ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

PLM-GAN: A Large-Scale Protein Loop Modeling Using pix2pix GAN

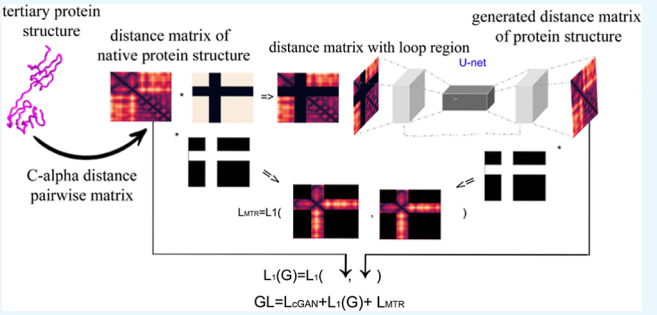
Mena Nagy A. Khalaf,* Taysir Hassan A Soliman, and Sara Salah Mohamed

Cite This: *ACS Omega* 2024, 9, 437−446

Read Online

ACCESS | Metrics & More | Article Recommendations

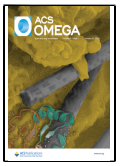
ABSTRACT: Revealing the tertiary structure of proteins holds huge significance as it unveils their vital properties and functions. These intricate three-dimensional configurations comprise diverse interactions including ionic, hydrophobic, and disulfide forces. In certain instances, these structures exhibit missing regions, necessitating the reconstruction of specific segments, thereby resulting in challenges in protein design, which encompasses loop modeling, circular permutation, and interface prediction. To address this problem, we present two pioneering models: pix2pix generative adversarial network (GAN) and PLM-GAN. The pix2pix GAN model is adept at generating and inpainting distance matrices of protein structures, whereas the PLM-GAN model incorporates residual blocks into the U-Net network of the GAN, building upon the foundation of the pix2pix GAN model. To bolster the models' performance, we introduce a novel loss function named the "missing to real regions loss" (L_{MTR}) within the GAN framework. Additionally, we introduce a distinctive approach of pairing two different distance matrices: one representing the native protein structure and the other representing the same structure with a missing region that undergoes changes in each successive epoch. Moreover, we extend the reconstruction of missing regions, encompassing up to 30 amino acids and increase the protein length by 128 amino acids. The evaluation of our pix2pix GAN and PLM-GAN models on a random selection of natural proteins (4ZCB, 3FJB, and 2REZ) demonstrated promising experimental results. Our models constitute significant contributions to addressing intricate challenges in protein structure design. These contributions hold immense potential to propel advancements in protein−protein interactions, drug design, and further innovations in protein engineering. Data, code, trained models, examples, and measurements are available on https://github.com/mena01/PLM-GAN-A-Large-Scale-Protein-Loop-Modeling-Using-pix2pix-GAN_.



INTRODUCTION

In the biological field, researchers are concerned with understanding biological molecules and their functions. Many molecules' functions, especially those of proteins, depend on their tertiary structures. Therefore, obtaining and understanding the tertiary structure of a protein have attracted huge attention in many important fields, such as drug design,[1] protein classification,[2] gene function annotation,[3] and immunotherapy.[4] Many methods are used to extract the tertiary structures of proteins in the laboratory, such as X-ray crystallography,[5] nuclear magnetic resonance (NMR),[6] and cryogenic electron microscopy (cryo-EM),[7] which can provide high accuracy but consume a lot of time and many resources. To overcome these restrictions, many methods have been developed for protein structure prediction. In silico, there are two main methods: ab initio and template-based.[8] The first is based on calculations of energy functions to predict the protein structure from the amino acid sequence, but this method becomes challenging when the number of amino acids (aa) is more than 150 aa.[9] The second is based on trying to find a template structure based on alignments with a similar sequence that can be found in various databases, such as Protein Data Bank (PDB)[10] and UniProt.[11] In many instances, some small regions of a protein structure are missing or need to be remodeled, and this problem is known as loop modeling (other names include interface prediction, circular permutation, and inpainting protein structure). Inpainting these missing regions is significant in determining the native protein structure, function, and its dynamics. In,[12] researchers utilized a self-supervised learning approach on 700 million unlabeled molecules, emphasizing the importance of extracting predictive representations. They employed diverse model combinations and an automated selection protocol, achieving superior performance, particularly leveraging pretrained models. In,[13] scientists trained a deep contextual language model on 250

Received: August 9, 2023
Revised: November 1, 2023
Accepted: November 22, 2023
Published: December 15, 2023



million protein sequences using unsupervised learning techniques. Their emphasis lies in understanding biological properties encoded in protein sequences, leading to representations capturing intricate details. This work highlights the potential of large-scale, unsupervised learning, particularly in the context of pretrained models, in deciphering complex biological information.

Recently in,[14,15] the researchers have highlighted the importance of protein inpainting in designing and predicting proteins. In,[16,17] the researchers introduced the first models that depended on the generative adversarial network (GAN) for inpainting the missing regions of the protein structure. The models provided promising results but used small distance matrices of the protein structures with the maximum length of the protein being only 50 aa. In addition, the missing region was constant and was restricted only to the middle of the distance matrix. The maximum length of the missing region was 12 aa. Finally, the models can generate only the missing region. In ref 18, the researchers used an autoencoder and a GAN to regenerate the distance matrix of a protein structure, which had a length of 64 aa with a missing region from 5 to 20 aa. Also, in ref 19, Rosette's research team developed the RosettaRemodel model to inpaint the missing region based on the energy score of the protein. However, RosettaRemodel consumed a lot of time to generate just one structure. For example, RosettaRemodel takes a mean of 20 min to generate a structure of 64 aa protein on a node with 16 CPU cores. Rosette's research team developed the RFDesign,[20] a state-of-art model to solve a loop modeling problem. It achieved impressive results, but it must run on a GPU RTX-2080 and takes several cycles to obtain the result. RFDesign was developed based on the RoseTTAFold model.

Latterly, the field of protein structure prediction has seen remarkable advancements, with the introduction of ground-breaking methods such as AlphaFold[21] and AlphaFold2.[22] AlphaFold, developed by DeepMind, has garnered significant attention for its unprecedented accuracy in predicting protein structures. This transformative development has reshaped the landscape of structural biology and holds great promise for various applications. In the context of this evolving field, our research aims to complement and contribute to the ongoing efforts in protein structure prediction. While AlphaFold represents a pioneering approach, our work introduces a novel methodology that combines Pix2Pix GANs and self-supervised learning techniques to address the challenge of modeling protein structures, particularly in regions with missing data. By doing so, we seek to expand the repertoire of tools available to researchers in this domain and provide an alternative approach that may prove valuable in various biological and biomedical applications.

In this paper, we introduce two models pix2pix GAN and PLM-GAN to generate the missing regions of the protein structure and preserve its original structural properties (backbone, local, and distal characteristics), which will give significant advancement in various applications, i.e., molecular inpainting, protein−protein interaction,[23] de novo protein design,[24] drug design,[25] and molecular dynamics.[26] PLM-GAN is based on the pix2pix GAN network that we trained on PDB data,[10] which contains an assortment of protein structures with various lengths of amino acids. We embedded residual blocks in the generator network of the pix2pix GAN architecture to enhance its performance.

