Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7628796

A HYBRID APPROACH USING RNN AND CNN FOR PREDICTING 1-4D STRUCTURE OF PROTEIN FROM AMINO ACID SEQUENCES

Atrakesh Pandey[1], Dr. Rakesh Kumar Saxena[2],Dr. Dinesh Kumar Singh[3]

[1]Department of Computer Engineering, Poornima University, Jaipur, Rajasthan, India
[2]Department of Computer Engineering, Poornima University, Jaipur, Rajasthan, India
[3]Department of Information Technology, Dr. Shakuntala Misra National Rehabilitation University Lucknow, UP, India

ABSTRACT

Protein structure expectation is a particularly mind boggling issue that it is frequently assaulted and disintegrated using four distinct levels and they are: 1-D forecast of under- lying highlights along the essential succession of amino acids sequences,2-D forecast of spatial connections between the sequence of amino acids,3- D forecast of a tertiary structure of protein and quaternary structure of protein. This paper also try to introduce some assessment tools for finding the accuracy of result from applying ML and DL tools. And try to analyses and compare various algorithms based on deep learning methods verses machine learning methods used for sequence prediction. This paper also examines the turn of events and utilization of concealed Markov model, uphold vector machines, Bayesian techniques, and grouping strategies. This investigation will be helpful in creating future strategies to improve the exactness of protein auxiliary structure expectation. In this paper, also introduce and summarize the problem of essential elements of : 1)VAE, Variational Auto-Encoder 2)GAN, Generative adversarial network 3) RNN, Recurrent Neural 4) CNN, Convolutional Neural Networks protein structure prediction. Later on also summarizes the evolution of predictive algorithms for 1-4D structure of protein from Amino Acid Sequences and summarize the Deep Learning Ideas to Prediction of Structure of Protein and learned algorithms of the last decade.

Index Terms—CNN, RNN, VAE,GAN ,FFNN.

1. INTRODUCTION

Protein is a peptide linkage which is combined strictly together to make the chain of amino acids named because the of the protein structure. A protein might be a polymeric macromolecule consisting of 20 building blocks of amino acids which organized in a sequential links of 20 alphabets that joined by the peptide bonds. The chain of linear polypeptides is named the primary protein structure. The different physiological capacities performed by proteins across all creatures are interceded by the remarkable three- dimensional constructions embraced by explicit amino corrosive arrangements. Given the value , both monetarily and time insightful, related with tentatively deciding a protein's construction and execution , broad exertion has been made to foster computational techniques equipped for displaying the constructions of regular protein arrangements or potentially planning new groupings with novel designs and capacities past proteins saw in nature. The procedure of profound AI, which has altered many fields of exploration, including PC vision, discourse acknowledgment, methodology games and analysis, hugely affects protein structure expectation and style. during this survey, we'll feature techniques utilized for protein structure forecast and protein plan, additionally in light of the fact that the effect brought about by profound learning on these fields, where a chose accentuation are having the opportunity to be put on improvements that have happened inside the beyond couple of years. Protein can categories the structure of the protein into four categories: • 1D Structure of Protein: The link of amino acids arrangement is proteins which are included as protein essential structure and the succession of amino acids introduced in the polypeptide linking. Essential protein structure is settled by quality comparable protein.

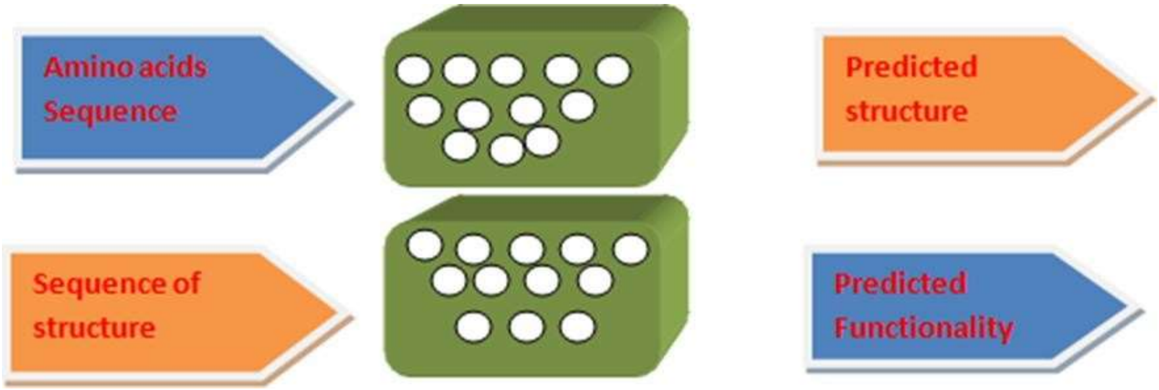


Fig. 1. Diagrammatic study of two tasks in structure prediction of protein: functionality predicted, structure predicted.

