Bioinformatics, 38(7), 2022, 1904–1910
https://doi.org/10.1093/bioinformatics/btac063
Advance Access Publication Date: 4 February 2022
Original Paper

Structural bioinformatics

A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers

Raj S. Roy, Farhan Quadir, Elham Soltanikazemi and Jianlin Cheng *

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on September 18, 2021; revised on January 17, 2022; editorial decision on January 29, 2022; accepted on January 31, 2022

Abstract

Motivation: Deep learning has revolutionized protein tertiary structure prediction recently. The cutting-edge deep learning methods such as AlphaFold can predict high-accuracy tertiary structures for most individual protein chains. However, the accuracy of predicting quaternary structures of protein complexes consisting of multiple chains is still relatively low due to lack of advanced deep learning methods in the field. Because interchain residue–residue contacts can be used as distance restraints to guide quaternary structure modeling, here we develop a deep dilated convolutional residual network method (DRCon) to predict interchain residue–residue contacts in homodimers from residue–residue co-evolutionary signals derived from multiple sequence alignments of monomers, intrachain residue–residue contacts of monomers extracted from true/predicted tertiary structures or predicted by deep learning, and other sequence and structural features.

Results: Tested on three homodimer test datasets (Homo_std dataset, DeepHomo dataset and CASP-CAPRI dataset), the precision of DRCon for top *L*/5 interchain contact predictions (*L*: length of monomer in a homodimer) is 43.46%, 47.10% and 33.50% respectively at 6 Å contact threshold, which is substantially better than DeepHomo and DNCON2_inter and similar to Glinter. Moreover, our experiments demonstrate that using predicted tertiary structure or intrachain contacts of monomers in the unbound state as input, DRCon still performs well, even though its accuracy is lower than using true tertiary structures in the bound state are used as input. Finally, our case study shows that good interchain contact predictions can be used to build high-accuracy quaternary structure models of homodimers.

Availability and implementation: The source code of DRCon is available at https://github.com/jianlin-cheng/DRCon. The datasets are available at https://zenodo.org/record/5998532#.YgF70vXMKsB.

Contact: chengji@missouri.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins fold into three-dimensional (3D) structures to carry out biological functions such as catalyzing chemical reactions and transporting nutrients. The 3D structure of a single protein chain is called tertiary structure. The tertiary structures of multiple protein chains usually interact to form a complex structure (i.e. quaternary structure). Both tertiary structure and quaternary structure are important for protein function. Because the experimental determination of protein structure is low-throughput and can be applied to only a small portion of proteins in the nature, the computational prediction of protein tertiary and quaternary structure is critical for obtaining structural information for most proteins to study their function.

The computational methods for predicting protein tertiary structures and quaternary structures are periodically evaluated in the Critical Assessment of Protein Structure Prediction (CASP) (Kryshtafovych *et al.*, 2014; 2019; Kwon *et al.*, 2021; Moult *et al.*, 2016) and the Critical Assessment of Protein Interaction (CAPRI) (Lensink *et al.*, 2016, 2018, 2021), respectively, or the joint experiment of the two. Driven by the application of deep learning methods to predicting residue–residue contacts and distances (Eickholt and Cheng, 2012; Adhikari *et al.*, 2018; Hou *et al.*, 2020; Jones and Kandathil, 2018; Li *et al.*, 2019; Senior *et al.*, 2020; Wang *et al.*, 2017; Wu *et al.*, 2021; Yang *et al.*, 2020) in the last several years, tertiary structure prediction has reached unprecedented high accuracy. In the 2020 CASP14 experiment, AlphaFold2 (Jumper *et al.*, 2021) predicted high-quality structures for most CASP14 targets with the accuracy equal to or close to that of the experimental structure determination. Recently, AlphaFold2 was applied to predict the structures for all the proteins in several species including human (Tunyasuvunakool *et al.*, 2021).

Despite the drastic advance in protein tertiary structure prediction, the prediction of quaternary structure has progressed slowly and still cannot reach high accuracy for most protein complexes.

One reason is more effort has been put into tertiary structure prediction than quaternary structure prediction because the former is needed as input for the latter. Another reason is the application of deep learning methods to protein quaternary structure prediction is still in the early stage and much fewer deep learning methods for quaternary structure prediction than tertiary structure prediction have been developed.

The most common approach to quaternary structure prediction is classic protein docking algorithms (Gray *et al.*, 2003; Johansson-Åkhe *et al.*, 2020; Li and Kihara, 2012; Lyskov and Gray, 2008; Pierce *et al.*, 2014; Venkatraman *et al.*, 2009), leveraging the geometric and electrostatic complementarity between protein tertiary structures. The residue–residue co-evolutionary methods such as the direct coupling analysis (DCA) (Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014) that were originally designed to predict intrachain residue–residue contacts in a protein chain were also used to predict interchain contacts from multiple sequence alignments (MSAs) of protein complex (e.g. protein heterodimers). The DCA-based methods require a large number of sequences in MSAs to generate accurate interchain contact predictions, which are not available for most protein complexes because there are not many known protein complexes available. The problem is alleviated for protein homodimers (a protein complex consisting of two identical chains) because the MSA of a monomer (a single chain) in a homodimer contains both intrachain and interchain residue–residue co-evolutionary signals (Quadir *et al.*, 2021a,b). The advantage of using the MSA of a monomer is that it is generally much deeper than the MSA of a protein complex. Recently several deep learning methods such as DNCON2_Inter (Quadir *et al.*, 2021a,b) and DeepHomo (Yan and Huang, 2021) use the MSA of a monomer in a homodimer to predict interchain contacts in homodimers, while another method, Glinter (Xie and Xu, 2021), uses the MSA of a dimer to perdict interchain contacts in both homodimers and heterodimers

