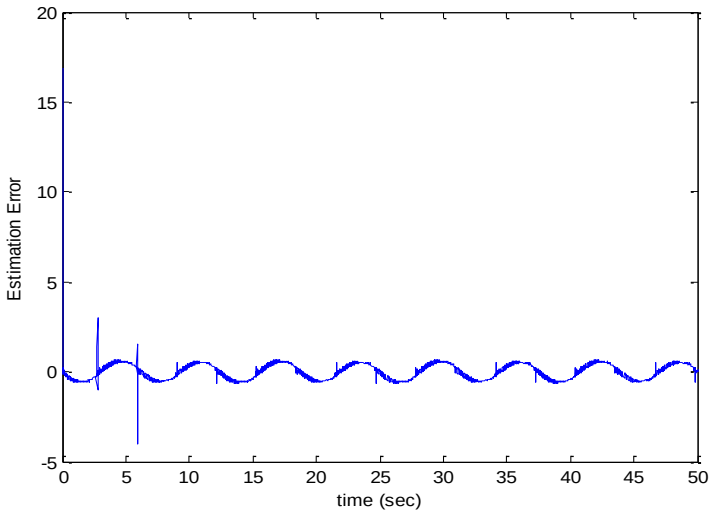


Figure 4Estimation error v/s time

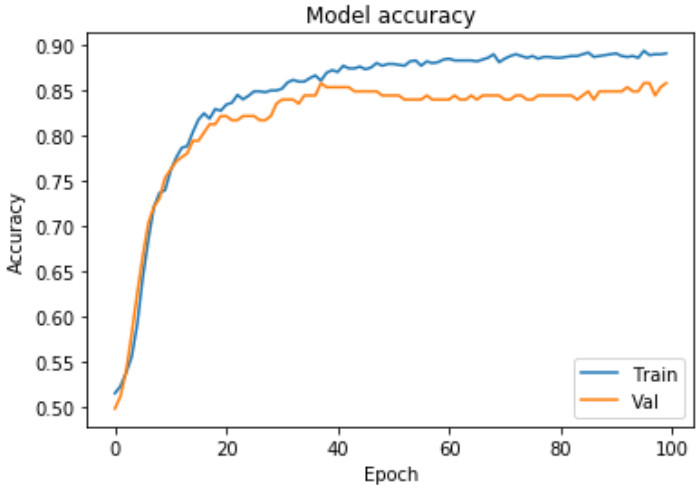
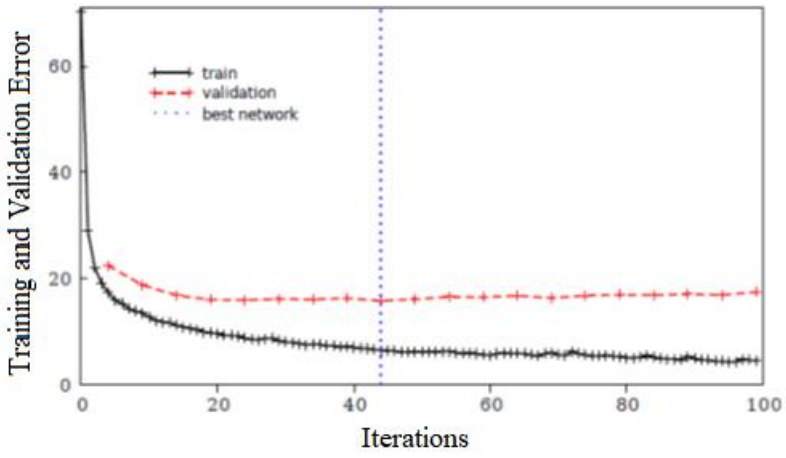


Figure 5Modeling Accuracy v/s number of iterations



b

Figure 6Training and Validation Error wrt number of iterations

For the training run on the internet database, Figures 5 and 6 show the accuracy and error of the models during validation and training over time.

VI.CONCLUSION

Here, we provide a novel approach to protein sequence categorization utilising AI and convolutional neural networks' (CNNs') prediction powers. We evaluated three distinct descriptors on both public and private datasets, two of which were devoted to the properties of individual amino acids and the third to three-dimensional structure data. Using an array of measures, the predictions' dependability was examined. Our findings suggest that CNN models, particularly those that incorporate amino acid properties, offer a robust solution to the challenges posed by protein redesign in the pharmaceutical industry. Unlike other methods that just examine altered sites, our CNN-based models use the sequential ordering of the whole protein sequence to capture high-level properties. This work not only demonstrates the effectiveness of CNNs in protein sequence estimation, but also illustrates potential applications for CNNs in drug discovery and protein engineering.

Competing Interests: *The authors declare that they have no conflict of interest.*
Funding Information: *The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.*
Author contribution: *All the authors have contributed equally.*
Data Availability Statement: *Not Applicable*
Research Involving Human and /or Animals: *Not Applicable*
Informed Consent: *Not Applicable*

REFERENCES

[1] Tillquist Richard C. Low-dimensional representation of biological sequence data. In: Proceedings of the 10th ACM International Conference onBioinformatics, Computational Biology and Health Informatics. BCB '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 555.

[2] Villmann Thomas, Schleif Frank-Michael, Kostrzewa Markus, Walch Axel, Hammer Barbara. Classification of mass-spectrometric data in clinical proteomicsusing learning vector quantization methods. Briefings Bioinform 2008;9(2):129–43.

[3] Schleif Frank-Michael, Villmann Thomas, Hammer Barbara. Prototype based fuzzy classification in clinical proteomics. Internat J Approx Reason2008;47(1):4–16.