2D Structure of Protein: Secondary protein organized to comprise of nearby collapsing consistencies which are up- held by hydrogen bond. • 3D Structure of Protein: Protein tertiary organized to allude to 3-dimensional protein atom structure. It alludes 3-D collapsing of the polypeptide. It included neighborhood 3-D connection with essential protein structure parts. • 4D (Quaternary) Structure of Protein: Protein quaternary structure is a game plan of different loop protein atoms in a multi subunit complex. Less research work is done. ML approaches for protein 1-4 D structure prediction in quality evaluation of protein structure the different ML ideas are utilized.Some of them are portrayed as underneath:

Random Forest algorithms:- Random timberland performs extremely escalated estimations. It gives great outcomes from precision. It runs proficiently on enormous information bases. It proposes an exploratory method from recognizing variables associations. Irregular information sources utilized for back- woods of trees.

• M5P algorithms:- A classifier of a tree is M5P.It is a relapse model. It provides a larger precise outcome in quality assessment of 1-4D structure of protein.

Decision Stump algorithms:- It is considered as a powerless student and It depends • on 1-hub of choice trees • Cubist algorithms:- It is an expansion model of the M5P relapse model.

• Linear Model algorithms:- Linear strategy depict a nonstop reaction variable as a component of at least one assessed factor. This methodology helps in agreement and assessment of the conduct of complex programs. It additionally investigated monetary, trial and natural dataset. In straight model creation measurable technique direct relapse is to be utilized. • Foba algorithms:- It is a usage of learning calcula- tion. • Decision Tree algorithms:- It is a choice emotionally supportive network. It used diagrams like trees or displaying of their choice and result, likewise incorporating odds of occasion result, cost of asset and utility. It's an approach to speak to deal with that contains restrictive control articulations.

Clustering:- It depends on unaided learning. • Boosting:- It is troupe meta-calculation for fundamentally diminishing inclination. Boosting procedure of AI used to improve the exactness of assessment. It is broadly utilized in information science. • SVM:- It is a statistical procedure. It is a managed figuring out how to show related learning calculations that dissected information onto arrangement and relapse. • GA:- It gives excellent answers for search issues and advancement by depending on bio-roused tasks. • NN:- Neural Networks is a sub part of AI methods which work around fake network of neurons models and disperse across at least three layers.

HMM:- It is utilized in discourse, penmanship, motion acknowledgment, grammatical feature labeling, melodic score and bioinformatics. It is spoken to by powerful Bayesian organization. • ANN:- It is a delicate figuring strategies and factual AI calculations. It is a registering framework which is organic roused neural organizations.

