CHAPTER 14

Generative adversarial networks in protein and ligand structure generation: a case study

Syed Aslah Ahmad Faizi[1], Nripendra Kumar Singh[2], Ashraf Kamal[3] and Khalid Raza[2]

[1]*Chennai Mathematical Institute, Chennai, India* [2]*Department of Computer Science, Jamia Millia Islamia, New Delhi, India* [3]*PayPal, Chennai, India*

14.1 Introduction

In all living organisms, proteins are essential and naturally occurring molecules that have numerous properties and functions. Proteins are made up of 20 naturally occurring amino acids that are joined together by polypeptide bonds, forming long chains (Raza, 2017). The arrangement of different amino acids in specific order determines the three-dimensional (3D) structure and functions of the protein. Why is a protein's 3D structure important? The answer to the questions is that a protein's biological function is dictated by its 3D structure, that is, the arrangement of the atoms in 3D space. A protein structure would provide an understanding of how a protein works such as allowing us to design site-directed mutations to change its function and help us to predict molecules that may bind to a protein (Alberts, Johnson, & Lewis, 2002).

The structure of the protein is too complex to determine. Its structure has been determined by structural biologists down to the atomic level. The structure of proteins is usually classified into four levels, including primary (amino acid sequences), secondary (helices, sheets, turns, coils), tertiary (3D structures), and quaternary (subunits) (Alberts, Johnson, Lewis, Raff, et al., 2002). Understanding these four levels is important to understand how proteins get their final shape. The primary structure comprises an amino acid sequence that makes a polypeptide chain. Although it is not really a structure, just a sequence of amino acids, however, protein structure depends on this sequence (Idicula-Thomas & Balaji, 2005). The *secondary structure* comes from interactions of amino acids where the polypeptide chain begins to fold into functional 3D form. There are two main types of protein secondary structures: *alpha-helix* and *beta-sheet*. In an alpha-helix, the protein chain coils into a spiral shape stabilized by hydrogen bonds between the carbonyl oxygen atom of one amino acid residue and the amino hydrogen atom of another amino acid residue further along the chain (Egli & Zhang, 2022). In a beta-sheet, the protein chain forms a flat, sheet-like structure with the amino acid residues arranged in rows, stabilized by hydrogen bonds between the carbonyl oxygen and amino hydrogen atoms in adjacent strands (Kim et al., 2015). The *tertiary structure* formed by a polypeptide chain is the overall 3D shape of the protein. The complex 3D tertiary structure is formed by the interactions between polar, nonpolar, acidic, and basic R groups (Godbey, 2022). In case the protein loses its 3D shape, it would no longer be a functional protein. The *quaternary*

Deep Learning Applications in Translational Bioinformatics. DOI: https://doi.org/10.1016/B978-0-443-22299-3.00014-1

structure refers to the spatial arrangement of multiple protein subunits that come together to form a larger, functional protein complex. The quaternary structure of a protein is important for its overall function, stability, and regulation. Many proteins, such as enzymes, receptors, and transporters, exist as multimeric complexes with distinct quaternary structures (Bhagavan, 2002).

14.1.1 Three-dimensional structure of a protein

The 3D structure of a protein refers to how the amino acid residues that make up the protein are arranged in 3D space (Ittisoponpisan et al., 2019). The structure of a protein is critical to its function, and different proteins can result in different biological activities. The 3D structure of a protein is typically determined using techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. Once the structure has been determined, it is often deposited in a publicly accessible database, such as the Protein Data Bank (PDB) (Burley et al., 2017). The PDB format is a standard file format for the representation of protein structures. It includes information about the amino acid sequence of the protein, the positions of the atoms in the protein, and the types of bonds and interactions between the molecules. The PDB file can be viewed and analyzed using various software tools, such as molecular visualization software or computational biology software (2023).

Understanding the 3D structure of a protein is important for understanding its function. The structure can help researchers to identify key regions of the protein that are involved in interactions with other molecules, and to design drugs or other therapies that can target these regions. The structure can also be used to study the evolution of proteins, as well as to engineer new proteins with specific functions (Dhanjal et al., 2018).

14.1.2 Computer representations of 3D structure of protein

The 3D structure of a protein can be represented inside a computer using a variety of software tools (2023). Many different software programs can be used to visualize, manipulate, and analyze protein structures, and they typically rely on a variety of algorithms and computational methods to represent the 3D structure of the protein (Schmidt et al., 2014). One common way to represent the 3D structure of a protein is through the use of molecular visualization software. These programs allow researchers to view and manipulate the 3D structure of the protein on a computer screen. The software typically uses a variety of graphics techniques to represent the protein, including wireframe models, ball-and-stick models, and space-filling models. These models can be colored and labeled to highlight specific regions of the protein, such as the active site or a particular domain (O'Donoghue et al., 2010).

Another way to represent the 3D structure of a protein is through the use of computational biology software (2023). These programs use algorithms and mathematical models to analyze the structure of the protein and to make predictions about its function. For example, computational biology software can be used to predict the binding affinity of a drug to a particular protein or to model the effects of a mutation on the protein's structure and function (Petukh et al., 2015).

In addition to visualizing and analyzing the 3D structure of proteins, computer representations of protein structures can also be used to store and share data about the protein. The PDB (https://www.rcsb.org/), for example, is a public database that stores 3D structural data for proteins and other biomolecules. The data in the PDB are stored in a standardized file format that can be read and analyzed using a variety of software tools. Hence, the representation of the 3D structure of proteins inside a computer relies on a combination of computational algorithms, mathematical models, and graphical techniques. These tools enable researchers to study and understand the complex structures and functions of proteins in a variety of biological systems (Meier et al., 2015).

14.1.3 Protein structure for machine-learning models

Protein 3D structure is a valuable input for machine-learning models that can be trained to predict new protein structure, protein function, stability, and interactions using information from the 3D structure of the protein (Baek et al., 2021). One way to use protein 3D structure for machine-learning models is through the use of protein structure prediction algorithms. These algorithms use computational methods to predict the 3D structure of a protein based on its primary sequence. Once the structure has been predicted, it can be used as input for machine-learning models that can predict protein function or interactions. Another way to use protein 3D structure for machine-learning models is through the use of feature engineering. Feature engineering involves identifying and extracting specific features or characteristics from the protein's 3D structure that are relevant to the machine-learning model. For example, features such as solvent accessibility, secondary structure, and binding pockets can be extracted from the protein's 3D structure and used as input for the machine-learning model (AlQuraishi, 2021).

