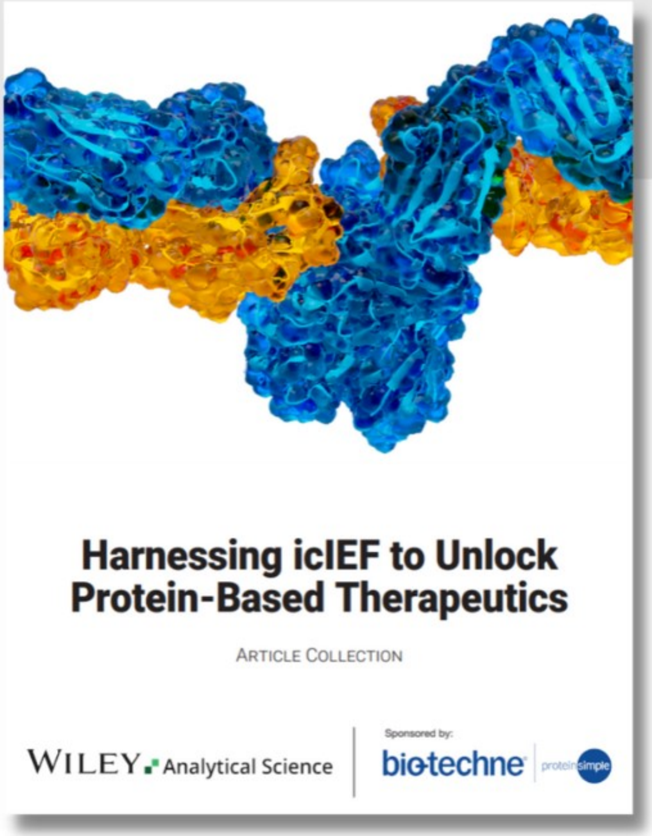
Harnessing icIEF to Unlock Protein-Based Therapeutics



Harnessing icIEF to Unlock Protein-Based Therapeutics

ARTICLE COLLECTION

WILEY Analytical Science

Sponsored by:

bio-techne proteinsimple

Read the new Article Collection

Keep up to date with the latest developments in biotherapeutics and the range of treatments for various diseases with our latest article collection. Find out how imaged cIEF (icIEF) technique is essential for quality control and analytical development of these drugs, as it accurately determines the surface charge of lipid nanoparticles and the charge heterogeneity of proteins and antibodies.

This article collection aims to provide you with more information on these techniques and technologies, helping you further your research in this field.

bio-techne® | proteinsimple

WILEY

Received: 17 October 2022 | Revised: 6 March 2023 | Accepted: 14 March 2023
DOI: 10.1002/prot.26493

RESEARCH ARTICLE

ComparePD: Improving protein–DNA complex model comparison with hydrogen bond energy-based metrics

Fareeha Kanwal Malik[1,2] | Jun-tao Guo[1]

[1]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223, USA

[2]Research Center of Modeling and Simulation, National University of Science and Technology, Islamabad 44000, Pakistan

Correspondence
Jun-tao Guo, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA.
Email: jguo4@uncc.edu

Funding information
National Science Foundation

Abstract

Computational modeling of protein–DNA complex structures has important implications in biomedical applications such as structure-based, computer aided drug design. A key step in developing methods for accurate modeling of protein–DNA complexes is similarity assessment between models and their reference complex structures. Existing methods primarily rely on distance-based metrics and generally do not consider important functional features of the complexes, such as interface hydrogen bonds that are critical to specific protein–DNA interactions. Here, we present a new scoring function, ComparePD, which takes interface hydrogen bond energy and strength into account besides the distance-based metrics for accurate similarity measure of protein–DNA complexes. ComparePD was tested on two datasets of computational models of protein–DNA complexes generated using docking (classified as easy, intermediate, and difficult cases) and homology modeling methods. The results were compared with PDDockQ, a modified version of DockQ tailored for protein–DNA complexes, as well as the metrics employed by the community-wide experiment CAPRI (Critical Assessment of PRedicted Interactions). We demonstrated that ComparePD provides an improved similarity measure over PDDockQ and the CAPRI classification method by considering both conformational similarity and functional importance of the complex interface. ComparePD identified more meaningful models as compared to PDDockQ for all the cases having different top models between ComparePD and PDDockQ except for one intermediate docking case.

KEYWORDS
complex similarity assessment, homology modeling, hydrogen bond energy, hydrogen bonds, protein–DNA complex, protein–DNA docking

1 | INTRODUCTION

Knowledge of protein–DNA complex structures is critical to understanding their roles in important biological processes such as regulation of gene expression. The structures of most protein–DNA complexes, however, remain unsolved due to technical challenges in experimental methods.[1–3] To address this issue, in silico prediction of three-dimensional structures of protein–DNA complexes is considered a valuable alternative in applications such as structure-based, computer aided drug discovery.[4,5] Despite efforts by the research community, computational modeling of complex macromolecular interactions remains a challenging problem.[6–11]

A key step in the development and evaluation of computational modeling methods is to assess the structural similarity between the predicted models and the experimentally solved reference structures. For individual proteins, several similarity assessment metrics such as RMSD (for root mean square deviation), TM-score (for template modeling score), and GDT_TS (for global distance test total score), have been developed for model comparison and have been used in the biennial CASP (Critical Assessment of protein Structure Prediction) competitions.[12–14] For comparison of complex structures such as protein–DNA complexes, similarity assessment is far more challenging because it not only needs to compare the structural similarity of the individual components, but more importantly, the similarity of the interfaces between interacting components. The community wide CAPRI (Critical Assessment of PRedicted Interactions) experiment assesses the prediction performance by grouping complex models into four discrete categories based on three distance-based metrics, iRMSD (interface RMSD), lRMSD (ligand RMSD), and F_{nat} (the fraction of the contacts in the native structure that is reproduced in the model). Model qualities are classified as high, medium, acceptable quality or incorrect based on different ranges of each of the three metrics (Table S1).[15]

The CAPRI's initial quality assessment method has served well as a standard protocol for evaluating protein–protein complexes at CAPRI experiments. However it is not suitable for other types of protein–ligand complexes including protein–peptide and protein–DNA complexes. For example, protein–peptide complexes have a smaller interface area than protein–protein complexes. As such, the CAPRI organizers modified the distance cut-offs to accurately reflect the smaller interface area of protein-peptide complexes.[15–17] Another problem lies in the intrinsic disadvantage of the RMSD-based metrics. RMSD is highly sensitive to conformational changes since each position is treated equally and the RMSD score can be misleading for complexes with larger flexible loops.[18–28] In addition, the classification of models into four groups is rather broad and not sensitive to minor critical differences in models. To overcome this issue, Basu and Wallner developed a continuous scoring function, DockQ, for protein–protein docking model similarity assessment with a score in the range of 0–1 by combining the three individual scoring metrics with scaled RMSD values.[18] While DockQ produces a continuous score that can be used for model ranking, a major limitation though is that the parameters in DockQ were derived for assessing protein–protein complex models and biologically relevant interface features of the complexes such as hydrogen bonds (HBs) were not considered.[18,20,29]

