Improved inter-residue contact prediction via a hybrid generative model and dynamic loss function

Mohammad Madani a,b, Mohammad Mahdi Behzadi a,b, Dongjin Song b, Horea T. Ilies a, Anna Tarakanova a,c,*

aDepartment of Mechanical Engineering, University of Connecticut, Storrs, CT, United States
bDepartment of Computer Science & Engineering, University of Connecticut, Storrs, CT, United States
cDepartment of Biomedical Engineering, University of Connecticut, Storrs, CT, United States

Abstract

Protein contact maps represent spatial pairwise inter-residue interactions, providing a protein’s translationally and rotationally invariant topological representation. Accurate contact map prediction has been a critical driving force for improving protein structure determination. Contact maps can also be used as a stand-alone tool for varied applications such as prediction of protein–protein interactions, structure-aware thermal stability or physicochemical properties. We develop a novel hybrid contact map prediction model, CGAN-Cmap, that uses a generative adversarial neural network embedded with a series of modified squeeze and excitation residual networks. To exploit features of different dimensions, we introduce two parallel modules. This architecture improves the prediction by increasing receptive fields, surpassing redundant features and encouraging more meaningful ones from 1D and 2D inputs. We also introduce a new custom dynamic binary cross-entropy loss function to address the input imbalance problem for highly sparse long-range contacts in proteins with insufficient homologs. We evaluate the model’s performance on CASP 11, 12, 13, 14, and CAMEO test sets. CGAN-Cmap outperforms state-of-the-art models, improving precision of medium and long-range contacts by at least 3.5%. As a direct assessment between our model and AlphaFold2, the leading available protein structure prediction model, we compare extracted contact maps from AlphaFold2 and predicted contact maps from CGAN-Cmap. The results show that CGAN-Cmap has a mean precision higher by 1% compared to AlphaFold2 for most ranges of contacts. These results demonstrate an efficient approach for highly accurate contact map prediction toward accurate characterization of protein structure, properties and functions from sequence.

1. Introduction

Protein structure determination from sequence has long been one of the most challenging problems in structural biology [1–3]. Experimental approaches to determine the 3D structure and properties of a protein are often time-consuming and costly [4,5]. For example, only about 200,000 protein structures have been identified through experimental methods such as X-ray crystallography compared to millions of existing protein sequences [6]. Progress in community-wide Critical Assessment of Structure Prediction (CASP) experiments has demonstrated that contact map prediction, used as an intermediary constraint, boosts the accuracy of ab initio folding simulations and improves the success rate for the prediction of protein targets [7]. Knowledge about contacts between pairs of residues provides a translationally and rotationally invariant topological representation of a protein that can capture the global topology of the protein fold [7,8]. In addition to protein structure prediction, recently, contact maps are widely used in many other predictive tools for elucidating protein properties and functions. For instance, structure-aware protein–protein interactions [9], thermal stability [10], and solubility predictors [11] have employed contact maps as key inputs within the training process to feed spatial and contextual information about the protein structure to the predictive tool without using 3D protein structure. Furthermore, notably, AlphaFold2 [12] made recent breakthroughs in sequence-based protein structure prediction.
predicting 98% of human proteins with accuracy close to experiment [12,13]. An important component of AlphaFold2’s model architecture is the Evoformer, utilizing various pairwise modules to work on pairwise relations, e.g., to determine the distance or contacts between pairs of residues within the protein sequence [12]. AlphaFold2’s superior performance suggests that accurate pairwise relations (distance maps or contact maps) are key determinants in protein structure prediction.

A number of computational tools to predict protein contact maps have been developed, mostly falling under two common classes of methods: Evolutionary Coupling Analysis (ECA) and machine learning (ML) [14] methods. ECA predicts contact maps by analyzing evolutionary correlations of proteins derived from multiple sequence alignments (MSAs). Direct evolutionary analysis (DCA) is most accurate among ECA methods, using MSAs to identify direct evolutionary residue-residue relationships and correlations. DCA considers the effect of residues in different positions on each pair of residues to quantify the correlations between them [15,16]. Popular DCA methods include CCMPred [17], FreeContact [18], GREMLIN [19], and PSICOV [20]. To find these relationships, DCA commonly utilizes graphical lasso [15] and pseudo-likelihood maximization [16]. Graphical lasso finds the graph structure from a covariance matrix, and pseudo-likelihood maximization is a probabilistic model to find the strength of interactions between residues [15,16]. Notably, DCA methods generally exhibit two major limitations. First, DCA-based methods have poor performance when the number of homologous sequences is lower than approximately 50 [21,22]. Second, these methods extract only linear relationships between pairs of residues [21]. Conversely, the relationships between pairs of residues are intrinsically non-linear [21].