[4] Alley Ethan C, Khimulya Grigory, Biswas Surojit, AlQuraishi Mohammed, Church George M. Unified rational protein engineering with sequence-based deeprepresentation learning. Nature Methods 2019;16(12):1315–22.

[5] A. E. W. Johnson, T. J. Pollard, L. Shen, H. Lehman, L. Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G Mark, "Mimic-III, a freely accessible critical care database, Scientific data, 3:160035, 2016.

[6] Nambiar Ananthan, Heflin Maeve, Liu Simon, Maslov Sergei, Hopkins Mark, Ritz Anna. Transforming the language of life: transformer neural networks for protein prediction tasks. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics.2020; P. 1–8.

[7] Heinzinger Michael, Elnaggar Ahmed, Wang Yu, Dallago Christian, Nechaev Dmitrii, Matthes Florian, Rost Burkhard. Modeling aspects of the language oflife through transfer-learning protein sequences. BMC Bioinformatics 2019;20(1):1–17.

[8] Madani Ali, McCann Bryan, Naik Nikhil, Keskar Nitish Shirish, Anand Namrata, Eguchi Raphael R, Huang Po-Ssu, Socher Richard. Progen: Languagemodeling for protein generation. 2020, arXiv preprint arXiv:2004.03497.

[9] Elnaggar Ahmed, Heinzinger Michael, Dallago Christian, Rehawi Ghalia, Wang Yu, Jones Llion, Gibbs Tom, Feher Tamas, Angerer Christoph,Steinegger Martin, et al. ProtTrans: towards cracking the language of life's code through self-supervised learning. 2021, BioRxiv, 2020-2007.

[10] Rives Alexander, Meier Joshua, Sercu Tom, Goyal Siddharth, Lin Zeming, Liu Jason, Guo Demi, Ott Myle, Zitnick C Lawrence, Ma Jerry, et al. Biologicalstructure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118(15).

[11] Rao Roshan, Bhattacharya Nicholas, Thomas Neil, Duan Yan, Chen Xi, Canny John, Abbeel Pieter, Song Yun S. Evaluating protein transfer learning withTAPE. In: Advances in Neural Information Processing Systems. 2019.

[12] Kimothi, D.; Soni, A.; Biyani, P.; Hogan, J. M. Distributed Representations for Biological Sequence Analysis. 2016, arXiv preprintarXiv:1608.05949.

[13] Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H. Learned Protein Embeddings for Machine Learning. Bioinformatics 2018, 34,2642–2648.

[14] Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. International conference on machine learning 2014,1188–1196.

[15] Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-GuidedDirected Evolution for Protein Engineering. Nat.Methods 2019, 16, 687.

[16] Wu, Z.; Kan, S. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with CombinatorialLibraries. Proc. Natl. Acad. Sci. U. S. A. 2019, 116, 8852–8858.

[17] Yang, K. K.; Wu, Z.; Arnold, F. H. Machine Learning in Protein Engineering. 2018,arXiv preprint arXiv:1811.10775.

[18] Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved pre-diction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med.2017;83:67–74.

[19] Zhou YZ, Gao Y, Zheng YY. Prediction of protein-protein inter- actions using local description of amino acid sequence. In: Zhou M, Tan H, editors. Advances in computer science and educationapplications. Berlin: Springer; 2011. p. 254–62.

[20] You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP, Ji Z. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. BMCBioinform. 2014;15(15):1–9.

[21] Xu H, Xu D, Zhang N, Zhang Y, Gao R. Protein-protein interaction prediction based on spectral radius and general regressionneural network. J Proteome Res. 2021;20(3):1657–65.

[22] Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: boosting prediction of protein–protein interactions with deep neuralnetworks. J Chem Inf Model. 2017;57(6):1499–510.

[23] Guo Y, Chen X. A deep learning framework for improving protein interaction prediction using sequence properties. bioRxiv,843755; 2019

[24] Wang X, Wu Y, Wang R, Wei Y, Gui Y. A novel matrix of sequence descriptors for predicting protein-protein interactionsfrom amino acid sequences. PLoS ONE. 2019;14(6): e0217312.

[25] Jha K, Saha S. Amalgamation of 3D structure and sequence information for protein–protein interaction prediction. Sci Rep.2020;10(1):1–14.

[26] Li H, Gong XJ, Yu H, Zhou C. Deep neural network based predictions of protein interactions using primary sequences. Molecules. 2018;23(8):1923.

[27] Gonzalez-Lopez F, Morales-Cordovilla JA, Villegas-Morcillo A, Gomez AM, Sanchez V. End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks. In 2018 IEEE International Conference on Bioinformatics andBiomedicine (BIBM). IEEE; 2018. p. 2344–2350.

[28] LeCun Y, Kavukcuoglu K, Farabet CC, others (2010) Convolutional networks and applications in vision. In: ISCAS. IEEE, pp 253–256.

[29] Lee C-Y, Gallagher PW, Tu Z (2016) Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: Artificial Intelligence and Statistics. pp 464–472.

[30] Li S, Liu Z-Q, Chan AB (2014) Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp 488–495.

[31] Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation Functions: Comparison of trends in Practice and Research for Deep Learning. arXiv Prepr arXiv181103378.

[32] Qureshi AS, Khan A (2018) Adaptive Transfer Learning in Deep Neural Networks: Wind Power Prediction using Knowledge Transfer from Region to Region and Between Different Task Domains. arXiv Prepr arXiv181012611.

[33] Shin H-CC, Roth HR, Gao M, et al (2016) Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 35:1285–1298.