To summarize, the central contributions of this work are as follows:

1. Applying pix2pix GAN to generate and inpaint distance matrix of protein structure.
2. Developing the PLM-GAN model to inpaint the missing region in the protein's tertiary structure. It was built by integrating the residual blocks into the U-Net network of the pix2pix GAN network.
3. Introducing the new loss function missing to real regions (L_{MTR}) in the pix2pix GAN loss functions to make PLM-GAN.
4. Maximizing the length of the missing regions from 5 to 30 aa.
5. Pairing two different distance matrices (one of the native protein structure and one of the same structure but with a missing region that changes in each successive epoch).

PROPOSED METHODOLOGY

The GAN[27] is a complex and significant network frequently used in many fields, such as computer vision,[28−30] natural language processing (NLP),[31] and cybersecurity.[32] Also, it has been used to generate protein structures that mimic the native protein structures.[18] To the best of our knowledge, our models are the first models that apply the pix2pix GAN network in the loop modeling problem.

Generative Adversarial Network. In brief, the GAN[27] involves two neural networks that work together as competitors to each other: generator and discriminator. The generator G is the network that tries to generate protein structures that mimic the natural protein structure and attempts to fool the discriminator. The discriminator D is the network in charge of differentiating between fake and real proteins. The loss function of the GAN network is shown in eq 1, as follows

$$\min_G \max_D \mathrm{GAN}(G, D) = E_{x p_r(x)}[\log\ D(x)] + E_{z p_z(z)}[\log 1 - D(G(z))] \quad (1)$$

where p_r denotes the real data distribution, p_z denotes the model distribution, and z denotes the input to the generator, which is randomly selected from some simple noise distribution.

pix2pix GAN Network. The pix2pix GAN network[33] is a type of GAN that is specifically designed for image-to-image translation tasks. It uses a U-net architecture for its generator, which is a type of convolutional neural network that has to skip connections to maintain high-resolution features throughout the network. The discriminator in the pix2pix GAN is responsible for distinguishing between the generated output and ground truth images. Unlike traditional GANs, the pix2pix GAN uses a conditional GAN loss, which is a type of adversarial loss that takes into account both the generated output and the input image. The conditional GAN loss ensures that the generated output is not only visually appealing but also relevant to the input image. The pix2pix GAN has been used for a variety of image-to-image translation tasks including semantic segmentation, style transfer, and super-resolution.

In our methodology, we employ the Pix2Pix GAN framework, which utilizes a loss function consisting of two key components: adversarial loss (LcGAN) and L_1 loss ($L_1(G)$). The adversarial loss, as described in eq 2, leverages the conditional GAN loss to measure the similarity between

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

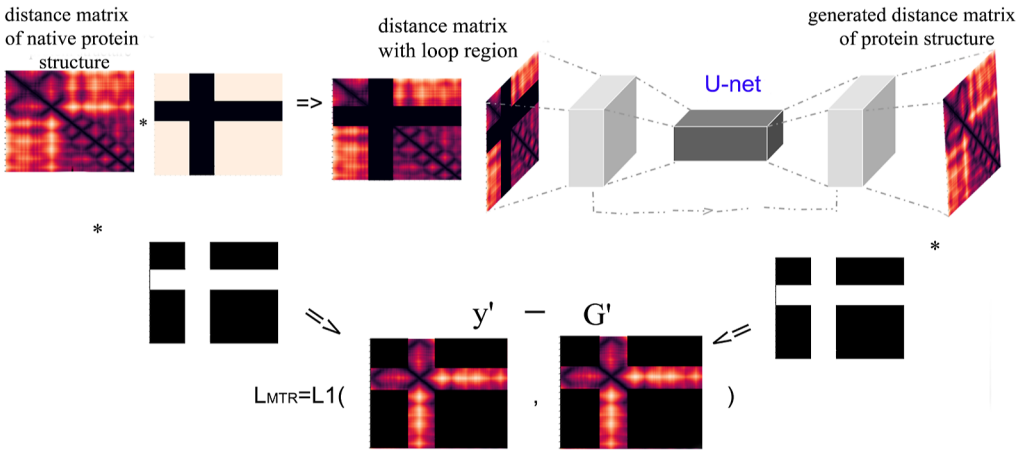


Figure 1. L_{MTR} loss function of the generator network on the PLM-GAN model. The asterisk denotes multiplication.