By contrast, machine learning, and more specifically deep learning, approaches, have been successfully utilized to find more accurate contact maps for proteins with few homologous sequences [23]. Most of the recently proposed models tested in CASP are based on deep learning, in particular residual neural networks (ResNets) [21–22,24], which resolve the classical machine learning vanishing gradient problem, and more importantly, show sufficient depth to accurately predict protein contact maps [25]. ResNets effectively encourage most important features and bypass low quality information within feature maps. Notably, two of the most commonly used and accurate contact map predictors (as confirmed by CASP 12 and 13), RaptorX-contact [24] and TripletRes [21], are ResNet-based. Besides residual networks, other deep learning models show comparable performance in predicting accurate contact maps, especially for long-range contacts [26,27]. Generative Adversarial Networks (GANs) are one of a powerful class of neural networks used for unsupervised learning, capable of learning low and high-level patterns [28]. GANs comprise two competitive networks, namely a generator network and a discriminator. During training, the generator learns to maximize the accuracy of generated samples to fool the discriminator. On the other hand, the discriminator tries to distinguish between the generated samples and real data [29]. Conditional GANs have been widely adopted for high-level resolution generation tasks including image-to-image translation [30]. Thus, this functionality enables GANs to be used in making predictions of optimal and more accurate contact maps [28,31,32]. Recently, GANCon [31] and contactGAN [32] have been introduced for contact map prediction utilizing the traditional GAN model based on the decoder and encoder-based architectures. GANCon [31] is an independent tool showing relatively weak performance on CASP datasets (9% less than RaptorX-Contact for CASP 13 and CASP 14) because of low depth of its architecture. ContactGAN [32] is a dependent tool used for improvement of predicted contact maps within existing models, rather than a direct prediction of contact maps, utilizing an architecture similar to GANCon [31] but with larger depth networks. ContactGAN [32] improved the precision of contact map prediction by 2% compared to RaptorX-Contact [24] for CASP 13, but it could not compete with TripletRes for CASP 13 and CASP 14 (with a mean precision at least 4.8% less than TripletRes [21]).

Despite recent progress in contact map prediction, a challenging outstanding problem remains: the prediction of medium and long-range contacts (i.e., where the distance between position indices of pairs of amino acid residues is more than 12 Å). These types of contacts are highly sparse, complicating accurate prediction [21]. Sparsity creates an imbalance problem, when the ratio of contacting and uncontacting residues is low (less than 3%). To tackle this, we build a new tool to accurately predict contact maps for medium and long-range contacts regardless of the number of homologous sequences available. Inspired by the capabilities of GANs for generating high-resolution images from input features, in this study we propose a contact map predictor, CGAN-Cmap, constructed via integration of a modified squeeze excitation residual neural network (SE-ResNet), SE-Concat, and a conditional GAN. In the architecture of the GAN generator, we define two novel subnets to extract feature representations during training and feed them to subsequent layers gradually. We also introduce a custom loss function for training the model, a dynamic weighted binary cross-entropy (BCE) loss function, which assigns a dynamic weight for classes based on the ratio of the uncontacted class to the contacted class in each iteration of the training process. This loss function emphasizes misclassified residue pairs to enhance the model training and tackle the imbalance problem. By taking advantage of the model architecture and custom loss function, the receptive field and feature reusability are significantly enhanced, which allows us to capture more complex non-linear relationships between amino acid residues even for proteins with few homologous sequences. We assess the performance of CGAN-Cmap on existing recent CASP datasets (CASP 11, CASP 12, CASP 13, CASP 14). The results reveal that CGAN-Cmap utilizing the dynamic BCE as a custom loss function consistently outperforms leading existing models [21,24] on medium and long-range contacts for all recent CASP datasets. Furthermore, as a direct assessment of our tool with AlphaFold, we compare the predicted contact map by CGAN-Cmap with extracted contact maps from AlphaFold1 and AlphaFold2. CGAN-Cmap consistently outperforms AlphaFold tools for different ranges of contacts. We also show that CGAN-Cmap has a moderate dependency on the number of homologous sequences available in MSAs, ensuring that it can accurately predict contact maps for proteins with insufficient homologous sequences. Overall, we show that CGAN-Cmap derives its superior performance from a custom loss function to overcome the sparsity problem, and a novel architecture, which increases the receptive field and feature reusability as well as encourages more important features within the feature map.

2. Materials and methods

2.1. Contact map definition

The protein contact map is symmetric and represents the probability of contact between two residues. If the Euclidean distance is less than 8 Å between two beta carbon (Cβ) atoms within two residues (for glycine alpha carbons (Cα) are used), these residues are considered to be contacting residues in the protein contact map. This definition for contact maps has been used consistently in all existing contact map predictors and verified by CASP [21,24,31–33]. Generally, contacts are divided into three categories based on the difference, Diff, in the position index of two residues, A and B, in the protein sequence, as follows:
Diff = Position index(Residue A) – Position index(Residue B)

\[
\text{if } 6 < \text{Diff} < 12 : \text{ short – range contact} \\
\text{if } 12 \leq \text{Diff} < 24 : \text{ medium – range contact} \\
\text{if } \text{Diff} > 24 : \text{ long – range contact} \\
\text{if } \text{Diff} > 50 : \text{ extra long – range contact}
\]

Note that because of the importance of long-range contacts, we introduce a new category termed long-range contacts.

2.2. Datasets

The training dataset for CGAN-Cmap is collected from SCOPe-2.07 [34] and a subset of the Protein Data Bank filtered at 25% sequence identity (PDB25) [35]. The protein sequence selection is conducted per the following criteria: 1) sequence length is between 25 and 700 residues; 2) the maximum sequence identity is set to 30%; 3) the native contact map resolution is more than 2.5 Å; 4) the protein contains a single chain. Based on these criteria, 7,128 protein sequences are selected as the training dataset. To evaluate the performance of our model, we use five different blind public test sets including CAMEO (76 proteins) [37], CASP 11 (110 protein targets) [38], CASP 12 (54 protein targets) [38], and CASP 13 (41 protein targets) [38]. In a second redundancy parameter tuning, the final model is the average of five models, which treat the feature maps equally, the SE block weighs information based on its effect on the final prediction, to boost model performance.

