Structural bioinformatics

Accurate self-assessment in 2D: protein contact map quality estimation by deep evolutionary reconciliation

Rahmatullah Roche[1], Sutanu Bhattacharya[2], Md. Hossain Shuvo[1], and Debswapna Bhattacharya[1,*]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA, [2]Department of Computer Science, Florida Polytechnic University, Lakeland, FL 33805, USA.

Abstract

Motivation: Protein contact maps have proven to be a valuable tool in the deep learning revolution of protein structure prediction, ushering in the recent breakthrough by AlphaFold2. However, self-assessment of the quality of predicted structures is typically performed at the granularity of 3D coordinates as opposed to directly exploiting the rotation- and translation-invariant 2D contact maps.

Results: Here we present rrQNet, a deep learning method for self-assessment in 2D by contact map quality estimation. Our approach is based on the intuition that for a contact map to be of high quality, the residue pairs predicted to be in contact should be mutually consistent with the evolutionary context of the protein. The deep neural network architecture of rrQNet implements this intuition by cascading two deep modules—one encoding the evolutionary context and the other performing evolutionary reconciliation. The penultimate stage of rrQNet estimates the quality scores at the interacting residue-pair level, which are then aggregated for estimating the quality of a contact map. This design choice offers versatility at varied resolutions from individual residue pairs to full-fledged contact maps. Trained on multiple complementary sources of contact predictors, rrQNet facilitates generalizability across various contact maps. By rigorously testing using publicly available datasets and comparing against several in-house baseline approaches, we show that rrQNet accurately reproduces the true quality score of a predicted contact map and successfully distinguishes between accurate and inaccurate contact maps predicted by a wide variety of contact predictors.

Availability: The open-source rrQNet software package is freely available at https://github.com/Bhattacharya-Lab/rrQNet.

Contact: dbhattacharya@vt.edu

1 Introduction

Deep neural networks have led to a paradigm shift in protein structure prediction (Xu, 2019; Senior *et al.*, 2020; Yang *et al.*, 2020; Baek *et al.*, 2021; Jumper *et al.*, 2021). Most notably, the AlphaFold2 method from DeepMind (Jumper *et al.*, 2021) has demonstrated unprecedented performance level in the 14th edition of the Critical Assessment of Structure Prediction (CASP14) experiment (Pereira *et al.*, 14) through successful application of deep neural networks, paving the way towards highly accurate prediction of protein structural models at proteome-wide scale (Tunyasuvunakool *et al.*, 2021). The core of AlphaFold2 consists of a neural embedding of the evolutionary profile coupled with pairwise relations of the various amino acid residues in the protein. This embedding is used to predict the inter-residue interactions between the amino acids by a collection of deep neural network modules trained end-to-end, leading to the final predicted three-dimensional (3D) structure. Inter-residue interactions can be represented as a binary two-dimensional (2D) matrix, also known as a contact map, which distinguishes interacting residue pairs from the non-interacting ones using a predefined distance threshold such as 8Å. A contact map has the advantage of being invariant to rotations and translations, while providing a 2D representation of a protein 3D structure. Intuitively, the choice of the distance threshold affects the granularity of a contact map, and the use of multiple distance thresholds results in finer-grained binned distance distributions. Regard-

less of their resolutions, contact maps encode inter-residue spatial proximity information that can be turned into Cartesian coordinates, leading to the predicted structure (Roche *et al.*, 2021). As such, a 2D contact map can be considered as a transitional state in the mapping from the 1D amino acid sequence to the 3D structure of a protein molecule.

In order for protein structure prediction to be practically useful to biologists, a reliable self-assessment of the prediction is critically important (Shuvo *et al.*, 2020; Alapati and Bhattacharya, 2018). Such a calibrated self-assessment of the quality of the prediction is an integral part of the AlphaFold2 system, which provides intrinsic model accuracy estimates in the form of predicted local-distance difference test (pLDDT) and predicted global superposition metric template modeling score (pTM). However, self-assessment of the quality of structure prediction using the state-of-the-art protein structure prediction systems typically rely on 3D coordinates as opposed to exploiting the residue pair representation captured by a 2D contact map. That is, the granularity of protein quality estimation is typically a predicted 3D structure rather than a 2D contact map.

Performing self-assessment at the granularity of 2D contact maps instead of 3D structures has several advantages. First, being invariant to rotations and translations, contact maps and their estimated confidence scores can be seamlessly integrated into geometric attention-based deep learning models while implicitly ensuring invariance (Jumper *et al.*, 2021). Second, the estimated quality score of a 2D contact map can be integrated into the loss function of an end-to-end structure prediction system. Third, such an integration enables error feedback (Carreira *et al.*, 2016) to propagate through the deep network for iterative optimization of the contact maps, similar to the recycling updates performed in AlphaFold2. Finally, a reliable quality estimate of a predicted contact map can inform contact-driven homology detection and reconstruction approaches (Roche *et al.*, 2021; Marks *et al.*, 2011; Bhattacharya *et al.*; Adhikari *et al.*, 2015; Bhattacharya and Bhattacharya, 2020; Bhattacharya *et al.*, 2021; Bhattacharya and Bhattacharya, 2019; Zheng *et al.*, 2019; Ovchinnikov *et al.*, 2017).