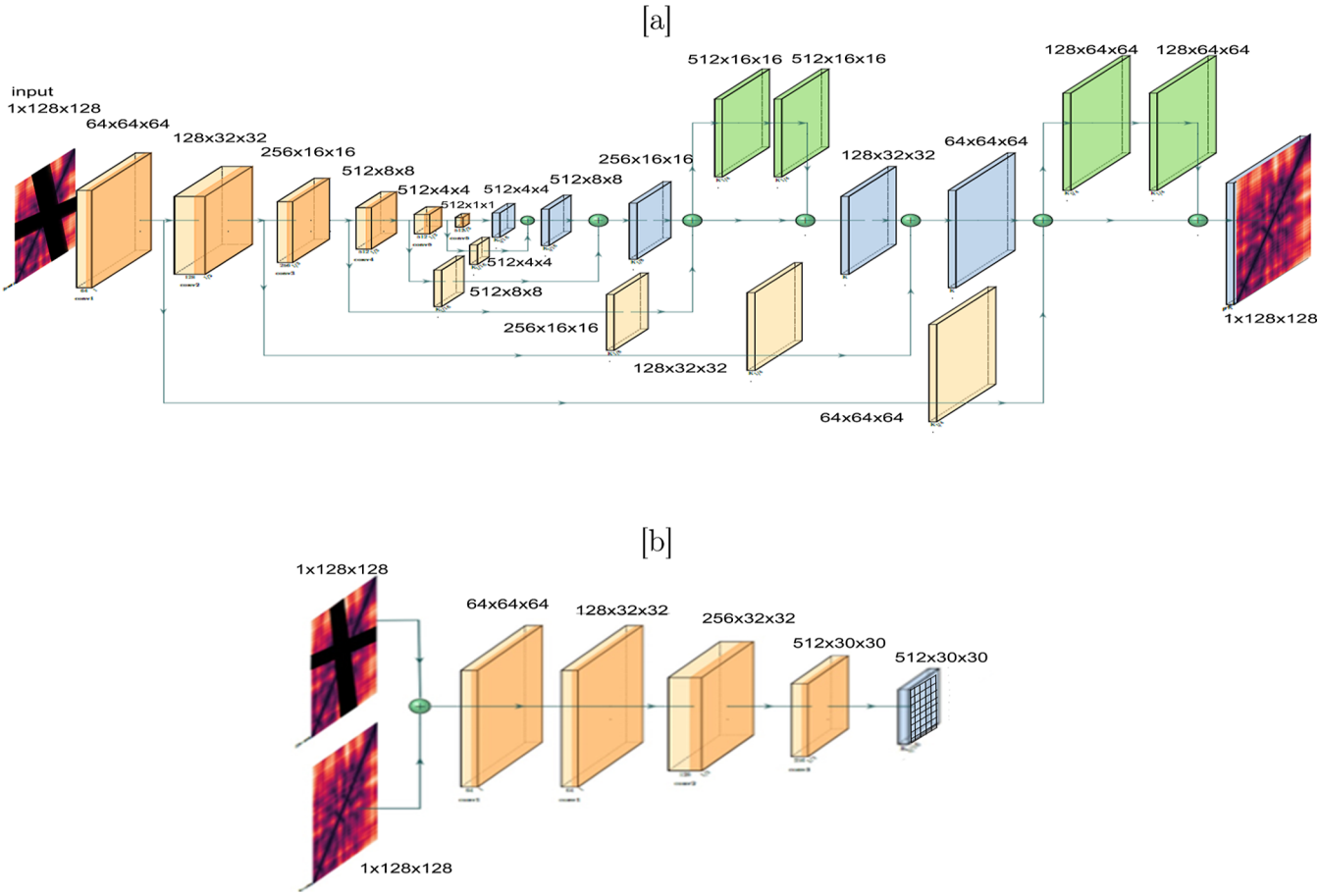


Figure 2. Proposed architecture PLM-GAN, where (a) represents the generator which consists of two residual blocks in U-net and (b) represents the discriminator.

the generated output and the ground truth image. This loss encourages the generator to produce outputs that closely resemble the ground truth images. On the other hand, the $L_1(G)$ loss, as presented in eq 3, quantifies the absolute pixel-wise difference between the generated output and the ground truth image. This component aims to ensure that the generated outputs align with the ground truth images in terms of the pixel values. The adversarial loss ensures overall structural realism, while the L_1 loss enforces precise similarity (pixel-wise) at the distance level. The final loss function, as shown in eq 4, combines these two losses.

$$L_{cGAN}(G, D) = E_{x,y}[\log\ D(x, y)] + E_{x,z}[\log 1 - D(x, G(x))] \quad (2)$$

$$L_1(G) = E_{x,y}[\|y - G(x)\|] \quad (3)$$

$$GL = L_{cGAN} + \gamma L_1(G) \quad (4)$$

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

Table 1. Layers of the Generator Network Architecture

layer	details	filter	strid	p	
input		1 × 128 × 128			
conv1		64 × 64 × 64	64 × 4 × 4	2	1
conv2		128 × 32 × 32	128 × 4 × 4	2	1
conv3		256 × 16 × 16	256 × 4 × 4	2	1
conv4		256 × 8 × 8	512 × 4 × 4	2	1
conv5		512 × 4 × 4	512 × 4 × 4	1	0
conv6		512 × 1 × 1	512 × 4 × 4	1	0
UpSample1	conv6	512 × 4 × 4	512 × 4 × 4	2	1
	conv5	512 × 4 × 4	512 × 4 × 4	2	1
UpSample2	UpSample1	512 × 4 × 4	512 × 4 × 4	2	1
	conv4	512 × 4 × 4	512 × 4 × 4	2	1
UpSample3	UpSample2	256 × 8 × 8	256 × 4 × 4	2	1
	conv3	256 × 8 × 8	256 × 4 × 4	2	1
Residual2	UpSample3	512 × 64 × 64	64 × 3 × 3	1	1
	UpSample3	512 × 64 × 64	64 × 3 × 3	1	1
UpSample4	Residual1	256 × 8 × 8	256 × 4 × 4	2	1
	conv2	256 × 8 × 8	256 × 4 × 4	2	1
UpSample5	UpSample4	256 × 8 × 8	256 × 4 × 4	2	1
	conv1	256 × 8 × 8	256 × 4 × 4	2	1
Residual1	UpSample5	128 × 64 × 64	128 × 3 × 3	1	1
	UpSample5	128 × 64 × 64	128 × 3 × 3	1	1
convT5		1 × 128 × 128	1 × 4 × 4	2	1

where x denotes the distance matrix of the protein structure involving the missing region and y denotes the native distance matrix of the protein structure.

PLM-GAN Model. As stated before, PLM-GAN is based on pix2pix GAN, which was harnessed to generate and paint the missing region of the distance matrix of the protein structure. The elucidation of the methodology of the loss function is delineated in Figure 1. Figure 2 presents the architectural configurations of both the generator and discriminator components within the PLM-GAN model.

To illustrate, the generator network uses a U-net architecture, which utilizes a skip-connection between symmetrical layers. We embedded it with residual blocks to improve the model performance and the convergence of deep neural networks. The generator maps between the distance matrices of protein structures with a missing region and the distance matrices of the native protein structure, allowing it to predict and fill in the missing region within the protein structure. The discriminator network works as a PatchGAN network,[33] a specialized type of discriminator composed of a deep convolution neural network (CNN). PatchGAN has been used to classify patches of an input distance matrix of the protein structure, rather than the entire distance matrix of the protein structure, as real or fake. The inputs to the PatchGAN discriminator network are pairs of two distance matrices of protein structure: the first is the source (the distance matrix of the protein structure with missing region) and the second is the target (the distance matrix of the native protein structure or the generated distance matrix of the protein structure). In the end, the PatchGAN maps the distance matrices of protein structures to a small dimensional matrix of 30 × 30. Then, it classifies whether the 30 × 30 patches in the input distance matrix of the protein structure are real or fake.

Missing to Real Loss (L_{MTR}). We introduced a new loss function called missing to real (L_{MTR}) in the pix2pix GAN loss functions. Instead of comparing the whole native distance matrix and the generated distance matrix, the L_{MTR} function focuses on the missing region that needs to be inpainted. For the PLM-GAN model, the overall generator loss function is the sum of L_{cGAN}, $L_1(G)$, and L_{MTR}. It can be calculated by eq 5. We have incorporated L_{MTR} to focus on the missing regions of the protein structure's distance matrix, as shown in eq 6.

$$GL = L_{cGAN} + \gamma L_1(G) + \partial L_{MTR} \quad (5)$$

where γ and ∂ are the hyper-parameters to determine the impact of L_1 and L_{MTR}.

$$L_{MTR}(G) = L_1[||y' - G'||] \quad (6)$$

where y' is the real region that the model tries to inpaint. G' is the generated inpainting region. The overall methodology of the loss function is explained in Figure 1.

Model Architecture. The PLM-GAN model architecture involves two parts: the generator network and the discriminator network, as shown in Figure 2.

Generator Network Architecture. The generator network consists of a U-net network boosted by two residual blocks to improve and increase model learning. Table 1 illustrates in detail the generator network parameters for 128 aa.

Discriminator Network Architecture. The patchGAN discriminator architecture is shown in Figure 2b. The inputs to the patchGAN discriminator are two distance matrices of the protein structure. It maps the input to a small dimension matrix of 30 × 30 patches. Additionally, it differentiates the real patches from the generated patches. The patchGAN discriminator involves five convolution layers. Table 2 illustrates the discriminator network parameters for 128 aa.