Machine-learning models can also be used to predict the effects of mutations on the protein's structure and function. These models can be trained using data from experimentally characterized mutations and can be used to predict the effects of novel mutations on the protein's stability, binding affinity, or enzymatic activity. Hence, the 3D structure of proteins can be a powerful input for machine learning models to predict protein function, stability, and interactions, as well as to identify mutations that can lead to disease or that may have therapeutic potential (Pandurangan & Blundell, 2020).

14.2 Generative adversarial networks: a brief overview

A generative adversarial network (GAN) is a deep-learning-based generative model (GM) developed by Goodfellow et al. (2014). It is a special kind of GM, which learns the GM of data distribution by applying an adversarial approach via a generator and discriminator. It consists of two neural networks working in tandem: a generator network and a discriminator network, illustrated in Fig. 14.1. The generator network creates new data samples while the discriminator network evaluates the authenticity of the generated samples. Both networks are trained together through an adversarial process, with the goal of the generator network learning to create data samples that are indistinguishable from real data by the discriminator network (Goodfellow et al., 2014). Recently, they are considered one of the most successful GMs. They are based on game theory as compared to other GMs, which are mainly based on optimizations. They are designed to solve the GM problem efficiently due to their ability to generate realistic, high-resolution images. The main applications of GANs include image generation, video generation, text generation, music generation, data

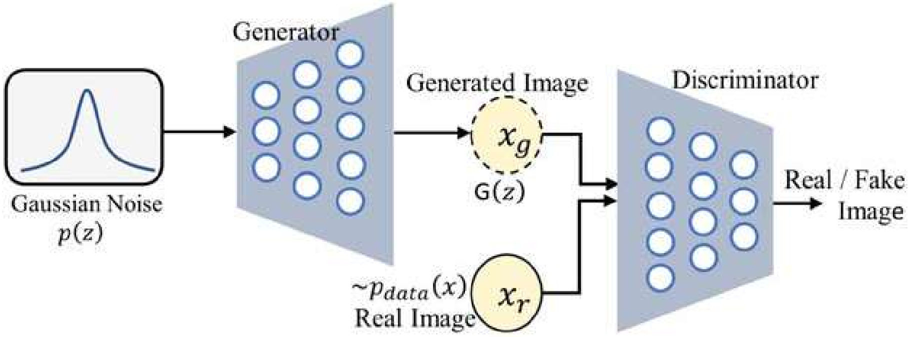

FIGURE 14.1

Illustration of GAN model having the capability of generating a synthetic image from priori distribution of input data. *GAN*, Generative adversarial network.

augmentation, text-to-image translation, generation of human poses, and 3D-object generation (Tang et al., 2021). Besides that, GANs also have wide applications in medical image processing and biomedical informatics (Raza & Singh, 2021; Singh & Raza, 2021). In other words, GANs can be used to create new, realistic-looking images or videos, generate realistic text or music, and create larger datasets for training other machine-learning models (Shahriar, 2022).

Unlike other unsupervised models, GANs are trained by a competitive game between two networks, a generator (G) that tries to map a random selection z, from a distribution, $P_z(z)$ such as Gaussian noise, to the distribution of a class of data $G(z)$ (e.g., an instance of real images); and a discriminator (D) whose job it is to determine whether the generated images are real or fake, that is, whether they belong to the real distribution $P_{data(x)}$ or not. The initial objective function used for training the generator and discriminator is represented by the mathematical equation given below, which is formulated by Goodfellow et al. (2014).

$$\min_G \max_D V(D,G) = E_{x_r \sim P_{data(x)}}[\log D(x_r)] + E_{x_g \sim P_z(z)}[1 - \log D(G(z))] \tag{14.1}$$

The differentiable loss function is trained through stochastic optimization. This initial GAN objective suffers from problems, such as mode collapse, vanishing gradient, problems with counting, and instabilities during training, despite the fact that it is straightforward. The Wasserstein GAN is a well-known extension of basic GAN that extends an alternative training scheme to the generator model and replaces the probabilistic discriminator with a critical score that realizes that the model is stable, faster, and reliable for training (Kurach et al., 2019).

There are several types of GANs, including conditional GANs, progressive GANs, CycleGANs, and InfoGANs. Conditional GANs allow the generation of samples conditioned on a specific input, while progressive GANs improve the quality of the generated samples over time. CycleGANs (Sandfort et al., 2019) enable the generation of samples in a different style or domain, while InfoGANs (Chen et al., 2016) aim to learn disentangled representations of the input data.

GANs can also be used for generating protein and drug ligand structures. In this case, the generator network is trained to produce novel molecular structures that meet specific criteria, such as having a certain biological activity or being structurally similar to known compounds. The discriminator network evaluates the generated structures based on their chemical properties, such as their solubility or binding affinity, to ensure that they are chemically valid. This approach has the potential to accelerate drug discovery by generating novel compounds with specific biological activities, which can then be synthesized and tested in the lab. However, it is important to note that the generated compounds still need to be validated through experimental testing, as GANs are not able to fully capture the complexity of biological systems (Gupta et al., 2018; Polykovskiy et al., 2018).

14.3 Machine learning in protein and ligand structure prediction

Protein structure prediction using Machine Learning involves developing algorithms that can predict the 3D structure of a protein and ligands based on complex multidimensional amino acid sequences and other relevant data. This is a challenging task due to the complexity of the problem and the substantial number of structures. Several types of Machine-Learning algorithms, such as deep residual networks (He et al., 2016), Attention networks (Vaswani et al., 2017), and various forms of GANs (Singh & Raza, 2021) have been used to tackle this problem. These algorithms learn from large datasets of known protein structures and use this knowledge to predict the structure of new proteins. However, the accuracy of these predictions is still limited, and improving it remains an active area of research in computational biology. Recently, several GAN-based approaches have been developed and have shown promising results in predicting protein structures. In this section, we are going to discuss some of the potential deep learning models that achieved state-of-the-art results.

14.3.1 AlphaFold

AlphaFold (https://www.nature.com/articles/s41586-021-03819-2) is a deep learning-based method developed by the DeepMind team. It uses a neural network to predict the 3D structure of a protein sequence. AlphaFold has been widely recognized for its high accuracy and was used to predict the structures of more than 350,000 proteins in the PDB in 2021.

14.3.2 RoseTTAFold

RoseTTAFold (https://www.science.org/doi/10.1126/science.abj8754) is another deep learning-based method for protein structure prediction. It was developed by a team of researchers from the University of Washington and uses a neural network that combines information from multiple sources, including evolutionary information and protein residue-residue contacts, to predict the 3D structure of a protein.