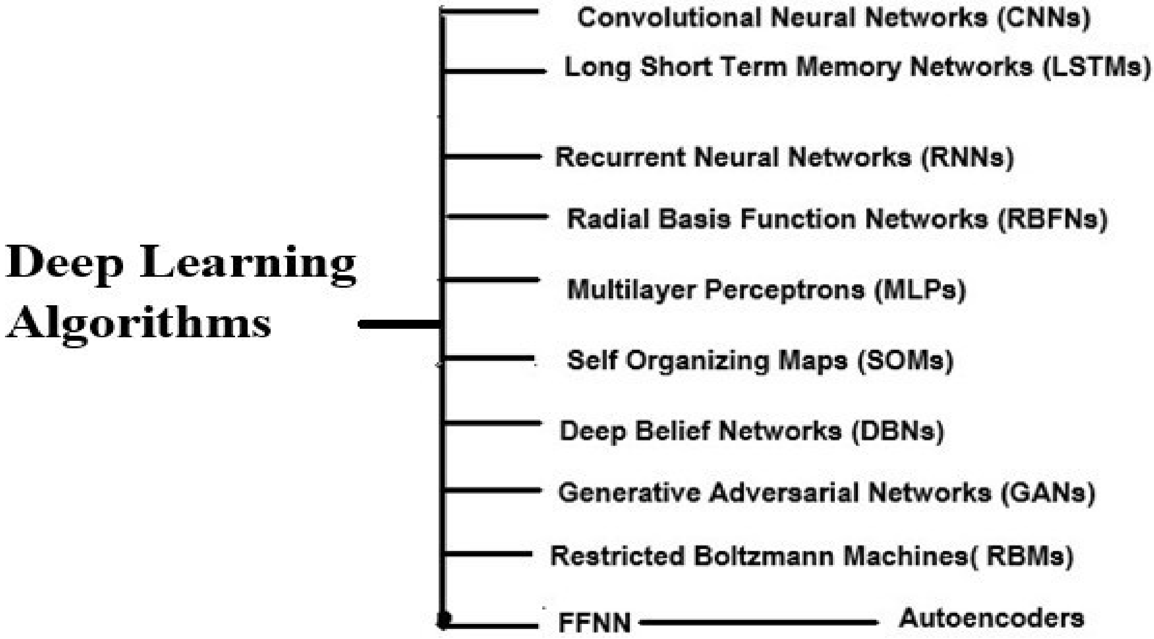

Deep Learning Algorithms
Convolutional Neural Networks (CNNs)
Long Short Term Memory Networks (LSTMs)
Recurrent Neural Networks (RNNs)
Radial Basis Function Networks (RBFNs)
Multilayer Perceptrons (MLPs)
Self Organizing Maps (SOMs)
Deep Belief Networks (DBNs)
Generative Adversarial Networks (GANs)
Restricted Boltzmann Machines(RBMs)
FFNN
Autoencoders

Fig. 2. Block Diagram of Deep Learning Algorithms

2. METHODOLOGY

This survey summed up the new advancement in applying all the strategies. DL to the issue of protein demonstrations and also talk about all the possible advantages and disadvantages and also limit the degree to capacity forecast and structure of protein, plan of protein with DL (see Fig. 1), and a widely large cluster of well-known systems utilized in these applications.

Collected and organized data of protein. After analyzing the data, it categorized into types and kind of structure of protein. [17]The protein is a polymeric macro-molecule consist- ing of building blocks of amino acids which organized in a sequential links that joined by the peptide bonds. The primary protein structure the linking of linear polypeptides.The primary structure is typically the twenty-letter alphabets represented by a sequence of letters and each letter is an amino acid. After categorization into types of structure of Protein different technologies used for predicting the structure of protein is studied and found an Artificial Intelligent approach, machine learning approach, Deep learning approach and some agent based methods which are described in table 1. In this table shows a matrix for existing methods for predicting all 1-4D structure of protein. Artificial Intelligent are large subject which make pc to act as human. By education a few records on it, it handled as device as gaining knowledge of. There are a few AI algorithms that imitate the operations of the human thoughts in making ready facts to be used in distinguishing objects, discourse perceiving, dialects deciphering, and selecting picks called Deep gaining knowledge of . In Deep gaining knowledge of man-made intelligence can research without human oversight, drawing from enter facts i.e each unlabelled and unstructured. Study movements like Amino acids sequences to shape prediction then via way of means of making use of it movements ahead to locating characteristic of protein and greater one leap forward offers sickness recognition.

In this paper all machine learning algorithms that are being used in predicting the 1-4 Dimensional structure of protein are collected. Also find available methods which mentioned in table 1, and also categories deep learning algorithms that are used any how in prediction of structure of protein which is shown in fig. 2, mathematical and statistical formulas and other tools that help you perform your research are essential in documenting your methodology. Further More, there are so many types of deep learning algorithms available in this literature survey found types of algorithm. The term profound comes from profound learning, a part of Machine Learning that centres on profound neural organizations. They have been canvassed broadly in the arrangement Understanding Deep Dreams, where they are acquainted with for an alternate (yet related) application. Neural organizations are computational framework inexactly roused by the manner by which the cerebrum measures data. Extraordinary cells called neurons are associated with one another in a thick organization (under- neath), permitting data to be prepared and sent. In Computer Science, counterfeit neural organizations are made out of thousands of hubs, associated in a particular design. Hubs are regularly organized in layers; the manner by which they are associated decides the sort of the organization and, eventually, its capacity to play out a specific computational errand over another. A conventional neural organization may resemble this. After analysing the machine learning as well as deep learning methods and applying algorithms for predicting structures, for ex: 3D structure prediction can be done by trRosetta. It is a computational procedure for quick and exact anew protein structure forecast. It fabricates the protein structure dependent on direct energy minimization with a controlled Rosetta. The restrictions incorporate between build-ups distance and direction appropriations, anticipated by a profound remaining neural organization. Then, at that point, contrast and machine v/s profound learning approaches and attempt to show how Protein structure expectation by profound learning and its effect achieved these learning. Deep computational procedures are shown in fig. 2, and process and concepts are shown in fig. 3.