■ ASSESSMENT OF THE PROPOSED MODEL

To assess the accuracy of the Pix2Pix-GAN and the PLM-GAN models in generating the missing region of the protein structure, in addition to the whole protein structure, we used a testing data set of 6200 proteins and compared our generated protein structures with the native protein structures and with those produced from the state-of-art algorithm RFDesign.[20]

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

Table 2. Layers of the Discriminator Network Architecture

layer	details	filter	strid	padding
input	1 × 128 × 128			
conv1	64 × 64 × 64	64 × 4 × 4	2	1
conv2	128 × 32 × 32	128 × 4 × 4	2	1
conv3	256 × 32 × 32	256 × 3 × 3	1	1
conv4	512 × 30 × 30	512 × 5 × 5	1	1
conv5	1 × 30 × 30	1 × 3 × 3	1	1

Our test data set, consisting of 6200 protein structures, is sourced separately from the training data and obtained from the PDB. This separation serves a crucial purpose in our study as it allows us to evaluate the generalization and predictive capabilities of our models on previously unseen protein structures. By utilizing a distinct test set, we ensure that our models' performance is assessed on data that were not part of the training process, thus providing a robust evaluation of their effectiveness. It is also worth mentioning that we used in this comparison several proteins not involved in our training data set, as shown in Figure 6. The root-mean-square deviation (rmsd)[34] is used to measure the similarity between the generated missing region and the native region of the protein structures by comparing their distance matrices, as shown in eq 7. To build the distance matrix, we used the distance between the CA (α carbon) atoms in the main chain of the protein structure. Finally, we compared the features of the generated proteins structures (backbone, local, and distal characteristics) with those of the native protein structures. A brief explanation of the assessment methodology is shown in Figure 3.

$$\mathrm{rmsd}(G, D) = \sqrt{\frac{\sum_{i=0}^{n} (x_i - y_i)^2}{n}} \tag{7}$$

where n denotes the size of the region and x and y refer to the inpainted and native regions of the distance matrix for the protein structure, respectively.

Assessment of Generated Missing Region of Protein Structure. Our evaluation methodology is based on assessing the similarity of the missing region in the natural, inpainted, and generated protein structures, utilizing the rmsd measure between them.

Assessment of the Whole Structure of the Generated Proteins. We compare the entire generated distance matrix of the protein structure with that of the native protein structure. Additionally, we perform a comparison based on the features of the distance matrices of the protein structure, which consist of backbone, short-range (local structure), and long-range (distal structure) distances.

Assessment of the Average Peptide Bond. The average peptide bond is a summation of the diagonal distance values of the protein distance matrix, which is then divided by the number of all entries along the main diagonal.[35,36] The generated protein tertiary structure distance matrix was evaluated by comparing its feature (distance of backbone, short-range, and long-range) to the natural feature protein.

Assessment of the Backbone Structure. The backbone in the regenerated protein distance matrix is the main diagonal. The diagonal of the distance matrix is formulated by every consecutive $(i,i + 1)$ CA atom pair, where $0 < i < n - 1$. According to the natural protein, the ideal distance is 3.79 Å between two consecutive amino acids in the natural protein for 128 aa.

Assessment of the Short-Range Structure and the Distal Structure. After calculating the backbone, we will calculate both the short-range (local structure) and long-range distances (distal structure). First, we compute the short-range, where we move forward from the backbone by every consecutive $(i,i + j)$ CA pair where $1 < j < 4$; the ideal short-range distance is 7.8 Å in the natural protein. Second, the long-range distance is calculated by expanding $j > 4$; the ideal long-range distance is 21.31 Å in the natural protein for 128 aa.

TRAINING DATA SET

The data set consists of 115 K proteins collected from the PDB site,[10] which holds various protein structures of various sizes. We have relied on the distance between the CA (α carbon)

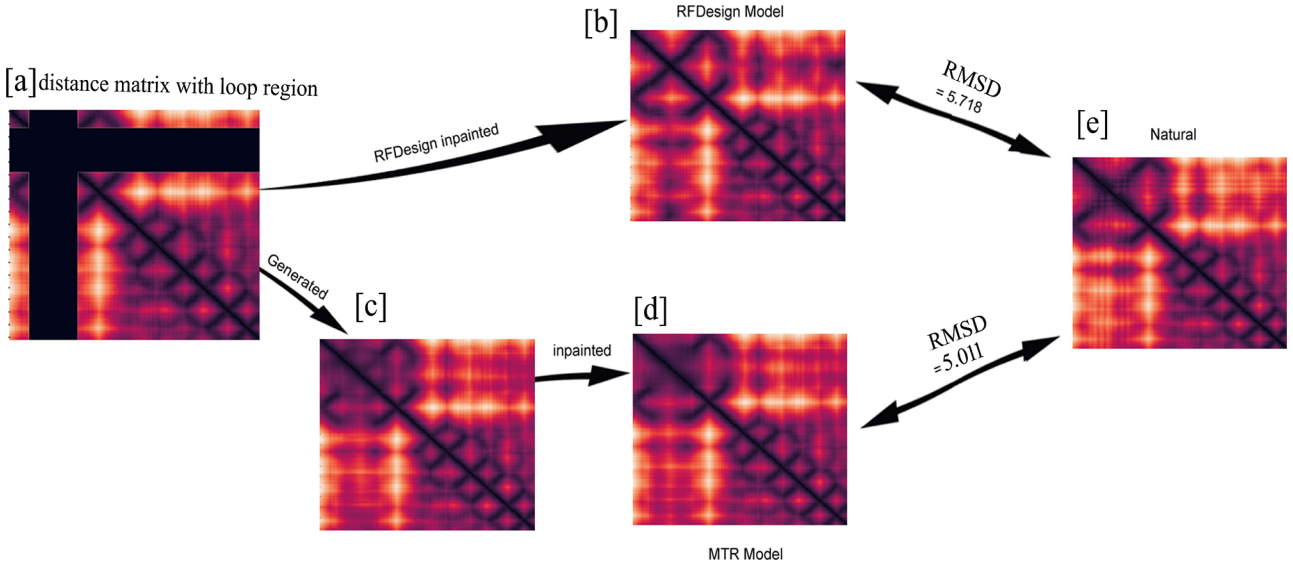


Figure 3. (a) Distance matrix of the 4ZCB protein structure that contains a missing region of length 25 aa. (b) Inpainted distance matrix of the protein structure by RFDesign model. (c) Regenerated distance matrix of the protein structure by the PLM-GAN model. (d) Distance matrix of the inpainted protein structure by the PLM-GAN model. (e) Native distance matrix of the protein structure.