14.3.3 RaptorX

RaptorX is a protein contact map prediction system based on deep residual networks. Wang et al. (2016) use a combination of template-based modeling and machine-learning methods to predict proteins' secondary and tertiary structures, distance, and contact maps by concentrating on the fundamental issue of transforming coevolutionary inputs into practical geometric constraints (Wang et al., 2016). The key innovation of RaptorX is its use of two deep residual neural networks, the first one-dimensional (1D) residual network performed a series of convolutional transformations of 1D sequential features. By using outer concatenation, the output of this 1D network is transformed into a two-dimensional (2D) matrix and input into the second module. The second module is a 2D residual network that transforms its input through several 2D convolutional operations. To estimate the pairwise distances between amino acids in a protein and the protein's 3D structure, the predictor module employs a deep neural network. RaptorX's deep neural network was able to discover complex correlations between amino acid sequences and protein structures since it was trained on a substantial dataset of protein sequences and structures. RaptorX has been utilized in numerous research to predict the structures of proteins with significant biological roles in addition to having high accuracy on protein structure prediction benchmarks. RaptorX was employed, for instance, in recent work to forecast the structure of the SARS-CoV-2 spike protein (Barbhuiya & Ahmad, n.d.; Jaimes et al., 2020) that is essential for virus entrance into host cells. The predicted structure offered insights into the mechanism of virus penetration and was shown to be in strong accord with experimental findings. Overall, RaptorX is a powerful protein structure prediction system that has been widely used in the scientific community.

14.3.4 SPARKS-X and SVM-fold

SPARKS-X is a protein structure prediction server that uses an SVM-based method to predict the 3D structure of a protein. The method combines sequence- and structure-based features to make its predictions (Yan et al., 2021).

14.3.5 CONFOLD and SPOT-ROD

CONFOLD is a protein structure prediction server that uses a random forest-based method. The method combines multiple sources of information, including coevolutionary information, to predict the 3D structure of a protein (Stansfield et al., 2018).

14.3.6 PEP-FOLD and CONFOLD

PEP-FOLD is a protein structure prediction server that uses a Bayesian-based method to predict the 3D structure of a protein. The method combines a variety of sources of information, including secondary structure prediction, solvent accessibility prediction, and homology modeling (Adhikari et al., 2018).

14.3.7 Rosetta and MODELLER

Rosetta is a software suite developed by the University of Washington that uses computational algorithms to predict the 3D structure of a protein (Leaver-Fay et al., n.d.). It utilizes a combination of physics-based energy calculations and machine-learning techniques to generate accurate protein structures. Rosetta can be used to predict the structure of proteins from scratch or to refine existing models. It is also capable of predicting protein—protein and protein—ligand interactions.

MODELLER, on the other hand, is a software package developed by the University of California, San Francisco, that uses comparative modeling techniques to generate protein structures (Webb & Sali, 2016). Comparative modeling involves building a protein model based on the known structure of a homologous protein. MODELLER utilizes a number of different algorithms to generate accurate models, including molecular dynamics simulations, energy minimization, and optimization of various structural parameters. It is particularly useful for modeling large proteins and protein complexes. Rosetta and MODELLER have been widely used in protein structure prediction and have contributed significantly to our understanding of protein structure and function. However, each software package has its own strengths and weaknesses, and researchers often use a combination of different methods to generate the most accurate protein structures.

14.4 Generative modeling for protein and ligand structures

Generative modeling is capable of learning the combined distribution of protein and ligand conformations that further enables principled sampling of diverse conformations and gives important insights into their ensemble attributes. Several GMs, as given below, have been used for protein design with their own limitations and trade-offs.

14.4.1 Autoregressive models

In *autoregressive models*, the outcome of the next token in textual data depends on its previous tokens. Likewise, protein sequences are taken as tokens using autoregressive models. Considering this, Alley et al. (2019) proposed protein sequences through recurrent neural networks (RNNs) and long short-term memory layers to predict the sequences of amino acids using the previous amino acid sequences. Ingraham et al. (2019) introduced Structured Transformer that follows an encoder–decoder architecture. The role of the encoder layer is to take the protein structure as input. The decoder is responsible for giving the amino acid sequences and self-attention to the preceding residues as output. Strokach et al. (2020) developed ProteinSolver that is based on a graph neural network. It generates novel sequences that refer to the stable proteins and desired topologies. It is better than transformers in terms of protein stability and affinity prediction.

14.4.2 Variational autoencoders

It consists of an encoder–decoder network, wherein the former is used to map the inputs to a low-dimensional hidden space and later is used to reconstruct those inputs by utilizing the sample from that low-latent space. Variational autoencoders (VAE) is applied for novel protein sequence generation by leveraging predetermined functions. In this direction, Greener et al. (2018) trained a conditional VAE, wherein a rough topology of the protein is considered as input with about 4000 short monomeric structures with their homologs. As a result, new protein sequences related to a specified topology are generated. Eguchi et al. (2020) considered a distance matrix as input and generated 3D coordinates using VAE. They have observed that the input matrix and the torsion angles are matched for a particular protein structure.

14.4.3 Normalizing flows

It learns a mapping between the inputs and the hidden representation in bidirectional mode. The modeling of protein dynamics is one of the most important applications of normalizing flows (NFs) with respect to protein design. Noé et al. (2019) developed Boltzmann generators that use a set of protein conformations and energies. It generates new conformations via molecular dynamics simulations and performs model transitions and energy differences.

14.4.4 Energy-based models

They have been applied to learn semantic representations of protein sequences and structures. Gainza et al. (2020) proposed a model called MaSIF. It is trained to map protein surface meshes into fingerprints to perform protein−protein interaction prediction. It is observed that protein docking has done significantly better than traditional approaches with respect to accuracy values.

14.4.5 Generative adversarial networks

It is considered a subset of energy-based models (EBMs). GANs are used to generate distance matrices and new protein sequences using folds and functions. Considering this, the existing method represents novel protein folds using GANs. Repecka et al. (2021) trained a GAN model on the malate dehydrogenase (MDH) sequences dataset. LiGANN (Skalic et al., 2019) is a structure-based de novo drug design tool, wherein for an input protein shape, a GAN gives complementary ligand shapes in a multimodal way. Anand and Huang (2018) trained a GAN model to generate distance matrices for novel protein folds. It consisted of 2D layers. Qiao et al. (2022) proposed NeuralPLexer, a deep GM framework. It is used for protein-ligand complex structure prediction. The fluctuation in it is measured via a protein backbone template and molecular graph inputs. Recently, Song et al. (2023) proposed DNMG, which is an amalgamation of deep GAN and transfer learning and provides a better binding ability to the target proteins.

14.5 Generative adversarial networks in protein and ligand structure generation: a case study

Case Study #1:GANs for generating protein structures (Anand & Huang, 2018).