2.3. Feature extraction

To build a machine learning model, we extract both pairwise (co-evolution) and sequential features in a process similar to that introduced in RaptorX-Contact [24]. Pairwise features are captured from multiple sequence alignments (MSAs). To generate the MSA for training and test sets, we use the modified version of DeepMSA first proposed by Zhang et al. [39]. Additional details on MSA generation are described in the Supplementary Information (SI). After MSA generation, pairwise features including the evolutionary coupling matrix generated by CCMPred [17], the mutual information matrix [40], the average product correction [41], and the pairwise contact potentials matrix [42] are extracted from MSAs. For sequential features, we use 3- and 8-state protein secondary structure (SS3, SS8), 3-state solvent accessibility (ACC), and features derived from MSAs in a 20-dimensional position-specific scoring matrix (PSSM) and 20-dimensional position-specific frequency matrix (PSFM) [43]. Because PSFM contains the frequency of amino acids in the protein sequence, we include it in the input features to complement the position-specific scoring matrix. In total, sequential features are represented in a two-dimensional matrix with a size of $L \times 54$, and pairwise features are represented in a three-dimensional matrix with a size of $L \times L \times 5$, where L is the sequence length. More detailed information about features extraction is provided in SI.

2.4. CGAN-Cmap architecture

We employ a conditional generative adversarial neural network (CGAN) which consists of a generator to produce the contact map and a discriminator to compare the generated contact map with the real contact map (Fig. 1). The overall structure of our GAN is inspired by the GANTL model [44]. In contrast to traditional GAN models [31,32], our generator does not have the decoder-encoder architecture. Rather, the generator contains three novel subnets to predict the contact map directly from input features without converting input features to a low dimensional space as an intermediate step to contact map prediction. The generator contains three novel subnets. The first subnet, the conversion subnet, includes a conversion block that converts 1D features to 2D matrices. Unlike the outer concatenation approach used in most existing models [22,24] to convert 1D features to 2D matrices, our conversion method eliminates the repetition of information, and more importantly, it is learned during the training process, rather than preprocessed. To do so, the model learns the best-converted matrix for each 1D feature (SI Fig. S2). The second subnet is the synthesis subnet and includes synthesis (SI Fig. S3A) and upsampling (SI Fig. S3C) blocks, as explained in the SI. This subnet is responsible for generating the contact map of proteins from the converted 2D matrix from the first subnet and a noise vector. The third subnet is the feature extraction subnet, which receives the 2D features (pairwise features) as input and extracts useful feature maps from 2D information. This subnet includes a series of SE-Concat blocks (SI Fig. S3B), that pass the extracted feature map to the synthesis subnet gradually to generate the protein contact map. [45,46]. SE-Concat is inspired from the SE-ResNet architecture, modified to use concatenation instead of summation, to reuse the extracted features and increase the information flow between layers [45,47,48]. The squeeze excitation (SE) block (SI Fig. S3D) within the SE-Concat block emphasizes the important information and improves channel interdependencies [46]. Unlike other models which treat the feature maps equally, the SE block weights information based on its effect on the final prediction, to boost model performance.

The discriminator consists of four convolutional layers and a dense layer that determine the probability of the image being real. The input to the discriminator is the original 2D features, converted 2D features derived from 1D features, and their corresponding contact map. It is worth noting that the discriminator uses the output of the dot layer in the conversion block as the converted 2D features, and ignores the remaining three convolutional layers in the conversion block. This ensures that all matrices have the same dimension, and the output of the dot layer has the same size as the contact map, so we can concatenate the converted 2D features and the original 2D features with the contact map.

2.5. Training process

We aim to train an end-to-end model in a single training process with the min–max objective function given by Eq. (2), where G and D are parameters (min_{G}, max_{D}) of the generator and the discriminator networks of our model, y_G^D is the real (correct) data distribution, y_G^E is the generated (fake) data distribution, D_r and D_f predicts probabilities of the real (x) and fake (z) data by discriminator to distinguish between real and fake data.

$$
\min_{G} \max_{D} [\log(D_r) + (1 - \log(D_f))]
$$

To do this, we use mini-batches, because the training samples have different size. We choose a batch of training samples and then consider the largest sample size in the batch; then, we pad the samples using the zero-padding method. The padding size is deter-
Hyperparameters are reported in SI Table S2. Stability is an important consideration when training generative models, meaning that the loss values for discriminator and generator should converge to a value after some iterations of training. We confirm the stability seen in the loss per iteration for both generator and discriminator.