Hydrogen bonds are weak interactions that are significantly more prevalent in protein–DNA complexes than those in protein–protein and protein–peptide complexes.[30] More importantly, the hydrogen bonds between protein sidechains and DNA bases are crucial for protein–DNA binding specificity.[31–33] We have previously demonstrated that incorporating a hydrogen bond energy term in a scoring function improves the prediction of transcription factor binding sites and considering the number of hydrogen bonds in the models can improve protein–DNA docking prediction.[34–37] The existing complex similarity assessment methods under-explore the role of interface hydrogen bonds. While Marcu et al. recently suggested $fnat_{hb}$, the number of conserved hydrogen bonds in models, for model comparison with the native complex structures,[38] there are two issues for methods that rely only on the number of annotated hydrogen bonds based on a single distance/angle or an energy cut-off. The first is that hydrogen bonds of different strength are treated equally.[39,40] Though this might work for complexes with mostly strong hydrogen bonds such as protein–protein and protein–peptide complexes, it is probably not suitable for protein–DNA complexes that have a unique, almost equal distribution of weak and strong hydrogen bonds as we demonstrated recently.[30] Secondly, since hydrogen bonds are identified based on a single cutoff threshold, either an energy cutoff[41] or a combination of distance/angle cutoff,[42] a small difference of hydrogen bond energy or distance/angle may result in different annotations. For example, an energy threshold of −0.6 kcal/mol is generally suggested for protein–DNA complexes by the author of FIRST.[41] However, two hydrogen bonds with very similar energy, say −0.595 and −0.605 kcal/mol, respectively, would result in 0 and 1 hydrogen bond, respectively.

In this study, we developed a novel continuous function, ComparePD, to assess the similarity of protein–DNA complexes with a weighted hydrogen bond energy-based term in combination with other distance-based metrics, which considers both conformational similarity and functional importance of the protein–DNA complex interface. To the best of our knowledge, this is the first approach that incorporates different strengths of hydrogen bonds into complex model comparison and assessment. ComparePD showed much improvement over PDDockQ, a modified version of DockQ tailored for protein-DNA complexes.

2 | MATERIALS AND METHODS

2.1 | Datasets of protein–DNA complex models

Protein–DNA complex models were generated using two methods; homology modeling and docking.[43] The homology modeling dataset comprises 75 models of 5 non-redundant homeodomain complexes, which share less than 35% protein sequence identity (Table S2). Templates of high structural quality and varying sequence similarity ranging between 35% and 70% were selected for each target complex and five models per template were generated for each target. Since the existing homology modeling methods do not model interfaces of protein–DNA complex structures, structurally aligned homology models of proteins to the native complex were combined with the native DNA to generate the complex models as shown in Figure S1. MODELLER was used for comparative protein structure modeling and TM-align was used for structural alignment.[44–46]

HADDOCK-MARTINI based protein–DNA docking models were obtained from the publicly available repository.[43,47] The original dataset comprises docking models of 43 complexes from the protein–DNA docking benchmark generated using MARTINI force field and ranked according to HADDOCK score.[47,48] These complex models were filtered using methods as described in our previous study.[30] Fifteen complexes with internal missing residues were discarded. Two

TABLE 1 Energy bins for each category of hydrogen bonds energy (HBE) and their corresponding weights.

Category	HBE range (kcal/mol)	Weights
I	$-0.6 \le \text{HBE} < -0.1$	0.5
II	$-1.5 \le \text{HBE} < -0.6$	0.8
III	$\text{HBE} < -1.5$	1

additional cases were removed because the hydrogen bond identification program (see next section) failed to annotate hydrogen bonds in the complexes. The filtering process resulted in a final dataset of 25 complexes. The top 20 models for each complex by the HADDOCK score were selected for comparison. Bonvin et al. have classified the complexes in the benchmark based on the docking difficulty as easy, intermediate and difficult cases.[43,44] The dataset used in this study comprises seven easy cases: 1by4, 1fok, 1hjc, 1h9t, 1mnn, 1rpe, and 3cro; twelve intermediate cases: 1azp, 1a74, 1ddn, 1ea4, 1f4k, 1g9z, 1kc6, 1r4o, 1vas, 1z9c, 2fio, and 2irf; and six difficult cases: 1qrv, 1rva, 2oaa, 2fl3, 3bam, and 7mht.

2.2 | Hydrogen bond energy

REDUCE was used to add hydrogen atoms to structural files for hydrogen bond calculations.[49] While adding hydrogen atoms, REDUCE also performs extensive optimization of the structures based on their local geometry. FIRST (Floppy Inclusion and Rigid Substructure Topography) was used to calculate the hydrogen bond energy using Equation (1).[40,41]

$$E_{\mathrm{HB}} = V_0 \left\{ 5\left(\frac{d_0}{d}\right)^{12} - 6\left(\frac{d_0}{d}\right)^{10} \right\} F(\theta,\phi,\varphi) \tag{1}$$

where d is the donor–acceptor distance. d_0 (2.8 Å) and V_0 (8 kcal/mol) represent the equilibrium distance and well-depth, respectively.[41] The angle term is estimated by exploring the hybridization state of the acceptor and donor atoms.[41] The hydrogen bonds were classified into three energy categories based on previous studies (Table 1).[30,41,50,51] Weights of 0.5, 0.8, and 1.0 are arbitrarily assigned to reflect the strength of hydrogen bonds due to the lack of sufficient number of protein–DNA complexes for training.

2.3 | PDDockQ

DockQ is a previously developed similarity assessment score for protein–protein complexes based on Fnat, iRMSD, and lRMSD.[18] We modified the parameters in DockQ for protein–DNA complexes to reflect the differences between protein–protein complexes and protein–DNA complexes with respect to different interface areas.[30] RMSD values in DockQ were scaled using an inverse square scaling method to account for two problems (Equation 2). First, a near-native model has higher Fnat and lower RMSD values. Second, arbitrarily large RMSD values can be misleading. Basu et al. have shown that the inverse square scaling of iRMSD and lRMSD provides a more sensitive discrimination between the qualities of protein–protein complex models.[18]

$$\mathrm{jRMSD}_{\mathrm{scaled}} = \frac{1}{1+\left(\frac{\mathrm{jRMSD}}{d_j}\right)^2} \tag{2}$$

where jRMSD is for iRMSD or lRMSD and d_j represents the corresponding scaling factors, d_i for iRMSD and d_l for lRMSD, respectively. The scaling factors in DockQ were optimized using grid search on a large number of protein–protein complexes.[18] We estimated $d_i \approx 1.04$ and $d_l \approx 2$ by comparing the differences of average interface areas between protein–protein and protein–DNA complexes as we used in our previous study[30] (see Appendix S1). PDDockQ is then defined as follows (Equation 3).

$$\mathrm{PDDockQ} = \frac{F_{\mathrm{nat}} + \mathrm{iRMSD}_{\mathrm{scaled}} + \mathrm{lRMSD}_{\mathrm{scaled}}}{3} \tag{3}$$

For F_{nat} calculation, a contact is defined between two heavy atoms if they are separated by a distance of 4.5 Å or less and the interface is defined as pairs of heavy atoms from the protein and DNA within 10 Å of each other. iRMSD is calculated with C_β atoms of proteins and N1 (for bases *C* and *T*) or N9 (for bases A and G) atoms of DNA of the interface. lRMSD is based on N1 (for bases C and T) and N9 (for bases A and G) interface atoms of DNA.