Another interesting recent development is the application of AlphaFold2 and RoseTTAFold (Baek *et al.*, 2021)—the two cutting-edge deep learning methods designed for prediction of tertiary structure to the prediction of the quaternary structures of several protein complexes, demonstrating the great potentials of deep learning methods for predicting protein quaternary structures. However, because the two methods are not specially designed for quaternary structure prediction and are not trained on the protein complex data, there is a significant need to develop more deep learning methods directly targeting quaternary structure prediction.

In this work, we develop a dilated convolutional residual neural network called DRCon to directly predict interchain contacts in homodimers from the MSA, intrachain contacts and other features of the monomers forming the homodimers. We test our method rigorously on the CASP-CAPRI dataset, DeepHomo test dataset and also on Homo_std test dataset. It performs better than two other deep learning methods (DeepHomo and DNCON2_Inter) for interchain contact prediction. The method works not only with true tertiary structures of monomers in the bound state as input but also predicted tertiary structures of monomers in the unbound state (e.g. tertiary structure models predicted by AlphaFold2). Moreover, we demonstrate that good interchain contact predictions can be used to build high-quality quaternary structures of homodimers.

2 Materials and methods

2.1 Datasets

Two residues from the two chains in a homodimer are considered an interchain contact if the Euclidean distance between any two heavy atoms of the two residues is less than or equal to 6 Å (Ovchinnikov *et al.*, 2014; Quadir *et al.*, 2021a,b; Zhou *et al.*, 2018). Multiple homodimer datasets with known quaternary structures and interchain contacts are used to develop DRCon. The Homo_std dataset used in DNCON2_Inter is used to train, validate and test DRCon. Homo_std was derived from the homodimers in the 3D Complex database (Levy *et al.*, 2006). All the complexes of the 3D Complex were released before October of 2005. The dimers in the database whose two chains have ≥95% sequence identity are treated as homodimers to create Homo_std. Homo_std has 8530 homodimers in total that has ≤30% pairwise sequence identity. It is split into a training dataset (5975 dimers), a validation dataset (853 dimers) and a test dataset (1702 dimers) according to the ratio of 7:1:2 to train, validate and test DRCon. Furthermore, in addition to the 6 Å threshold, 8 Å is also used as a threshold to generate inter-chain contacts to train a network in order to compare it with two inter-chain contact predictors [DeepHomo and Glinter (Xie and Xu, 2021)] trained at the threshold.

Moreover, two independent datasets (the CASP-CAPRI dataset and DeepHomo dataset) are used to test DRCon. The CASP-CAPRI dataset contains 40 homodimers collected from CASP-CAPRI-11, 12, 13 and 14 experiments that are publicly available. We discarded homodimers whose monomer has more than 500 residues in order to make a fair comparison with Glinter as it has a limitation of about 1024 residues for the combined length of the two monomers in a dimer.

The DeepHomo dataset used here contains 218 homodimers out of the 300 homodimers in its original version (Yan and Huang, 2021). 82 homodimers in the original DeepHomo dataset that are present in the Homo_std training dataset are removed to avoid the evaluation bias.

The statistics of the number of the dimers, the length of the dimers (i.e. the length of the monomer sequence in a homodimer) and the contact density of the dimers (i.e. the number of true interchain contacts divided by the length of the monomer in a homodimer) of the three test datasets above is reported in Table 1.

2.2 Input features

The input features for DRCon are stored in $L\times L\times d$ tensors (L: length of the sequence of the monomer in a homodimer; d is the number of features for each pair of interchain residues) that describe the features of all pairs of interchain residues. Since the two chains in a homodimer are identical and interchain residue–residue coevolution features are also preserved in the multiple sequence alignment (MSA) of one chain (monomer), only the sequence of a monomer is used to generate the input features for interchain contact prediction in this work.