There are various traditional methods for predicting the 3D structure of a protein, such as homology modeling and energy-based methods, but they do not produce accurate results. Anand and Huang (2018) use generative modeling for protein structures, which propose the use of GMs such as GANs and VAEs as promising alternative approaches. Anand and Huang (2018) then present several case studies that demonstrate the potential of GMs for protein structure prediction. They show how GANs and VAEs can be used to generate realistic protein structures with novel folds, and how these models can be used to improve the accuracy of existing protein structure prediction methods. They introduced a convex formulation to recover corruption-robust 3D structures from generated pairwise distance maps using the alternating direction method of multipliers (ADMM). The authors utilized a deep convolutional GAN (DCGAN) as their GM. The PDB, an online repository of experimentally determined structures, was considered. The 3D structures were encoded to pairwise 2D distances between α-carbons on the protein backbone. Once the pairwise distance maps are generated, recovering or folding the corresponding 3D structure is required. The best way to do this is by using the ADMM algorithm (Anand & Huang, 2018). The protein structure data representation, working pipeline, and used DCGAN model architecture are shown in Fig. 14.2.

Case Study #2:ProteinGAN: A GM that learns natural protein to generate protein sequences under biological constraints (Repecka et al., 2021).

Proteins are essential molecules that carry out many functions in living organisms. However, there are still many unknown protein sequences that could potentially have important functions. The authors in this work proposed ProteinGAN to generate novel protein sequences that are likely to have functional properties, this is specially designed for learning patterns of amino acids in biological sequences. It is difficult to determine the merit of the artificially generated sequences or structures, that is, whether the created sequence reflects a natural protein contrary to the synthesis of images or music-related data. So, we checked to see if ProteinGAN (illustrated in Fig. 14.3) could replicate key sequence characteristics found in the MDH protein family.

The ProteinGAN model consists of two separate modules of specialized neural networks: a generator that generates fake data samples, and a discriminator that tries to distinguish between actual and artificially generated data. Both of the neural networks use ResNet blocks (He et al., 2016) with specialized configurations. In addition to ResNet, there are three 1D convolution layers with a filter size of 3, and leaky ReLU activations function were present in each block of the discriminator while a residual block of generator contained two transposed convolutional layers for upsampling and one convolution layer with similar filter size and activation function (Barbhuiya et al., 2022). In addition, one self-attention layer with each network calculates the loss with R1 regularization and implements spectral normalization in all layers to ensure the stable training of the network (Ahmad et al., 2023). Training these networks together with the ratio 1:1 step using Adam optimizer to optimize both of the networks. To evaluate the performance of protein GAN, generated sequences during training were validated with the BLAST dataset and various scores like standard deviation (SD) of the discriminator layer.

Training data: The dataset of 16,898 MDH sequences was obtained (Bateman, 2019), after filtering the sequence with the criterion that the sequence length of the amino acids or noncanonical acids should be less or equal to 512 sequences. In this experiment, a total of 16,706 natural sequences were used for training and the remaining 192 sequences were used for the validation set.

The proposed work performed the following approach to analyze the generated sequence:

Multiple sequence alignment (MSA): They created MSA by combining equal numbers of synthetic and natural datasets. The Shannon entropy used to measure the alignment of natural and generated sequences as

$$SE = -\sum_{i=1}^{20} p(x_i)\log_{20} p(x_i) \tag{14.2}$$

Where $p(x_i)$ represents amino acid sequence frequency.

