Received: 16 July 2021 | Revised: 25 September 2021 | Accepted: 12 October 2021

DOI: 10.1002/prot.26269

RESEARCH ARTICLE

PROTEINS WILEY

# Distance-based reconstruction of protein quaternary structures from inter-chain contacts

Elham Soltanikazemi | Farhan Quadir | Raj S Roy | Zhiye Guo | Jianlin Cheng

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

**Correspondence**
Jianlin Cheng, Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA.
Email: chengji@missouri.edu

**Funding information**
Department of Energy, USA, Grant/Award Numbers: DE-AR0001213, DE-SC0020400, DE-SC0021303; National Institutes of Health, Grant/Award Number: R01GM093123; National Science Foundation, USA, Grant/Award Numbers: DBI1759934, IIS1763246

## Abstract

Predicting the quaternary structure of protein complex is an important problem. Inter-chain residue-residue contact prediction can provide useful information to guide the ab initio reconstruction of quaternary structures. However, few methods have been developed to build quaternary structures from predicted inter-chain contacts. Here, we develop the first method based on gradient descent optimization (GD) to build quaternary structures of protein dimers utilizing inter-chain contacts as distance restraints. We evaluate GD on several datasets of homodimers and heterodimers using true/predicted contacts and monomer structures as input. GD consistently performs better than both simulated annealing and Markov Chain Monte Carlo simulation. Starting from an arbitrarily quaternary structure randomly initialized from the tertiary structures of protein chains and using true inter-chain contacts as input, GD can reconstruct high-quality structural models for homodimers and heterodimers with average TM-score ranging from 0.92 to 0.99 and average interface root mean square distance from 0.72 Å to 1.64 Å. On a dataset of 115 homodimers, using predicted inter-chain contacts as restraints, the average TM-score of the structural models built by GD is 0.76. For 46% of the homodimers, high-quality structural models with TM-score ≥ 0.9 are reconstructed from predicted contacts. There is a strong correlation between the quality of the reconstructed models and the precision and recall of predicted contacts. Only a moderate precision or recall of inter-chain contact prediction is needed to build good structural models for most homodimers. Moreover, GD improves the quality of quaternary structures predicted by AlphaFold2 on a Critical Assessment of Techniques for Protein Structure Prediction–Critical Assessments of Predictions of Interactions dataset.

**KEYWORDS**
distance-based modeling, gradient descent optimization, inter-chain contact prediction, protein complex, protein quaternary structure modeling

## 1 | INTRODUCTION

Determination of interactions between protein chains in a protein complex is important for understanding protein function and cellular processes and can play significant roles in designing and discovering new drugs.[1] Detailed protein–protein interactions are represented by the three-dimensional shape of a complex consisting of interacting proteins (i.e., quaternary structure). Experimental techniques such as

X-ray crystallography and nuclear magnetic resonance (NMR) can determine the quaternary structure of protein complexes with high accuracy. However, these experimental approaches are costly and time-consuming, and therefore cannot be applied to most protein complexes. Therefore, computational modeling approaches, which provide a faster and inexpensive way to predict quaternary structures, have become increasingly popular and important.[2]

Computational protein docking, currently the most widely used approach for modeling complex structures, takes the tertiary structures of individual proteins as input to build the quaternary structure of the complex as output.[3–9] Docking methods can be largely divided into two categories including template-based modeling, in which known protein complex structures in the Protein Data Bank (PDB) are used as templates[10–17] to guide modeling, and template-free modeling (ab initio docking), which does not use any known structure as template, and instead searches through a large conformation space for relative orientations of protein chains with minimum binding energy. The binding energy is often roughly approximated by geometric and electrostatic complementarity, inter-chain hydrogen binding, hydrophobic interactions, and residue–residue contact potentials.[18–23]

Although template-based docking works well if a good structural template is available, it cannot be applied to most protein complexes that lack suitable templates.[2,24] Ab initio docking methods can predict the quaternary structure of acceptable quality for some protein complexes, but according to several rounds of Critical Assessments of Predictions of Interactions (CAPRI), they still cannot achieve adequate accuracy for most protein complexes.[24,25] One main reason for the low accuracy is that the ab initio docking methods need to search through a huge conformation space, which is usually not feasible with limited time and computing resources. To reduce the search space, several methods started to use the interface contacts between proteins to constrain conformation search[26–30] and were able to enhance docking accuracy,[30] showing inter-chain (inter-protein) contacts can provide valuable information to build protein quaternary structures as what had happened in protein tertiary structure prediction.

The major advances of ab initio tertiary structure prediction of a single protein chain have been largely driven by accurate prediction of intra-chain residue–residue contact prediction and the development of methods of reconstructing tertiary structures from the contacts.[31–36] However, there are still very few methods available to reconstruct protein quaternary structures from predicted inter-chain residue–residue contacts. With the emergency of inter-chain contact prediction enhanced by residue–residue co-evolutionary analysis and deep learning,[37–41] it is crucial to create robust methods to efficiently and effectively use inter-chain contacts to directly reconstruct protein quaternary structures. Despite both reconstructing protein tertiary structures from intra-chain contacts and reconstructing quaternary structures from inter-chain contacts use contacts as distance restraints to build three-dimensional (3D) structures, they have some significant difference. On one hand, the quaternary structure reconstruction depends on the quality of the tertiary structure input as well as the accuracy of inter-chain contact prediction because it keeps the tertiary structure of monomers largely unchanged in the modeling process. Therefore, the accuracy of the quaternary structure reconstruction is low if the quality of tertiary structures is low, but the reconstruction of tertiary structure from intra-chain contact prediction only depends on the accuracy of intra-chain contact prediction. On the other hand, the reconstruction of quaternary structure mostly needs to orient the tertiary structures of a limited number of monomers (e.g., two for dimer) correctly, which has much less degree of freedom and likely requires fewer accurate distance restraints than the reconstruction of tertiary structures that needs to accurately position many (e.g., hundreds of) amino acids. Finally, a unique challenge for the quaternary structure reconstruction is to account for the potential change of tertiary structures of monomers upon protein–protein interaction.

Gradient descent optimization has become a popular method to build the tertiary structure of proteins using intra-protein (intra-chain) residue–residue contacts or distances. AlphaFold,[31] which was ranked first in 13th Critical Assessment of Techniques for Protein Structure Prediction (CASP13), developed a gradient descent-based folding method to generate protein tertiary structure from intra-chain distances. trRosetta,[32] a powerful tool for protein tertiary structure modeling, uses a gradient descent-like method (MinMover from pyRosetta) to build the structure of individual proteins from predicted residue-residue distances. A recent protein folding framework based on gradient descent, GDFOLD,[42] uses intra-chain contacts as input constraints to directly optimize the positions of $C_\alpha$ atoms of a protein.

Motivated by the recent success of applying gradient descent to protein tertiary structure prediction, in this study, we develop an ab initio gradient descent optimization-based method (GD) to construct quaternary structures of protein dimers from inter-chain contacts. We first test if the proposed method can generate high quality structures of protein dimers using true contacts. Then, we apply it to construct quaternary structures of homodimers from predicted, noisy, and incomplete contacts. To rigorously benchmark its performance, we also implement a Markov Chain simulation method (MC) based on RosettaDock[43] and apply a simulated annealing method based on Crystallography and NMR System (CNS)[41] to reconstruct protein complex structures from inter-protein contacts and compare them with GD. We evaluate the three methods on several in-house datasets consisting of 233 homodimers and heterodimers in total as well as on a standard dataset of 32 heterodimers[40,44] with true or predicted contacts. GD consistently performs better than MC and CNS on all the datasets. It can reconstruct high-quality structures from true inter-chain contacts and good structures for most homodimers when predicted contacts are only moderately accurate. Finally, we also apply GD to 28 homodimers used in the several recent CASP–CAPRI experiments and seven homomeric targets of the latest 2020 CASP14–CAPRI experiment to investigate how the quality of input (i.e., tertiary structure of monomers predicted by AlphaFold2 and inter-chain contact prediction) influences its performance.