The number of features (d) for each interchain residue pair is 592. 49 features are the same kind of features used by DNCON2 (Adhikari *et al.*, 2018) for intrachain contact prediction, including solvent accessibility of residues as well as interchain residue–residue coevolution features calculated from MSAs of a monomer by CCMpred (Seemayer *et al.*, 2014) and PSICOV(Jones *et al.*, 2012). 526 features generated from MSAs by trRosetta (Yang *et al.*, 2020) are also used. The 8-state secondary structure prediction for each residue (i.e. 16 features for a pair of residues) made by SCRATCH (Cheng *et al.*, 2005) is also included. Finally, a binary feature indicating if two residues form an intrachain contact (i.e. Cb–Cb atom distance is less than or equal to 8 Å (Adhikari *et al.*, 2015; Wu *et al.*, 2021) is also used as input, which is useful for the neural network to distinguish interchain contacts from intrachain contacts. It worth noting that all the features except secondary structure, solvent accessibility, intrachain contact and CCMpred features used in DRCon are different from DeepHomo. In the training phase, the intrachain contacts are derived from the true tertiary structures of monomers in the dimers. In the test phase, the intrachain contacts may be either derived from true tertiary structures of monomers in the bound state or predicted from sequences/tertiary structure models of monomers in the unbound state, depending on the experimental setting. Specifically, for the training and validation datasets, the intrachain contacts are derived from the known tertiary structures of the monomers in the homodimers (the bound state). For the test datasets, either true intrachain contacts or predicted intrachain contacts made by trRosetta or extracted from AlphaFold2 tertiary structure models in the unbound state are used to generate intrachain contact features.

Most of the 592 features above are generated from the MSAs of the monomers in the homodimers. The DNCON2's MSA generation procedure is used to generate the MSAs for all the datasets by using

Table 1. The statistics of the Homo_std test dataset, DeepHomo test dataset and CASP-CAPRI test dataset

Name	Number of dimers	Range of length	Average length	Range of contact density	Average contact density
Homo_std test dataset	1702	30 to 600	254.94	0.003 to 4.54	0.67
DeepHomo	218	48 to 498	235.9	0.210 to 4.5	1.06
CASP-CAPRI	40	73 to 480	248.97	0.346 to 4.96	2.04

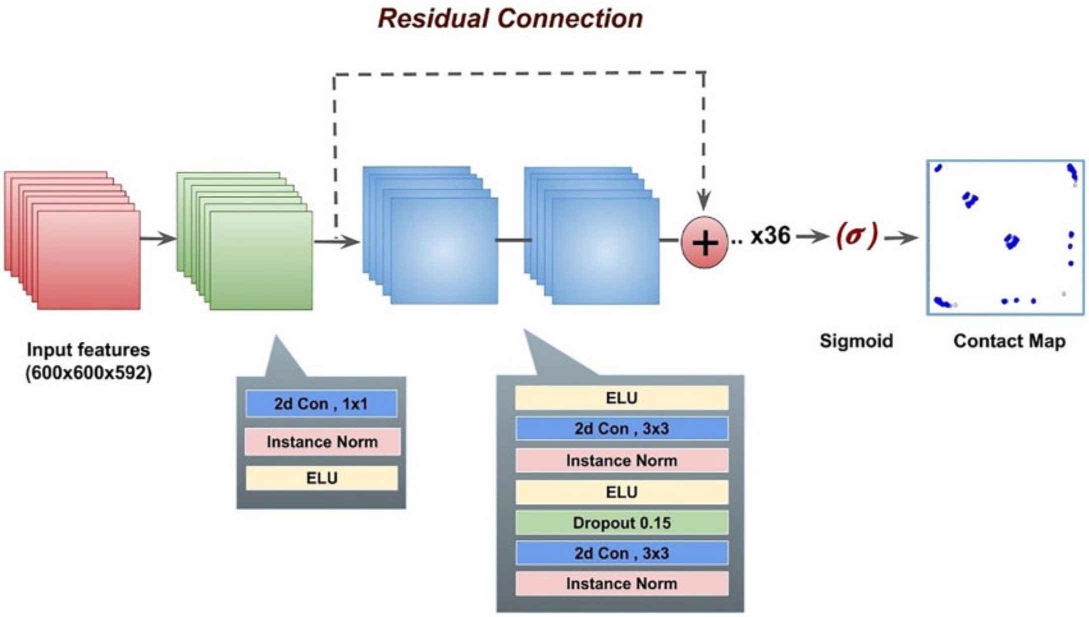

Fig. 1. The deep learning architecture of DRCon for interchain contact prediction in homodimers. For a homodimer in which the length of the monomer sequence is L, the input is a $L \times L \times 592$ tensor. The number of input features for each pair of residues is 592. For convenience, L is set to a fixed number—600. 0 padding is applied if L is less than 600. It is worth noting that in the prediction phase, no zero padding is used in generating the input tensor if L is greater than 600. The input is transformed to a $600 \times 600 \times 48$ tensor using a 2D-convolutional layer which has a kernel size of 1 and uses Exponential Linear Unit (elu). The output of the convolution layer is passed through 36 residual blocks with kernel size of 3x3. Each residual block uses a 2D-convolution layer with a kernel size of 3, instance normalization and dropout of 15% probability of a neuron being ignored, followed by a dilated convolution layer without dropout. The step of the dilation in the dilated convolution layers in these blocks changes from 1, 2, 4, 8, 16 periodically. The sigmoid activation function is applied to the output of the last residual block to calculate the contact probability of each interchain residue–residue pair. The probabilities for residue pair (i, j) and residue pair (j, i) are averaged to a symmetric final contact map

HHBlits (Remmert *et al.*, 2011) to search UniRef30_2020_02 database (Suzek et al., 2015) and Jackhmmer (Johnson *et al.*, 2010) to search Uniref90. In addition, DeepMSA (Zhang *et al.*, 2020) is used to generate MSAs for the CASP-CAPRI dataset. The MSAs with more sequences generated by DNCON2 or DeepMSA are selected for the proteins in this dataset.