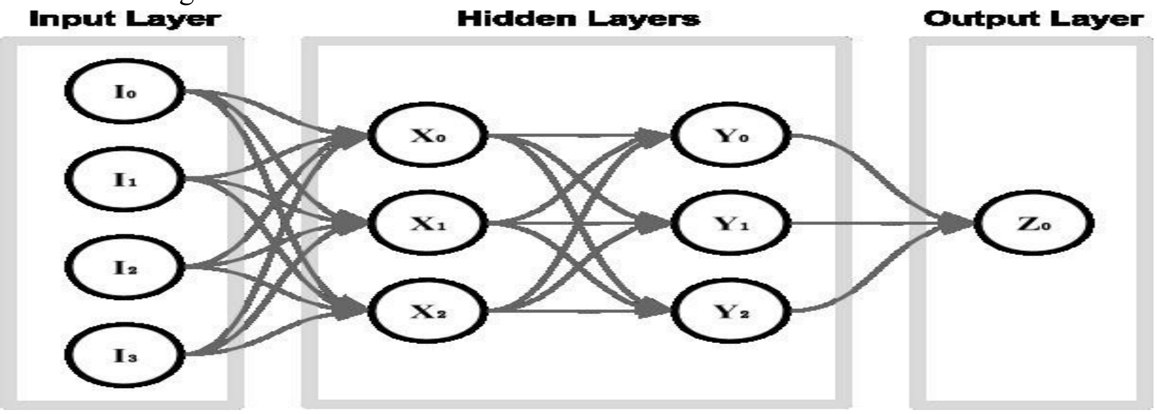


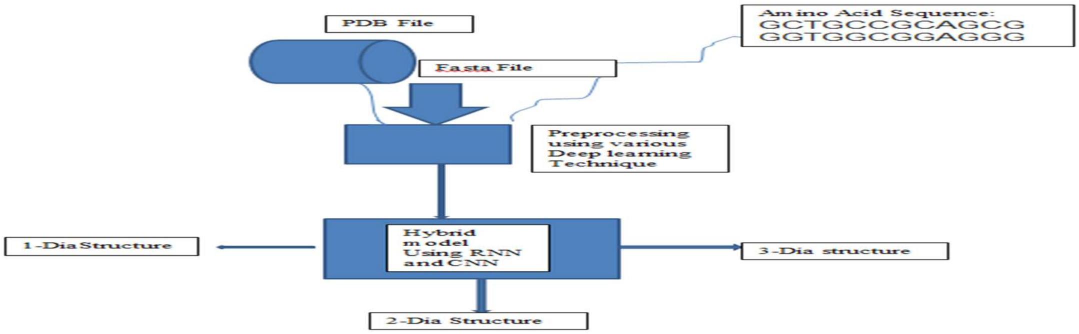
Fig. 3. Block Diagram of Deep Learning concept

Proposed Methodology:

In proposed method only the basic knowledge of Python is required and various m/c learning and deep learning algorithms may be applied for classification and regression. The data is collected from SwissProt and protein database(PDB) website and prepared labels as well as validate and annotated these dataset. Features are extracted & Classified through convolutional neural network. A significant variation of the CNN is the lingering organization which fuses skip-associations between the layers. Fig. 8 shows the complete process.

RNN: A RNN approach with regards to protein forecast is utilizing a N-terminal .

VAE: Encoded Images, applying VAE in the protein displaying field.



GAN: Regenerate the images, GAN objective can experience the ill effects of issues.