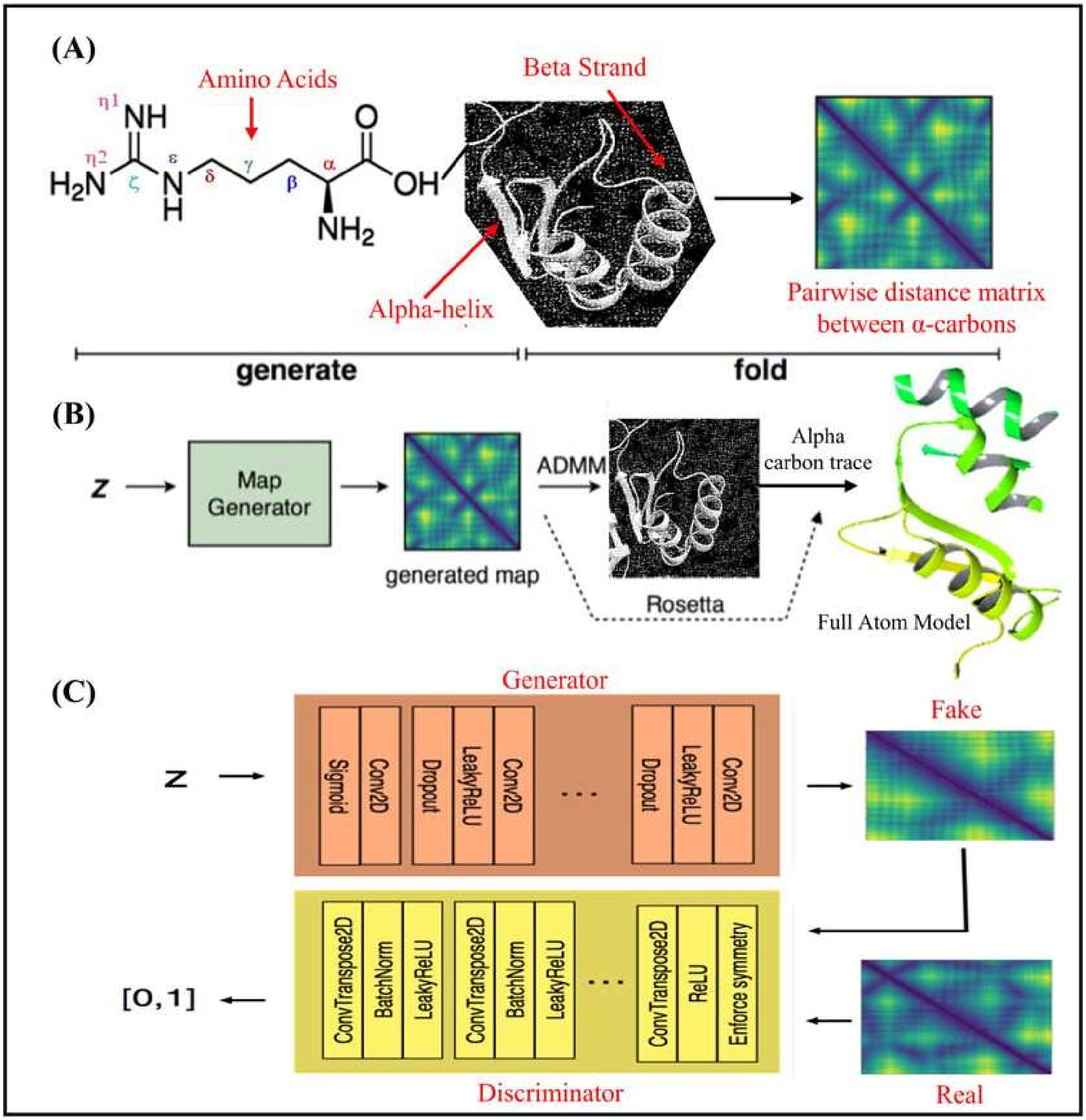

FIGURE 14.2

(A) Data representation. Proteins consist of chains of amino acids and have their secondary structure conformation such as alpha helices, beta sheets, turns, and coils. Protein structures were represented to pairwise distance matrices. (B) Working Pipeline. A pairwise distance matrix is generated by a GAN, which is "folded" into a 3D structure by the ADMM algorithm to get α-carbon coordinate positions. Structures were folded directly using pairwise distances using Rosetta. (C) Model. The architecture of the DCGAN that generates pairwise distance maps. Here, the generator takes in a random vector z ← N (0, 1) and gives a fake distance map as an output to fool the discriminator. Further, the discriminator predicts whether inputs are real or fake.

Adopted from Anand and Huang (2018).

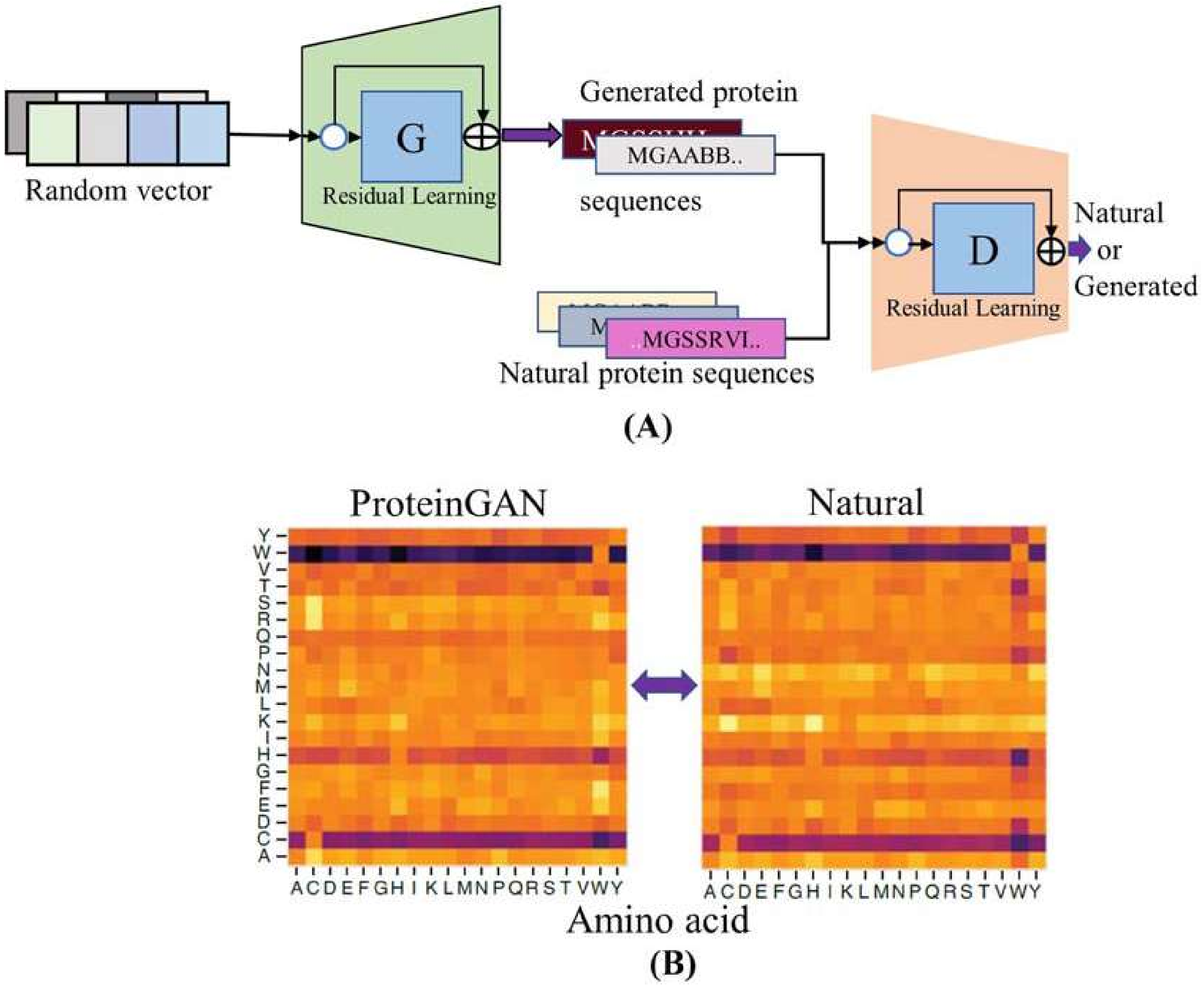

FIGURE 14.3

(A) Illustrate the training scheme of the proteinGAN model, (B) showcase the visual comparison between natural MDH sequences and amino acid distribution that are accurately captured by ProteinGAN. Shannon entropies are used to express sequence variability for synthesized and training data derived from MSA. Here, high entropy suggests substantial amino acid variety at a particular site, and low entropy indicates high similarity and consequently functionally relevant locations. *GAN*, Generative adversarial network; *MDH*, malate dehydrogenase; *MSA*, multiple sequence alignment.

Correlation matrices: For each potential pair of amino acids in a sequence, an amino acid pair correlation matrix was constructed and normalized over the entire dataset. The correlation value is represented as

$$Z_m(a,b) = \frac{p_m(a,b) - p_m(a, Rand(b))}{\sigma_{p_m(a,Rand(b))}} \tag{14.3}$$

The average and SD of the same amino acid pair's arbitrarily generated sequence correlation score is given as $\sigma_{p_m(a,Rand(b))}$ and $p_m(a,Rand(b))$, respectively. The correlation function for scoring was calculated as minimum range capability:

$$p_m(a,b)_{=}\frac{1}{n}\sum_{i=1}^{n}\text{min}_{j=1,.....,m}\left[\left|x_i-y_i\right|\right] \tag{14.4}$$

Here, the nearest occurrence of amino acid (b) at the place of y_i is found for each position x_i of amino acid (a), and the average distance between the pairings is determined.