## 2 | RESULTS AND DISCUSSIONS

### 2.1 | Reconstruction of quaternary structure from native (true) contacts and true tertiary structures of monomers in the bound state

To check if the methods can work well when the perfect input is provided, we first apply GD, MC and CNS to generate quaternary structures for 44 homodimers in the Homo44 dataset using true inter-chain contacts as constraints and true tertiary structures of monomers in homodimers (i.e., in the bound state) as input. The models reconstructed by the methods are evaluated by five complementary metrics against known experimental structures of the homodimers: root-mean-square deviation (RMSD), TM-score, the percentage of native contacts existing in predicted models (f_nat), interface RMSD (I_RMSD), and ligand RMSD (L_RMSD) widely used in the field. The TM-score between the reconstructed complex structure and the native complex structure is computed using TM-align.[45] RMSD is calculated using CA-RMSD in PyRosetta. I_RMSD, L_RMSD, and f_nat are calculated using Dock-Q[46] and our in-house programs.

The detailed results of GD on the Homo44 dataset in terms of TM-score, RMSD, f_nat, I_RMSD, and L_RMSD) and the length and number of contacts of the homodimers are reported in Table S1. GD is able to generate high-quality structural models for all the dimers when true inter-chain contacts are provided as constraints. For instance, TM-score of the models ranges from 0.936 to 0.999 and I_RMSD from 0.204 Å to 1.85 Å. The average of RMSD, TM-score, f_nat, I_RMSD, and L_RMSD of GD, MC and CNS is compared in Table 1 (see the per-dimer comparison of the three methods in terms of each metric in Figures S1–S5). GD performs best in terms of all the metrics, while MC performs better than CNS. The average RMSD of GD is 0.63 Å, which is lower than 0.76 Å of MC and 1.16 Å of CNS. The average TM-score of GD is 0.99—an almost perfect score, which is higher than 0.98 of MC and 0.91 of CNS. Moreover, GD realizes 92.19% of native contacts (f_nat = 92.19%), higher than 91.39% of MC and 82.49% of CNS. The average I_RMSD and L_RMSD of GD are 0.77 Å and 1.38 Å, lower than those of the other two methods. Figure 1 illustrates high-quality structural models reconstructed by GD, MC, and CNS that are superimposed with the true structure of a dimer (PDB code: 1XDI) in Homo44.

We then evaluate GD with MC and CNS on 73 heterodimers in the Hetero73 dataset using true inter-chain contacts as constraints and true tertiary structures of monomers in homodimers (i.e., in the bound state) as input. The detailed per-dimer results of GD are shown in Table S2. A comparison of the three methods is shown in Table 2. The average RMSD, I_RMSD, and L_RMSD of GD are lower than the other two methods, while its average TM_score and f_nat are higher than the other two methods, indicating that GD performs best, while MC works better than CNS. A per-dimer comparison of RMSD and TM-score of the models reconstructed by the three methods is depicted in Figures S6 and S7, respectively. The models reconstructed by GD for the heterodimers have high quality on average (e.g., mean RMSD = 1.23 Å and TM-score = 0.92). However, in comparison with the results on homodimers in Table 1, the average accuracy on heterodimers is lower than that on homodimers. A main reason is that heterodimers tend to have lower inter-chain contact density (i.e., # of inter-chain contacts/sum of the sequence lengths of two chains in a dimer)[41,47] on average, leading to fewer distance restraints available for structure reconstruction.

Moreover, we evaluate the three methods on 32 heterodimers in the Std32 dataset. The detailed results of GD are presented in Table S3. The average TM-score, RMSD, f_nat, I_RMSD, and L_RMSD of the models reconstructed by GD, MC, and CNS are reported in Table 3. Similar to the results on the other datasets, GD generates high-quality models on average and performs best in terms of all the metrics, while MC performs substantially better than CNS.

### 2.2 | Analysis of two key factors impacting the quality of models reconstructed by GD from native contacts

We have observed that the quality of the generated structures is affected by two factors: initial structure in the optimization and inter-chain contact density in a dimer.

Figure 2 compares TM-score and I_RMSD of 20 models reconstructed by GD with the corresponding 20 different start models for a dimer 1Z3A in Homo115 using noisy predicted inter-chain contacts and the true tertiary structure of the monomer in the bound state as input. The quality of the initial models is generally poor (TM-score ranging from 0.5 to 0.61 and I_RMSD ranging from 12 Å to 26 Å). In 19 out of 20 cases, GD improves TM-score of the models and in all 20 cases, it reduces I-RMSD. The quality of the models reconstructed by GD varies a lot (e.g., TM-score ranging between 0.5

**TABLE 1** Mean and SD of root mean square distance (RMSD), TM-score, f_nat, I_RMSD, and L_RMSD of the three methods on 44 homodimers in Homo44

| Evaluation metric | GD | MC | CNS |
|---|---|---|---|
| RMSD (mean, SD) | 0.63 ± 0.3788 | 0.76 ± 0.361 | 1.16 ± 1.0043 |
| TM-score (mean, SD) | 0.99 ± 0.0132 | 0.98 ± 0.014 | 0.91 ± 0.0102 |
| f_nat (mean, SD) | 92.19 ± 8.64 | 91.39 ± 9.08 | 82.49 ± 22.02 |
| I_RMSD (mean, SD) | 0.77 ± 1.05 | 1.35 ± 3.98 | 12.46 ± 8.46 |
| L_RMSD (mean, SD) | 1.38 ± 0.8 | 1.7 ± 0.9 | 11.18 ± 14.51 |