How can we get the representational benefits of self-assessment of the prediction at the 2D contact map level, without necessarily turning the contact map into 3D coordinates? Here we explore a potential solution of estimating the contact map quality by *deep evolutionary reconciliation*. Our approach is based on the intuition that for a contact map to be of high quality, the residue pairs predicted to be in contact should be mutually consistent with the evolutionary context encoded in the multiple sequence alignment of the target protein sequence. The deep neural network architecture of rrQNet directly implements this intuition by cascading two deep modules—one encoding the evolutionary context and the other performing evolutionary reconciliation. The penultimate stage in rrQNet is a rough estimate of the predicted confidence score at the level of a pair of interacting residue pairs, which is subsequently aggregated for scoring a predicted contact map. We rigorously test rrQNet on publicly available datasets by comparing against several baseline approaches and conducting pseudo-blind assessment using protein targets from the latest rounds of CASP experiments for several hundred predicted contact maps generated by a wide variety of contact predictors. The experimental results show that our method can accurately reproduce the true quality score of a predicted contact map and can successfully distinguish between accurate and inaccurate contact prediction. Our generalizable deep learning model is versatile for confidence estimation at varying granularities from individual residue-pair to full-fledged contact map. The open-source rrQNet software package is freely available at https://github.com/Bhattacharya-Lab/rrQNet.

2 Methods

2.1 Model Architecture

As shown in **Figure 1**, our deep learning model for protein contact map quality estimation consists of two major modules, each being a residual neural network (ResNet) (He *et al.*, 2016).

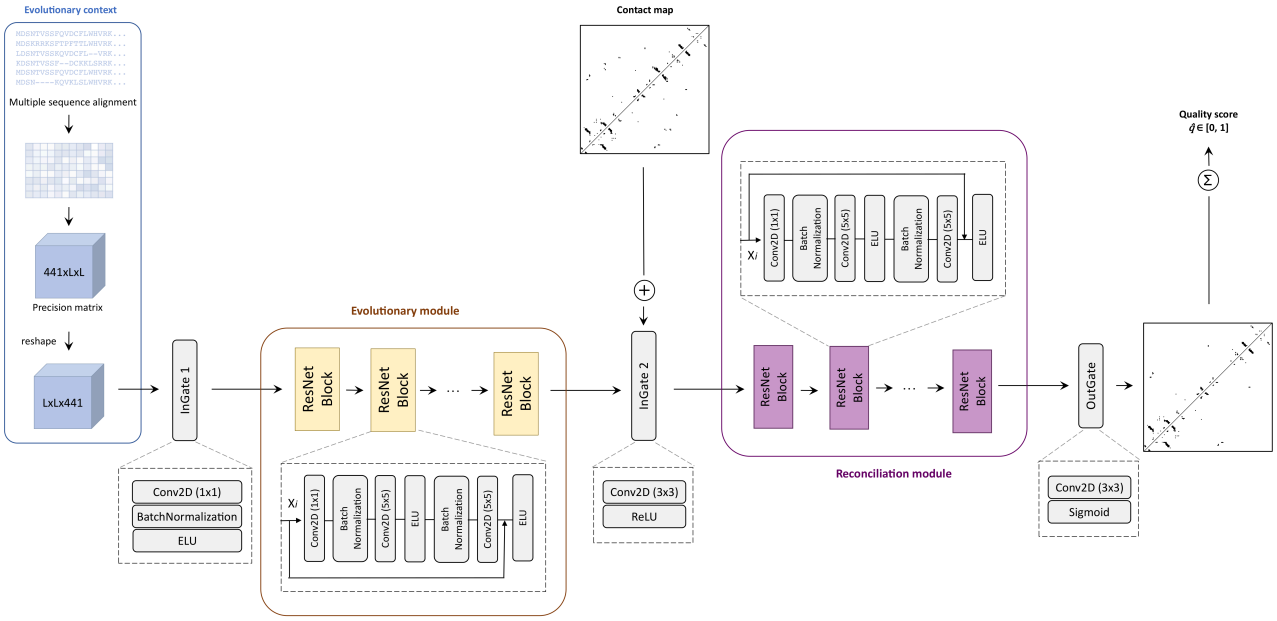

Figure 1. rrQNet architecture overview. *Left to right*: The evolutionary context extracted from the covariational signal encoded in the multiple sequence alignment (MSA) is fed to the *evolutionary module*, which conducts a series of 2D convolutional transformations. The output of the evolutionary module is then converted to a 2D matrix and then fed into the *reconciliation module* together with the predicted contact map to perform *evolutionary reconciliation*. Finally, the output of the reconciliation module is utilized to predict the correctness of contacts at the residue pair level, which can be subsequently aggregated for estimating the overall quality of the full-fledged contact map.

The first module is the *evolutionary module*, which conducts a series of 2D convolutional transformations on the evolutionary context extracted from the covariational signal encoded in the multiple sequence alignment (MSA). The output of the evolutionary module is then converted to a 2D matrix and then fed into the second module together with the predicted contact maps. The second module is the *reconciliation module*, which performs a series of 2D convolutional transformations of the evolutionary context-integrated contact maps. Finally, the output of the reconciliation module is utilized to predict the correctness of contacts at the residue pair level, which can be subsequently aggregated for estimating the overall quality of the predicted contact map.