Instead of the above approach for analyzing generated sequences, here, multiple strategies are adopted to validate the synthetically generated sequences. Sequence clustering uses MMseq2 to normalize the dataset and generate validation set data. Domain search for all generated sequences to classify each representative cluster and domain diversity controls like checkpoint introduced to analyze generated sequences. Throughout ProteinGAN, equivalent controls were created by randomly selecting 64 sequences at each fixed training step.

Overall, the result in this paper demonstrates that the generated MDH sequences in different clusters 45%–98% of amino acids are in functional positions when experimentally tested and validated in in vitro conditions. In a comparison of the discriminator, the decision with the self-attention layer improved by 66% to distinguish the position of amino acids in natural and synthetic sequences.

Case Study #3: GAN-based protein secondary structure prediction (PSSP).

Proteins are significantly important in human life activities. The functional mechanism of a protein relies on its 3D structure. In a living cell, it is done via protein sequence and folding activities. However, the 3D structure of a protein is generally received using X-ray, magnetic resonance, etc., which are expensive, slower, and available by PPDB. Considering this, it is important for researchers to consider 3D structures of protein prediction using faster sequences and at less cost. The protein secondary structure maintains the gap between 3D structures and sequences, and that is decided by the effect of hydrogen bonds available in the polypeptide chain. In this direction, existing relevant studies emphasize that 3D structures can be learned using their secondary structures and it improves the overall accuracy of 3D-structure prediction.

Levitt and Chothia (1976) introduced the PSSP which consists of three stages. The methods used in the first stage depend on the statistical probability of the individual residue. The second stage considers the neighboring residue information of the protein. It is done using a sliding window. Finally, the third stage uses MSA profiles for PSSP to increase prediction accuracy. Earlier, the secondary structures of a protein referred to helix, strand, and coil as three states. Later, it is extended to eight states to explain proteins in more detail with respect to the local structure information. Considering this, earlier methods give a better prediction on three states but perform poorly for eight states due to the increased complexity at the local level. In this line, several neural-network methods are taken into consideration for eight-state prediction like RaptorX (Wang et al., 2016) and deep convolutional and RNN (DCRNN) (Li & Yu, 2016).

Recently, deep learning has gained popularity over traditional methods due to its immense classification accuracy. Also, in PSSP, the eight states have received impressive prediction accuracy using deep neural networks. Moreover, GANs, a deep learning-based GM have achieved better performance. It is quite effective to extract features and perform signal reconstruction using GAN. It is mainly used in image generation and classification problems. In this line, Jin et al. (2022) introduce the conditional GAN-based PSSP (CGAN-PSSP) model, a new novel PSSP model. It predicts protein secondary structures using both three and eight states. This study describes generative adversarial learning achieved using GAN as quite effective for protein structure prediction in PSSP. In CGAN-PSSP, the role of the generator is to predict the secondary structure of proteins. It is achieved by giving input in terms of the position-specific scoring matrix and protein sequences. The generator learns the complex features of protein sequences. On the other hand, a discriminator conflicts with the generator. Besides that, a new multiscale convolution (MSC) is introduced, which has a modified improved channel attention (ICA) module. It is used to generate the secondary structure. The role of MSC is to extract the features of protein sequences. The proposed ICA module is added to the MSC classification modules for the proposed module to automatically recognize several functional channels.

The training of the proposed model is done on Nvidia's Titan RTX GPU, implemented in Keras (a popular neural network library). Mish and Softmax activation functions, MSRA for weight initialization, and Adam as optimizer are used in the proposed model. In this study, four publicly available datasets—CB513 (Cuff & Barton, 1999), CullPDB (Wang & Dunbrack, 2003), CASP10 (Kryshtafovych et al., 2014), and CASP11 (Moult et al., 2014) are considered. The CullPDB dataset is further split into training, validation, and testing in terms of sequences, while rest three datasets are used for testing the proposed model. Jiang et al. (2017) proposed the Q-score, which measures the correctly identified amino acid residues in terms of percentage. It is taken as an evaluation metric for empirical evaluation of the proposed model in terms of accuracy values of residues in three and eight states. The experimental results show that the proposed methods perform significantly better as compared to the traditional models. Moreover, the proposed MSC and ICA modules also show significant performance. Furthermore, the experimental results reveal that the GAN-based method remarkably minimizes the dependency of the training dataset in the proposed model as compared to other training dataset-dependent methods that are generally hard to collect due to their limitations.

In this study, the authors consider that the structure of GAN is very effective for the performance of PSSP tasks. However, they also mention that there is further scope for improvement in GAN-based PSSP. The authors highlight that the ability to learn features and pattern classification of adversarial learning is quite effective and fits the problem. They also emphasize that the consideration of GAN in image generation tasks is an important investigation problem on PSSP.

14.6 Conclusion

The applications of GAN and other GMs for protein and ligand structure prediction have shown great promise in the field of computational biology and bioinformatics. These models have the potential to generate novel protein and ligand structures, improve the accuracy of existing structure prediction methods, and aid in drug discovery. While there are still challenges that need to be overcome, such as the need for larger and more diverse datasets, advances in generative modeling techniques and computing power continue to push the boundaries of what is possible in protein and ligand structure prediction. Moreover, the applications of GAN and other GMs have expanded beyond the prediction of individual structures to include the generation of entire protein−protein and protein−ligand complexes, which has important implications for drug discovery and design.

Hence, GMs for protein and ligand structure prediction are a rapidly evolving research area with significant potential for advancing our understanding of protein and ligand structures and their role in biological processes.

There are several exciting future research directions in the field, a few possibilities are as follows:

Development of new GMs: While GANs and VAEs have shown promise in protein and ligand structure prediction, there is still room for the development of new and improved GMs that can better capture the complexity and diversity of protein and ligand structures.

Integration with experimental data: The incorporation of experimental data such as NMR spectroscopy, X-ray crystallography, and cryo-EM can provide valuable constraints for GMs and improve their accuracy.

Applications to larger systems: The current applications of GAN and other GMs for protein and ligand structure prediction have mostly focused on individual structures or small complexes. Future research could explore the generation of larger systems such as protein—protein complexes or entire cellular pathways.

Impact on drug discovery: GMs have the potential to revolutionize drug discovery by generating novel ligand structures that can be used as starting points for drug design. Future research could focus on the use of GMs to generate ligands with specific properties or to optimize existing ligands.

Exploration of GMs in other areas of translational bioinformatics: The applications of GMs are not limited to protein and ligand structure prediction. Future research could explore the use of these models in other areas of translational bioinformatics such as gene expression prediction, protein folding, drug toxicity prediction, medical image and signal processing, synthetic DNA sequences generation, and patient-specific models generated to predict treatment outcomes in personalized treatment settings, and so on.