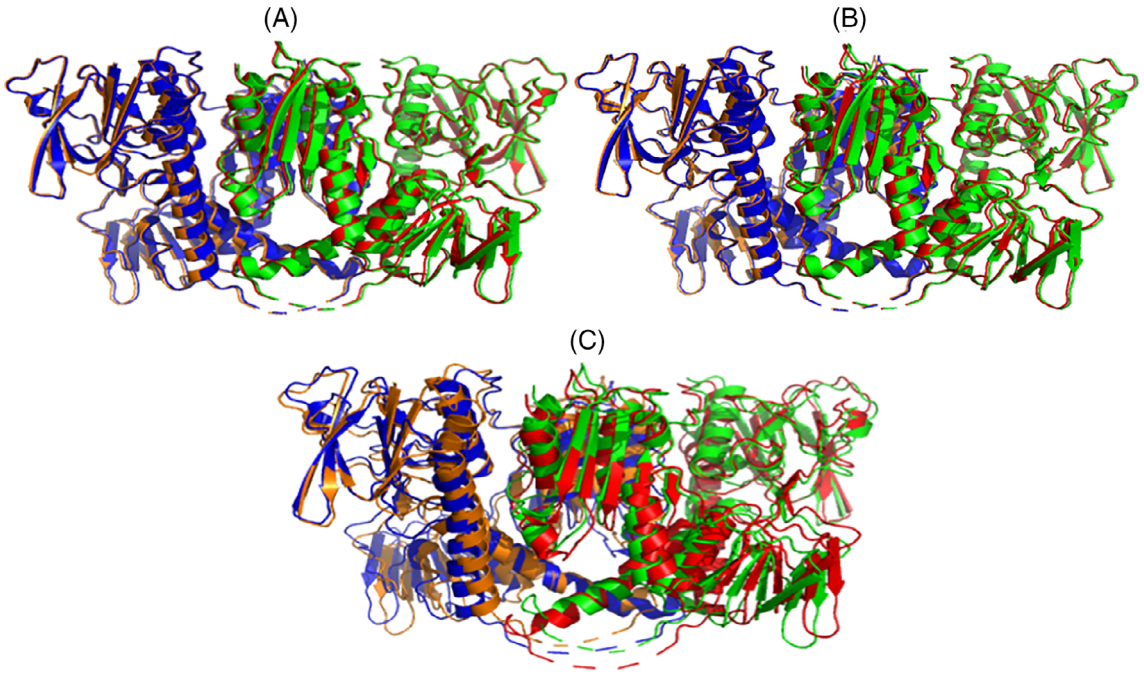


**FIGURE 1** The superposition of the native structure of 1XDI and the models reconstructed by three methods (i.e., green and orange denoting the true dimer structure and blue and red the reconstructed dimer structure): (A) GD, (B) MC, and (C) CNS. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by GD are 0.99, 0.56 Å, 94.52%, 0.24 Å, and 0.74 Å, respectively. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by MC are 0.99, 0.61 Å, 93.15%, 0.45 Å, and 1.29 Å, respectively. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by CNS are 0.88, 2.25 Å, 74.79%, 1.49 Å, and 5.18 Å, respectively. CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance

**TABLE 2** Mean and SD of RMSD, TM-score, f_nat, I_RMSD, and L_RMSD results of the three methods on 73 heterodimers in the Hetero73 dataset

| Evaluation metric | GD | MC | CNS |
|---|---|---|---|
| RMSD (mean, SD) | 1.23 ± 1.91 | 4.76 ± 8.01 | 7.7 ± 12.99 |
| TM-score (mean, SD) | 0.92 ± 0.12 | 0.85 ± 0.16 | 0.79 ± 0.23 |
| F_nat (mean, SD) | 90.31 ± 16.77 | 82.59 ± 26.68 | 84.43 ± 23 |
| I_RMSD (mean, SD) | 0.72 ± 1.02 | 1.58 ± 1.7 | 1.65 ± 4.51 |
| L_RMSD (mean, SD) | 3.75 ± 6.15 | 7.78 ± 11.8 | 9.21 ± 14.05 |

Abbreviations: CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance.

**TABLE 3** Average RMSD, TM-score, f_nat, I_RMSD, and L_RMSD of GD, MC and CNS on 32 dimers in the Std32 dataset

| Evaluation metric | GD | MC | CNS |
|---|---|---|---|
| TM-score | 0.96 | 0.95 | 0.82 |
| RMSD | 1.95 | 2.9 | 10.04 |
| f_nat | 92.78 | 92.43 | 69.13 |
| I_RMSD | 1.64 | 1.99 | 3.71 |
| L_RMSD | 4.65 | 7.16 | −14.99 |

Abbreviations: CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance.

and 0.88 and I_RMSD between 2.5 Å and 20 Å). Given a reasonable initial structure, GD converges to a high-quality local minimum. But starting from a poor initial model, the algorithm can get stuck in a bad local minimum, producing a low-quality model. However, the correlation between the quality of the randomly initialized models and the reconstructed models is low. Therefore, it is useful to run GD multiple times with different start models.

Based on the experiment on Homo44 and Hetero73 datasets, using 20 different start models to run GD 20 times can build almost perfect quaternary structural models with TM-score = 0.99 and an RMSD less than 1 Å from true inter-chain contacts and true tertiary structures in the bound state for most dimers (see Tables S1 and S2 for details).

In addition to initial models, the contact density of a dimer strongly influences the quality of the models reconstructed from native contacts. Figure 3 illustrates how TM-score and RMSD of the models reconstructed for 73 heterodimers change with respect to the density of true contacts. When the contact density is above ~0.25, almost all the models have a very low RMSD (< 1 Å) and a very high TM-score (close to 1). When contact density is lower than ~0.25,

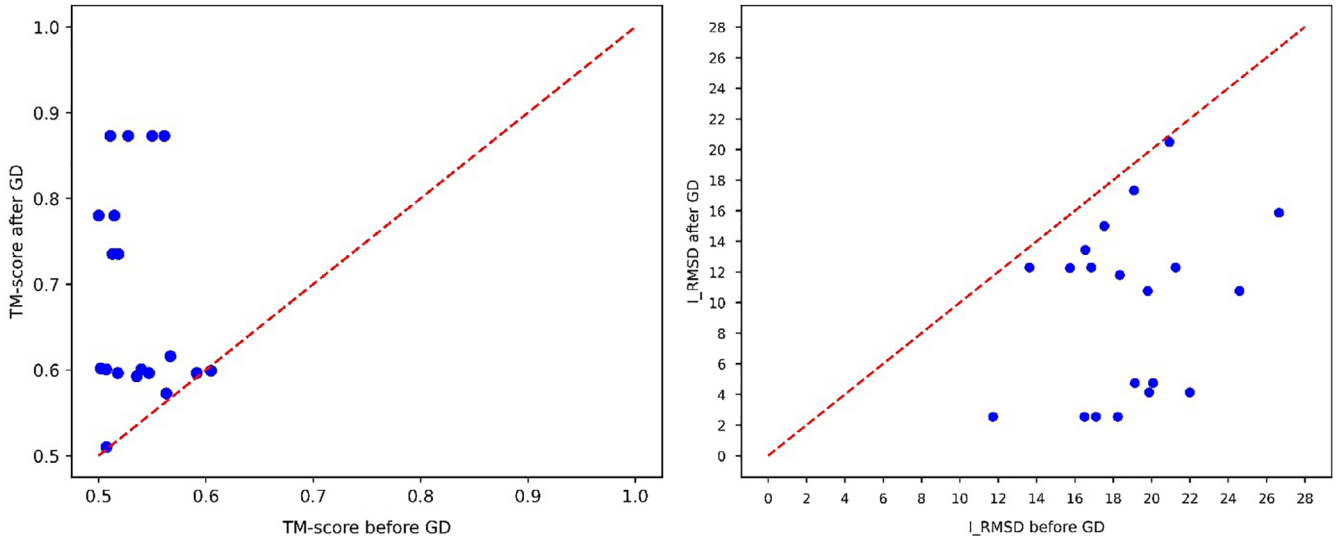


**FIGURE 2** TM-score and I_RMSD of quaternary structure models for a homodimer 1Z3A before applying GD and after applying GD during 20 runs. The 20 start models are initialized from the true tertiary structure of the monomer in the dimer before GD is applied. GD is then used to reconstruct the quaternary structures from predicted inter-chain contacts. The *x*-axis denotes the quality (TM-score or I_RMSD) of 20 initial quaternary structure models and *y*-axis the quality of 20 final models built by GD from the initial models. In 19 out of 20 cases, GD improves the TM-score of the models. In all 20 cases, GD reduces the I_RMSD of the models. GD, gradient descent optimization; RMSD, root mean square distance