For the dynamic weighted BCE loss function, $L_{\text{wbce}}$, the second one is a dynamic weighted BCE ($L_{\text{wbce}}$). The loss functions are defined as follows:

$$L_{\text{bce}} = -\frac{1}{N} \sum_{n=1}^{N} [y_n^t \times \log (y_n^p) + (1 - y_n^t) \times \log (1 - y_n^p)]$$

$$L_{\text{wbce}} = -\frac{1}{N} \sum_{n=1}^{N} [w_{fn} \times y_n^t \times \log (y_n^p) + w_{fp} \times (1 - y_n^t) \times \log (1 - y_n^p)]$$

where $N$ is the number of samples, $y_n^t$ is the $n_{th}$ ground truth sample, $y_n^p$ is the $n_{th}$ predicted sample, $w_{fn}$ is the marginal cost of a false negative over true positive, and $w_{fp}$ is the marginal cost of a false positive over true negative [49]. The weights control the contribution of each part of the loss function in the training process. For the dynamic weighted binary cross-entropy loss function, we allow the model to choose the weights of each sample dynamically based on the false negative and false positive values in that sample. Finally, we use fivefold cross-validation for training and the overall performance is calculated by averaging from all folds. Hyperparameters are reported in SI Table S2. Stability is an important consideration when training generative models, meaning that the loss values for discriminator and generator should converge to a value after some iterations of training. We confirm the stability seen in the loss per iteration for both generator and discriminator in SI Fig. S5. For strict evaluation, we report results for our cross validation in SI Table S3.

### 2.6. Model evaluation

We compare the performance of our model to state-of-the-art contact map predictors, including TripletRes [21], RaptorX-contact [24], DeepCon [33], DeepCov [50], GANcon [31], contactGAN [32], AlphaFold1 [51], and AlphaFold2 [12]. For reference, we provide the important characteristics of these existing models in SI Table S4. We use mean precision of top $L/k$ contacts ($k = 10, 5, 2, 1$; $L$ is the sequence length) for short, medium, long, and extra long-range contacts to evaluate the prediction performance of the protein contact map tools.

### 3. Results & discussion

#### 3.1. Model performance on CASP 11, 12, and CAMEO targets

Table 1 summarizes the performance of our proposed tool on CASP 11, CASP 12, and CAMEO test sets, with standard BCE and dynamic weighted BCE, compared with five existing models. The results show that our models with both standard BCE and dynamic weighted BCE outperform state-of-the-art predictors when performance is evaluated based on mean precision. For instance, on CASP 12, mean precision of CGAN-Cmap for long-range top $L/10, L/5$, and $L/2$ contacts is higher by at least $4\%$, $5\%$, and $4\%$, respectively, compared to close competitors. These results confirm that the combination of the GAN model with our modified version of residual network blocks (SE-Concat blocks) and dynamic BCE loss function can be employed to improve model performance for protein contact map prediction. For the CAMEO dataset, CGAN-Cmap shows performance slightly inferior (by less than $1\%$) than TripletRes for some contact ranges. We expect this result to reflect the nature of proteins within the CAMEO dataset: membrane proteins found in the CAMEO dataset are more difficult prediction targets compared to globular proteins.

#### 3.2. Model performance on CASP 13 and CASP 14 targets

We also evaluate the performance of our model on recent CASP targets, CASP 13 and CASP 14, and compare to existing models (Table 2). As a note, we extract input features for all models from the same MSA for a fair comparison. In CASP 13, a new range of contact, termed extra long-range, was introduced. So, to be consistent with the recent evaluation protocol in CASP, we evaluate the performance of all models by calculating the mean precision of top $L, L/2, L/5$, and $L/10$ contacts in the (medium + long)-range, long-range, and extra long-range. We find that both versions of the CGAN-Cmap model consistently outperform all existing mod-
Model performance of CGAN-Cmap, TripletRes [21], RaptorX-contact [24], DeepCon [33], DeepCov [50], contactGAN [32], and GANcon [31] based on the mean precision of short, medium, and long-range top L/10, L/5, L/2, and L contacts in predicted contact maps for CASP 11 [38], CASP 12 [38], and CAMEO [37] test sets. The highest accuracy in each category is highlighted in bold font.

| Test set | Model               | Short-range (L/10) | Medium-range (L/5) | Long-range (L) |
|----------|---------------------|--------------------|--------------------|----------------|
| CASP 11  | DeepCon             | 0.21               | 0.70               | 0.77           |
|          | GNCon               | 0.20               | 0.37               | 0.70           |
|          | ContactGAN          | 0.24               | 0.45               | 0.73           |
|          | DeepCov             | 0.21               | 0.39               | 0.69           |
|          | RaptorX-contact     | 0.27               | 0.46               | 0.74           |
|          | TripleRes           | 0.26               | 0.44               | 0.72           |
|          | CGAN-Cmap (standard)| 0.25               | 0.42               | 0.72           |
|          | CGAN-Cmap (dynamic) | 0.27               | 0.49               | 0.74           |
| CASP 12  | DeepCon             | 0.25               | 0.58               | 0.69           |
|          | GNCon               | 0.24               | 0.60               | 0.69           |
|          | ContactGAN          | 0.48               | 0.71               | 0.78           |
|          | DeepCov             | 0.23               | 0.62               | 0.70           |
|          | RaptorX-contact     | 0.44               | 0.69               | 0.75           |
|          | TripleRes           | 0.47               | 0.70               | 0.77           |
|          | CGAN-Cmap (standard)| 0.53               | 0.68               | 0.80           |
|          | CGAN-Cmap (dynamic) | 0.58               | 0.72               | 0.81           |
| CAMEO    | DeepCon             | 0.18               | 0.33               | 0.48           |
|          | GNCon               | 0.20               | 0.34               | 0.47           |
|          | ContactGAN          | 0.22               | 0.35               | 0.56           |
|          | DeepCov             | 0.19               | 0.31               | 0.49           |
|          | RaptorX-contact     | 0.24               | 0.39               | 0.58           |
|          | TripleRes           | 0.26               | 0.42               | 0.60           |
|          | CGAN-Cmap (standard)| 0.21               | 0.37               | 0.58           |
|          | CGAN-Cmap (dynamic) | 0.29               | 0.41               | 0.61           |