Fig 8: Flow graph

3. RESULTS ANDDISCUSSION

Deep learning algorithms of prediction of structure of protein from amino acid sequences: Expectation of protein 3- D structure this might be a stupendous test in computational biochemistry quit file the combination of peptide bonds may be a major idea in biochemistry and nuclear labor architecture of protein and interfaces are the frequently this part to get their capacity and to balancing and designing them on account of the new advances in leading is of Sequencing the amino acids innovations they are presently a quiet 180 million millions of protein grouping recorded in data Bank of protein that is PDB has decided that there are 1540000 the computational structure for caste may be a basic issue of both comments and hypo- thetical interest of the Mole as lately the advances in the three dimensional structure of protein for cost is prompted. All the more as lately, the structural advances forecast have prompted an expanding interest within the protein plan issue. In plan, the goal is to urge a completely unique protein arrangement which will crease into a perfect structure or play out a specific capacity, for instance, catalysis. Normally happening proteins speak to only a minuscule subset of all conceivable amino corrosive arrangements chose by the transformative cycle to play out a specific natural capacity. [2] Proteins with more vigor (higher warm steadiness, protection from corruption) or improved properties (quicker catalysis, more tightly official) may dwell the space that has not been investigated commonly, yet is conceivably open by again plan. The present methodol- ogy for computational again configuration depends on physical and developmental standards and requires huge space mastery.

Conventional computational approaches:

The latest technique for the computation of protein structure forecast is typically founded on speculation based on Anfinsen's thermodynamic, [7] which attempt to express all the local structures of a 1protein should be with the small amount energy that are free which is represented by all energy scene of potential adaptations which are related with its succession.[8]Furthermore, the methodology used to precise free energy capacity to portray the protein energy scene as well as rank various adaptations hooked in to their energy, alluded because the "scoring issue". Considering these difficulties, current computational strategies depend intensely on multi- scale draws near.[9]Notwithstanding criticality in advancement of the foremost recent a really long while in the area of protein structure computational forecasting and plan, [2,6] exact expectation of structure and testing. Regular method- ologies depend intensely on the energy capacities exactly to portray protein material science and hence the proficiency of examining calculations to research the protein arrangement and space in the structure.

Deep learning algorithms:

In regular computational methodologies, expectations from information are made by methods for actual conditions and demonstrating. AI advances an alternate worldview during which calculations naturally induce – or learn – a connection among data sources and yields from a bunch of theories. The decision of the organization decides how the theory class is defined. Profound neural organizations regularly actualize a non-straight capacity as the synthesis of relative guides, normally completed by means of stochastic slope plummet calculations or varieties thereof, productively actualized through back-proliferation. Or maybe, in this segmentsummed up the absolute most well-known models broadly utilized in protein primary demonstrating. Significant level graphs of the significant models are appeared in Fig. 4,5,6,7.

Convolutional neural networks (CNN):

There exists a discriminative Convolutional neural network (CNN) that is profound design made out of various pooling layers and convolutional. Process of CNN is enlivened by the straightforward and the model that contains complex cell for visualization. The information is convolved with teachable channels and inclinations to register neighborhood highlights like straightforward cells in the convolutional layer. At that point, the pooling layer and like complex cells consolidates the neighboring highlights through light weighted entirety and a nonlinear change, delivering a more theoretical arrangement of highlights of diminished measurement. The pooling layers and convolutional can be exchanged on various occasions in a CNN, shaping a progressive profound design. The yields of the last pooling layer become the contribution to a standard neural organization that makes the CNN's last yield. [10]There are some change required in designing of CNN to look at the various picture issue where more requirement of in variance is or co change. [3]CNNs embrace convolution bits for the layer-wise relative change to catch this translational invariance remains be in similar article. A significant variation of the CNN is the lingering organization which fuses skip- associations between the layers. [11] A model is AlphaFold5 in which the info is given and the yield is a relating buildup distance map (see Fig. 4). CNNs are built to accomplish incredible powerful and accurate.