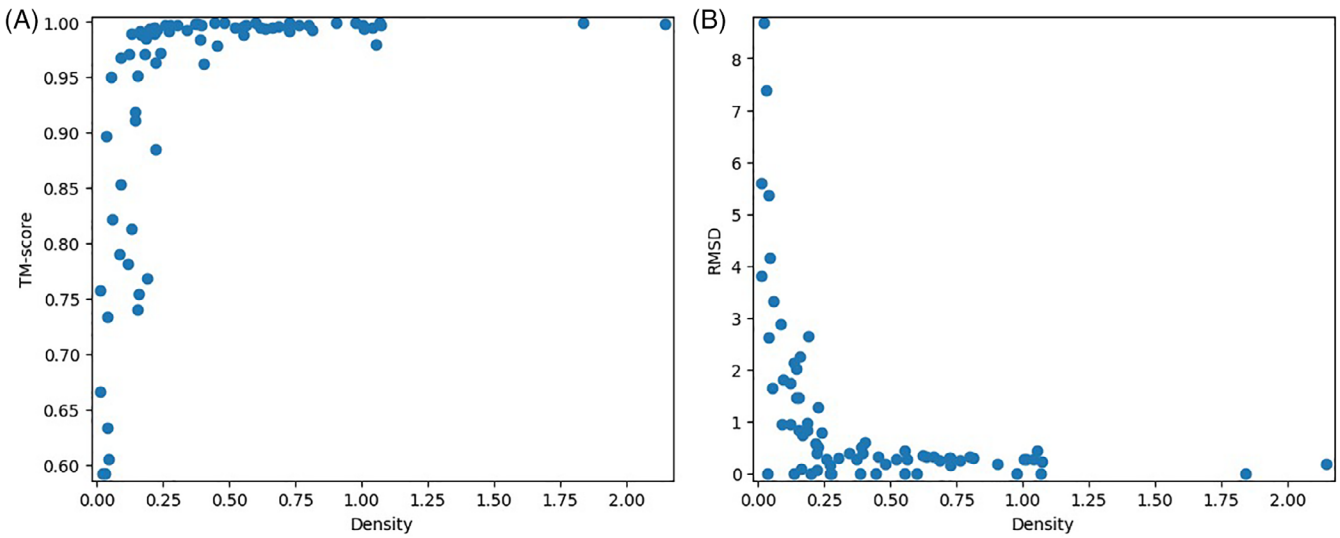


**FIGURE 3** TM-scores and root mean square distance (RMSD) of the models versus the inter-chain contact density of 73 heterodimers

there are both good-quality and low-quality models. Overall, with an increase of the contact density, the quality of the reconstructed structure increases in terms of all the metrics: RMSD, TM-score, f_nat, I_RMSD, and L_RMSD (results for f_nat, I_RMSD, and L_RMSD not shown).

We also investigate if the number of residues of the heterodimers affects the quality of the generated models. According to the plot of TM-score versus the dimer length in Figure S8, there is no direct relationship between the length of complexes and the quality of models (the correlation between the two $= -0.013$).

## 2.3 | Reconstruction of quaternary structures of homodimers from predicted inter-chain contacts and true tertiary structures of monomers in the bound state

We evaluate the performance of the three optimization methods on homodimers using predicted inter-chain contacts because the newly developed deep learning methods such as ResCon can make inter-chain contact prediction with reasonable accuracy for a large portion of homodimers. The three methods are compared on three subsets

(Set A, Set B, and Set C) of homodimers in the Homo115 dataset. Set A consists of 40 dimers with small interaction interfaces. Set B has 37 dimers with medium interaction interfaces. Set C contains 38 complexes with large interaction interfaces. The detailed results of GD (average TM-score, RMSD, f_nat, I_RMSD, and L_RMSD) as well as the precision and recall of the predicted contacts for sets A, B, and C are shown in supplemental Tables S4–S6, respectively. The precision of predicted inter-chain contacts is measured by $\frac{\#\text{correctly predicted contacts with probability} \geq \text{cut-off probability}}{\#\text{predicted contacts with probability} \geq \text{cut-off probability}}$, and the recall of predicted inter-chain contacts by $\frac{\#\text{correctly predicted contacts with probability} \geq \text{cut-off probability}}{\#\text{native contacts}}$, where cut-off probability of selected contacts is set to 0.5. The predicted inter-chain contacts and the true tertiary structures of the monomers in the dimers in the bound state are used as input.

**TABLE 4** Average RMSD, TM-score, f_nat, I_RMSD, and L_RMSD of the best models reconstructed by the three methods for the homodimers in Set A using predicted contacts as input

| Evaluation metrics | GD | MC | CNS |
|---|---|---|---|
| TM-score | 0.68 | 0.66 | 0.58 |
| RMSD | 10.81 | 11 | 17.48 |
| f_nat | 22.47 | 18.38 | 14.67 |
| I_RMSD | 9.93 | 10.03 | 12.37 |
| L_RMSD | 25.46 | 27.81 | −30.35 |

Abbreviations: CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance.

**TABLE 5** Average RMSD, TM-score, f_nat, I_RMSD, and L_RMSD of the best models reconstructed by the three methods for Set B with predicted inter-chain contacts as input

| Evaluation metrics | GD | MC | CNS |
|---|---|---|---|
| TM-score | 0.8 | 0.77 | 0.64 |
| RMSD | 6.78 | 8.3 | 12.89 |
| f_nat | 32.18 | 28.66 | 22.19 |
| I_RMSD | 6 | 7.6 | 13.3 |
| L_RMSD | 14.87 | 18.46 | −20.69 |

Abbreviations: CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance.

**TABLE 6** Average RMSD, TM-score, f_nat, I_RMSD, and L_RMSD of the best models reconstructed by the three methods for Set C with predicted contacts as input

| Evaluation metrics | GD | MC | CNS |
|---|---|---|---|
| TM-score | **0.81** | 0.80 | 0.76 |
| RMSD | **6.26** | 6.77 | 9.5 |
| f_nat | 37.43 | 35.07 | **42.3** |
| I_RMSD | **5.01** | 5.46 | 7.41 |
| L_RMSD | **12.73** | 13.96 | −16.3 |

*Note*: The highest average TM-score, f_nat, and lowest mean RMSD, I_RMSD and L_RMSD are marked in bold.
Abbreviations: CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance.

The average performance of GD, MC, and CNS on Sets A, B, and C is compared in Tables 4–6, respectively. Similar as observed on models reconstructed from true inter-chain contacts, GD performs best here, MC second, and CNS third in terms of almost all the evaluation metrics. Moreover, the average accuracy generally increases with the increase of the size of the interaction interfaces (i.e., accuracy of Set C > accuracy of Set B > accuracy of Set A), showing that it is easier to reconstruct quaternary structures with larger interaction interfaces. The average TM-score of the structural models built for the three datasets by GD is 0.68, 0.80, and 0.81, respectively, higher than the models predicted by MC and CNS. GD generates models with higher TM-score for most dimers. The average TM-score of the models reconstructed by GD for all 115 homodimers in Set A, Set B, and Set C is 0.76. Moreover, for 53 out of 115 (46%) homodimers, the models reconstructed by GD have high TM-scores (≥0.9) (see Tables S4–S6), suggesting that GD is able to reconstruct high-quality models for a large portion of dimers using only predicted inter-chain contacts as input. Figure 4 illustrates a high-quality model reconstructed for dimer 1C6X (precision of contact prediction = 40.24% and recall of contact prediction = 49.28%, TM-score = 0.99, f_nat = 84.61%).

We investigate the relationship between the quality of the models generated by GD and the precision and recall of predicted contacts. Figure 5A plots the TM-score of the models constructed for the dimers in Homo115 against the precision of contacts predicted for them. The correlation between the two is 0.78, indicating that the quality of the structural models increases with respect to the precision of predicted contacts. It is worth noting that if the precision is >20%, most reconstructed models have good quality (e.g., with TM-score > 0.8 or even close to 1). If the precision is >40%, all the models have good quality (TM-score > 0.8). The results demonstrate that there is no need to get a very high accuracy of contact prediction for GD to obtain high-quality structural models for homodimers as long as its accuracy reaches a specific threshold. GD is robust against the noise in predicted contacts. This result is encouraging news for the community to develop more methods to predict inter-chain contacts in protein complexes. Figure 5B also reveals the strong positive correlation between the percent of true contacts existing in the reconstructed structural models (f_nat) and the precision of predicted contacts. Pearson's correlation between the two is 0.94. Moreover, when the precision of predicted contacts is >40%, a higher percent (>50%) of native contacts are realized in the models.