2.3 Deep learning architecture for interchain contact prediction

Figure 1 illustrates the deep learning architecture for interchain contact prediction. The input tensor ($L \times L \times 592$) is first transformed by a block consisting of a convolutional layer and instance normalization. The instance normalization instead of the batch normalization is used because the former is better at dealing with a small batch size (Lian and Liu, 2019). The transformed tensor is then processed by 36 residual blocks containing regular convolutional layers, instance normalization, dilated convolutional layers and residual connections. The residual connection makes the learning of deep networks more efficient and effective. The dilated convolution can capture a larger input area than the regular convolution with the same number of parameters, which has been shown to improve intrachain residue–residue distance prediction in AlphaFold1 (Senior *et al.*, 2019). This dilated residual architecture is different from that in DeepHomo.

The network is trained on the Homo_std training dataset with 0.0001 learning rate and optimized with Adam (Kingma and Ba, 2015) optimizer using a batch size of 2 and the binary cross entropy as loss function. Each epoch of training the network on six 32 GB NVIDIA V100 GPUs takes around 2 h. The deep network is implemented on Pytorch and horovod (Sergeev and Del Balso, 2018) to leverage the distributed deep learning training. The deep learning model with the highest precision for top $L/5$ interchain contact predictions on the Homo_std validation dataset is selected as the final model for testing.

3 Results and discussions

DRCon has been extensively benchmarked on three datasets: Homo_std test dataset, DeepHomo test dataset and CASP14-CAPRI dataset. The contact-level precision and the target-level accuracy rate at the various thresholds (i.e. Top 10, top $L/10$, top $L/5$, top L interchain contact predictions) are used to compare DRCon with existing methods, where L is the length of the monomer sequence in a homodimer. The contact-level precision is the number of correctly predicted contacts divided by the total number of contact predictions. And the target-level accuracy rate (Zhao and Gong, 2019) is defined as the percentage of dimers (targets) with non-zero correct interchain contact prediction when a certain number of predicted interchain contacts are evaluated.

3.1 Evaluation on Homo_std test dataset

We compare DRCon with DNCON2_Inter on the Homo_std test dataset. DRCon is run in the three settings. In one setting, the true intrachain contacts extracted from known tertiary structures of a monomer in each homodimer are used as input. In another setting, the intrachain contacts predicted by trRosetta are used as input. Predicted intrachain contacts are converted from the distance probabilities predicted by trRosetta. A cutoff probability of 0.5 is applied to make the conversion. The precision of top *L* and top 2*L* intrachain contact predictions made by trRosetta is 86% and 78%, respectively, indicating the quality of the intrachain contact prediction is good. Lastly, the intrachain contacts extracted from Alphafold2 predicted tertiary structures which had an average TM-score of 0.948 were used as input.

The precision of the interchain contact prediction on the Homo_std test dataset is reported in Table 2. The precision of DRCon in all the settings is more than twice that of DNCON2_Inter in most cases. For instance, the precision of DRCon with intrachain contact prediction made by trRosetta as input for top *L*/10 interchain contact prediction is 37.25%, higher than 17.32% of DNCON2_Inter. The difference is largely because DRCon is specially designed and trained to predict interchain contacts, but DNCON2_Inter is adapted from a deep learning method designed and trained to predict intrachain contacts. It is worth noting that the interchain contact prediction accuracy of using AlphaFold predicted tertiary structure as input (DRCon_alpha) is very close to that of using true tertiary structures as input (DRCon_true), indicating that tertiary structures predicted by AlphaFold are sufficiently accurate for interchain contact prediction.

In contrast, the precision of DRCon with trRosetta predicted intrachain contacts (DRCon_pre) as input is worse than that of DRCon with true contacts by about 6 to 11 percentage points for Top 10, Top *L*/10, Top *L*/5 and Top *L* interchain contact predictions, indicating that more precise intrachain contact prediction (or tertiary structure prediction) of monomer leads to the higher accuracy of the interchain contact prediction.

Because the predicted intrachain contacts represent the tertiary structures of monomers in the unbound state (i.e. in the free state without a binding partner) while the true intrachain contacts represent the tertiary structures in the bound state (i.e. in the state of binding with a partner in complex), the reasonable performance of DRCon_pre and DRCon_alpha shows that DRCon trained on the dimers and the true tertiary structures of monomers in the bound state can work well on the predicted input intrachain contacts (or predicted tertiary structures) in the unbound state. The similar trend is also observed in the target-level prediction accuracy rate on the dataset (Table 3).