2.4 | ComparePD: a continuous function for assessing protein–DNA complex similarity

ComparePD is a linear continuous function for comparing protein–DNA complex similarity by calculating the mean value of the distance-based features, Fnat, the scaled iRMSD and lRMSD, and a novel weighted hydrogen bond energy-based score, $\mathrm{Composite}_{\mathrm{HBE}}$ (Equation 4). There are two major considerations for assigning equal weights to the four terms. One is for fair performance comparison with the modified DockQ scoring function PDDockQ, which adopts equal weights for the three distance-based metrics.[18] The other is that currently there are not sufficient data for training optimal weights.

$$\mathrm{ComparePD} = \frac{F_{\mathrm{nat}} + \mathrm{iRMSD}_{\mathrm{scaled}} + \mathrm{lRMSD}_{\mathrm{scaled}} + \mathrm{Composite}_{\mathrm{HBE}}}{4} \tag{4}$$

$\mathrm{Composite}_{\mathrm{HBE}}$ is a weighted hydrogen bond energy-based score between a protein–DNA complex model and a reference complex. The key part of this score is the calculation of F_{HBnat} and F_{HBSPnat}, for the fraction of total hydrogen bonds and the fraction of the specific (SP) sidechain-base hydrogen bonds reproduced in the models, respectively (Figure 1 and Equation 5). The weight for each reproduced hydrogen bond in the model and native complex is determined based on their HB energy as shown in Table 1. W_{HB} is the ratio of weights between the corresponding hydrogen bonds in the native

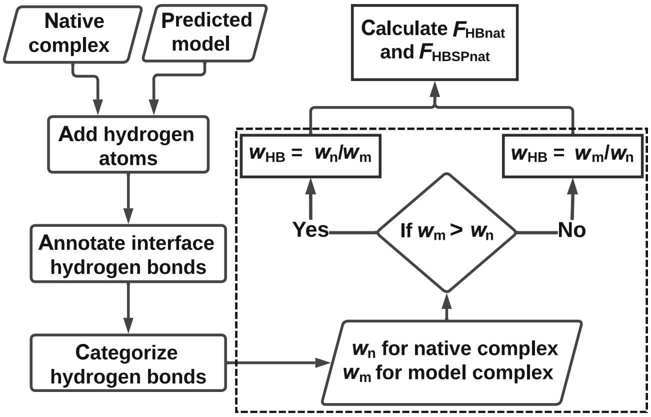


FIGURE 1 A flowchart for calculating F_{HBnat} and $F_{HBSPnat}$ based on weighted hydrogen bond energy. The box with dashed line represents calculation for each captured native hydrogen bond in the model.

(w_n) and the model (w_m) complexes, respectively (Figure 1). W_{HB} equals w_n/w_m if w_m is larger than w_n, otherwise it equals w_m/w_n to ensure that the value of W_{HB} falls between 0 and 1. F_{HBnat} is the sum of W_{HB} for the reproduced hydrogen bonds across the interface of the model normalized by the weighted sum of total number of hydrogen bonds C_{HB} in each energy category (Table 1) in the reference complex (Equation 5).

$$F_{HBnat} = \frac{\sum W_{HB}}{\sum_{i=1}^{3} W_i (C_{HB})_i} \tag{5}$$

$F_{HBSPnat}$ is calculated similarly for capturing the reproduced sidechain-base hydrogen bonds in the models. Finally, a composite score of F_{HBnat} and $F_{HBSPnat}$ is calculated to reflect overall capture of hydrogen bonds in the model when compared to the native complex (Equation 6).

$$\text{Composite}_{HBE} = \frac{w_1 C_{HB} F_{HBnat} + w_2 C_{HBSP} F_{HBSPnat}}{w_1 C_{HB} + w_2 C_{HBSP}} \tag{6}$$

where C_{HB} and C_{HBSP} represent the total number of interface hydrogen bonds and sidechain-base hydrogen bonds in the native complex respectively, $w_1 = 0.3$ and $w_2 = 1 - w_1 = 0.7$. w_1 is estimated by comparing the average number of the sidechain-base hydrogen bonds to all interface hydrogen bonds in a pooled non-redundant dataset of highly specific (HS) and multi-specific (MS) protein–DNA complexes.[32] The distributions of the ratios in the HS and MS individual datasets are also similar as shown in Figure S2. Higher weight is assigned to sidechain-base hydrogen bonds because of their important role in protein–DNA binding specificity. The sum of the weighted F_{HBnat} and $F_{HBSPnat}$ is normalized by the weighted total number of interface hydrogen bonds and sidechain-base hydrogen bonds in the native complex.

2.5 | Statistical tests

To compare the reproduced number of hydrogen bonds between the top models from ComparePD and PDDockQ, Wilcoxon rank sum test for independent samples and Wilcoxon signed rank test for paired samples were employed to assess if there are significant differences between the ComparePD and PDDockQ top model selections.

3 | RESULTS

In complex structure similarity assessment, one important question is how to determine one complex model is better than the other. Similar to protein structure comparison methods, it is hard to define "better" in similarity measures without considering the context. Different scoring functions, such as RMSD, TM-score, and GDT_TS, have been developed for protein structure comparison with a specific focus or for different application purposes. Each scoring function has its pros and cons. For example, while RMSD values can indicate the overall similarity or difference between two structures, TM-score was developed for detecting core structure similarity at fold level, which RMSD sometimes fails to reveal due to the large RMSD from the flexible regions.[12,52] In terms of protein–DNA complex structure comparison, while the distance-based approaches, including iRMSD, lRMSD, and F_{nat}, can indicate the similarity of interface conformation and contacts between the two complexes but they do not fully capture the functionally important features on the interface, such as hydrogen bonds. This is because distance-based metrics treat each position at the interface equally regardless of the functional importance. Since we know that hydrogen bonds play a significant role in the binding specificity of protein–DNA complexes, in addition to looking for a smaller interface conformational difference, a near-native model should also capture the hydrogen bond interactions in terms of both the number of hydrogen bonds and the strength for useful downstream applications. Therefore, if two models, model A and model B, selected by two different scoring functions, have similar distance measures, but model A better captures the hydrogen bonds in the reference structure, we consider model A is a better model. In this section, we performed detailed analyses by considering both the distance-based similarity and hydrogen bond conservation to determine which method, PDDockQ or ComparePD, picks a better model for each case in the docking and homology modeling datasets.