Model performance of CGAN-Cmap, TripletRes [21], RaptorX-contact [24], DeepCon [33], DeepCov [50], contactGAN [32], and GANcon [31] based on the mean precision of short, medium, and long-range top L/10, L/5, L/2, and L contacts in predicted contact maps for CASP 13 and CASP 14 test sets. The highest accuracy in each category is highlighted in bold font.

| Test sets | Model               | (Medium + Long-range) (L/10) | Long-range (L/5) | Extra long-range (L) |
|-----------|---------------------|-------------------------------|-----------------|----------------------|
| CASP 13   | DeepCon             | 0.56                          | 0.70            | 0.81                 |
|          | GNCon               | 0.48                          | 0.61            | 0.72                 |
|          | ContactGAN          | 0.55                          | 0.69            | 0.79                 |
|          | DeepCov             | 0.46                          | 0.58            | 0.69                 |
|          | RaptorX-contact     | 0.62                          | 0.72            | 0.81                 |
|          | TripleRes           | 0.62                          | 0.72            | 0.81                 |
|          | CGAN-Cmap (standard)| 0.60                          | 0.69            | 0.80                 |
|          | CGAN-Cmap (dynamic) | 0.63                          | 0.73            | 0.82                 |
| CASP 14   | DeepCon             | 0.28                          | 0.37            | 0.45                 |
|          | GNCon               | 0.22                          | 0.28            | 0.34                 |
|          | ContactGAN          | 0.27                          | 0.37            | 0.51                 |
|          | DeepCov             | 0.21                          | 0.27            | 0.32                 |
|          | RaptorX-contact     | 0.32                          | 0.42            | 0.57                 |
|          | TripleRes           | 0.33                          | 0.42            | 0.56                 |
|          | CGAN-Cmap (standard)| 0.32                          | 0.41            | 0.55                 |
|          | CGAN-Cmap (dynamic) | 0.35                          | 0.46            | 0.62                 |

The highest accuracy in each category is highlighted in bold font.
72.4% by TripletRes, respectively. As shown in Fig. 3A and 3B, RaptorX-Contact and TripletRes do not predict any long-range contacts in Region 1, a critical loop-loop contact region. RaptorX-Contact fails to predict contacts in Region 2 (Fig. 3A). By contrast, CGAN-Cmap predicts both Regions 1 and 2, with high accuracy (Fig. 3C).

The performance of CGAN-Cmap is superior to existing state-of-the-art tools on CASP datasets, exceeding the best current methods. While some models such as contactGAN [32] and GANcon [31] have used the GAN model architecture, they did not show comparable performance with CGAN-Cmap or other top models, due to traditional decoder-encoder based GANs with low depth networks which could not capture highly sparse residues. In fact, this improved performance of CGAN-Cmap compared to other existing models originates from its dynamic weighted BCE loss function and its novel architecture. To summarize the effect of the architecture on the performance of the model, we highlight various blocks within the generator that are deterministic compo-
3.3. Direct assessment of predicted contact map via CGAN-Cmap by extracted contact map prediction from AlphaFold1 and AlphaFold2

Pairwise relationships between residues or protein contact maps can be used as a determinative constraint for computational protein structure prediction [1]. AlphaFold2 [12] is the most accurate protein structure predictor to date, utilizing MSA and pairwise distance maps as inputs for 3D protein structure prediction. The architecture of AlphaFold2 is constructed via 3 main modules, including the Evoformer module, structural module and recycling module [12]. To capture the last updated pairwise representations (pairwise distance maps) of proteins, we extract the output of the Evoformer module from the pre-trained AlphaFold2 model [12]. By considering the distance threshold of 8 Å, we convert extracted pairwise distance maps to contact maps, for a direct comparison with CGAN-Cmap. For additional assessment, we also consider the distance map converted to contact map from the first version of AlphaFold [51] (AlphaFold1). We compare contact map prediction performance of CGAN-Cmap with AlphaFold1 and AlphaFold2 for CASP 13 and CASP 14 (Table 3). Our model shows the mean precision of at least 6 % and 1 % higher than AlphaFold1 [51] and AlphaFold2 [12], respectively. Thus, we expect that our tools can improve the distance map converted to contact map from the first version of AlphaFold2 by adding some SE-ResNet, respectively, on CASP 13 and CASP 14 and summarize the performance in Table 4. Our CGAN-Cmap with SE-Concat blocks has at least 4 %, and 6 % mean precision improvement, for CASP 13 and CASP 14, respectively, on various top L/k long-range contacts compared to CGAN-Cmap with SE-ResNet blocks. We also calculate the average precision on the training set over the training epoch (Fig. 4). These results confirm that CGAN-Cmap with SE-Concat displays a consistently better performance than CGAN-Cmap with SE-ResNet in most epochs of the training process. SE-Concat blocks help to extract more meaningful patterns from the input features, since the concatenation layer reuses the extracted features and helps to increase the information flow between layers [45,46]. The enhancement in information flow alleviates the vanishing-gradient problem and strengthens feature propagation to more accurately predict sparse contacts within the contact maps.