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

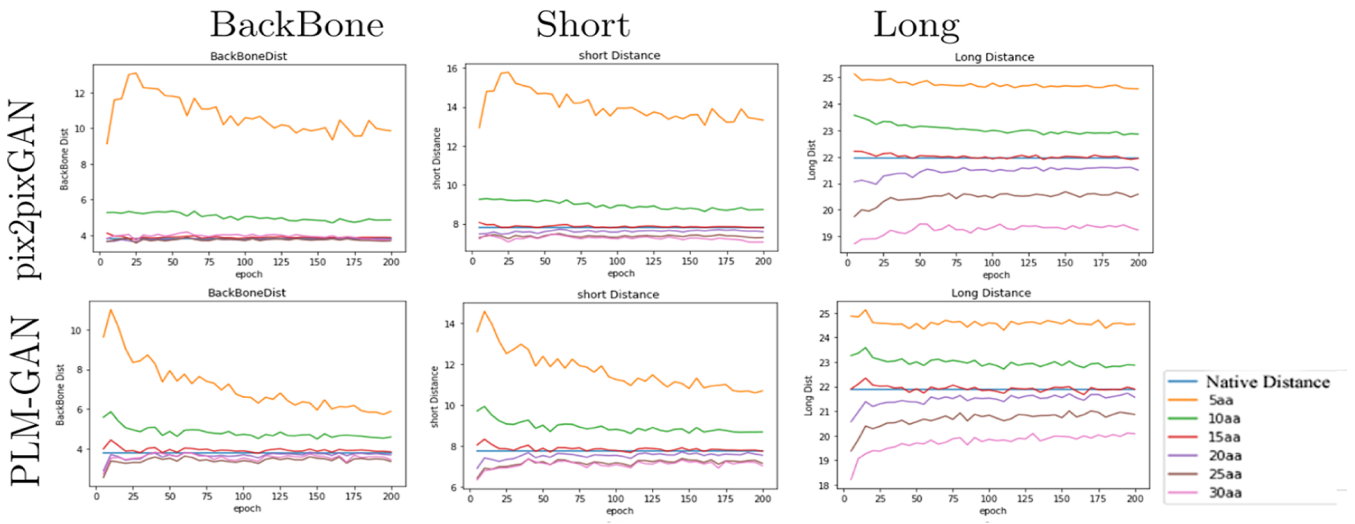


Figure 4. Comparison between the native and the generated proteins' structure distance matrix features (backbone, local, and distal characteristics) in each epoch.

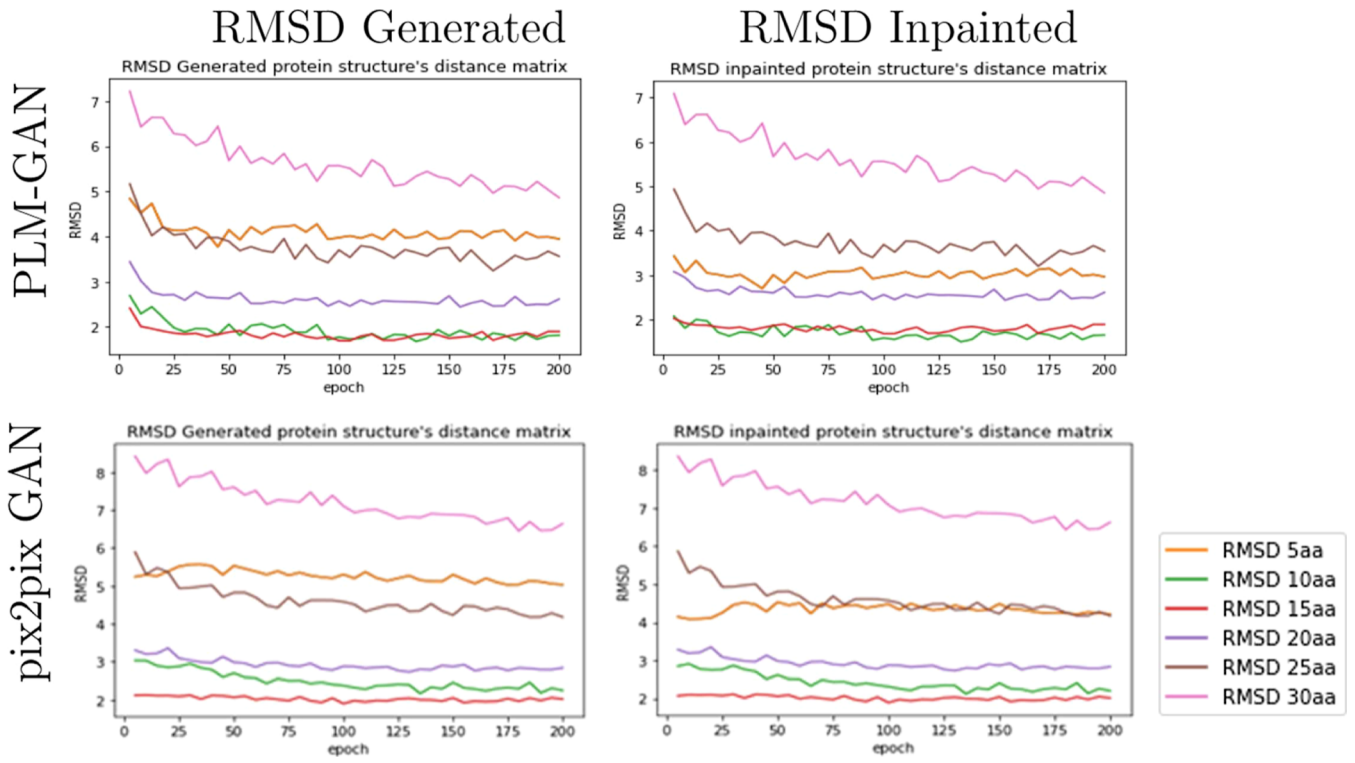


Figure 5. Comparison between the native and the generated and inpainted proteins' structure distance matrix by pix2pix GAN and PLM-GAN models through the rmsd along each epoch.

atoms in the main chain of the protein structure to build the distance matrix. The dimensions of the matrices are $n*n$, where n is equal to 128 aa. To ensure the diversity of the missing regions, the proposed models paired native protein structure's distance matrices with changing the position of the missing regions in each epoch, which increases the accuracy of the models.

IMPLEMENTATION DETAILS

The PyTorch framework was employed to conduct all experiments, utilizing an RTX2080 GPU and 128GB of RAM. For each discriminator and generator, the learning rate was set to 0.001. We employed the Adam optimizer, configuring the values of β_1 and β_2 to 0.9 and 0.999, respectively. Following a series of trial experiments, we determined that the hyperparameters γ and ∂ for the PLM-GAN model should be set to 200 and 2000, respectively. The training duration for each epoch was approximately 9 min, with a total of 200 epochs.

RESULTS AND DISCUSSION

We depended on a diversity of measurements to evaluate the performance of our models, which shows the quality and efficiency of our models.