References

Adhikari, U. K., Tayebi, M., & Mizanur Rahman, M. (2018). Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging oropouche virus. *Journal of Immunology Research*, *2018*, 1—22. Available from https://doi.org/10.1155/2018/6718083.

Ahmad, N., Sumedha, A., & Deepak, S. (2023). Predicting Risky Environment for Child Inside House using Deep Learning. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE.

Alberts, B., Johnson, A., & Lewis, J. (2002). *Analyzing protein structure and function*. Garland Science.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The shape and structure of proteins. *Molecular Biology of the Cell*.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315—1322. Available from https://doi.org/10.1038/s41592-019-0598-1, http://www.nature.com/nmeth/.

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1—8. Available from https://doi.org/10.1016/j.cbpa.2021.04.005, http://www.elsevier.com/locate/cbi.

Anand, N., & Huang, P. S. (2018). Neural information processing systems foundation United States Generative modeling for protein structures. *Advances in Neural Information Processing Systems*, 7494–7505, 10495258. Available from https://papers.nips.cc/2018.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., . . . Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science (New York, N.Y.)*, *373*(6557), 871–876. Available from https://doi.org/10.1126/science.abj8754, https://science.sciencemag.org/content/373/6557/871.full.

Barbhuiya, R. K., Ahmad, N., & Akram, W. (2022). *Application of convolutional neural networks in cancer diagnosis* (pp. 95–109). Springer Science and Business Media LLC. Available from https://doi.org/10.1007/978-981-16-9221-5_5.

Barbhuiya, R. K., & Ahmad, N. (2021). IoT applications in translational bioinformatics. In *Translational bioinformatics in healthcare and medicine*. Academic Press.

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(1), D506–D515. Available from https://doi.org/10.1093/nar/gky1049, https://academic.oup.com/nar/issue.

Bhagavan, N. V. (2002). *Three-dimensional structure of proteins* (pp. 51–65). Elsevier BV. Available from https://doi.org/10.1016/b978-012095440-7/50006-8.

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). *Protein Data Bank (PDB): The single global macromolecular structure archive Methods in Molecular Biology* (pp. 627–641). Humana Press Inc.. Available from http://www.springer.com/series/7651, https://doi.org/10.1007/978-1-4939-7000-1_26.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Neural information processing systems foundation United States InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, *10495258*, 2180–2188.

Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, *34*(4), 508–519. Available from https://doi.org/10.1002/(SICI)1097-0134(19990301)34:4 < 508::AID-PROT10 > 3.0.CO;2–4.

Dhanjal, J. K., Malik, V., Radhakrishnan, N., Sigar, M., Kumari, A., & Sundar, D. (2018). *Computational protein engineering approaches for effective design of new molecules Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (pp. 631–643). Elsevier. Available from https://doi.org/10.1016/B978-0-12-809633-8.20150-7.

Egli, M., & Zhang, S. (2022). How the α-helix got its name. *Nature Reviews. Molecular Cell Biology*, *23*(3), 165. Available from https://doi.org/10.1038/s41580-021-00449-4.

Eguchi, R. R., Choe, C. A., & Huang, P. S. (2020). *IG-VAE: Generative modeling of protein structure by direct 3D coordinate generation*. bioRxiv. Available from: https://www.biorxiv.org. https://doi.org/10.1101/2020.08.07.242347.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184–192. Available from https://doi.org/10.1038/s41592-019-0666-6, http://www.nature.com/nmeth/.

Godbey, W. T. (2022). *Proteins* (pp. 47–72). Elsevier BV. Available from https://doi.org/10.1016/b978-0-12-817726-6.00003-4.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Neural information processing systems foundation Canada Generative adversarial nets. *Advances in Neural Information Processing Systems*, *3*, 2672–2680, 10495258 January.

Greener, J. G., Moffat, L., & Jones, D. T. (2018). Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, *8*(1). Available from https://doi.org/10.1038/s41598-018-34533-1, http://www.nature.com/srep/index.html.

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., & Schneider, G. (2018). Generative recurrent networks for de novo drug design. *Molecular Informatics*, *37*(1−2), 1700111. Available from https://doi.org/10.1002/minf.201700111.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition. IEEE Computer Society United States*. https://doi.org/10.1109/CVPR.2016.90, 9781467388504 770−778.

Idicula-Thomas, S., & Balaji, P. V. (2005). Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Science*, *14*(3), 582−592. Available from https://doi.org/10.1110/ps.041009005.

Ingraham, J., Garg, V. K., Barzilay, R., & Jaakkola, T. (2019). *Deep generative models for highly structured data, DGS@ICLR 2019 Workshop international conference on learning representations, ICLR United States Generative models for graph-based protein design*. https://deep-gen-struct.github.io/index.html

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. E. (2019). Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *Journal of Molecular Biology*, *431*(11), 2197−2212. Available from https://doi.org/10.1016/j.jmb.2019.04.009, https://www.journals.elsevier.com/journal-of-molecular-biology.

Jaimes, J. A., André, N. M., Chappie, J. S., Millet, J. K., & Whittaker, G. R. (2020). Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. *Journal of Molecular Biology*, *432*(10), 3309−3325. Available from https://doi.org/10.1016/j.jmb.2020.04.009, https://www.journals.elsevier.com/journal-of-molecular-biology.

Jiang, Q., Jin, X., Lee, S. J., & Yao, S. (2017). Protein secondary structure prediction: A survey of the state of the art. *Journal of Molecular Graphics and Modelling*, *76*, 379−402. Available from https://doi.org/10.1016/j.jmgm.2017.07.015, http://www.elsevier.com/inca/publications/store/5/2/5/0/1/2/index.htt.

Jin, X., Guo, L., Jiang, Q., Wu, N., & Yao, S. (2022). Prediction of protein secondary structure based on an improved channel attention and multiscale convolution module. *Frontiers in Bioengineering and Biotechnology*, *10*. Available from https://doi.org/10.3389/fbioe.2022.901018, http://journal.frontiersin.org/journal/bioengineering-and-biotechnology#archive.

Kim, S., Kim, J. H., Lee, J. S., & Park, C. B. (2015). Beta-sheet-forming, self-assembled peptide nanomaterials towards optical, energy, and healthcare applications. *Small (Weinheim an der Bergstrasse, Germany)*, *11* (30), 3623−3640. Available from https://doi.org/10.1002/smll.201500169, http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1613-6829.