3.4. SE-Concat block improves CGAN-Cmap performance to predict sparse contacts

One of the key novelties of our model is the use of the SE-Concat block instead of the standard ResNet or SE-ResNet layer. To show the effect of SE-Concat on the performance of CGAN-Cmap, we calculate the performance of the model with SE-Concat and SE-ResNet, respectively, on CASP 13 and CASP 14 and summarize the performance in Table 4. Our CGAN-Cmap with SE-Concat blocks has at least 4 %, and 6 % mean precision improvement, for CASP 13 and CASP 14, respectively, on various top L/k long-range contacts compared to CGAN-Cmap with SE-ResNet blocks. We also calculate the average precision on the training set over the training epoch (Fig. 4). These results confirm that CGAN-Cmap with SE-Concat displays a consistently better performance than CGAN-Cmap with SE-ResNet in most epochs of the training process. SE-Concat blocks help to extract more meaningful patterns from the input features, since the concatenation layer reuses the extracted features and helps to increase the information flow between layers [45,46]. The enhancement in information flow alleviates the vanishing-gradient problem and strengthens feature propagation to more accurately predict sparse contacts within the contact maps.

Input features are one of the key determinants of model performance. To clarify the impact of individual groups of input features on the performance of the CGAN-Cmap model, we examine the efficiency and importance of the ensembled feature collection on the contact predictions. To do so, we categorize all input features into four groups: 1) sequential features (L × 54), 2) only pairwise features (L × L × 5), 3) pairwise features (L × L × 5) + SS3 & SS8 (L × 11) + ACC (L × 3), and 4) pairwise features (L × L × 5) + PSFM (L × 20) + PSSM (L × 20). In all groups, except the first, we consider pairwise features since we expect these to be essential input features for protein contact map prediction. Fig. 5 presents the average top L/5 long-range contacts precision values over the training epochs. All models become stable after 300 epochs of training and reach a precision value of 65.4 %, 71.9 %, 80.4 %, 82.3 %, and 86.1 %, for groups 1, 2, 3, 4, and an ensemble of all input features, respectively. CGAN-Cmap trained with an ensemble of all input features achieves consistently better performance than CGAN-
Cmap trained with one of groups 1 to 4 only. In contrast to the model which uses sequential features (group 1), CGAN-Cmap with pairwise features (group 2) achieves higher mean precision for protein contact prediction, confirming that pairwise features are key elements for input. We note that when using group 4, our model achieves a mean precision close to the original model with an ensemble of all input features (lower by only 4%), suggesting that features derived directly from MSA such as PSSM and PSFM are critically important co-evolution features for improving performance of CGAN-Cmap. This analysis highlights the importance of pairwise features and specifically MSA-derived features (PSSM, PSFM) for contact map prediction, because these co-evolutionary features distinguish correlations resulting from direct or indirect effects of residue interactions, and thereby improve prediction accuracy of contacts. Result with group 3 suggest that other sequential features such as ACC and SS3 & SS8 can be used for further improvement of the model.

3.6. Effect of MSA protocol on CGAN-Cmap performance

The quality of the generated MSA has an essential impact on predicted contact maps because of its role in input feature generation. To show the type of dependency of CGAN-Cmap performance on the MSA extraction protocol and, more importantly, on the number of homologous sequences within the MSA, we train our model on input features extracted from MSA generated via only the HHblits approach [54] without any constraint on Neff value and compare its performance with original CGAN-Cmap trained with DeepMSA (Table 5). The model’s performance with DeepMSA is only slightly better than with HHblits on different ranges of contacts (by at most 2.5%), suggesting that our model has a modest dependency on the MSA building protocol and even with HHblits or other MSA generation approach, the model still maintains its performance. To further validate this modest dependency, we com-

### Table 4

Model performance of the original CGAN-Cmap and CGAN-Cmap with SE-ResNet blocks instead of SE-Concat based on the mean precision of the predicted contact map on CASP 13 and CASP 14 test sets.

| Test set | Model (Medium + Long)-range | Long-range | Extra long-range |
|----------|-----------------------------|------------|------------------|
|          | (L | L/2 | L/5 | L/10) | L | L/2 | L/5 | L/10 | L | L/2 | L/5 | L/10 |
| CASP 13  | SE_ResNet-CGAN-Cmap         | 0.58    | 0.66 | 0.76 | 0.84 | 0.55 | 0.64 | 0.76 | 0.81 | 0.34 | 0.49 | 0.61 | 0.68 |
|          | CGAN-Cmap (dynamic)         | 0.63    | 0.72 | 0.81 | 0.88 | 0.62 | 0.71 | 0.81 | 0.84 | 0.40 | 0.56 | 0.68 | 0.72 |
| CASP 14  | SE_ResNet-CGAN-Cmap         | 0.31    | 0.39 | 0.54 | 0.61 | 0.23 | 0.32 | 0.40 | 0.39 | 0.20 | 0.24 | 0.25 | 0.33 |
|          | CGAN-Cmap (dynamic)         | 0.35    | 0.46 | 0.62 | 0.69 | 0.32 | 0.38 | 0.44 | 0.48 | 0.24 | 0.29 | 0.33 | 0.39 |

Fig. 4. Comparison of the performance of the original CGAN-Cmap and CGAN-Cmap with SE-ResNet blocks by evaluation of the mean precision of the validation set over training epochs A) for top L/2 long-range contacts, and B) top L/5 long-range contacts.