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

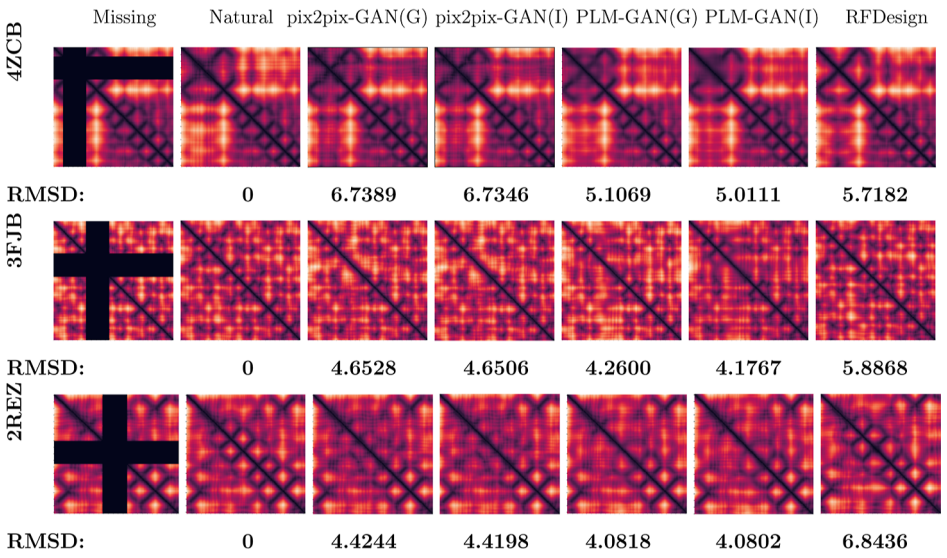


Figure 6. Heatmaps visually represent the distance matrices of protein tertiary structures generated using different models: pix2pix GAN, PLM-GAN hybrid, and RFDesign. Additionally, the rmsd between each generated or inpainted protein structure's distance matrix and the native one is indicated as (G) for generated and (I) for inpainted.

Average Peptide Bond. As previously stated, the distance matrix of protein structures is estimated by considering the average length of the peptide bond for the backbone as well as the short-range (local) and long-range (distal) distances. As shown in Figure 4, when using a test set of 6200 proteins with different masks ranging from 5 to 30 aa, our models were trained over multiple epochs, resulting in distance matrices for the proteins that closely resemble the native distance matrices. The stability of the models is reflected in the proximity of the generated distance matrices to their native counterparts, even in cases in which the models alter the length and positioning of the missing regions.

Evaluation of rmsd. As mentioned before, we used rmsd to compare the generated and the inpainted distance matrices of protein structure with the native distance matrix of the protein structure. As illustrated in Figure 5, our models are stable and effective in inpainting the missing region and generating the whole protein structure with a small rmsd value from as early as 25 epochs to the end of the training.

Although both models achieved great performance, the PLM-GAN model had better results in rmsd in both the generated and inpainted distance matrices. There was a clear superiority of PLM-GAN in rmsd between both the generated and inpainted distance matrix of protein structure. In addition, the features of the distance matrix of the protein structure are close to the native features, and this is evident, especially in the missing regions of small (5−10 aa) and medium (10−20 aa) length.

Comparison between Pix2Pix GAN and PLM-GAN Models with State-of-the-Art-Methods. In this study, we conducted extensive training on our model to tackle the challenging task of handling missing regions in protein structures. Our training data set covered a diverse range of lengths for these missing regions, spanning from 5 to 30 amino acids. To ensure a fair and rigorous comparison, we deliberately maintained a consistent missing region length of 25 aa across the proteins when comparing our model with the state-of-the-art model RFDesign.[20] This thoughtful approach allowed us to directly evaluate the efficacy of both models under identical conditions, shedding light on their performance in handling the same inpainting task. By standardizing the missing region length and considering various regions of the proteins, we significantly bolstered the reliability and validity of our comparative analysis. This allowed us to gain valuable insights into how each model performed when confronted with inpainting challenges of the same missing length in different protein regions.

Incorporating RFDesign as a benchmark in our comparison enabled us to present a comprehensive assessment of our model's painting capabilities within the context of cutting-edge techniques. The outcomes of this comparison contribute significantly to the advancement of protein structure inpainting research, providing a clearer understanding of the strengths and limitations of both models under these specific conditions. To ensure impartial evaluation, we consciously excluded proteins 4ZCB, 3FJB, and 2REZ from our model's training data set. Additionally, we randomly selected these proteins from the PDB for the purpose of comparing them with RFDesign. This careful selection ensured that these proteins served as unseen data, guaranteeing an unbiased evaluation of the performance of our model's performance. Through this approach, we effectively assessed how well our model could be generalized to novel protein structures. By keeping the evaluation process consistent across all compared methods, we maintained the integrity of our findings. This allowed us to draw meaningful conclusions about its inpainting capabilities on diverse and previously unseen protein structures. In our model, the separation of the training and testing data ensured the reliability and validity of our comparative analysis.

The RFDesign algorithm does not generate distance matrices. However, it generates the protein structures as PDB files, which we converted to a distance matrix to compare it with the generated and the inpanted distance matrices of our models and the native distance matrix. Figure 6 demonstrates the results of three models (the pix2pix GAN, the PLM-GAN, and the RFDesign model) in each of the three proteins compared with the native protein. Our two models (pix2pix GAN and PLM-GAN) have two results in each protein: one for generating the whole protein (p2p-generated and PLM-GAN-generated) and the other for inpainting only the missing