Inputs for training and inference. The inputs to rrQNet comprise of a predicted contact map and multiple sequence alignment (MSA) for the target protein sequence. MSA is generated by searching against diverse sequence sources after merging sequences from whole-genome sequence databases and from metagenome database (Zhang *et al.*, 2020). From the MSA, an inverse co-variation matrix (or precision matrix) (Fan *et al.*, 2016) is generated which captures dependent conditional correlations among pairwise variables. A precision matrix contains L×L blocks (L is the length of the protein sequence) and each block contains 21×21 matrix for direct couple correlation for a particular pair of residues (Li *et al.*, 2019) (i.e., having dimension of 21×21=441 for each pair of residues). We reshape the 441×L×L precision matrix to L×L×441 before feeding into the evolutionary module. During training, the input MSA is used for generating a diverse array of contact maps using seven different methods: CCMpred (Seemayer *et al.*, 2014), FreeContact (Kaján *et al.*, 2014), PSICOV (Jones *et al.*, 2012), MetaPSICOV (Jones *et al.*, 2015), DeepConPred2 (Ding *et al.*, 2018), DNCON2 (Adhikari *et al.*, 2018), and ResPRE (Li *et al.*, 2019). CCMpred is a pseudo-likelihood maximization (PLM)-based approach (Balakrishnan *et al.*, 2011; Ekeberg *et al.*, 2013) for contact prediction optimized for performance. FreeContact is a speed-optimized implementation of EVfold-mfDCA (Marks *et al.*, 2011). PSICOV uses sparse inverse co-variation estimation (Meinshausen and Bühlmann, 2006) for predicting protein contact map. MetaPSICOV is a two-stage contact predictor that employs a combination of approaches for obtaining MSA-derived co-variation signal and various statistics computed from the MSA for contact prediction. DeepConPred2 is an improved re-optimized version of the deep learning-based approach DeepConPred (Xiong *et al.*, 2017). DNCON2 employs a two-level deep convolutional neural network for contact prediction. ResPRE leverages residual neural network architecture coupled with precision matrix to predict contact maps. All methods are locally installed and run with default parameter settings to generate contact maps and top L contacts (L is the length of protein sequence) are fed to the reconciliation module. Following CASP standard of contact map assessment and to ensure a fair comparison between the contact predictors, we consider top L contacts for all methods. Our approach can be easily adapted for any number of contact pairs including all residue pairs. During inference, rrQNet takes only one contact map as input predicted by any method (including but not limited to the seven contact predictors used for training) along with the multiple sequence alignment.

Evolutionary module. The evolutionary module is preceded by a gate (InGate 1) comprising of a 2D convolutional layer with a kernel size of 1×1, batch normalization, and a nonlinear transformation called exponential linear unit (ELU). The input to this gate is the reshaped precision matrix of size L×L×441, which is transformed to a size of L×L×64, using a 1×1 convolutional layer with a filter size of 64. This input gate is followed by a stack of 55 residual neural network (ResNet) blocks. Each residual block in this ResNet stack consists of three 2D convolutional layers with a filter size of 64. The number of ResNet blocks is determined through validation on the independent CASP11 dataset to avoid overfitting, as discussed later. The first convolutional layer with kernel size 1×1 is used to restore the dimensionality of the feature vector and is followed by batch normalization. The second convolutional layer with kernel size of 5×5 is followed by ELU activation unit and batch normalization. The third convolutional layer with kernel size of 5×5 is the output layer of a ResNet block having a shortcut connection to the input layer for skipping the intermediate layers. This output propagates to the next ResNet block through a nonlinear transformation via the ELU unit.

Reconciliation module. Similar to the evolutionary module, the reconciliation module is preceded by a gate (InGate 2) consisting of a 2D convolutional layer having a kernel size of 3×3 and filter of 1, followed by a different transformation called rectified linear unit (ReLU) to transform the size to L×L×1, before concatenating with the input contact map. The reconciliation module leverages a stack of 40 residual neural network blocks having architecture similar to the evolutionary module determined through validation on the independent CASP11 dataset to avoid overfitting, as discussed later.

Quality estimation. The reconciliation module is followed by a gate (OutGate) comprising of a 2D convolutional layer having a kernel size of 3×3 followed by sigmoidal transformation, whose output leads to a 2D matrix of predicted confidence score at the level of a pair of interacting amino-acid residues from the input contact map. The 2D matrix of predicted confidence score is then subsequently aggregated for estimating the final quality score $\hat{q} \in [0, 1]$ of the full-fledged input contact map as: $\hat{q} = \Sigma p_{ij}/L$, where L is the number of residues, p_{ij} is 1 if the average predicted likelihood of residue pair (i,j) and the residue pair (j,i) is greater than the threshold of 0.5, and 0 otherwise. As such, $\hat{q}$ is analogous to the precision of the input contact map.

2.2 Training and implementation details

Our method is implemented using Keras with TensorFlow (Abadi *et al.*, 2016) backend and trained using Quadro RTX 5000 GPU with 16GB memory. We train our model for 50 epochs and use checkpoint for saving the best model on each epoch. Models are trained on a batch size of 2, shuffling the training data, using ADAM optimizer (Kingma and Ba, 2017) with a learning rate of 0.001 and binary cross-entropy loss function. While training, we crop (or pad) the input dimension to 256, but the trained model can estimate the quality score for any input length.

2.3 Datasets for training and performance evaluation

For training the deep learning model, we use a total of 3,073 contact maps of varying qualities predicted by seven different contact prediction methods previously discussed using a non-redundant dataset of 439 proteins including 421 protein targets from the DeepCov (Jones and Kandathil, 2018) dataset and 18 free modeling (FM) target domains from the 7th through 10th editions of the Critical Assessment of Protein Structure Prediction (CASP7-CASP10) experiments. We consider protein chains having no more than 35% pairwise sequence identity to remove redundancy in the training set.