Fig. 5. Mean precision of top L/5 long-range contacts over training epoch for the validation set showing the impact of various input features (ensemble of all input features, sequential features, pairwise, pairwise + PSSM + PSFM, and pairwise + SS3 + SS8 + ACC) on CGAN-Cmap performance.
compare the precision of top L/2 and top L/5 long-range contacts in predicted models with and without deep MSAs for all targets in CASP 12, 13, and 14 test sets (Fig. 6). The performance of CGAN-Cmap with both MSA building approaches is similar for most targets within the CASP test sets. In another analysis, we characterize the dependency of our model on the Neff value. To that end, first we divide targets into low Neff targets (Neff < 10) and high Neff targets (Neff > 10) and then define a precision threshold of 50%. Then, we calculate the precision of Top L/5 long-range contacts for all individual targets within the CASP 12, 13, and 14 by CGAN-Cmap, TripletRes [21], and Raptorx-Contact [24](Fig. 7).

Our model predicts most of the high Neff targets with a precision greater than 50% (Fig. 7A), confirming that higher Neff targets are predicted accurately with higher likelihood based on the higher quality information in their MSA. Interestingly, for low Neff targets, our model predicts about 67% (28/44) of them with the precision of more than 50%, suggesting that CGAN-Cmap can predict protein contact maps with high accuracy for proteins with a low Neff value. For comparison, we perform the same analysis for RaptorX-contact and TripletRes models (Fig. 7B and 7C). TripletRes has good performance on high Neff targets, but it cannot predict low Neff targets with high accuracy, in contrast to our model that demonstrates its low dependency on number of homologous sequences within the MSAs for contact map prediction.

Our tool displays high confidence in prediction of contact maps from sequence, especially for those sequences whose 3D structures are not experimentally identified. To validate our results, we calculate mean absolute error for different ranges of contacts for all CASP test sets (SI Table S5). CGAN-Cmap shows minimum mean absolute error compared to existing models [21,24]. As another validation of our model’s performance on blind test sets, we observe low dependency of CGAN-Cmap on the number of homologous sequences in MSA. Even for Neff < 10, CGAN-Cmap predicts contact maps with high confidence. Furthermore, we analyze our training and test sets for bias on specific ranges of sequence length, identifying that CGAN-Cmap can predict contact maps for proteins with sequence length from 46 to 583 with high accuracy. For example, the mean precision of long-range top L/10 contacts for proteins with sequence length more than 400 is about 0.85 and for proteins with sequence length less than 300 is 0.86. This analysis highlights that CGAN-Cmap can be effectively applied to diverse protein sequences which lack a verified 3D structure.

### 4. Conclusion

Accurate protein contact maps have been widely used for de novo protein property and structure prediction to guide construction of protein-based materials by e.g. identification of appropriate protein–protein interfaces. In this study, we propose a novel GAN-based architecture, CGAN-Cmap, for contact map prediction. During the adversarial learning process, the generator network of CGAN-Cmap captures the underlying contact information from versatile protein features by employing a dedicated generator including various blocks such SE-Concat and synthesis blocks.
while the discriminator network learns the difference between generated contact maps and real ones and automatically transfers them back to the generator network. By taking advantage of the novel architecture and training approach, our model outperforms all existing protein contact map predictors when the evaluation is assessed on different CASP targets. This capability of CGAN-Cmap can open a route to use more accurate contact maps for many structure-aware predictive tools beyond protein structure predictors. Furthermore, our proposed model exhibits a modest dependency on the number of homologous sequences in the MSA, resulting in accurate predictions for proteins with low Neff value. Although CGAN-Cmap shows promising performance for protein contact map prediction, there is still room for further improvement. In future studies advanced GAN training methods, such as WGAN [55], can improve training stability. A symmetric loss functions such as the focal loss [56] may also boost the predictive power of GAN-based architectures. The presented GAN-based deep learning architecture for contact map prediction can efficiently improve overall prediction performance and serves as a powerful new tool for contact map prediction.

5. Funding sources

This work utilized the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant No ACI-1548562. XSEDE resources Stampede 2 and Ranch at the Texas Advanced Computing Center and Bridges at the Pittsburg Supercomputing Center through allocation TG-MCB180008 were used.

Author contributions

The manuscript was written with contributions of all authors. All authors have given approval to the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.020.

Fig. 7. The precision of top 1/5 long-range predicted contacts versus Neff of MSAs for A) CGAN-Cmap, B) RaptorX-contact [24], and C) TripleletRes [21]. Horizontal dotted line is precision threshold (50%) and vertical dotted line represents Neff threshold (10).