Furthermore, there is a strong correlation between the quality of reconstructed models (e.g., TM-score and f_nat) and the recall of the predicted inter-chain contacts as shown in Figure 6. Pearson's correlation between TM-score and recall is 0.78 and between f_nat and recall is 0.93, showing that a higher recall of predicted contacts leads

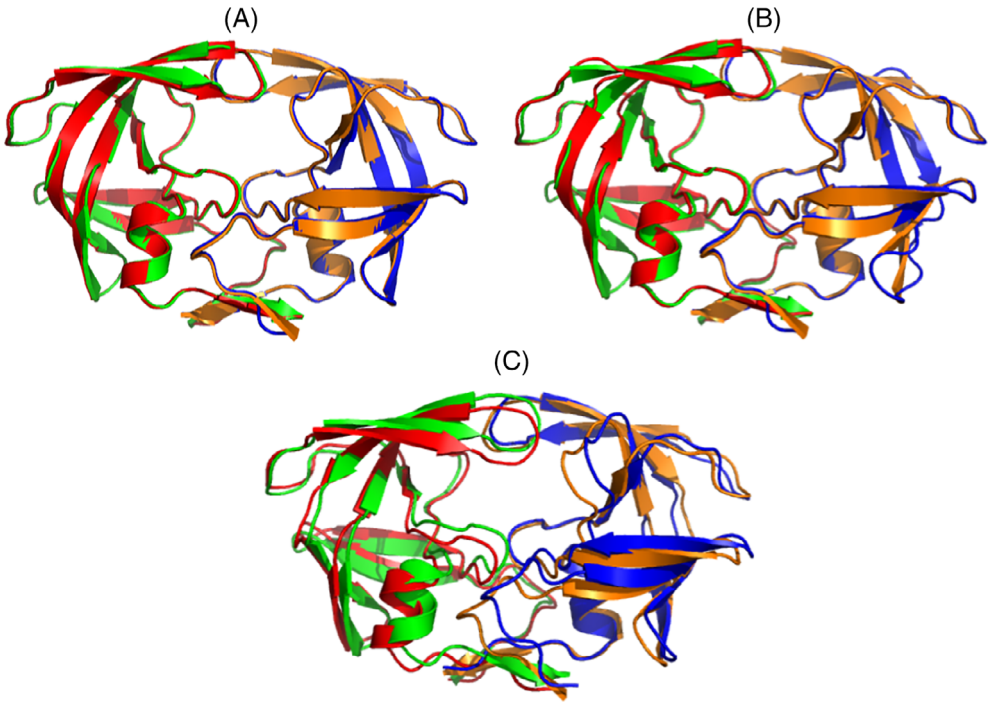


**FIGURE 4** The superposition of the native structure of 1C6X and the models generated by three methods (i.e., green and orange representing the true dimer structure, blue, and red the generated models): (A) GD, (B) MC, and (C) CNS. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by GD are 0.99, 0.4 Å, 84.61%, 0.4 Å, and 0.91 Å, respectively. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by MC are 0.98, 0.6 Å, 78.84%, 0.6 Å, and 1.6 Å, respectively. TM-score, RMSD, f_nat, I_RMSD, L_RMSD of the model predicted by CNS are 0.86, 2.02 Å, 41.6%, 2.14 Å, and 5.68 Å, respectively. CNS, crystallography and NMR system; GD, gradient descent optimization; MC, Markov chain; RMSD, root mean square distance

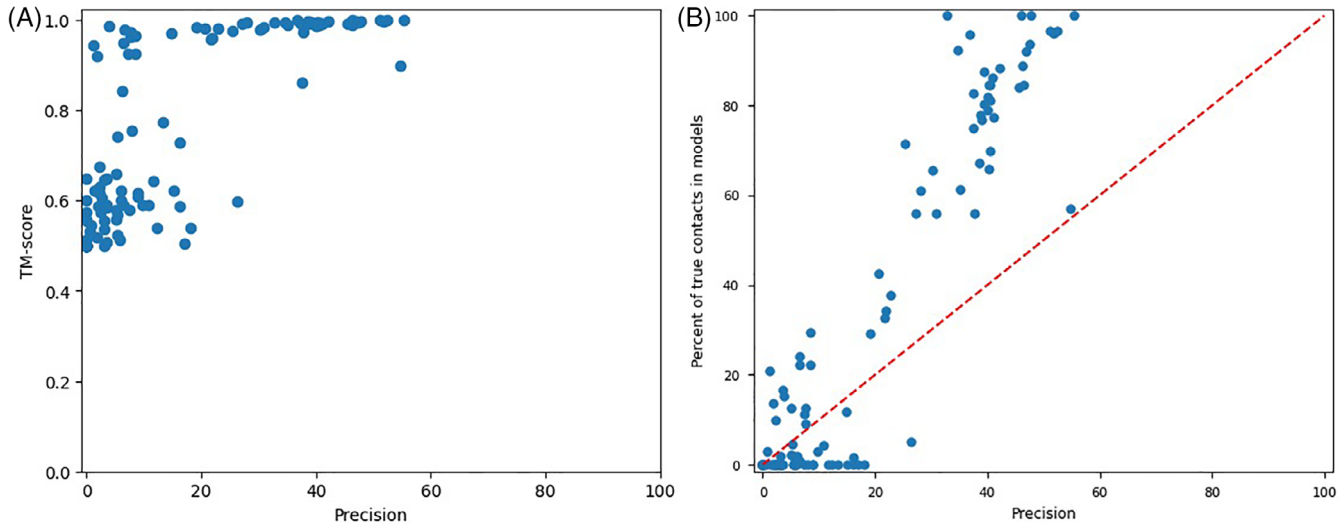


**FIGURE 5** The plot of TM-score and percent of native contacts of the models (f_nat) against the precision of predicted contacts on Homo115 dataset. (A) Pearson's correlation between TM-score and precision is 0.78. (B) Pearson's correlation between f_nat and precision is .94

to better reconstructed models. As shown in Figure 6A, when the recall of predicted contacts is >20%, all the reconstructed models except a few cases have good quality, that is, their TM-score is >0.8 and even close to 1, indicating only a small portion of true contacts are needed to build good quaternary structural models for most homodimers. Even if the recall of predicted contacts is <20%, good models (TM-score > 0.8) can still be reconstructed for some dimers. Moreover, as shown in Figure 6B, when the recall of predicted contacts is >20%, the percent of true contacts (f_nat) in the models reconstructed for all but a few dimers is higher than the recall of