3.1 | Overview of the docking and homology protein–DNA complex models

The docking models from HADDOCK protein–DNA docking benchmark were first evaluated and categorized according to the CAPRI criteria. The best model category for most of the cases is medium (76%) and a few cases have only acceptable quality models (24%) (Figure 2A, Table S3). In terms of docking difficulty, all of the easy cases, 58.3% of intermediate cases and 83.3% of difficult cases have

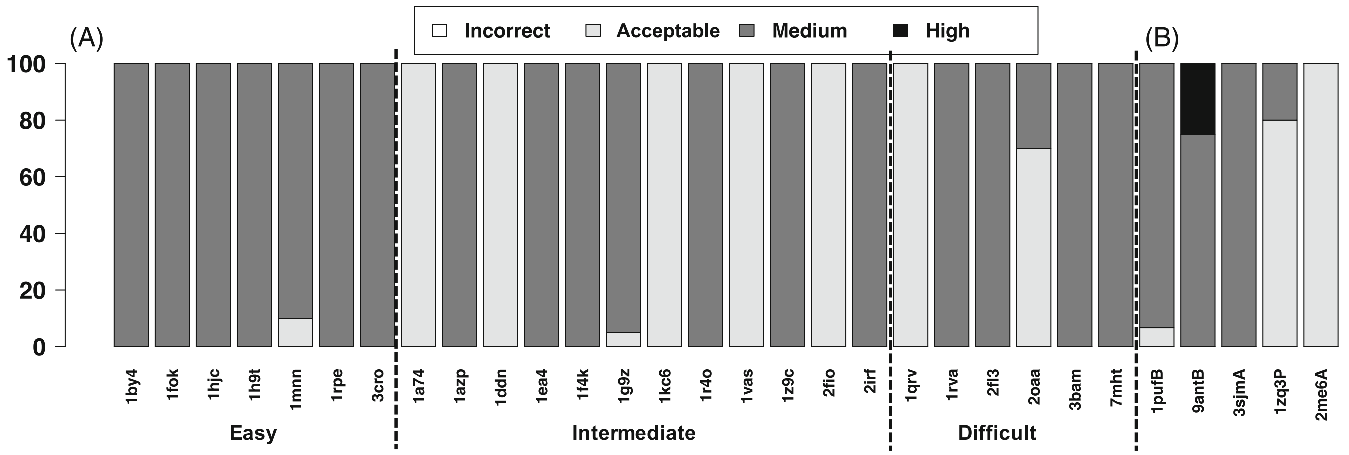


FIGURE 2 CAPRI-based classification of (A) all the docking models classified as easy, intermediate and difficult cases; and (B) all homology models.

best models in the medium quality while 41.7% of intermediate cases and 16.7% of difficult cases have top models in the acceptable category (Table S3). There are no cases with any models in the high or incorrect category (Figure 2A). For homology protein–DNA complex models, one of the five case, 9antB, has high quality models, three cases have a number of medium quality models (1zq3P, 3sjmA, and 1pufB) and one case (2me6A) only has acceptable quality models (Figure 2B, Table S3).

ComparePD identifies the same top model as PDDockQ in 48% of the docking cases (Figure 3A). Figure S3 shows three such examples for one easy (3cro), one intermediate (1z9c), and one difficult (7mht) case. In 3cro, even though both methods identify model 6 as the top model, ComparePD helps improve confidence in selection by clearly distinguishing model 6 from other models with similar PDDockQ values to model 6 (Figure S3A). As for homology protein-DNA complex models, ComparePD picks the same top model as PDDockQ in two (40%) cases (Figure 3B). The two scoring metrics agree on the selection of the top models in 1zq3P and 2me6A (Figure S4). The top model in 1zq3P by both scores is the third model generated using 4rduA as a template that has 45% sequence identity with the target (Figure S4A). For 2me6A, the top model is the third model generated using 1fjlA as a template that has 40% sequence identity with the target protein sequence (Figure S4B). The two scores have similar trends for all the 20 models indicating stronger agreement between them.

3.2 | Comparison of the different top models selected by ComparePD and PDDockQ

There are a total of 16 cases (13 docking cases and 3 homology modeling cases) that have different top models between ComparePD and PDDockQ (Figure 3). ComparePD selects a different top model from PDDockQ in 52% of docked protein–DNA complexes, 28.6% of easy, 75% of intermediate and 33.3% of difficult cases (Figure 3A). In three of the five homology modeling cases (9antB, 1pufB, and 3sjmA), ComparePD and PDDockQ pick different top models. We compared the number of interface hydrogen bonds, including all hydrogen bonds (Figure 4A) and the sidechain-base interface hydrogen bonds (Figure 4C) that are reproduced in the top models selected by ComparePD and PDDockQ, respectively. Independent Wilcoxon rank sum tests show that the top models from ComparePD capture significantly more total hydrogen bonds (p-value = .022, Figure 4A) and sidechain-base hydrogen bonds (p-value = .038, Figure 4C) than those from the top PDDockQ models. We also performed paired Wilcoxon signed rank tests, which reveal that the differences of the number of hydrogen bonds reproduced between the top models selected by ComparePD and PDDockQ, respectively for each complex are also significantly different for all hydrogen bonds (p-value = .0016, Figure 4A) and the sidechain-base hydrogen bonds (p-value = .0103, Figure 4C).

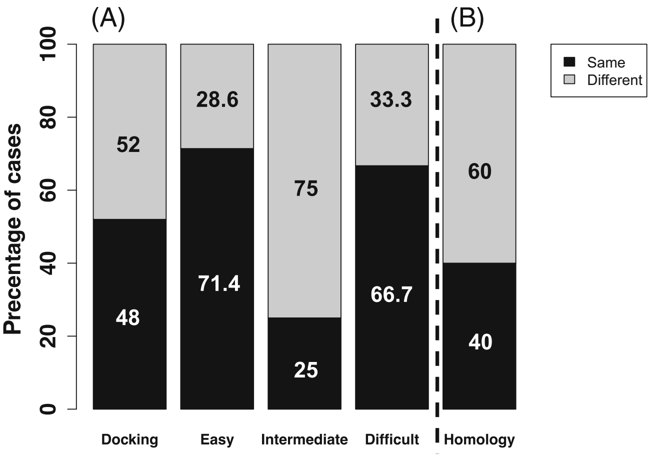


FIGURE 3 Comparison of scoring methods for the selection of top models in (A) docking dataset with easy, intermediate and difficult cases; and (B) homology modeling dataset.