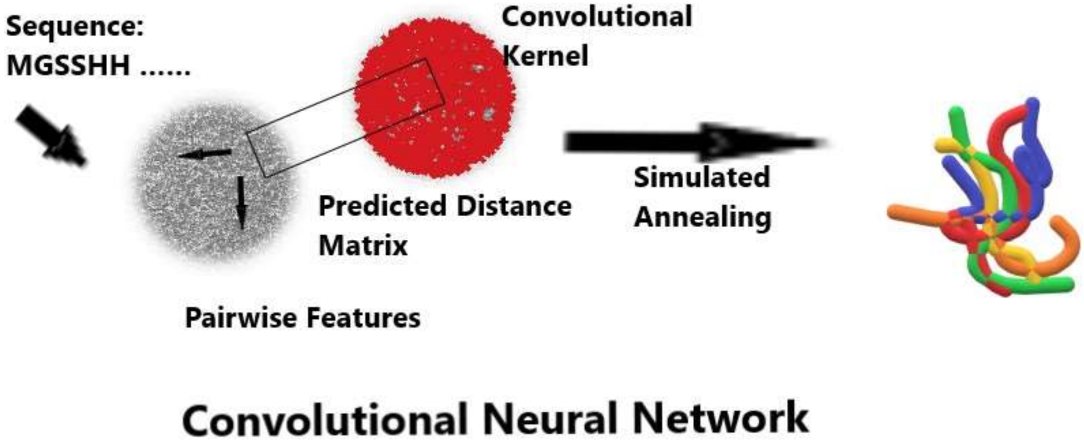


Fig. 4. CNN

Recurrent neural networks (RNN):

Intermittent designs depend on applying a few emphases of a similar capacity along successive information. [12] This can be viewed as an unfurled engineering, and has been broadly used to deal with consecutive information, for example, composed content and time arrangement information. An illustration of a RNN approach with regards to protein forecast is utilizing a N-terminal aftereffect of a protein and foreseeing the following amino corrosive in the protein (Fig. 5; for example Muller et al. [13]). As comparative game plan of limits can be applied discontinuously along the back to back data, a commitment of variable length can be dealt with to a RNN. As a result of the incline vanishing and impact issue (the bumble signal decreases or additions drastically during planning), later varieties of standard RNN, explicitly Long Short-Term Memory (LSTM)[14]and Gated Recurrent Unit (GRU)[14]are largely the more comprehensively used.

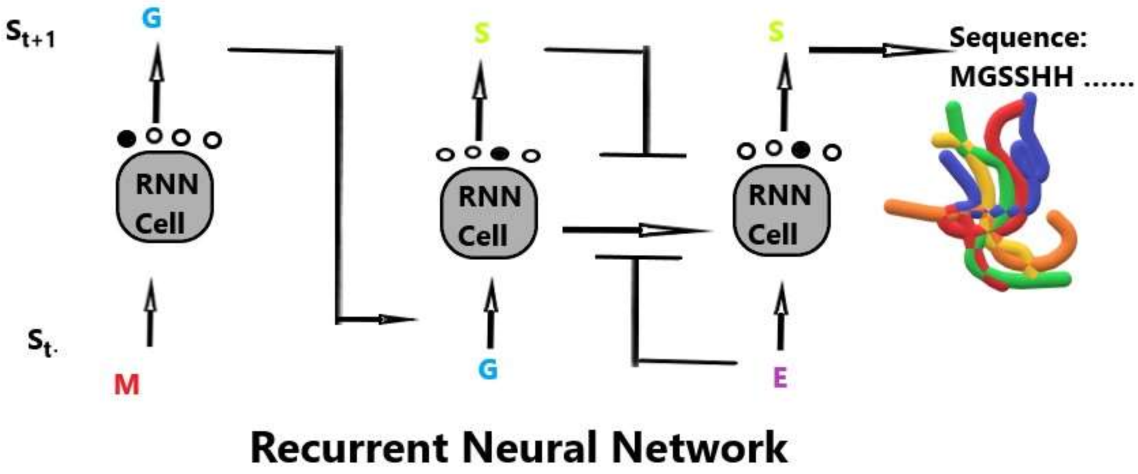


Fig. 5. RNN

Variational auto-encoder (VAE):

Auto-Encoders [15](AEs), is dissimilar to the ones talked about up until this point, give a model to solo learning. Inside this solo system, an auto-encoder doesn't learn marked yields yet rather endeavors to get familiar with some portrayal of the first info. Variation Auto- Encoders (VAEs),specifically, give a stochastic guide between the inactive space and the info space. The dispersion of the portrayal z can be a lot less difficult; e.g., Gaussian. These strategies are gotten from variation surmising, a strategy from AI densities improvement. Kullback–Leibler (KL).One can utilize variation factorization and probability dispersion For the productive improvement of these models .An illustration of applying VAE in the protein displaying field is learning a portrayal of against microbial protein arrangements (Fig. 6;e.g., Das et al.[15]).