Table 2. The interchain contact prediction precision of DNCON2_Inter, the DRCon with true intrachain contacts as input (DRCon_true), DRCon with AlphaFold2 predicted tertiary structure's intrachain contacts as input (DRCon_alpha) and DRCon with trRosetta predicted intrachain contacts as input (DRCon_pre) on Homo_std test set

Predictor	Top10 (%)	Top *L*/10 (%)	Top *L*/5 (%)	Top *L* (%)
DNCON2_Inter	16.9	17.32	16.31	13.69
DRCon_pre	40.20	37.25	33.75	18.92
DRCon_alpha	49.71	46.21	42.12	24.04
DRCon_true	50.61	47.21	43.46	25.05

Note: The precision of DNCON2_Inter is reported with its best parameter setting (relax_removal = 2).

L, length of a monomer in a dimer.

Table 3. Target-level accuracy rate of DNCON2_Inter, DRCon_pre, DRCon_alpha and DRCon on the Homo_std test dataset

Predictor	Top 10 (%)	Top *L*/10 (%)	Top *L*/5 (%)	Top *L* (%)
DNCON2_Inter	23.52	30.19	37.50	43.36
DRCon_pre	58.28	63.40	67.74	73.85
DRCon_alpha	67.09	70.38	74.14	83.31
DRCon_true	67.39	70.86	75.56	80.38

3.2 Evaluation on DeepHomo test dataset

We compare DRCon, DNCON2_Inter and DeepHomo on the DeepHomo test dataset (see the contact-level precision and target-level accuracy rate in Tables 4 and 5, respectively). For a fair comparison, we use the same tertiary structures of monomers in the homodimers provided by the DeepHomo server to extract the intrachain contacts as input for DRCon and for the DeepHomo server itself to make interchain contact predictions. The interchain contacts predicted by DeepHomo consist of only the upper triangle of the interchain contact map. They are converted to a diagonally symmetric full contact map for evaluation as DeepHomo assumes the contact map is of C2-symmetry. DRCon performs better than DeepHomo in terms of contact-level precision and target-level accuracy rate at all the thresholds except for the target-level accuracy rate of top *L* contact predictions. For instance, the contact-level precision and target-level accuracy of DRCon for top *L*/10 interchain contact prediction is 50.17% and 76.15%, higher than 38.74% and 70.77% of DeepHomo. The result of this experiment with 8 Å contact threshold is presented in Supplementary Tables S1 and S2 in Supplementary Material. However, it is worth pointing out that the sequence redundancy between the DeepHomo test dataset and the training data of DRCon is not filtered by the 30% sequence identity threshold, which may lead to an overestimate the performance of DRCon.

3.3 Evaluation on CASP-CAPRI dataset using true or predicted tertiary structures as input

We compare DRCon, DeepHomo, Glinter and DNCON2_inter on the CASP-CAPRI dataset. Only the contact-level precision is used to evaluate them because only 40 targets are not sufficient to reliably estimate the target-level accuracy rate. DRCon is run in the two settings (the ideal setting and the realistic setting). In the ideal setting (DRCon_true), the known tertiary structures of the monomers in the homodimers are used to generate the true intrachain contacts as input for DRCon. In the realistic setting (DRCon_alpha), the tertiary structures of the monomers predicted by AlphaFold2 (Jumper *et al.*, 2021) are used to generate the interchain contacts for DRCon. The AlphaFold2 model with the highest confidence is used for each target. The average TM-scores (Zhang and Skolnick, 2004) of the tertiary structures for the 40 targets predicted by AlphaFold2 is 0.931.

The precision of interchain contact predictions of the four methods are shown in Table 6. The precision of both DRCon_true and DRCon_alpha is substantially higher than that of DeepHomo and DNCON2_Inter at all the thresholds. For instance, for top *L*/10 interchain contact predictions, the precision of DRCon_true and DRCon_alpha is 38.65% and 36.88% in comparison with 23.88% of DeepHomo and 2.36% for DNCON2_Inter. DRCon_true performs better than DRCon_alpha, indicating that more accurate

Table 4. The interchain contact prediction precision of DRCon and DeepHomo, DNCON2_Inter on the DeepHomo test dataset

Predictor	Top 10 (%)	Top *L*/10 (%)	Top *L*/5 (%)	Top *L* (%)
DRCon	**53.66**	**50.17**	**47.10**	**27.81**
DNCON2_Inter	7.43	7.59	7.95	7.67
DeepHomo	43.80	38.74	34.10	21.35

Bold fold is used to denote the highest precision in terms of each metric.

Table 5. The target-level accuracy rate of DRCon, DNCON2_Inter and DeepHomo on the DeepHomo test dataset

Predictor	Top 10 (%)	Top *L*/10 (%)	Top *L*/5 (%)	Top *L* (%)
DRCon	**72.95**	**76.15**	**80.73**	86.69
DNCON2_Inter	22.01	28.44	34.40	64.22
DeepHomo	66.67	70.77	77.17	**87.61**

Bold fold is used to denote the highest precision in terms of each metric.