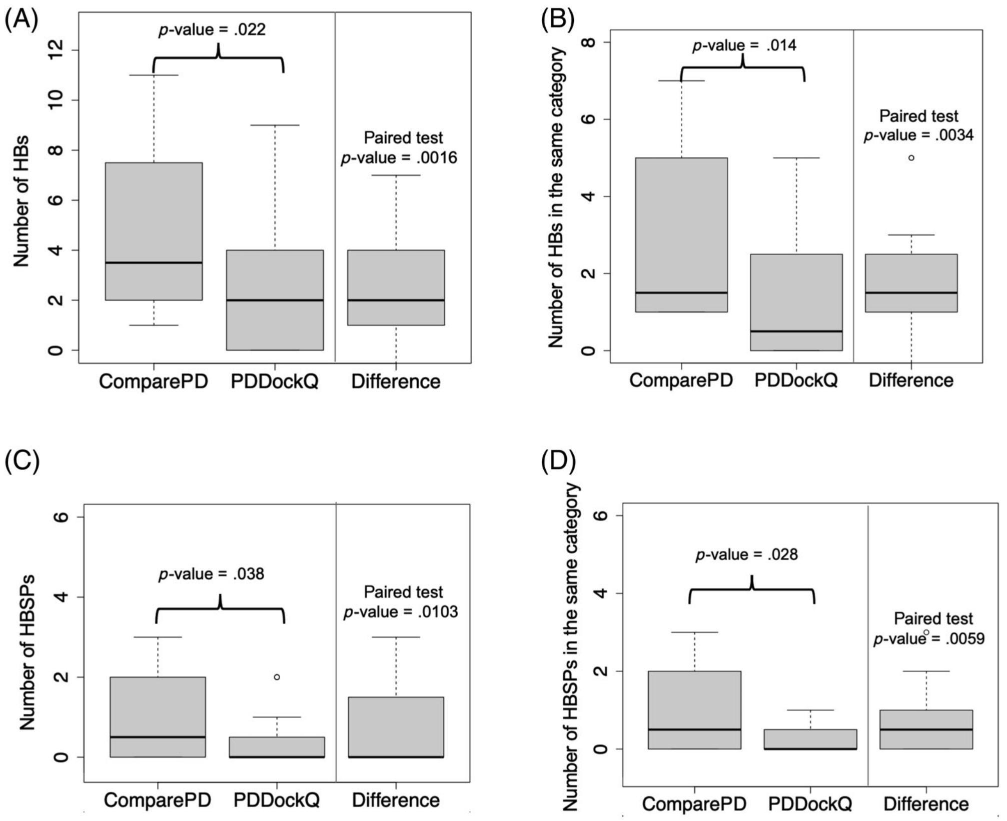


FIGURE 4 Comparison of the number of reproduced hydrogen bonds in the top models based on ComparePD and PDDockQ scores. (A) The number of total interface hydrogen bonds; (B) the number of total interface hydrogen bonds in the model that are in the same hydrogen bond energy category as those in the reference structure; (C) the number of total interface sidechain-base hydrogen bonds; (D) the number of total interface sidechain-base hydrogen bonds in the model that are in the same hydrogen bond energy category as those in the reference structure. *p*-Values from the Wilcoxon rank sum tests between ComparePD and PDDockQ and from Wilcoxon signed rank tests for the paired differences are shown.

TABLE 2 Comparison of the protein–DNA docking models by ComparePD and PDDockQ.

Category	The same top model from ComparePD and PDDockQ	ComparePD is better than PDDockQ	PDDockQ is better or comparable to ComparePD
Easy	1fok, 1h9t, 1hjc, 1rpe, 3cro	1by4, 1mnn	
Intermediate	1r4o, 1z9c, 2irf	1a74, 1azp, 1ddn, 1ea4, 1g9z, 1kc6, 1vas, 2fio	1f4k
Difficult	2fl3, 2oaa, 3bam, 7mht	1qrv, 1rva	

To further evaluate the performances between ComparePD and PDDockQ, we compared the number of hydrogen bonds that are reproduced in the models and are in the same hydrogen bond energy category as those in the reference complexes (Figure 4B for total hydrogen bonds and Figure 4D for sidechain-base hydrogen bonds). Both Wilcoxon rank sum tests and signed rank tests show that the top ComparePD models pick up more hydrogen bonds in the same energy category as the reference structures than the top PDDockQ models. Overall, the top docking models selected by ComparePD are better than PDDockQ in 100% of easy cases and difficult cases, and 88.9% of intermediate cases because they capture more hydrogen bonds in the reference complex interface while having similar or comparable distance-based measures (Table 2).

Below we describe detailed analyses of three docking examples, 1mnn, 1ea4, and 1rva for easy, intermediate and difficult cases, respectively (Figure 5), and one homology modeling case 9antB (Figure 6) by comparing the distance-based metrics, the number of hydrogen bonds and hydrogen bond energy between the top models selected by ComparePD and PDDockQ, respectively.

1mnn: Complex 1mnn is an easy docking case in the protein–DNA docking benchmark. ComparePD selects model 9 as the top model whereas PDDockQ considers model 2 as the most similar to the native complex (Figure 5A). Model 9 and model 2 have similar distance-based metrics, which results in similar PDDockQ scores for these two models. However, model 9 has a much higher ComparePD score than model 2 as it reproduces seven interface hydrogen bonds that appear in the native complex interface, whereas model 2 selected by PDDockQ captures no hydrogen bonds (Figures 5A and 7). Moreover, four of the seven hydrogen bonds (ARG277-G4, ARG79-G25 and the two hydrogen bonds between ARG65-G23 pairs) in model 9 also have similar hydrogen bond energy and are placed in the same hydrogen bond energy category as those on the native interface. The combination of the distance-based data and the hydrogen bond energy-based analysis clearly demonstrates that model 9 is better than model 2.