Kryshtafovych, A., Barbato, A., Fidelis, K., Monastyrskyy, B., Schwede, T., & Tramontano, A. (2014). Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function and Bioinformatics*, *82*(2), 112−126. Available from https://doi.org/10.1002/prot.24347.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., & Gelly, S. (2019). *A large-scale study on regularization and normalization in GANs. In Proceedings of the 36th international conference on machine learning, ICML 2019 International Machine Learning Society (IMLS) undefined*. 2019−9781510886988 6350−6367.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., & Bradley, P., Chapter 19—Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Computer methods* (pp. 545−574).

Levitt, M., & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, *261*(5561), 552−558. Available from https://doi.org/10.1038/261552a0.

Li, Z., & Yu, Y. (2016). Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2560−2567.

Meier, A., Söding, J., & Ben-Tal, N. (2015). Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. *PLoS Computational Biology*, *11*(10), e1004343. Available from https://doi.org/10.1371/journal.pcbi.1004343.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) — Round x. *Proteins: Structure, Function and Bioinformatics*, *82*(2), 1—6. Available from https://doi.org/10.1002/prot.24452.

Noé, F., Olsson, S., Köhler, J., & Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science (New York, N.Y.)*, *365*(6457). Available from https://doi.org/10.1126/science.aaw1147, https://science.sciencemag.org/content/365/6457/eaaw1147/tab-pdf.

O'Donoghue, S. I., Goodsell, D. S., Frangakis, A. S., Jossinet, F., Laskowski, R. A., Nilges, M., Saibil, H. R., Schafferhans, A., Wade, R. C., Westhof, E., & Olson, A. J. (2010). Visualization of macromolecular structures. *Nature Methods*, *7*(3), 1427. Available from https://doi.org/10.1038/nmeth.1427.

Pandurangan, A. P., & Blundell, T. L. (2020). Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Science*, *29*(1), 247—257. Available from https://doi.org/10.1002/pro.3774, http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1469-896X.

Petukh, M., Li, M., Alexov, E., & MacKerell, A. (2015). Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. *PLoS Computational Biology*, *11*(7) e1004276. Available from https://doi.org/10.1371/journal.pcbi.1004276.

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., Bozdaganyan, M., Aliper, A., Zhavoronkov, A., & Kadurin, A. (2018). Entangled conditional adversarial autoencoder for de novo drug discovery. *Molecular Pharmaceutics*, *15*(10), 4398—4405. Available from https://doi.org/10.1021/acs.molpharmaceut.8b00839, http://pubs.acs.org/journal/mpohbp.

Qiao, Z., Nie, W., Vahdat, A., Miller, T., & Anandkumar, A. (2022). *Dynamic-backbone protein-ligand structure prediction with multiscale generative diffusion models*. arXiv. Available from: https://arxiv.org, https://doi.org/10.48550/arXiv.2209.15171.

Raza, K., & Singh, N. K. (2021). A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, *17*(9), 1059—1077. Available from https://doi.org/10.2174/1573405617666210127154257.

Raza, K. (2017). *Protein features identification for machine learning-based prediction of protein-protein interactions*. *Communications in Computer and Information Science* (750, pp. 305—317). Springer Verlag. Available from https://doi.org/10.1007/978-981-10-6544-6_28 18650929, http://www.springer.com/series/7899.

RCSB-PDB 2023 3 20 Molecular Graphics Software https://www.rcsb.org/docs/additional-resources/molecular-graphics-software.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M. K. M., & Zelezniak, A. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, *3*(4), 324—333. Available from https://doi.org/10.1038/s42256-021-00310-5, https://www.nature.com/natmachintell/.

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, *9*(1). Available from https://doi.org/10.1038/s41598-019-52737-x, http://www.nature.com/srep/index.html.

Schmidt, T., Bergner, A., & Schwede, T. (2014). Modelling three-dimensional protein structures for applications in drug design. *Drug Discovery Today*, *19*(7), 890—897. Available from https://doi.org/10.1016/j.drudis.2013.10.027, http://www.elsevier.com/locate/drugdiscov.

Shahriar, S. (2022). GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays*, *73*, 102237. Available from https://doi.org/10.1016/j.displa.2022.102237.

Singh, N. K., & Raza, K. (2021). *Medical image generation using generative adversarial networks: A review studies in computational intelligence* (pp. 77–96). Springer Science and Business Media Deutschland GmbH. Available from http://www.springer.com/series/7092, https://doi.org/10.1007/978-981-15-9735-0_5.

Skalic, M., Sabbadin, D., Sattarov, B., Sciabola, S., & De Fabritiis, G. (2019). From target to drug: Generative modeling for the multimodal structure-based ligand design. *Molecular Pharmaceutics*, *16*(10), 4282–4291. Available from https://doi.org/10.1021/acs.molpharmaceut.9b00634, http://pubs.acs.org/journal/mpohbp.

Song, T., Ren, Y., Wang, S., Han, P., Wang, L., Li, X., & Rodriguez-Patón, A. (2023). DNMG: Deep molecular generative model by fusion of 3D information for de novo drug design. *Methods (San Diego, Calif.)*, *211*, 10–22. Available from https://doi.org/10.1016/j.ymeth.2023.02.001, http://www.elsevier.com/inca/publications/store/6/2/2/9/1/4/index.htt.

Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., & Dozmorov, M. G. (2018). HiCcompare: An R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, *19*(1). Available from https://doi.org/10.1186/s12859-018-2288-x, http://www.biomedcentral.com/bmcbioinformatics/.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., & Kim, P. M. (2020). Fast and flexible protein design using deep graph neural networks. *Cell Systems*, *11*(4), 402–411. Available from https://doi.org/10.1016/j.cels.2020.08.016, e4, Available from, http://www.journals.elsevier.com/cell-systems/.

Tang, H., Liu, H., Xu, D., Torr, P. H., & Sebe, N. (2021). Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A.N. (2017). Attention is all you need. In *Advances in neural information processing systems*.

Wang, G., & Dunbrack, R. L. (2003). PISCES: A protein sequence culling server. *Bioinformatics (Oxford, England)*, *19*(12), 1589–1591. Available from https://doi.org/10.1093/bioinformatics/btg224, http://bioinformatics.oxfordjournals.org/.

Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: A web server for protein structure property prediction. *Nucleic Acids Research*, *44*(W1), W430–W435. Available from https://doi.org/10.1093/nar/gkw306.

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *54*(1). Available from https://doi.org/10.1002/cpbi.3.

Yan, K., Wen, J., Liu, J. X., Xu, Y., & Liu, B. (2021). Protein fold recognition by combining support vector machines and pairwise sequence similarity scores. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(5), 2008–2016. Available from https://doi.org/10.1109/TCBB.2020.2966450, http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber = 8857.