References

[1] Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–6.
[2] Zheng Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18:342–8.
[3] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 2019;20:881–97.
[4] Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-EM. Nature 2020;587:157–61.
[5] Smyth M, Martin J. X Ray crystallography. Mol Pathol 2000;53:53:8.
[6] Berman HM et al. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907.
[7] Zheng W et al. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins Struct Funct Bioinf 2019;87:1149–64.
[8] Aknata LA, Tamb G, Monastyrskyy B, Krzhizhanovsky A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins Struct Funct Bioinf 2018;86:97–112.
[9] Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics 2022;38:125–32.
[10] Chu, S. K. & Siegel, J. Predicting single-point mutational effect on protein stability, growth 16, 35 (2021).
[11] Chen J, Zheng S, Zhao H, Yang Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. J Cheminf 2021;13:1–10.
[12] Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[13] Tunyasuvunakool K et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:596–60.
[14] Li J, Lin Y, Elhossain A, Yao Y. Protein contact map prediction based on ResNet and DenseNet. Biomed Res Int 2020:2020.
[15] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9:432–41.
[16] Korber BT, Farber BM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci 1993-90:7176–80.
[17] Seemayer S, Gruber M, Söding J. CCMpred--fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30:3128–30.
[18] Kaján L, Hopf TA, Kalaf M, Marks D5, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinf 2014;15:1–6.
[19] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolutionary matrices through deep residual networks and singularity enhanced loss function. BioRxiv 2021.
[20] Jones DT, Buchan DW, Cozzetto D, Ponti M. PSCOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–90.
[21] Li Y et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS Comput Biol 2021;17:e1008865.
[22] Si Y, Yan C. Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. bioRxiv 2021.
[23] Pinheiro, P. O. & Collobert, R. in Proceedings of the IEEE conference on computer vision and pattern recognition. 1713–1721.
[24] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 2017;13: e1005324.
[25] He, K., Zhang, X., Ren, S. & Sun, J. in Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[26] Chen C, Wu T, Guo Z, Cheng J. Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. Proteins Struct Funct Bioinf 2021;89:697–707.

[27] Zeng H et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res 2018;46:W432–7.

[28] Creswell A et al. Generative adversarial networks: an overview. IEEE Signal Process Mag 2018;35:53–65.

[29] Goodfellow, I. et al. Generative adversarial nets. Advances in neural information processing systems 27 (2014).

[30] Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. in Proceedings of the IEEE conference on computer vision and pattern recognition. 1125–1134.

[31] Yang H, Wang M, Yu Z, Zhao X-M, Li A. GANcon: Protein contact map prediction with deep generative adversarial network. IEEE Access 2020;8:38089–907.

[32] Subramaniya MV, S.R., Terashi G, Jain A, Kagaya Y, Kihara D. Protein contact map refinement for improving structure prediction using generative adversarial networks. Bioinformatics 2021;37:3368–74.

[33] Adhikari B. DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout. Bioinformatics 2020;36:470–7.

[34] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42:D304–9.

[35] Burley SK et al. Protein Data Bank (PDB): the single global macromolecular structure archive. Protein Crystallography 2017;627–641.

[36] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-MET: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[37] Robin X et al. Continuous Automated Model EvaluatiOn (CAMEO)—Perspectives on the future of fully automated evaluation of structure prediction methods. Proteins Struct Funct Biol 2021;89:1977–86. https://www.predictioncenter.org/.

[38] Zhang C, Zheng W, Mortuza S, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020;36:2105–12.

[39] Duncan TE. On the calculation of mutual information. SIAM J Appl Math 1970;19:215–20.

[40] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–40.

[41] Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–8.

[42] Xie S, Li Z, Hu H. Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. Gene 2018;642:74–83.

[43] Behzadi, M. M. & Ilies, H. T. GANTL: Towards Practical and Real-Time Topology Optimization with Conditional GANs and Transfer Learning. arXiv preprint arXiv:2105.03045 (2021).

[44] Madani, M., Lin, K. & Tarakanova, A. DSResSol: A sequence-based solubility predictor created with Dilated Squeeze Excitation Residual Networks. bioRxiv (2021).

[45] Hu, J., Shen, L. & Sun, G. in Proceedings of the IEEE conference on computer vision and pattern recognition. 7132–7141.

[46] Huang, C., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in Proceedings of the IEEE conference on computer vision and pattern recognition. 4700–4708.

[47] Kunkel G, Madani M, White SJ, Verardi PH, Tarakanova A. Modeling coronavirus spike protein dynamics: implications for immunogenicity and immune escape. Biophys J 2021;120:5592–618.

[48] Ho Y, Wooskey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access 2019;8:4806–13.

[49] Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018;34:3308–15.

[50] Senior AW et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577:706–10.

[51] Jiang W et al. MrpH, a new class of metal-binding adhesin, requires zinc to mediate biofilm formation. PLoS Pathog 2020;16:e1008707.

[52] Park, H., Yoo, Y. & Kwak, N. Mc-gan: Multi-conditional generative adversarial network for image synthesis. arXiv preprint arXiv:1805.01123 (2018).

[53] Rennert M, Biegert A, Hauser A, Söding J. Hflb1ps: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–5.

[54] Madani, M., Lin, K. & Tarakanova, A. DSResSol: A sequence-based solubility predictor created with Dilated Squeeze Excitation Residual Networks. bioRxiv (2021).

[55] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–40.

[56] Zhou F-Q, Waterman-Storer CM, Cohin CS. Focal loss of actin bundles causes microtubule redistribution and growth cone turning. J Cell Biol 2002;157:839–49.