1ea4: Figure 5B shows the comparison of different scores for an intermediate case 1ea4. Model 9 is the top model according to the ComparePD score whereas model 12 is the top model based on

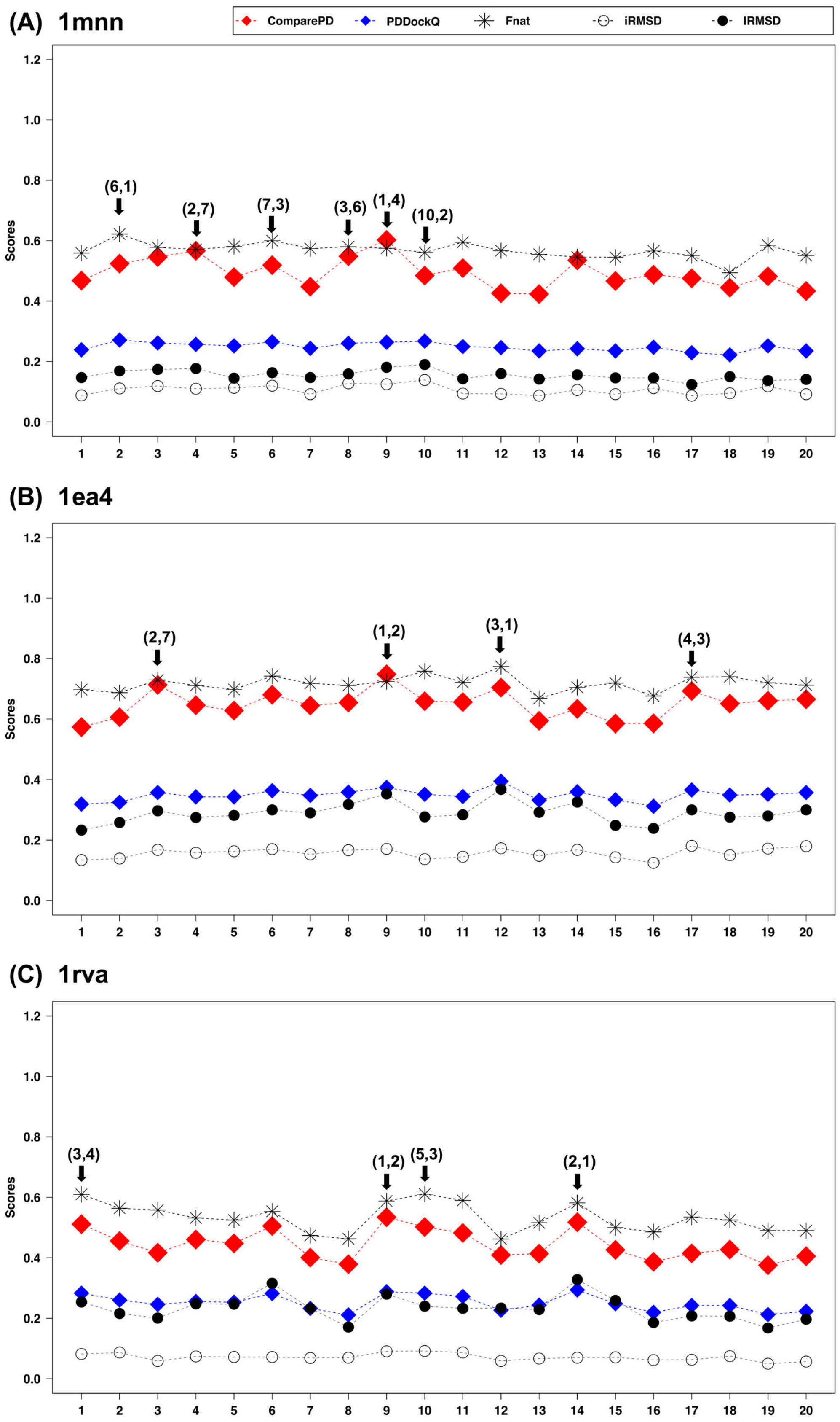


FIGURE 5 Comparison of docking models scored by ComparePD and PDDockQ along with their individual Fnat, scaled iRMSD and lRMSD scores for (A) an easy case 1mnn, (B) an intermediate case 1ea4; and (C) a difficult case 1rva. Top three models from ComparePD and PDDockQ are highlighted, and the corresponding ranks are reported as (ComparePD, PDDockQ) for each complex. iRMSD and lRMSD are scaled values.

PDDockQ. Model 12 has very similar iRMSD and slightly better Fnat and lRMSD, and therefore has a slightly better PDDockQ score than model 9. However, model 9 retains 11 out of 12 (91.7%) native interface hydrogen bonds whereas model 12 only captures 5 (41.6%) native hydrogen bonds (Figure S5). The first five rows show hydrogen bonds captured in the model selected by ComparePD but not in

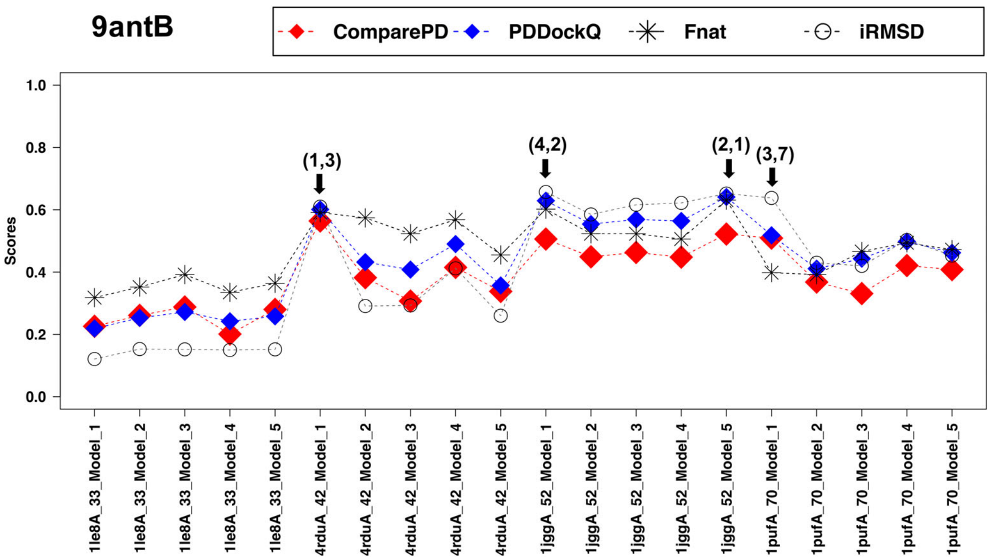


FIGURE 6 Comparison of homology models scored by ComparePD and PDDockQ along with their individual Fnat, scaled iRMSD, and lRMSD scores for 9antB. iRMSD represents scaled values.

PDDockQ. There are two hydrogen bonds between SER29-T37, only one of them is captured by model 12 whereas both are found in model 9. The energy categories of 9 out of 11 reproduced hydrogen bonds in model 9 do not change. In model 12, however, only 3 out of 5 captured hydrogen bonds stays in the same energy category. Model 9 captures 75% (9 out of 12) of the native complex interface in the same hydrogen bond energy category while model 12 only gets 25% (3 out of 12). Taken together, model 9 is a better complex model than model 12.