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

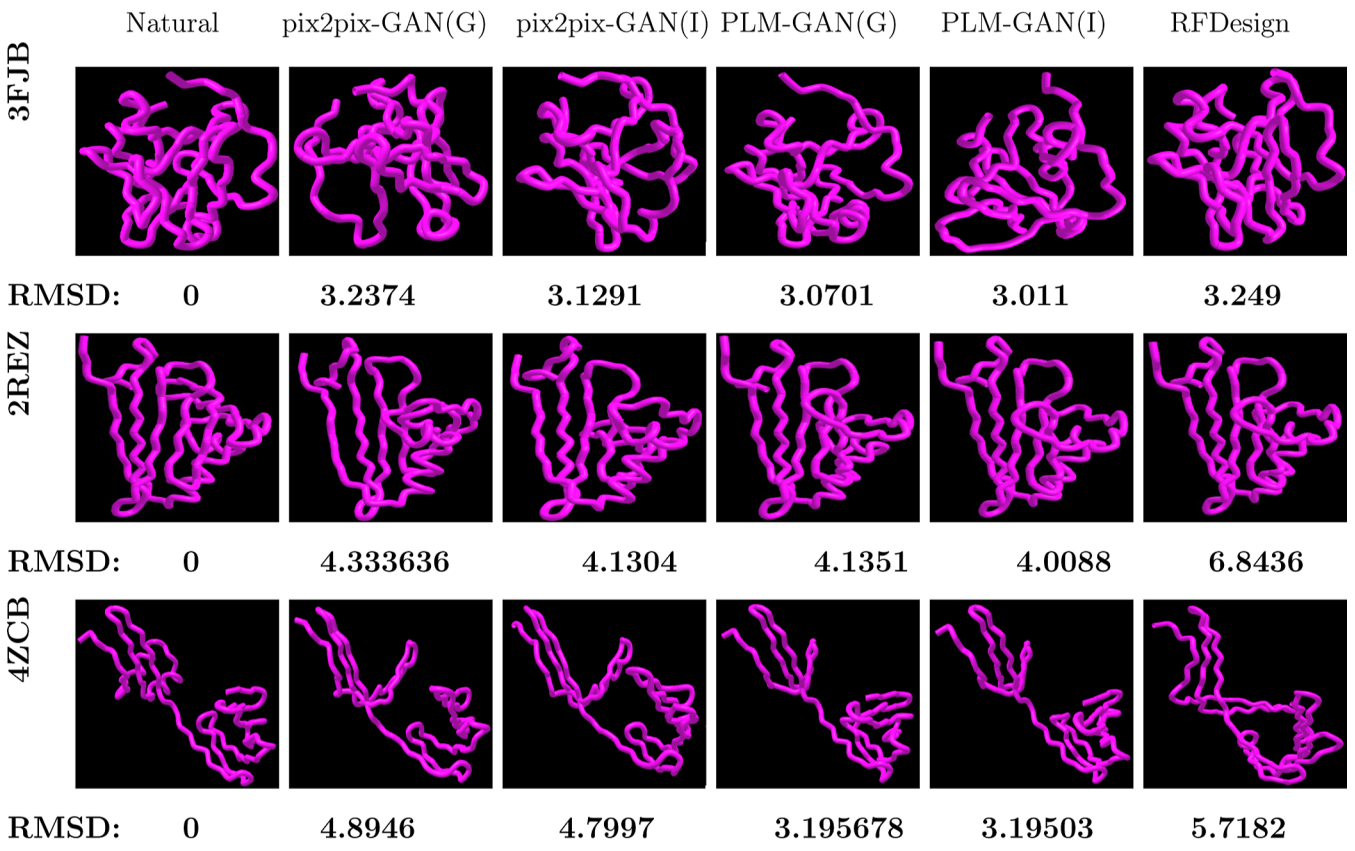


Figure 7. rmsd between each generated or inpainted protein structure and the native protein structure for the three models: (G) for generated and (I) for inpainted.

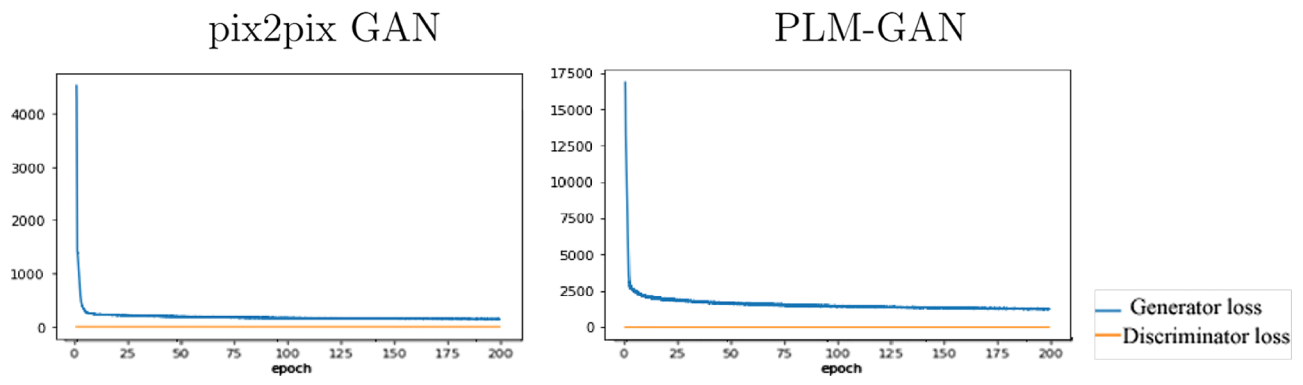


Figure 8. Convergence and stability analysis of pix2pix GAN and PLM-GAN losses.

region (p2p-inpainted and PLM-GAN-inpainted). We found the rmsd between the inpainted distance matrix of protein structure and the distance matrix of native protein structure for 4ZCB (6.7346, 5.0111, and 5.7182), 3FJB (4.6506, 4.1767, and 5.886), and 2REZ (4.4198, 4.0802, and 6.8436) for pix2pix GAN, PLM-GAN, and RFDesign, respectively. In all the proteins, PLM-GAN achieved better results than those of the RFDesign. In two of the proteins (3FJB and 2REZ), the pix2pix GAN was better than the RFDesign. Moreover, we observed that PLM-GAN outperforms RFDesign with an average processing time of under a second, making it superior for time-critical applications. In contrast, RFDesign requires several seconds and does necessitate GPU, thereby the PLM-GAN offers a more accessible solution for users without high-end hardware. These findings provide valuable insights for researchers and practitioners in selecting the most suitable methodology based on their specific needs. When comparing the performance of the models on the 3D protein structure, all three matrices were turned to 3D using metric multidimensional scaling (MMDS).[37] The generated and inpainted 3D protein structures for 4ZCB, 3FJB, and 2REZ by pix2pix GAN, PLM-GAN, and RFDsign models are shown in Figure 7. Both models pix2pix GAN and PLM-GAN obtained smaller rmsd than that of the RFDesign, and PLM-GAN achieved the best result of the three models.

Convergence Analysis. We examined the convergence and reduction of loss for both the generator and the discriminator curves during training. In Figure 8, we illustrate the performance loss of our models, pix2pix GAN and PLM-GAN, on the 128 aa training data set. The results demonstrate the stability and convergence of the loss curves over the epochs.

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446

CONCLUSIONS

In this article, we have focused on the loop modeling problem, in which the protein tertiary structures may have missing regions or regions that need to be reconstructed. Solving this problem represents a major step toward simplification of protein design and protein prediction models. In addition, it will help other models of protein−protein interactions and drug design achieve better results. The pix2pix GAN and PLM-GAN models were developed to generate and inpaint protein distance matrices and use MMD to "fold" the protein structure. Our models were developed through five contributions: (I) applying pix2pix GAN to generate and inpaint distance matrix of protein structure. (II) Developing the PLM-GAN model based on the pix2pix GAN by integrating the residual blocks in the U-Net network of the GAN network. (III) Adding a new loss function missing to real (L_{MTR}) loss in pix2pix GAN to make PLM-GAN. (IV) Pairing two different distance matrices (one of the native protein structure and one of the same structure but with a missing region that changes in each successive epoch). (V) Increasing the length of the missing region up to 30 aa and the length of the protein to 128 aa. We applied the pix2pix GAN and PLM-GAN models on the natural proteins 4ZCB, 3FJB, and 2REZ, obtaining promising experimental results for inpaint in 2D: rmsd of 6.7346, 4.6506, and 4.4198 for pix2pix GAN and 5.0111, 4.1767, and 4.0802 for PLM-GAN. The distance matrix was converted to inpaint in 3D, obtaining an rmsd of 3.1291, 4.1304, and 4.7997 for pix2pix GAN and 3.0110, 4.0088, and 3.1953 for PLM-GAN.

In future work, we can increase the length of the missing region to more than 30 aa. Also, we can create a graphical user interface for the user to generate the missing region in different protein structures. We may also extend our models to work on 3D proteins directly without needing MMD to convert them. In addition, we will focus on the functional sites, protein−ligand binding sites, and protein−protein interactions to be handled by the GAN models.