For performance evaluation, we use 280 contact maps predicted by the seven locally installed contact prediction methods for 40 FM target domains with publicly available experimental structures from the 12th and 13th editions of CASP experiments (CASP12 and CASP13) to perform benchmark assessment; and 660 contact maps predicted by 30 fully automated server predictors participating in the contact prediction cate-

gory of the 14th edition of CASP (CASP14) for 22 FM targets to perform pseudo-blind assessment. Additionally, 210 contact maps predicted by the locally installed seven contact prediction methods for 30 FM target domains from the 11th edition of CASP (CASP11) are used as a validation set. Both the test and validation sets are structurally independent from the training set, having TM-align score < 0.5 (Zhang and Skolnick, 2005). For the benchmark assessment on CASP12 and CASP13 sets as well as validation on CASP11 set, we run the seven locally installed contact predictors with default parameter settings using multiple sequence alignments (MSA) generated by searching against diverse sequence sources including whole-genome sequence databases and metagenome database (Zhang *et al.*, 2020). Top L predicted contacts (L is the length of the protein sequence) are then fed into rrQNet for contact map quality estimation. For CASP14 pseudo-blind assessment, we download the contact predictions submitted by the server predictors directly from the CASP website and feed them into rrQNet along with the corresponding MSAs for quality estimation.

2.4 Comparison against other baseline approaches

To evaluate the effectiveness of our method in the absence of other competing approaches, we compare its performance against three in-house baseline contact quality estimation methods: consensus, shallow neural network, and naïve Bayes. The consensus-based baseline estimates the quality of a predicted interacting residue-pair by a contact predictor based on the average likelihood values of the predictions made for that residue-pair by the other contact predictors. That is, the average estimated confidence score across the other contact predictors for a residue-pair is its consensus-based quality score. Following this approach, we calculate a consensus score of each predicted contact in a contact map to estimate the quality at the residue-pair level and subsequently aggregate the resulting 2D matrix containing the consensus residue-pair confidence scores for estimating the final quality score of the full-fledged input contact map. The shallow neural network baseline consists of two fully connected layers. The precision matrix and the input contact map are concatenated and passed to the first layer having ten neurons. The second layer consists of one neuron followed by a sigmoid activation function whose output leads to a 2D matrix of predicted confidence score at the residue-pair level for subsequent aggregation and the final quality estimation of the full-fledged input contact map. The naïve Bayes baseline implements a Gaussian naïve Bayes using the precision matrix concatenated with the input contact map for contact map correctness estimation of at the residue-pair level to be subsequently aggregated for the final quality estimation of the full-fledged input contact map. For a fair comparison against the controls, the shallow neural network and naïve Bayes baselines are trained on the same training dataset as used in rrQNet.

2.5 Evaluation metrices

We employ twofold evaluation criteria for the performance assessment: (i) capability to reproduce the true quality score of a predicted contact map and (ii) capability to distinguish between accurate and inaccurate contact prediction. Ground truth is quantified by computing the precision of a predicted contact map against the contact map derived directly from the experimental 3D structure as: precision = TP/ (TP+FP), where TP (true positive) is the number of correctly predicted contacts and FP (false positive) is the number of incorrectly predicted contacts.

For the first evaluation criterion, we calculate Pearson (r), Spearman (ρ), and Kendal's tau (τ) correlation coefficients between the estimated contact map quality and ground truth precision. We compute the per-target average correlation coefficients (Per-target average r, ρ, τ) by pooling together all predicted contact maps for a specific protein target, as well as the overall global correlation coefficients (Overall global r, ρ, τ) by considering all predicted contact maps for all targets. The higher the correlation is, the better the performance is. The second evaluation criterion is measured by the area under the receiver operating characteristic curve (ROC) with a ground truth precision cutoff of 0.6. The AUC (Area Under Curve) assesses how well the predicted quality score may distinguish accurate prediction from inaccurate ones. The larger the area under the curve (AUC) of ROC, the better the capability is to distinguish between accurate and inaccurate contact prediction.

3 Results

3.1 Quality estimation performance on CASP12 and CASP13 benchmark sets

Table 1 reports the performance of rrQNet as well as the baseline approaches in reproducing the true quality scores of predicted contact maps for 154 predicted contact maps for 22 FM target domains from CASP12 and 126 predicted contact maps for 18 FM target domains from CASP13. rrQNet consistently attains high positive correlations between the estimated contact map quality and ground truth precision that are significantly better than all the baseline approaches.

Table 1. rrQNet performance in reproducing true quality scores of predicted contact maps on CASP12 and CASP13 targets compared to our in-house baseline approaches implementing consensus, shallow neural network (NN), and naïve Bayes (NB).

Dataset	Method	Avg. r	Avg. ρ	Avg. τ	Global r	Global ρ	Global τ
CASP 12	rrQNet	**0.87**	**0.80**	**0.65**	**0.81**	**0.79**	**0.60**
	Consensus	-0.49	-0.51	-0.41	0.72	0.72	0.53
	NN	0.09	0.09	0.11	0.75	0.74	0.53
	NB	0.75	0.69	0.51	-0.11	-0.06	-0.04
CASP 13	rrQNet	**0.90**	**0.84**	**0.71**	**0.81**	**0.79**	**0.59**
	Consensus	-0.54	-0.46	-0.39	0.49	0.55	0.40
	NN	0.14	0.12	0.14	0.58	0.57	0.39
	NB	0.64	0.57	0.44	-0.23	-0.22	-0.14

Note: Values in bold represent the best performance.