1rva: 1rva is an example of difficult docking cases in the HADDOCK protein–DNA docking benchmark. The top selection by ComparePD is model 9 whereas the top PDDockQ choice is Model 14 (Figure 5C). Model 14 and model 9 are the top two models by PDDockQ scores with highly similar PDDockQ scores. Therefore, it is very difficult to distinguish between the two models based on PDDockQ scores alone. Both models have similar iRMSD and Fnat. Model 14 actually has a better lRMSD than model 9. However, when hydrogen bond energy is taken into account, model 9 clearly has a better ComparePD score than model 14. Figure S6 shows a comparison of hydrogen bonds and their energies in both models and the native complex structure. Model 9 captures 8 out of 9 (88.9%) native interface hydrogen bonds whereas model 14 only has 2 of them (22.2%). Seven native hydrogen bonds, TYR338-C21, ASN312-C9, ASN184-G16, GLY183-G16, GLY427-G4, ASN67-T23, and SER111-T6 are captured in model 9 but not in model 14. The energy category of these hydrogen bonds except for GLY427-G4 is also conserved in model 9. One strong native hydrogen bond, TYR338-T22, is reproduced in both models in the same energy category in both models. Hydrogen bond between LYS118 and A5 only appears in model 14. Overall, model 9 is considered a better model because it reproduces more hydrogen bonds with the same strength (Figure S6) while the distance-based scores are indistinguishable (Figure 5C).

9antB: 9antB is a homology modeling case. Both ComparePD and PDDockQ pick one of the high quality models as classified by CAPRI. But a detailed examination of hydrogen bonds and the distance-based scores in each of these models indicate that ComparePD selects a better overall model than the model selected by PDDockQ (Figures 6 and S7). ComparePD ranks homology model 1 generated using template 4rduA (sequence identity 42%) as the top model, whereas PDDockQ selects model 5 generated using template 1jggA (sequence identity 52%) as the top model (Figure 6). The top model selected by ComparePD captures 7 of the 8 native interface hydrogen bonds. Five of these hydrogen bonds are also in the same hydrogen bond energy categories (5/8 = 62.5%), including four strong hydrogen bonds: ARG55-C406, ARG33-A404, ARG7-T518, and ASN53-A520. The two hydrogen bonds whose energy category changes in the model selected by ComparePD (GLN8-A519 and ARG55-C406) are of high energy, weaker hydrogen bonds in the native complex. Sidechain conformations of residues involved in hydrogen bonding are similar between the top ComparePD model and the native complex structure, resulting in similar hydrogen bond energy. The top model from PDDockQ captures only 5 of 8 hydrogen bonds. Four of the hydrogen bonds (ARG33-A404, ARG45-T521, and two hydrogen bonds in ASN53-520) are of low energies. And only one of these four hydrogen bonds, ARG45-T521, has the same hydrogen bond energy category as that in the native complex (1/8 = 12.5%). Even though the Fnat and iRMSD values of the top PDDockQ model (0.631 and 0.878 Å) are slightly better than the top ComparePD model (0.591 and 0.959 Å), poor recovery of hydrogen bonds energy (62.5% vs. 12.5% capture of hydrogen bonds in the same energy category) results in lower overall score for the top PDDockQ model. We also note that in complex model generation using homology techniques, there is some weak correlation between sequence identity (between the target sequence and the homology template sequence) and the PDDockQ or ComparePD score indicating that the templates of higher sequence identity in homology modeling can result in better complex models (Figure S8).

Similar detailed structural analysis for other cases shows similar results that the top models selected by ComparePD have captured more hydrogen bonds in terms of both the number of hydrogen bonds and the conservation of energy categories despite having similar

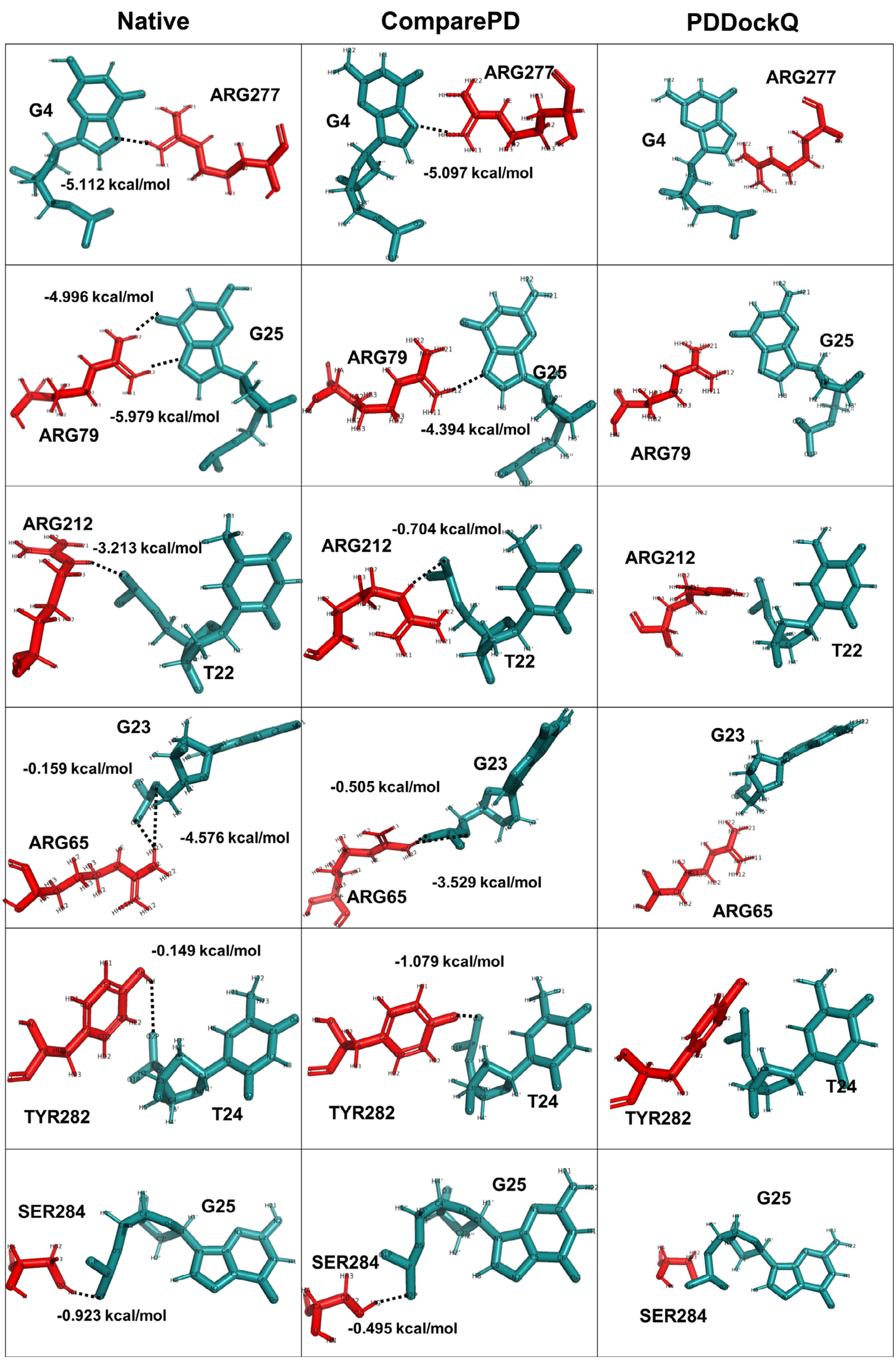


FIGURE 7 Detailed analysis of the interface hydrogen bonds and their energy in the native complex 1mnn and the top models selected by ComparePD and PDDockQ, respectively.