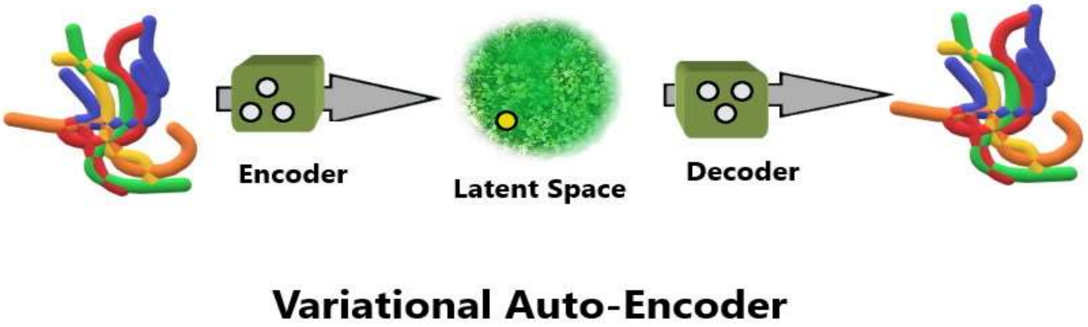


Fig. 6. VAE

Generative adversarial network (GAN):

[16]Generative adversarial networks (GANs) are another class of unaided (generative) models. Not at all like VAEs, GANs are prepared by an ill-disposed gap between two models: a generator and a discriminator, D, whose assignment is to identify whether the pictures are genuine Preparing is performed by stochastic streamlining of this differentiable misfortune work. While natural, this unique GAN objective can experience the ill effects of issues, for

example, mode breakdown and insecurities during preparing. [16] The Wasser- stein GAN (WGAN) is a well-known augmentation of GAN which presents a Wasserstein-1 distance measure between disseminations, prompting simpler and more hearty preparing practically speaking. An illustration of GAN, with regards to protein 4demonstrating, is crafted by Anand et al., to gain proficiency with the dissemination of protein spine removes and produces novel protein-like folds (Fig 7).

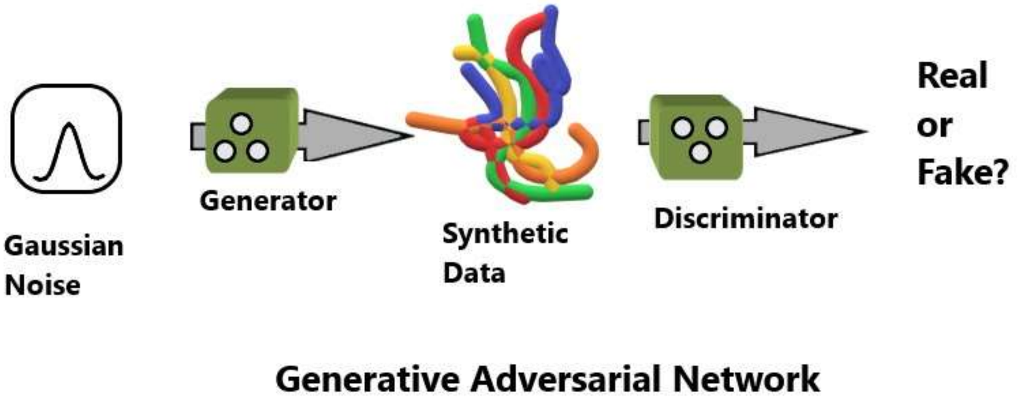


Fig.7. GAN

Work	Ideas	AlgorithmLearned
Secondary StructurePrediction	Extraction of connected Featurelayersto adeep stackofCNN. Deep learned networkMulti-layerconvolution generativestochasticnetwork	Multi-task learningapplyingback-propagation Unsupervised pre-training +supervised fine-tuning +applyingback-propagation forMulti-tasklearning
Residue-Residue ContactsPrediction	Multi-tasklearningusingback-propagation	applying back-propagationforLevel-wise,hierarchicalsupervisedleaning and Unsupervised pre-training +supervised fine-tuning +boosting
DisorderedRegionPrediction	Boostedensemblesofdeepnetworks	Unsupervised pre-training +supervised fine-tuning +boosting

Backbone Dihedral AnglePrediction	Stackedsparseauto-encoder	Unsupervised pre-training +supervisedefine-tuning
Turnerystructureprediction	RNN,CNN,RBMS,FNN	Unsupervised pre-training +applying back-propagationforsupervisedefine-tuning
Primarystructureprediction	RELU	Supervisedtrainingondataset

Tab1e 2.1: A Summary of Deep Learning Ideas to Prediction of Structure of Protein

4. Conclusion

RNN and VAE and GAN combination make more accurate. It is additionally seen that still presentation should be improved. Design forecast is still will be as yet open for research with extent of precision improvement. Mix of on the web and disconnected provisions could be way out. Pre prepared CNN can expand precision. Analyses can be performed on open dataset with more number of pictures and with more number of Epochs.