For instance, rrQNet attains the highest per-target average Pearson correlation (r) > 0.85, Spearman correlation (ρ) > 0.8, and Kendal's tau correlation (τ) > 0.65; as well as the highest global Pearson and Spearman correlations ~0.8 and Kendal's tau correlation ~0.6. Consistently better performance of rrQNet compared to the baselines demonstrates the effectiveness of the deep neural network architecture adopted in rrQNet. For the baseline approaches, it is interesting to note the performance tradeoff between the per-target average correlations and the global corre-

lations. For example, the consensus-based baseline performs reasonably well in terms of global correlations, but attains negative per-target average correlations for both CASP12 and CASP13. The shallow neural network baseline achieves weak positive per-target average correlations on both CASP12 and CASP13, but shows reasonable performance in terms of global correlations. On further inspection, we find that the consensus and the shallow neural network baselines often misclassify most of the true positive contacts and estimate low-quality scores for the full-fledged contact maps when the ground truth quality scores are in fact much higher, leading to poor per-target average correlations. For example, the consensus approach misclassifies most of the true positive contacts for contact map predicted by ResPRE for the target T0864-D1, resulting in a final quality score of 0.04 (lowest in the pool of estimated quality scores for target T0864-D1), whereas the ground truth score is 0.9 (highest in the pool of ground truth scores for the target T0864-D1). On the other hand, for the contact map predicted by PSICOV for target T0864-D1, the final quality estimation score by consensus approach is 0.1 (highest in the pool of estimated quality scores for the target T0864-D1), whereas the ground truth score is 0.28 (second lowest in the pool of ground truth scores for the target T0864-D1). As such, the consensus-based baseline results in negative correlations (r = -0.92, ρ = -0.88, τ = -0.78) for the target T0864-D1. Similarly, for the target T0968s1-D1, the consensus-based method misses most of the true positive contacts predicted by ResPRE and estimates ResPRE as lower quality (consensus-based quality estimation score is 0.27, whereas the ground truth is 0.82) than FreeContact (consensus-based quality estimation score is 0.29, whereas the ground truth is 0.35), leading to negative correlations (r = -0.91, ρ = -0.82, τ = -0.69). Likewise, the shallow neural network baseline misclassifies most of the true positive contacts for contact map predicted by MetaPSICOV for the target T0862-D1, resulting in negative correlations (r = -0.83, ρ = -0.68, τ = -0.45). By contrast, the naïve Bayes baseline attains reasonable per-target average correlations, but negative global correlations. Closer inspection reveals that the negative global correlations are caused by the naïve Bayes baseline frequently estimating higher quality scores (>0.7) for contact maps predicted by ResPRE for several targets including T0859-D1, T0862-D1, T0863-D1, T0990-D2, and T0990-D3 when the ground truth scores are low (<0.3); and lower quality scores (<0.4) for targets including T0864-D1, T0912-D3, T0968s2-D1, and T1022s1-D1 when the ground truth scores are high (>0.8). While the baseline methods exhibit performance tradeoff, rrQNet consistently delivers high performance on both global and per-target correlations. That is, rrQNet is able to accurately reproduce the true contact map quality score across all assessment metrics.

To investigate the capability of rrQNet against the baseline approaches in distinguishing between accurate and inaccurate contact prediction, we perform ROC analysis using all 280 contact predictions from combined CASP12 and CASP13 set containing 40 FM target domains. **Figure 2** shows the rrQNet ROC curve with a high AUC value of 0.94, which is better than all the baseline methods including consensus, shallow neural network and naïve Bayes. Except for naïve Bayes, which has very low AUC value since it frequently estimates higher quality scores for when the ground truth scores are low and lower quality scores when the ground truth scores are high, the other baselines achieve reasonable AUC values, albeit much lower than rrQNet. Overall, the results demonstrate that rrQNet attains superior performance in separating accurate and inaccurate contact prediction compared to the baseline approaches.

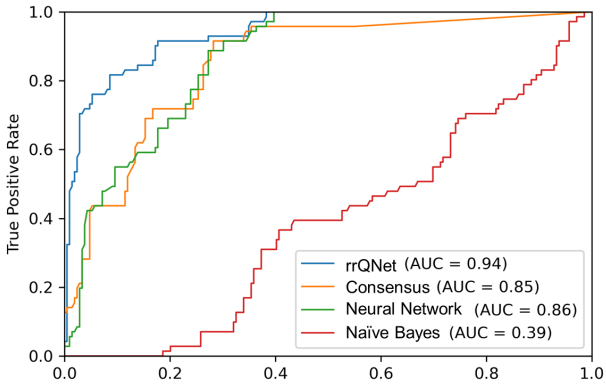

Figure 2. Receiver operating characteristic (ROC) analysis. The ROC curve and the corresponding AUC (Area Under Curve) value are used to assess rrQNet performance compared to our in-house baseline approaches implementing consensus, shallow neural network, and naïve Bayes in distinguishing between accurate and inaccurate contact prediction with a ground truth precision cutoff of 0.6 on 280 contact maps predicted by the seven locally installed contact prediction methods for 40 FM targets from CASP12 and CASP13.

3.2 Confidence prediction of contacts at residue-pair resolution

While rrQNet ultimately provides an estimated quality score of a full-fledged input contact map, the penultimate stage of the rrQNet pipeline outputs a 2D matrix of predicted confidence score at the level of individual residue-pairs from the input contact map. That is, rrQNet intrinsically performs confidence prediction of individual contacts at the residue-pair resolution to classify a contact as true positive if the average predicted likelihood of the residue pair (i,j) and the residue pair (j,i) is greater than the threshold of 0.5. In **Figure 3 (A)-(F)**, we show the true positive contacts predicted by rrQNet (lower triangles, green dots) out of the top L predicted contacts (lower triangles, red dots) against the ground truth (upper triangle, blue dots) for six representative targets from the CASP12 and CASP13 benchmark sets. For targets T0864-D1, T0968s2-D1, and T0969-D1, the contact maps predicted by ResPRE, MetaPSICOV, and DeepConPred2, respectively, are shown as representative examples of accurate contact prediction. For each of the three contact predictions shown in **Figure 3 (A)-(C)**, true positive contacts predicted by rrQNet (green dots) cover almost all the contact patterns present in the ground truth (blue dots). As such, the final estimated quality scores predicted by rrQNet for the three targets are 0.87, 0.74, and 0.7, respectively, which are close approximations of the corresponding ground truth precision scores of 0.9, 0.71, and 0.65, respectively. On the other hand, for target domains T0904-D1, T0950-D1, and T0980s1-D1 shown in **Figure 3 (D)-(F)**, contact maps predicted by FreeContact and PSICOV are examples of noisy contact prediction, having ground truth precisions of 0.02, 0.01, and 0.09, respectively. Here, rrQNet correctly recognizes most of these contacts as false positives, leading to the final estimated quality scores of 0.03, 0.01, and 0, respectively. In summary, rrQNet is useful in predicting the confidence of contacts at the residue-pair resolution.