Table 6. The precision of DRCon, DeepHomo, Glinter and DNCON2_inter on the CASP-CAPRI test dataset

Predictor	Top 10 (%)	Top *L*/10 (%)	Top *L*/5 (%)
DRCon_true	**41.75**	**38.65**	**33.50**
DRCon_alpha	40.25	36.88	31.67
DeepHomo	26.0	23.88	20.27
Glinter	37.5	35.16	30.61
DNCON2_Inter	2.36	2.33	2.39

Note: DRCon_true and DeepHomo use the true tertiary structures of monomers in the bound state to extract intrachain contacts as input. DRCon_alpha uses the tertiary structures predicted by AlphaFold2 in the unbound state to extract intrachain contacts as input. Bold fold is used to denote the highest precision in terms of each metric.

intrachain contact input leads to better interchain contact prediction. However, because the quality of the AlphaFold predicted tertiary structure is very high, the difference between DRCon_true and DRCon_alpha is small. In addition, both DRCon and DrCon_alpha performs better than Glinter by up to a few percentage points. The result of this experiment with 8 Å contact threshold is presented in Supplementary Table S3. At 8 Å contact threshold, the precision of top *L*/10 and *L*/5 contact predictions of Glinter is slightly higher than DRCon_true, but its precision of top 10 contact predictions is slightly lower.

3.4 Effect of contact density on interchain contact prediction

We investigate the interchain contact density in a dimer with the precision of the interchain contact prediction on the Homo_std test dataset. The Pearson's correlation coefficient between the precision of top *L*/5 interchain contacts and contact density is 0.4211, indicating a moderate correlation between the two. The lowest average precision (a little over 4%) is recorded for targets with the low contact density between 0 and 0.25, indicating that when the interchain contact map is very sparse, the prediction is generally difficult. According to Figure 2, there is an uptrend of the average contact prediction precision with the increasing contact density until the density of 1.5.

3.5 Impact of sequence similarity on interchain contact prediction

A common threshold—30% sequence identity is used to remove the redundancy between the Homo_std test dataset and the training dataset. However, low sequence identity between two sequences does not always mean they do not significant similarity/homology. To investigate how sequence similarity may influence the accuracy of interchain contact prediction, we build a hidden Markov model profile database for the sequences in the training dataset from their multiple sequence alignments using Hhsuite, and then use hhsearch in Hhsuite to search each sequence in Homo_std test dataset, DeepHomo test dataset and CASP-CAPRI dataset against the profile database. The ranges of e-values of the top hits for the test sequences measuring the similarity between the test sequences and the training data for the three test datasets are reported in Supplementary Table S4. All the test datasets including the CASP-CAPRI dataset whose proteins were released in the Protein Data Bank (PDB) much later than the proteins in the training dataset and were carefully selected by CASP organizers to rigorously test protein structure prediction methods contain some proteins having significant similarity with the training data. On all the three test datasets, the average top *L*/5 precision of interchain contact prediction for 6 and 8 Å thresholds largely increases as the sequence similarity increases between test sequence and the training data (i.e. e-value decreases). The similar trends are observed with Glinter and DeepHomo (Supplementary Table S5) on the CASP-CAPRI test dataset.

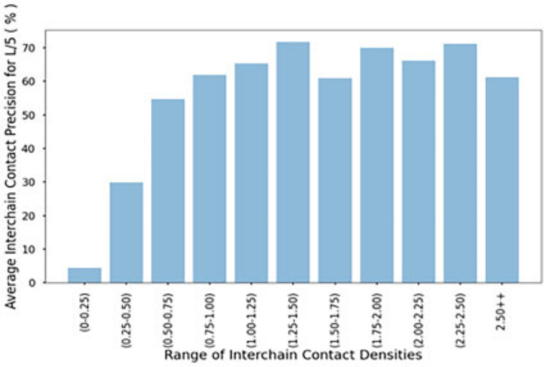

Fig. 2. Illustrating the effect of contact density on interchain contact prediction precision on the Homo_std test dataset

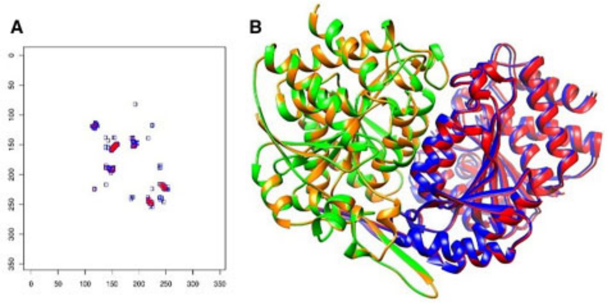

Fig. 3. (**A**) The predicted and true contact maps of target 1DR0. The top *L*/5 predicted contacts (red dots) and true contacts (blue dots) are plotted. Most predicted contacts overlap with the true contacts, indicating a high contact prediction precision. (**B**) The superimposition of the true quaternary structure (chain A in red and chain B in green) and the predicted quaternary structure (chain A in blue and chain B in orange). The two quaternary structures are quite similar

Moreover, we select all the test sequences in the CASP-CAPRI dataset that has e-value > 0.1 (no similarity) with all the sequences in the training sets of DRCon, Glinter and DeepHomo to compare their performance on the common de novo targets. Nine such test sequences are obtained. The precision of top *L*/5 interchain contact prediction at 8 Å threshold is reported in Supplementary Table S6. Glinter performs better than DRCon, while DRCon performs better than DeepHomo and DNCON_Inter.