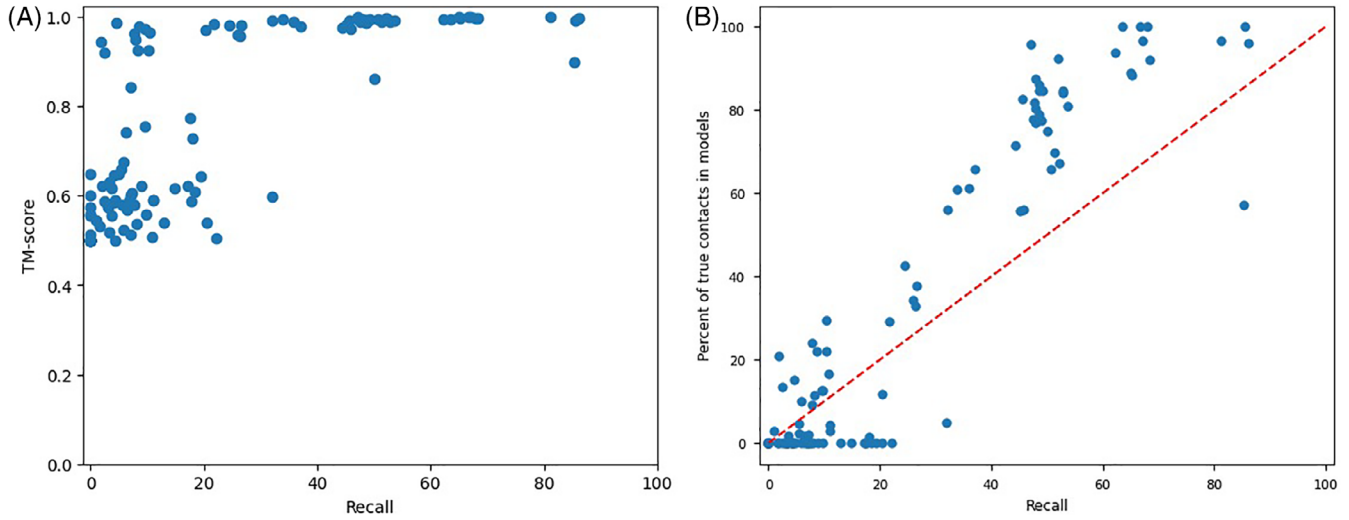


**FIGURE 6** TM-score and percent of native contacts of the predicted models (f_nat) reconstructed by GD versus the recall of the predicted inter-chain contacts on the Homo115 dataset. (A) Pearson's correlation between TM-score and recall is 0.78. (B) Pearson's correlation between f_nat and recall is 0.93. GD, gradient descent optimization

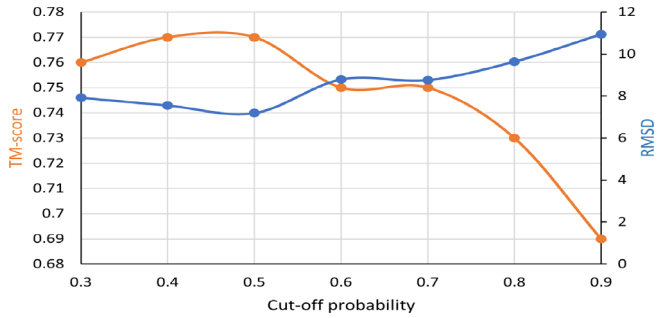


**FIGURE 7** The average root mean square distance (RMSD) and TM-score of models reconstructed for homodimers in the Homo115 dataset versus the cut-off probability of selecting predicted inter-chain contacts as restraints

predicted contacts that are used as input, indicating that the optimization process of GD can realize (recall) more true contacts than what is provided in the predicted input contacts.

Moreover, we investigate how the cut-off probability of selecting predicted inter-chain contacts as input affects the quality of reconstructed structural models. To determine good cut-off probabilities for selecting predicted contacts, we test different cut-off values in the range [0.3, 0.9], with a step size of 0.1. Figure 7 shows how the average TM-score and RMSD of reconstructed models change with respect to the cut-off probabilities on the Homo115 dataset. The best model quality (lowest RMSD and highest TM-score) is reached at the cut-off probability of 0.5 on the dataset. We imagine that the best cut-off probability can be data- and method-dependent. Therefore, it can be useful to try different cut-off probabilities to reconstruct models and then select good ones from them on different datasets.

## 2.4 | Performance of GD on CASP and CAPRI targets using predicted inter-chain contacts and true/predicted tertiary structures of monomers as input

We have further validated the performance of our method on the CASP–CAPRI dataset of 28 homodimers[48] used by DeepHomo (called CASP–CAPRI dataset) and the most recent CASP14–CAPRI test dataset of seven homodimers (called CASP14–CAPRI dataset). The average precision of top L/5 contact predictions (L: the length of monomer in homodimer) made by a deep learning method[49] for CASP14–CAPRI and CASP–CAPRI datasets is 24.82% and 32.29%, respectively. Using the true tertiary structures of monomers in the bound state and predicted interchain-contacts as input, the average TM-score of the quaternary structures of homodimers reconstructed by GD is 0.75 and 0.74 on the CASP14–CAPRI and CASP–CAPRI datasets, respectively (see the detailed results in Tables S7 and S8).

Using the tertiary structures of monomers predicted by AlphaFold2[50] in the unbound state and the same predicted inter-chain contacts as input, the average TM-score of quaternary structures reconstructed by GD for CASP–CAPRI dataset is slightly decreased to score 0.69, but the average TM-score for the CASP14–CAPRI dataset is decreased more to 0.63. The results indicate that the quality of tertiary structure input is important for the quality of the reconstructed quaternary structures. One reason for the substantial decrease on the CASP14–CAPRI dataset is a few tertiary structures of the monomers predicted by AlphaFold2 are not of high quality (see the relatively low TM-score of predicted tertiary structures for T1070, T1052, T1032 in Table S9). The detailed results of GD (TM-score of quaternary structures, RMSD, f_nat, I_RMSD, L_RMSD, the precision of top L/5 inter-chain contact predictions, and TM-score of the monomer structures predicted by AlphaFold2) on CASP14–CAPRI and CASP–CAPRI datasets are shown in Tables S9 and S10.

### 2.5 | Comparison of GD with AlphaFold2 on CASP–CAPRI dataset

AlphaFold2 was developed to predict tertiary structures of proteins. However, it can be used to predict structures of protein complex by joining the sequences of multiple chains into one sequence and adding a linker sequence (e.g., 20 glycines) to separate the sequences of two adjacent chains. Adding a sufficiently long linker is necessary to account for the possible large distance between the last amino acid of a chain and the first amino acid of the following chain. We apply this approach to use AlphaFold2 to predict quaternary structures for the 28 homodimers in the CASP–CAPRI dataset. Unlike AlphaFold2, GD is not a fully fledged quaternary structure predictor and needs inter-chain contacts and tertiary structures of monomers as input to build quaternary structure. Therefore, we can only compare GD with the final step of quaternary structure model of AlphaFold2. To fairly compare them, we use the inter-chain contacts extracted from the quaternary structure models predicted by AlphaFold2 and the tertiary structures of monomers predicted by AlphaFold2 in the unbound state as input for GD to reconstruct quaternary structures. The results of AlphaFold2 and our approach of using GD with AlphaFold2 (AlphaFold2 + GD) are reported in Table S11. AlphaFold2 achieves a high average TM-score of 0.82 on this dataset, while AlphaFold2 + GD obtains an even higher TM-score of 0.84. The results shows that GD can build better quaternary structures models than AlphaFold2 if the quality of the input (inter-chain contacts and tertiary structure models of monomers) is similar, demonstrating that GD can add value on top of the current most sophisticated protein structure prediction tool—AlphaFold2 for quaternary structure prediction. Using AlphaFold2 predictions as input for GD generates better quaternary structures on this dataset (i.e., TM-score = 0.84) than using the inter-chain contact prediction of the deep learning predictor as input for GD in the previous experiment (i.e., TM-score = 0.69) is because the inter-chain contact predictions in the quaternary structure models built by AlphaFold2 are more accurate than the deep learning predictor on the dataset.