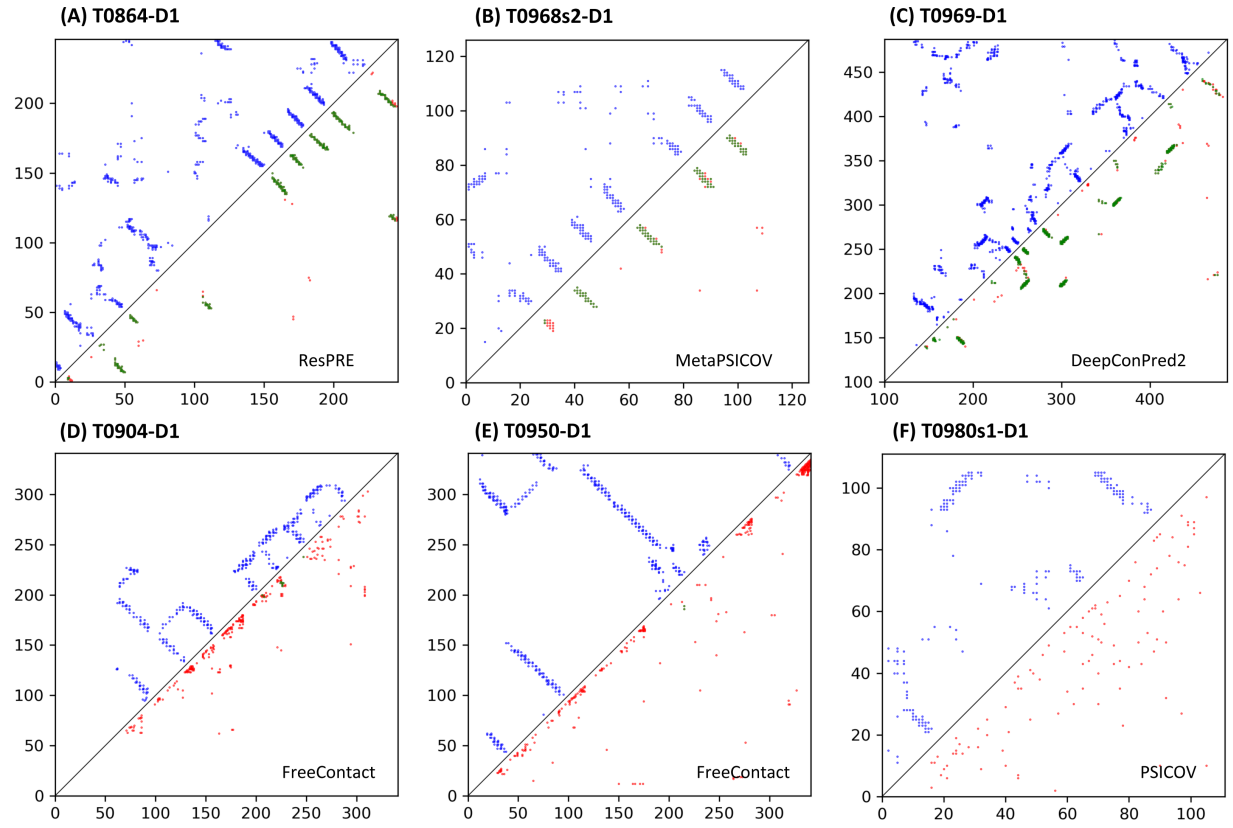

Figure 3. Confidence prediction of contacts. Six representative examples are shown. The upper triangle with blue dots represents the ground truth, the lower triangle with red dots represents the top L predicted contacts using various predictors, and the lower triangle with green dots represents the contacts classified by rrQNet as true positives.

3.3 Pseudo-blind assessment on CASP14 contact prediction experiment

To investigate the generalizability of rrQNet in accurately estimating the qualities of contact maps beyond the seven locally installed contact prediction methods used for training and benchmarking, we perform a pseudo-blind assessment by estimating the qualities of the contact maps submitted by 30 server predictors participating in CASP14 contact prediction category. **Figure 4** shows the rrQNet estimated quality score versus the CASP reported ground truth precision, averaged over all targets submitted by various server predictors for long-range, medium- and long-range, and short-range contacts. rrQNet consistently achieves high positive correlations between the estimated contact map quality and ground truth precision. For instance, the Pearson correlation (r) for long-range only contacts is 0.83, combined medium- and long-range contacts is 0.88, and short-range only contacts is 0.92. Spearman correlation (ρ)