3.6 A case study of applying interchain contact prediction to build quaternary structure

Figure 3 visualizes the top *L*/5 interchain contact predictions for a target (PDB code: 1DR0) from the Homo_std test dataset and the quaternary structure reconstructed from the interchain contacts predicted by DRcon and the known tertiary structure of a chain in the dimer. The quaternary structure is built by GD (Soltanikazemi *et al.*, 2022), which applies the gradient descent optimization to build quaternary structures by using interchain contacts as distance restraints.

It is shown in Figure 3A that most of the interchain contact predictions overlap with the true interchain contacts, indicating a high prediction precision. Indeed, the precision of top *L*/5 and top *L* contact predictions is 100% and 75%, respectively. The quaternary structure reconstructed from the predicted interchain contacts is also very similar to the native structure (Fig. 3B). The TM-score of the predicted quaternary structure in comparison with the true quaternary structure is 0.99. TMalign (Zhang and Skolnick, 2005) is used to calculate the TM-score. The predicted quaternary structure has a fraction of the native contacts (F_{nat}) of 0.88, interface RMSD (iRMS: root mean square displacement of inter-protein heavy atoms that are within 10 Å) of 0.3 Å, ligand RMS (LRMS) of 0.83 Å and a DockQ score of 0.95. F_{nat}, iRMS, LRMS and DockQ score of the

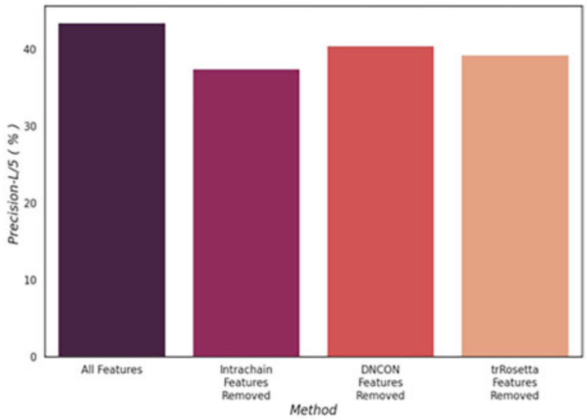

Fig. 4. Impact of different groups of features on the average top *L*/5 precision on the Homo_std test dataset

predicted quaternary structure are calculated against the true quaternary structure by DockQ (Basu and Wallner, 2016). A DockQ score of 0.8 indicates a high-quality quaternary structure prediction. The scores (e.g. TM-score = 0.99) of the structures reconstructed by GD with interchain contact predictions are much better than the scores (e.g. TM-score = 0.55) of the structures reconstructed by GD without contact predictions (see Supplementary Table S7 for details), indicating that the interchain contact prediction plays a critical role in reconstructing correct quaternary structures.

3.7 Ablation studies

The DRCon utilizes a variety of features to make interchain contact prediction. Collectively the features can be grouped into 3 divisions, i.e. the DNCON2, trRosetta and intrachain features. Using these groups, we conducted an ablation study to find out their impact on the precision of DRCon. The findings of the investigation are illustrated on Figure 4. In each run we leave out one group of features and then compare the prediction precision with the DRCon using all the features. The results show that removing any group of the features decreases the performance. Among the 3 groups of features DNCON2 feature group appear the least impactful. After removing it, the average *L*/5 precision drops from 43.46% to 40.44%. TrRoseetta feature group is the second most significant as its absence causes the precision to be dropped to 39.27% from 43.46. Finally, the intrachain feature group is the most impactful feature group. Leaving it out causes the average accuracy to sharply fall to 37.47%. Even though each group of features has different impacts, combining them together works best.

4 Conclusion and future work

In this work, we develop a deep network (DRCon) consisting of residual connections, regular and dilated convolutions and instance normalizations to predict interchain homodimers from sequence and structural features of monomers in homodimers. DRCon trained on known homodimer structures can predict interchain contacts well. Moreover, DRCon is robust against the errors in input tertiary structures or intrachain contacts of monomers. It maintains the reasonable prediction precision when predicted tertiary structures of monomers in the unbound state instead of true tertiary structures in the bound state are used as input. The work demonstrates that deep learning methods specially designed for interchain contact prediction can be trained on known homodimer structures to substantially improve the prediction of interchain residue–residue contacts as what had happened in protein tertiary structure prediction. In the future, we plan to further improve the deep learning architecture, input features and training strategies to improve interchain contact prediction. We will generalize the method to predict interchain residue–residue distances. Moreover, we plan to develop similar methods to predict interchain contacts and distances in heterodimers and generalize them to multimers consisting of more than two chains.

Funding

The work was partly supported by the Department of Energy, USA [DE-AR0001213, DE-SC0020400 and DE-SC0021303], National Science Foundation [DBI1759934 and IIS1763246], National Institutes of Health [R01GM093123] and the Thompson Missouri Distinguished Professorship. The project leveraged the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

Conflict of Interest: none declared.