### 2.6 | Limitation and future development

The experiments in this work use true or predicted inter-chain contacts as input. However, contact is a coarse description of the distance between residues. The more accurate quantification of the inter-chain residue–residue distances such as the fine-grained or real-value distance between residues will likely further improve the performance of the distance-based reconstruction of quaternary structures as the intra-chain residue–residue distance prediction improved tertiary structure prediction in the last several years.[31,51] The GD method can be readily adapted to take in inter-chain residue–residue distances to reconstruct quaternary structures. As deep learning methods for predicting inter-chain residue-residue distances are developed and available in the field, we will assess how well GD may reconstruct quaternary structures from predicted inter-chain distances in the future. Moreover, a major trend in tertiary structure prediction exemplified by AlphaFold2 is to use the end-to-end deep learning model to directly predict tertiary structure from sequence input without intermediate steps. We envision that the similar end-to-end model will be developed for quaternary structure prediction soon. The distance-based reconstruction of protein quaternary structure in this work is complementary to the upcoming end-to-end deep learning model. It can be used to further refine the quaternary structure predicted by the end-to-end model as shown in our study that GD improves the quality of the quaternary structures of dimers predicted by AlphaFold2.

## 3 | CONCLUSION

We design and develop the first gradient descent distance optimization (GD)-based method to reconstruct quaternary structure of protein dimers from inter-protein contacts and compare it with the Markov Chain Monte Carlo and simulated annealing optimization methods adapted to address the problem. GD performs consistently better than the other two methods in reconstructing quaternary structures of dimers from either true or predicted inter-chain contacts. GD can reconstruct high-quality structures for almost all homodimers and heterodimers from true inter-chain contacts and can build good structural models for many homodimers from only predicted inter-chain contacts, demonstrating distance-based optimizations are useful tools for predicting the quaternary structures. Moreover, we show that the contact density, size of interaction interface, precision and recall of predicted contacts, and threshold of selecting contacts as restraints influence the accuracy of reconstructed models. Particularly, when the precision and recall of predicted contacts reach a moderate level (e.g., >20%), GD can construct good models for most homodimers, demonstrating that predicting inter-chain contacts or even distances and distance-based optimization are a promising ab initio approach to predicting the quaternary structures of protein complexes.

## 4 | MATERIALS AND METHODS

### 4.1 | Inter-chain contacts and dimer datasets

Two residues from two protein chains in a dimer are considered an inter-chain contact if any two heavy atoms from the two residues have a distance less than or equal to 6 Å.[41,47] True contacts of a dimer with the known quaternary structure in the PDB are identified according to the coordinates of atoms in the PDB file of the dimer.

We use several in-house datasets of protein homodimers and heterodimers with true and/or predicted inter-protein contacts as well as a standard datasets consisting of 32 heterodimers (Std32)[52] to evaluate the methods. The first in-house dataset has 44 homodimers randomly selected from the Homo_Std[41] curated from the 3D Complex database,[53] each of which have 39–621 true contacts (called

Homo44). The second in-house data includes 115 homodimers (called Homo115) selected from Homo_Std, each of which has at least 21 predicted inter-chain contacts with a probability of ≥0.5. Our in-house deep learning method—ResCon[52] is applied to predict inter-chain contacts for the dimers in Homo115. Homo115 is divided into three subsets (Set A, Set B, and Set C) according to the size of interfaces. Set A has 40 protein complexes with small interaction interfaces consisting of 14–68 true inter-chain contacts. Set B consists of 37 complexes having medium interaction interfaces with 69–129 true contacts. Set C consists of 38 complexes having large interaction interfaces with 131–280 true contacts. The third in-house dataset contains 73 heterodimers (called Hetero73)[52] curated from the PDB, in which the sum of the lengths of the two chains is less than or equal to 400. The heterodimers in Hetero73 have 2 to 255 true inter-protein contacts.

Moreover, GD is also tested on 28 homodimers from the recent CASP–CAPRI joint experiment used in DeepHomo[48] (called CASP–CAPRI dataset), and seven homodimers from the latest CASP14–CAPRI experiment collected from five homodimeric targets (T1054, T1078, T1032, T1083, T1087) and two homotrimers (T1050, and T1070; called CASP14-CAPRI dataset).

## 4.2 | Gradient descent cost function and optimization

The inter-chain contacts are used as distance restraints for the gradient descent method to build the structures of protein dimers. The cost function to measure the satisfaction of the distance between any two residues in contact to guide the structural modeling is defined as follows:

$$f(x) = \begin{cases} \left(\frac{x-lb}{sd}\right)^2 & x < lb \\ 0 & lb \le x \le ub \\ \left(\frac{x-ub}{sd}\right)^2 & ub < x < ub + sd \\ \frac{1}{sd}(x - (ub + sd)) & x > ub + sd \end{cases}$$

Here, *lb* and *ub* represent the lower bound and upper bound of the distance ($x$) between two residues that are assumed to be in contact. As mentioned earlier, two residues are considered in contact if the distance between their heavy atoms is less than 6 Å. However, to simplify the process of restraint preparation, two residues are considered in contact if the distance between their $C_\beta$ atoms ($C_\alpha$ for glycine) is less than 6 Å. The lower bound (*lb*) is empirically set to 0 and the upper bound (*ub*) to 6 Å. *sd* is the standard deviation, which is set to 0.1. Based on this cost function, if the distance between two residues in contact is ≤6 Å, that is, the contact restraint is satisfied, and the cost is 0.

The complete contact cost function for a structural model of a dimer to be minimized is the sum of the costs for all contacts used in modeling (called contact energy). For simplicity, all restraints have equal weights and play equally important roles in modeling. The contact energy function is differentiable with respect to the distances between residues and coordinates of atoms of the residues, and therefore it can be minimized by a gradient descent iterative algorithm (GD), that is, Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS)[31,54]) used in this study.

We implement GD on top of pyRosetta. The total energy function for the structural optimization is the combination of the contact energy and the talaris2013 potentials,[31] which works better than the contact energy alone in our experiments. The input to the algorithm includes inter-chain contacts and an initial random conformation of a dimer initialized from the tertiary structures of individual protein chains (monomers) of the dimer. The tertiary structure of monomer can be either in the bound state or unbound state. It can be either an experimentally determined true structure or a predicted structure. A predicted structures is considered in the unbound state because the existing methods generally predicts the tertiary structure of a chain without considering its partner. An initial conformation of a protein dimer is generated by making 40 random, rigid rotations and translations ranging from $1^\circ - 360^\circ$ and $1\,\text{Å} - 20\,\text{Å}$ of the tertiary structures of the two chains in a dimer after putting them in the same coordinate system. Specifically, the tertiary structure of each protein chain is rotated and translated arbitrarily along the line connecting the centers of the two chains, aiming to make the two protein chains facing each other.

Then, 6000 iterations of the gradient descent optimization (i.e., L-BFGS) are carried out to generate new structural models. In most cases tested, it converged after only 1000 iterations. Since the quality of the final structure is influenced by the initial structure, the optimization process is carried out 20 times, each with a random structure as the start point. The optimized structure with the lowest energy is selected as the final predicted structure of a dimer. For 20 runs, the total number of iterations of the gradient descent optimization is $1.2 \times 10^5$.

## 4.3 | Markov chain Monte Carlo optimization

We apply a Rosetta protocol in pyRosetta based on Metropolis–Hasting sampling[55] to implement a Markov chain Monte Carlo (MC) optimization to reconstruct complex structures according to the Boltzmann distribution. An initial conformation of a dimer is generated in the same way as in the GD algorithm. Starting from the initial conformation, a low-resolution rigid-body search is employed to rotate and translate one chain around the surface of the other chain to generate new structures in the MC optimization. Five hundred Monte Carlo moves are attempted. Each move is accepted or rejected based on the standard Metropolis acceptance criterion.[56]

After the low-resolution search, back-bone and side-chain conformations are further optimized with the Newton minimization method in a high-resolution refinement process, in which the gradient of the

scoring function dictates the direction of the starting point in the rigid-body translation/rotation space. This minimization process is repeated 50 times to detect the local minimum of the energy function that may have similar performance as the global minimum.[6]

We implement the MC method above using high-resolution and low-resolution docking protocols in RosettaDock to optimize the same energy function used in the GD method. Low-resolution docking is performed using the DockingLowRes protocol, whereas DockMCMProtocol is used to perform high-resolution docking. For a dimer, $10^5$ to $10^7$ rounds of MC optimization with different initial conformations are executed to generate structural models. At the end, $10^5$ to $10^7$ models are generated, among which the model with the lowest energy is selected as the final prediction.