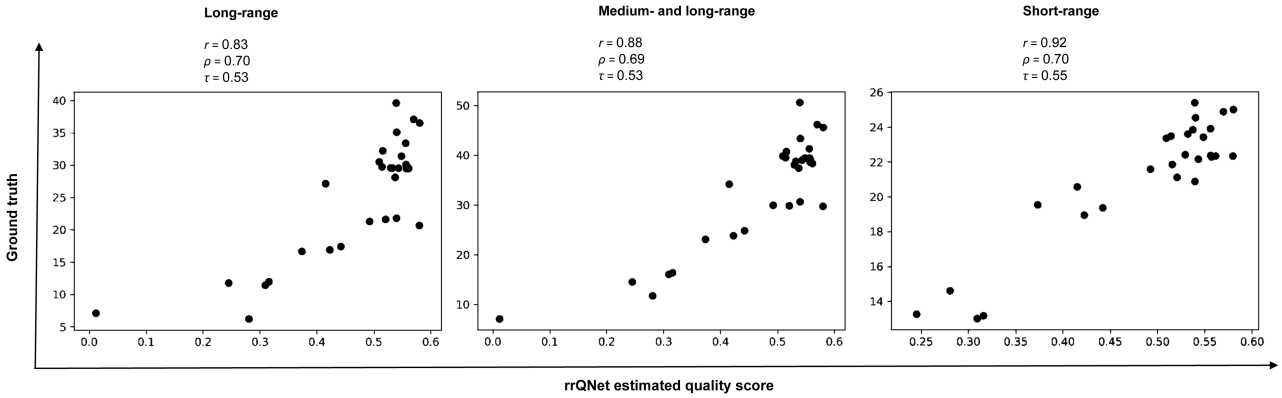

Figure 4. Estimated quality score versus ground truth precision for CASP14*. rrQNet estimated quality score versus the CASP reported ground truth precision, averaged over all targets submitted by various server predictors are shown for long-range, medium- and long-range, and short-range contacts. *BrainFold method is discarded for the calculation of short-range contacts due to missing prediction.

(and Kendal's Tau correlation (τ)) for long-, medium- and long-, and short-range contacts are 0.7 (and 0.53), 0.69 (and 0.53), and 0.7 (and 0.55), respectively. It is interesting to note that rrQNet continues to deliver good quality estimation performance for a wide variety of methods participating in CASP14, some of which are significantly improved versions of contact predictors than those used in rrQNet training. The results demonstrate that rrQNet can generalize well for estimating the qualities of contact maps predicted by a diverse array of methods. This suggests that the quality score of rrQNet can be potentially used as a loss function of an end-to-end structure prediction system as well as for iterative optimization of the contact maps by propagating the feedback through the end-to-end architecture. Furthermore, contact-driven homology detection and reconstruction approaches can benefit from such a generalizable quality estimate of a contact map predicted by external methods.

3.4 Ablation study

To examine the contribution of various components used in our deep learning-based contact map quality estimation method such as the deep neural network architecture adopted and the contact map prediction methods utilized for the training, we perform an ablation study using an independent validation set of 30 FM targets from CASP11. We follow the same multiple sequence alignment (MSA) generation protocol and the same training and inference procedure as described before to perform head-to-head comparison.

Contribution of deep neural network architecture. rrQNet consists of the evolutionary and reconciliation modules, each being a stack of residual neural network (ResNet) blocks. We study the effect of the network architecture on performance by gradually varying the number of ResNet blocks in the evolutionary and reconciliation modules while utilizing the same training dataset and evaluating quality estimation performance on the CASP11 validation set. As shown in **Figure 5**, the global Pearson correlation steadily increases with the increasing number of the ResNet blocks and attains a value of 0.8 for 55 ResNet blocks in the evolutionary module and 40 ResNet blocks in the reconciliation module (hereafter called the 55/40 architecture), before sharply dropping, possibly due to overfitting when deeper architecture beyond 55/40 is used. The per-target average Pearson correlation remains relatively steady while attaining the highest value of 0.91 for the 55/40 architecture. As such, this architecture is empirically chosen as the *de facto* model used in rrQNet. When we replace the ResNet blocks with standard convolutional blocks in the de facto model, the global Pearson correlation drops from 0.8 to 0.77, thus justifying the choice of ResNet architecture used in rrQNet.

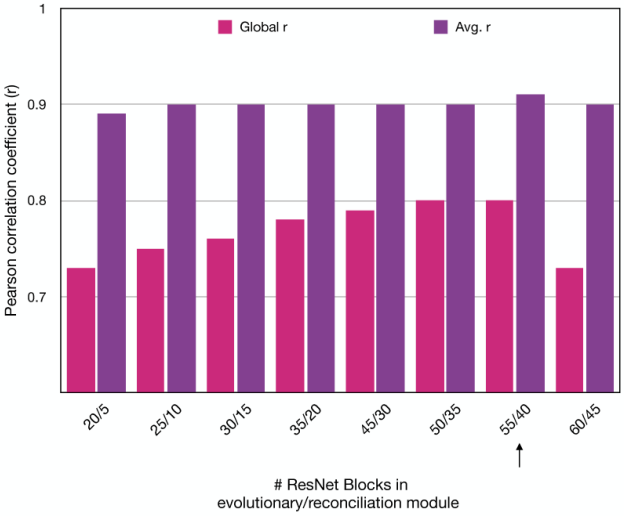

Figure 5. Effect of deep neural network architecture. Global and per-target average Pearson correlations are shown for a gradually increasing number of ResNet blocks in the evolutionary and reconciliation modules. The arrow indicates the *de facto* model chosen in rrQNet.