This current paper, have summed up the all scenario of the DL procedures applied to the issue of structure of protein forecast and plan. As in numerous different zones, DL shows the possibility to reform the field of protein displaying. In table 2 summarizes of Deep Learning Ideas to Prediction of Structure of Protein. The proposed methodology i.e the hybrid approach definitely improves accuracy.

The exhibition of DL strategies depends vigorously on the nature of information; however the openly accessible datasets may not cover significant example space on account of exploratory availability at the hour of tests.

At last, with the expectation of acquiring understanding into the major study of bimolecular, there is a craving to connect man-made consciousness (AI) ways to deal with the hidden rule. In this manner, while endeavors carefully restricted to groupings are plentiful, and accepted that models with primary experiences will assume a more basic job later on.

At last this straight audit sketching out the latest job of Deep Learning methods in the more extensive stacks to foresee structures of protein as well as attempting the envision whatever the opportunities as well as challenges are emerging away is finished up. Deep computational procedures have altogether affected protein structure expectation and protein plan over AI.

References

[1] Ranjan, A.; Fahad, M. S.; Fernandez-Baca, D.; Deepak, A.; Tripathi, S. Deep Robust Framework for Protein Function Prediction using Variable-Length Protein Sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019, 1–1.

[2] Huang, P. S.; Boyken, S. E.; Baker, D. The coming of age of de novo protein design. Nature 2016, 537, 320–327.

[3] Bohr, H.; Bohr, J.; Brunak, S.; J. Cotterill, R.; Fredholm, H.; Lautrup, B.; Petersen, S. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks.FEBS letters 1990, 261, 43–46.

[4] Sutton, R. S.; Barto, A. G. Reinforcement learning: An introduction; MIT press, 2018.

[5] Li, Y.; Huang, C.; Ding, L.; Li, Z.; Pan, Y.; Gao, X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods 2019, 166, 4–21.

[6] King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; Andr´e, I.; Gonen, T.; Yeates, T. O.; Baker, D. Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 2012, 336, 1171–1174.

[7] Jordan, M. I. Serial order: A parallel distributed processing approach. Advances in Psychology 1997, 121, 471–495. (47) M¨uller, A. T.; Hiss,
J. A.; Schneider, G. Recurrent Neural Network Model for Constructive Peptide Design. Journal of Chemical Information and Modeling 2018, 58, 472–479.

[8] Cho, K.; Van Merri¨enboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares,
F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation.arXiv preprint 2014, 1406.1078.

[9] Hinton, G. E.; Zemel, R. S. Autoencoders, minimum description length and Helmholtz free energy. Advances in Neural Information Processing Systems 1994, 3–10.

[10] Kingma, D. P.; Welling, M. Auto-encoding variationalbayes. arXiv preprint 2013, 1312.6114.

[11] Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational inference: A review for statisticians. Journal of the American Statistical Association 2017, 112, 859–877.

[12] Hamel, L., Sun, G., Zhang, J. (2005, November). Toward protein struc- ture analysis with self-organizing maps. In 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (pp. 1-8). IEEE.

[13] Das, P.; Wadhawan, K.; Chang, O.; Sercu, T.; Santos, C. D.; Riemer, M.; Chenthamarakshan, V.; Padhi, I.; Mojsilovic, A. Pepcvae: Semi- supervised targeted design of antimicrobial peptide sequences. arXiv preprint 2018, 1810.07743.

[14] Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., ... Cao, R. (2019). Survey of machine learning techniques in drug discovery. Current drug metabolism, 20(3), 185-193.

[15] Nguyen, S. P., Li, Z., Xu, D., Shang, Y. (2017). New deep learning methods for protein loop modeling. IEEE/ACM transactions on compu- tational biology and bioinformatics, 16(2), 596-606.

[16] ul Hassan, F. (2017, December). Deep and Self-Taught Learning for Pro- tein Accessible Surface Area Prediction. In 2017 International Conference on Frontiers of Information Technology (FIT) (pp.264-269)

[17] Pandey, A., Jain, R. (2023). 1–4D Protein Structures Prediction Using Machine Learning and Deep Learning from Amino Acid Sequences. In Proceedings of the Third International Conference on Information Man- agement and Machine Intelligence (pp. 615-621). Springer, Singapore.