### 4.4 | Simulated annealing optimization based on crystallography and NMR system

This structure optimization method, Con_Complex[41] in the DeepComplex package, is implemented on top of the CNS[57,58] that uses a simulated annealing protocol to search for quaternary structures that satisfy inter-chain contacts.[52] This method takes the PDB files of monomers (protein chains) in a protein multimer (e.g., homodimer) and the true or predicted inter-protein contacts as input to reconstruct the structure of the multimer without altering the shape of the structure of the monomer. The inter-protein contacts are converted into distance restraints used by CNS. This process generates 100 structural models and then picks five models with lowest CNS energy. It is worth noting that this method can handle the reconstruction of the quaternary structure of any multimer consisting of multiple identical or different chains. Because inter-chain contacts are the main restraints to guide structure modeling, the performance of this method mostly depends on the quality of the inter-protein contact predictions.

## ACKNOWLEDGMENTS

Research reported in this publication was supported in part by Department of Energy grants (DE-AR0001213, DE-SC0020400, and DE-SC0021303), two NSF grants (DBI1759934 and IIS1763246), and an NIH grant (R01GM093123).

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26269.

## DATA AVAILABILITY STATEMENT

The source code of the methods and test data sets are available in the DeepComplex2 package at: https://github.com/jianlin-cheng/DeepComplex2.

## ORCID

*Jianlin Cheng* https://orcid.org/0000-0003-0305-2853

## REFERENCES

1. Hadarovich A, Kalinouski A, Tuzikov AV. Deep learning approach with rotate-shift invariant input to predict protein homodimer structure. *Bioinformatics Research and Applications*. Springer; 2020:296-303.
2. Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*. 2014;42:W252-W258.
3. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein−protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*. 2003;125:1731-1737.
4. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins*. 2003;52:80-87.
5. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*. 2004;20:45-50.
6. Gray JJ, Moughon S, Wang C, et al. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*. 2003;331:281-299.
7. Smith GR, Sternberg MJ. Prediction of protein–protein interactions by docking methods. *Curr Opin Struct Biol*. 2002;12:28-35.
8. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res*. 2006;34:W310-W314.
9. Hwang H, Vreven T, Janin J, Weng Z. Protein–protein docking benchmark version 4.0. *Proteins*. 2010;78:3111-3114.
10. Mukherjee S, Zhang Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*. 2011;19:955-966.
11. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins*. 2002;49:350-364.
12. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Res*. 2014;42:W285-W289.
13. Källberg M, Wang H, Wang S, et al. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc*. 2012;7:1511-1522.
14. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. *Proteins*. 2010;78:3235-3241.
15. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci*. 2012;109:9438-9441.
16. Negroni J, Mosca R, Aloy P. Assessing the applicability of template-based protein docking in the twilight zone. *Structure*. 2014;22:1356-1362.
17. Vakser IA. Low-resolution structural modeling of protein interactome. *Curr Opin Struct Biol*. 2013;23:198-205.
18. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS One*. 2011;6:e24657.
19. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*. 2007;67:1078-1086.
20. Zacharias M. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*. 2003;12:1271-1282.
21. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*. 1997;272:106-120.
22. Neveu E, Ritchie DW, Popov P, Grudinin S. PEPSI-dock: a detailed data-driven protein–protein interaction potential accelerated by polar Fourier correlation. *Bioinformatics*. 2016;32:i693-i701.
23. Kastritis PL, Bonvin AM. Are scoring functions in protein− protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res*. 2010;9:2216-2225.
24. Lensink MF, Velankar S, Kryshtafovych A, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and

template-based modeling: a CASP-CAPRI experiment. *Proteins*. 2016; 84:323-348.
25. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Sci*. 2005;14:278-283.
26. Mintseris J, Weng Z. Atomic contact vectors in protein-protein recognition. *Proteins*. 2003;53:629-639.
27. Wodak SJ, Méndez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol*. 2004;14:242-249.
28. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*. 2005;21:1487-1494.
29. De Vries SJ, Bonvin AM. Intramolecular surface contacts contain information about protein–protein interface regions. *Bioinformatics*. 2006;22:2094-2098.
30. Chelliah V, Blundell TL, Fernández-Recio J. Efficient restraints for protein–protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *J Mol Biol*. 2006;357:1669-1682.
31. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706-710.
32. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci*. 2020;117:1496-1503.
33. Hou J, Wu T, Cao R, Cheng J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*. 2019;87:1165-1178.
34. Liu J, Wu T, Guo Z, Hou J, Cheng J. Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *Proteins*. 2021. doi:10.1002/prot.26186
35. Xu J. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci*. 2019;116:16856-16865.
36. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun*. 2019;10:1-13.
37. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci*. 2009;106:67-72.
38. Hopf TA, Schärfe CPI, Rodrigues JPGLM, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*. 2014;3: e03430.
39. González AJ, Liao L, Wu CH. Prediction of contact matrix for protein–protein interaction. *Bioinformatics*. 2013;29:1018-1025.
40. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*. 2014;3:e02030.
41. Quadir F, Roy RS, Halfmann R, Cheng J. DNCON2_Inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning. *Sci Rep*. 2021;11:12295.
42. Mao W, Ding W, Xing Y, Gong H. AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nat Mach Intell*. 2020;2:25-33.
43. Lyskov S, Gray JJ. The RosettaDock server for local protein–protein docking. *Nucleic Acids Res*. 2008;36:W233-W238.
44. Zeng H, Wang S, Zhou T, et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res*. 2018;46:W432-W437.
45. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33:2302-2309.
46. Basu S, Wallner B. DockQ: a quality measure for protein-protein docking models. *PloS One*. 2016;11:e0161879.
47. Zhou T-M, Wang S, Xu J. Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis. *bioRxiv*. 2018;240754.
48. Yan Y, Huang S-Y. Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes. *Brief Bioinform*. 2021;22(5):bbab038. https://doi.org/10.1093/bib/bbab038
49. Roy RS, Quadir F, Soltanikazemi E, Cheng J. A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. *bioRxiv*. 2021. 2021.2009.2019.460941. doi: 10.1101/2021.09.19.460941
50. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.
51. Wu T, Guo Z, Hou J, Cheng J. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinform*. 2021;22:1-17.
52. Quadir F, Roy RS, Soltanikazemi E, Cheng J. DeepComplex: a web server of predicting protein complex structures by deep learning inter-chain contact prediction and distance-based modeling. *Front Mol Biosci*. 2021;8:827-831.
53. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. *PLoS Comput Biol*. 2006;2:e155.
54. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program*. 1989;45:503-528.
55. Zhang Z, Schindler CE, Lange OF, Zacharias M. Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta. *PLoS One*. 2015;10:e0125941.
56. Allen M, Tildesley D. *Computer Simulation of Liquids*. Oxford University Press; 1989.
57. Brunger AT. Version 1.2 of the crystallography and NMR system. *Nat Protoc*. 2007;2:2728-2733.
58. Brünger AT, Adams PD, Clore GM, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*. 1998;54:905-921.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Soltanikazemi E, Quadir F, Roy RS, Guo Z, Cheng J. Distance-based reconstruction of protein quaternary structures from inter-chain contacts. *Proteins*. 2022;90(3):720-731. doi:10.1002/prot.26269