Contribution of contact prediction methods used for training. Our deep learning model is trained on seven contact predictors ranging from pure co-evolutionary methods to deep learning approaches. A natural question to ask is can we improve the performance even further by discarding the low-accuracy co-evolutionary contact prediction methods for training and instead exploiting cutting-edge approaches for predicting finer-grained binned distance distributions, which can lead to contact maps as a byproduct? To examine such question, we retrain rrQNet after discarding pure co-evolutionary methods PSICOV and FreeContact and evaluate quality estimation performance on the CASP11 validation set. The model retrained without PSICOV and FreeContact underperforms (global Pearson correlation 0.78) the original rrQNet (global Pearson correlation 0.8). We also independently retrain rrQNet after adding two recent distance map predictors trRosetta (Yang *et al.*, 2020) and RaptorX (Xu, 2019) and converting the binned distance distributions into contact maps by summing up the likelihood values of the predicted distance bins up to 8Å. The model retrained with trRosetta and RaptorX still underperforms (global Pearson correlation 0.77) the original rrQNet (global Pearson correlation 0.8). The results suggest that pure co-evolutionary methods used in rrQNet are useful, possibly by calibrating the training data in terms of false positives, whereas finer-grained distance map predictors may not bring additional value for contact map quality estimation.

4 Discussion

We have introduced rrQNet, a deep learning method for contact map quality estimation that can accurately perform self-assessment of protein structures at the residue pair representation captured by a 2D contact map, without explicitly turning the contact map into 3D coordinates. By performing deep neural network-based reconciliation of the evolutionary context-integrated contact maps, our trained classifier distinguishes true positive from false positive contacts, ultimately approximating the precision of an input contact map. We have shown that its predictions achieve much better performance than our in-house baselines, generalize well for contact maps generated by a wide variety of contact predictors, and offer versatility for contact quality estimation at varying granularities from individual residue-pair to full-fledged contact map. The contact quality score predicted by our method may help improve end-to-end protein structure prediction.

We may further improve the sensitivity of our method by extending the deep learning model to identify false negative contacts, which can be probabilistically combined with the estimated true positives in order to estimate the recall of a predicted contact map. Estimated recall can be used together with the predicted precision for a more rigorous evaluation of the input contact map. This will likely improve the robustness of our method by approximating the F-score metric, which can be calculated as the harmonic mean of the estimated precision and recall values. We may also improve the quality estimation performance by combining both evolutionary and physical constraints. For example, we may enforce a set of realistic physical constraints on the contact map (Wang and Xu, 2013), along with the recognition of protein-like contact patterns

(Skwark *et al.*, 2014). Such integration of both evolutionary and physical constraints using a deep learning model could result in improved quality estimation. Finally, instead of quality estimation of predicted contact maps, our deep learning model can be extended to estimate the quality of an inter-residue distance matrix, which encodes finer-grained information than contact maps and provides more physical constraints of a protein structure and thus, potentially benefits protein structure prediction more than the quality estimation of predicted contacts.

Funding

This work was partially supported by the National Institute of General Medical Sciences [R35GM138146 to DB] and the National Science Foundation [IIS-2030722, DBI-1942692 to DB].
Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]*.

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

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

Alapati,R. and Bhattacharya,D. (2018) clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons. In, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18. Association for Computing Machinery, New York, NY, USA, pp. 307–314.

Baek,M. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.

Balakrishnan,S. *et al.* (2011) Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, **79**, 1061–1078.

Bhattacharya,S. *et al.* DisCovER: distance- and orientation-based covariational threading for weakly homologous proteins. *Proteins: Structure, Function, and Bioinformatics*, **n/a**.

Bhattacharya,S. *et al.* (2021) Recent Advances in Protein Homology Detection Propelled by Inter-Residue Interaction Map Threading. *Frontiers in Molecular Biosciences*, **8**, 377.

Bhattacharya,S. and Bhattacharya,D. (2019) Does inclusion of residue-residue contact information boost protein threading? *Proteins: Structure, Function, and Bioinformatics*, **87**, 596–606.

Bhattacharya,S. and Bhattacharya,D. (2020) Evaluating the significance of contact maps in low-homology protein modeling using contact-assisted threading. *Sci Rep*, **10**, 2908.

Carreira,J. *et al.* (2016) Human Pose Estimation with Iterative Error Feedback. In, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, pp. 4733–4742.

Ding,W. *et al.* (2018) DeepConPred2: An Improved Method for the Prediction of Protein Residue Contacts. *Comput Struct Biotechnol J*, **16**, 503–510.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.

Fan,J. *et al.* (2016) An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, **19**, C1–C32.

He,K. *et al.* (2016) Deep Residual Learning for Image Recognition. In, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, pp. 770–778.

Jones,D.T. *et al.* (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999–1006.

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.

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

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kaján,L. *et al.* (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, **15**, 85.

Kingma,D.P. and Ba,J. (2017) Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

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

Marks,D.S. *et al.* (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*, **6**, e28766.

Meinshausen,N. and Bühlmann,P. (2006) High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, **34**, 1436–1462.

Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.

Pereira,J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, **n/a**.

Roche,R. *et al.* (2021) Hybridized distance- and contact-based hierarchical structure modeling for folding soluble and membrane proteins. *PLOS Computational Biology*, **17**, e1008753.

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

Shuvo,M.H. *et al.* (2020) QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. *Bioinformatics*, **36**, i285–i291.

Skwark,M.J. *et al.* (2014) Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLOS Computational Biology*, **10**, e1003889.

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

Wang,Z. and Xu,J. (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, **29**, i266–i273.

Xiong,D. *et al.* (2017) A deep learning framework for improving long-range residue–residue contact prediction using a hierarchical strategy. *Bioinformatics*, **33**, 2675–2683.

Xu,J. (2019) Distance-based protein folding powered by deep learning. *PNAS*, **116**, 16856–16865.

Yang,J. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *PNAS*, **117**, 1496–1503.

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.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**, 2302–2309.
Zheng,W. *et al.* (2019) Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Computational Biology*, **15**, e1007411.