ASSOCIATED CONTENT

Data Availability Statement

Code, data, and models are available on https://github.com/mena01/PLM-GAN-A-Large-Scale-Protein-Loop-Modeling-Using-pix2pix-GAN_. The data used to train models is from the PDB.[10] The id of proteins and code used to download proteins and generate the distance matrices are provided at https://github.com/mena01/PLM-GAN-A-Large-Scale-Protein-Loop-Modeling-Using-pix2pix-GAN_/tree/main/data%20set.

AUTHOR INFORMATION

Corresponding Author

Mena Nagy A. Khalaf − *Information System Department, Faculty of Computer and Information, Assiut University, Assiut 71515, Egypt;* orcid.org/0000-0002-6207-3075; Email: Mena.Nagy@aun.edu.eg

Authors

Taysir Hassan A Soliman − *Information System Department, Faculty of Computer and Information, Assiut University, Assiut 71515, Egypt*

Sara Salah Mohamed − *Information System Department, Faculty of Computer and Information, Assiut University, Assiut 71515, Egypt; Mathematics and Computer Science Department, Faculty of Science, New Valley University, New Valley 71511, Egypt*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05863

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

I would like to express my gratitude for the invaluable assistance provided by the Data Science Lab Department of Information System, Faculty of Computers and Information, Assiut University, Egypt.

REFERENCES

(1) Yu, W.; MacKerell, A. D. Computer-aided drug design methods. *Antibiotics* **2017**, *1520*, 85−106.

(2) Renaud, N.; Geng, C.; Georgievska, S.; Ambrosetti, F.; Ridder, L.; Marzella, D. F.; Réau, M. F.; Bonvin, A. M.; Xue, L. C. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nat. Commun.* **2021**, *12*, 7068.

(3) Xia, W.; Zheng, L.; Fang, J.; Li, F.; Zhou, Y.; Zeng, Z.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. P. F.D. L. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput. Biol. Med.* **2022**, *145*, 105465.

(4) Chen, S.; Zhang, E.; Jiang, L.; Wang, T.; Guo, T.; Gao, F.; Zhang, N.; Wang, X.; Zheng, J. Robust prediction of prognosis and immunotherapeutic response for clear cell renal cell carcinoma through deep learning algorithm. *Front. Immunol.* **2022**, *13*, 798471.

(5) Smyth, M. S.; Martin, J. H. x Ray crystallography. *Mol. Pathol.* **2000**, *53*, 8−14.

(6) Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **2001**, *8*, 923−925.

(7) Carroni, M.; Saibil, H. R. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods* **2016**, *95*, 78−85.

(8) Vangaveti, S.; Vreven, T.; Zhang, Y.; Weng, Z. Integrating ab initio and template-based algorithms for protein−protein complex structure prediction. *Bioinformatics* **2020**, *36*, 751−757.

(9) Kumar, P.; Halder, S.; Bansal, M. *Biomolecular Structures: Prediction, Identification and Analyses*; Elsevier, 2019.

(10) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(11) The UniProt Consortium. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.

(12) Chen, D.; Zheng, J.; Wei, G.-W.; Pan, F. Extracting Predictive Representations from Hundreds of Millions of Molecules. *J. Phys. Chem. Lett.* **2021**, *12*, 10793−10801.

(13) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.

(14) Ding, W.; Nakai, K.; Gong, H. Protein design via deep learning. *Briefings Bioinf.* **2022**, *23*, bbac102.

(15) Ovchinnikov, S.; Huang, P.-S. Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **2021**, *65*, 136−144.

(16) Li, Z.; Nguyen, S. P.; Xu, D.; Shang, Y. Protein loop modeling using deep generative adversarial network. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017; pp 1085−1091.

(17) Nguyen, S. P.; Li, Z.; Xu, D.; Shang, Y. New deep learning methods for protein loop modeling. *IEEE ACM Trans. Comput. Biol. Bioinf.* **2019**, *16*, 596−606.

(18) Anand, N.; Huang, P. Generative modeling for protein structures. *Advances in neural information processing systems*; Curran Associates, Inc., 2018; Vol. *31*.
(19) Huang, P.-S.; Ban, Y.-E. A.; Richter, F.; Andre, I.; Vernon, R.; Schief, W. R.; Baker, D. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **2011**, *6*, No. e24109.
(20) Wang, J.; Lisanza, S.; Juergens, D.; Tischer, D.; Watson, J. L.; Castro, K. M.; Ragotte, R.; Saragovi, A.; Milles, L. F.; Baek, M.; et al. Scaffolding protein functional sites using deep learning. *Science* **2022**, *377*, 387−394.
(21) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(22) Yang, Z.; Zeng, X.; Zhao, Y.; Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Targeted Ther.* **2023**, *8*, 115.
(23) Khan, A. A.; Khan, Z. Comparative host−pathogen protein−protein interaction analysis of recent coronavirus outbreaks and important host targets identification. *Briefings Bioinf.* **2021**, *22*, 1206−1214.
(24) Pan, X.; Kortemme, T. Recent advances in de novo protein design: principles, methods, and applications. *J. Biol. Chem.* **2021**, *296*, 100558.
(25) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Briefings Bioinf.* **2021**, *22*, bbaa161.
(26) T Magalhães, B.; Lourenço, A.; Azevedo, N. F. Computational resources and strategies to assess single-molecule dynamics of the translation process in S. cerevisiae. *Briefings Bioinf.* **2021**, *22*, 219−231.
(27) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*; Curran Associates, Inc., 2014; Vol. *27*.
(28) Huang, R.; Zhang, S.; Li, T.; He, R. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, 2017; pp 2439−2448.
(29) Vondrick, C.; Pirsiavash, H.; Torralba, A. Generating Videos with Scene Dynamics. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2016.
(30) Yeh, R. A.; Chen, C.; Yian Lim, T.; Schwing, A. G.; Hasegawa-Johnson, M.; Do, M. N. Semantic Image Inpainting With Deep Generative Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
(31) Thirumagal, E.; Saruladha, K. *Generative Adversarial Networks for Image-to-Image Translation*; Elsevier, 2021; pp 17−57.
(32) Chen, D.; Wawrzynski, P.; Lv, Z. Cyber security in smart cities: a review of deep learning-based applications and case studies. *Sustain. Cities Soc.* **2021**, *66*, 102655.
(33) Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2016**, arXiv:1611.07004. https://arxiv.org/abs/1611.07004
(34) Maiorov, V. N.; Crippen, G. M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* **1994**, *235*, 625−634.
(35) Rahman, T.; Du, Y.; Zhao, L.; Shehu, A. Generative adversarial learning of protein tertiary structures. *Molecules* **2021**, *26*, 1209.
(36) Khalaf, M. N. A.; Soliman, T. H. A.; Mohamed, S. S. Generating Nature-Resembling Tertiary Protein Structures with Advanced Generative Adversarial Networks (GANs). *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 1078.
(37) Salkind, N. J. *Encyclopedia of measurement and statistics*; SAGE publications, 2006.

https://doi.org/10.1021/acsomega.3c05863
ACS Omega 2024, 9, 437−446