References

Adhikari,B. *et al.* (2015) CONFOLD: residue–residue contact-guided ab initio protein folding. *Proteins*, **83**, 1436–1449.

Adhikari,B. *et al.* (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics (Oxford, England)*, **34**, 1466–1472.

Baek,M. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 876.

Basu,S. and Wallner,B. (2016) DockQ: a quality measure for protein-protein docking models. *PLoS One*, **11**, e0161879.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Eickholt,J., and Cheng,J. (2012) Predicting protein residue-residue contacts using deep networks and boosting. Bioinformatics (Oxford, England), **28**, 3066–3072.

Gray,J.J. *et al.* (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Hopf,T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *ELife*, **3**, e03430.

Hou,J. *et al.* (2020) The MULTICOM protein structure prediction server empowered by deep learning and contact distance prediction. *Methods Mol. Biol. (Clifton, N.J.)*, **2165**, 13–26.

Johansson-Åkhe,I. *et al.* (2020) InterPep2: global peptide–protein docking using interaction surface templates. *Bioinformatics*, **36**, 2458–2465.

Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Jones,D.T. and Kandathil,S.M. (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics (Oxford, England)*, **34**, 3308–3315.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–511.

Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015).

Kryshtafovych,A. *et al.* (2014) CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins Struct. Funct. Bioinf.*, **82**, 7–13.

Kryshtafovych,A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinf.*, **87**, 1011–1020.

Kwon,S. *et al.* (2021) Assessment of protein model structure accuracy estimation in CASP14: old and new challenges. *Proteins Struct. Funct. Bioinf.*, **89**, 1940–1948.

Lensink,M.F. *et al.* (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins*, **84**, 323–348.

Lensink,M.F. *et al.* (2018) The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins Struct. Funct. Bioinf.*, **86**, 257–273.

Lensink,M.F. *et al.* (2021) Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. *Proteins Struct. Funct. Bioinf.*, **89**, 1800–1823.

Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.

Li,B. and Kihara,D. (2012) Protein docking prediction using predicted protein–protein interface. *BMC Bioinformatics*, **13**, 7.

Li,Y. *et al.* (2019) ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics (Oxford, England)*, **35**, 4647–4655.

Lian,X. and Liu,J. (2019) Revisit batch normalization: new understanding and refinement via composition optimization. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3254–3263. PMLR, 2019.

Lyskov,S. and Gray,J.J. (2008) The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.*, **36**, W233–W238.

Moult,J. *et al.* (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins*, **84**, 4–14.

Ovchinnikov,S. *et al.* (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *ELife*, **3**, e02030.

Pierce,B.G. *et al.* (2014) ZDOCK server: interactive docking prediction of protein—protein complexes and symmetric multimers. *Bioinformatics*, **30**, 1771–1773.

Quadir,F. *et al.* (2021a) DNCON2_Inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning. *Sci. Rep.*, **11**, 12295.

Quadir,F. *et al.* (2021b) DeepComplex: a web server of predicting protein complex structures by deep learning inter-chain contact prediction and distance-based modelling. *Front. Mol. Biosci.*, **8**, 716973.

Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Seemayer,S. *et al.* (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.

Senior,A.W. *et al.* (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinf.*, **87**, 1141–1148.

Senior,A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.

Sergeev,A. and Del Balso,M. (2018) Horovod: fast and easy distributed deep learning in TensorFlow. ArXiv, *1802.05799 [Cs, Stat]*. https://github.com/horovod/horovod.

Soltanikazemi,E. *et al.* (2022) Distance-based reconstruction of protein quaternary structures from inter-chain contacts. Proteins, **90**(3), 720–731. https://doi.org/10.1002/prot.26269

Suzek,B.E. *et al.*; The UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Tunyasuvunakool,K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.

Venkatraman,V. *et al.* (2009) Protein–protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*, **10**, 407.

Wang,S. *et al.* (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.

Wu,T. *et al.* (2021) DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics*, **22**, 30.

Xie,Z. and Xu,J. (2021) Deep graph learning of inter-protein contacts. *Bioinformatics*, **38**(4), pp. 947–953, https://doi.org/10.1093/bioinformatics/btab761.

Yan,Y. and Huang,S.-Y. (2021) Accurate prediction of inter-protein residue–-residue contacts for homo-oligomeric protein complexes. *Brief Bioinform.*, **22**, bbab038.

Yang,J. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA*, **117**, 1496–1503.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinf*. doi: 10.1002/prot.20264.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**(7), pp. 2302–2309, https://doi.org/10.1093/nar/gki524.

Zhang,C. *et al.* (2020) DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, **36**, 2105–2112.

Zhao,Z. and Gong,X. (2019) Protein–protein interaction interface residue pair prediction based on deep learning architecture. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16**, 1753–1759.

Zhou,T. *et al.* (2018) Deep learning reveals many more inter-protein residue–-residue contacts than direct coupling analysis. *BioRxiv*, 22nd International Conference on Research in Computational Molecular Biology, RECOMB 2018, 295–296. https://doi.org/10.1101/240754.