distance-based scores to the top models selected by PDDockQ. One exception is 1f4k where PDDockQ identified a comparable or slightly better top model than that from ComparePD in terms of hydrogen bonds and other distance related metrics (Table 2, Figures S9–S10). The top models identified by PDDockQ and ComparePD have similar Fnat and RMSDs and differ only by one hydrogen bond (Figure S10). A detailed comparative structural analysis of hydrogen bonds in the top models and native complexes reveals that the conformations from the two selections are very close to each other. Even though the number of hydrogen bonds in the PDDockQ selection is more than that in the ComparePD selection (3 to 2), the energy conservation of the former is worse. The top model selected by ComparePD for 1f4k, model 14, captures two hydrogen bonds and more importantly, both are in the same hydrogen bond energy category as those in the reference complex (Figure S10). While the top model selected by PDDockQ, model 5, captures one more hydrogen bond than model 14, it only has one of them in the same hydrogen bond energy range when compared to the hydrogen bonds on the native complex interface. Therefore, the top models selected by ComparePD and PDDockQ are comparable since there is no clear indication that one is better than the other.

4 | DISCUSSION

Computational methods for predicting protein–DNA complex structures not only can fill the sequence-structure gap, they can also help time-sensitive applications such as computer aided structure-based drug design. Complex structure modeling, a very challenging task, requires accurate capture of interface features and reliable assessment of the model quality. Unlike protein structure prediction where new models and the corresponding methods can be assessed by scores such as TM-score, GDT_TS, and RMSD score, no standard criteria exist for comparing protein–DNA complexes.[12,53,54]

We present here ComparePD, a novel scoring function to assess the similarity of protein–DNA complex structures by incorporating hydrogen bond energy. Hydrogen bonds are important to the binding specificity between protein and DNA. Recent studies indicated that incorporating biologically relevant measures in the development of in silico structure prediction of complexes can help improve performance.[55] We demonstrated that a combination of conventional interface features F_{nat}, iRMSD and lRMSD with the hydrogen bond energy in ComparePD captures both structural similarity and some functional similarity and therefore better describes the similarity of protein–DNA complex models. Our approach by comparing the hydrogen bond energy or strength overcomes the limitations of simply comparing the number of reproduced hydrogen bonds in model interface. This is important because protein–DNA complexes are intrinsically dynamic and interface hydrogen bonds have similar distributions between strong and weak ones.[56–58] ComparePD treats native hydrogen bonds of varying strengths differently by assigning higher weights to stronger hydrogen bonds and lower weights to weaker ones. Unlike the cutoff-based identification of hydrogen bonds, which would result in a gain or a miss for a small hydrogen bond energy difference, our approach takes account of the dynamic nature into consideration without penalizing small shifts in hydrogen bond energy. On the other hand, a large change in hydrogen bond energy compared to the native hydrogen bond will score relatively lower. While the use of hydrogen bond numbers for similarity assessment of complexes has previously been suggested, to the best of our knowledge, this is the first time a similarity assessment method based on different hydrogen bond energy ranges has been explored to compare protein–DNA complexes.

The key element in our new scoring function is $\text{Composite}_{\text{HBE}}$, which is calculated by the weighted combination of F_{HBnat} and F_{HBSPnat} with the latter having a higher weight due to its importance in specific protein–DNA binding (Equation 6). F_{HBnat} and F_{HBSPnat} are normalized by the weighted sum of total hydrogen bonds C_{HB} or C_{HBSP} in each energy category for two considerations (Equation 5). One considers the strength of hydrogen bonds in contributions. Another consideration is that the three categories have different ranges, from the narrow ranges (weak hydrogen bonds) to larger ranges (strong hydrogen bonds). Therefore reproducing the weak hydrogen bonds in the same category between the model and the reference structure is more difficult or challenging than the stronger ones that have a bigger range and better chance to be in the same category. The value distributions of F_{HBNat} and F_{HBSPNat} in top models selected by ComparePD are shown in Figure S11.

The performance of ComparePD for comparison of models generated using two different computational methods, homology modeling and docking, has been demonstrated. ComparePD has consistently identified a better model than existing metrics for models generated from both approaches indicating its general capability in assessing complex model similarity. The benefit of the combination of the distance-based and hydrogen bond energy-based metrics is that it captures both the overall interface structural similarity as well as the important functional feature. The relationship between ComparePD and PDDockQ can be analogized to the difference between RMSD score and TM-score for protein structure comparison. While both RMSD and TM-score provide overall similarity scores for two protein structures, TM-score is more useful to identify two structures with fold level similarity by weighting more for the core similarity. ComparePD is designed to score the conformational similarity as well as functionally important interface features, the hydrogen bonds between protein and DNA.

Our results show that the improved performance of ComparePD over PDDockQ is better for the intermediate docking cases than those in the easy and difficult sets (Table 2). This is not surprising. As shown in Figure 2, there are more cases with lower quality top models in the intermediate set (41.7%) than those in the easy (0%) and difficult (16.7%) sets. In most cases that have higher quality, near native models, PDDockQ is able to identify the same best model as ComparePD because PDDockQ may capture these reproduced hydrogen bonds implicitly with the distance-based metrics. However, for lower quality models, adding hydrogen bond information can help identify the near native models (Figure 3).

Because ComparePD is a continuous scoring function that has better performance, it can facilitate the development of new methods for modeling and evaluating protein–DNA complex models, and can be applied to machine learning based methods for assessing the quality of protein–DNA complexes. There is a potential that the performance of ComparePD can be further improved. For example, due to the limited availability of data, it is not practical to perform weight training and optimization for the four metrics in ComparePD. When larger datasets become available in the future, the weights can be optimized to further improve the accuracy. In addition, it has been demonstrated that π–π and cation–π interactions play important roles in protein–DNA interactions.[32,34,37,59–63] Not only can π–π interactions contribute to the binding affinity and complex stability, they also play a role in conferring specific protein–DNA recognition. We previously demonstrated that considering protein–DNA π–π interactions explicitly helps improve structure-based prediction of transcription factor binding sites.[34] A recent study also suggested the implication of incorporating π–π interactions in the development of scoring functions for docking.[63] As for assessing the interface similarity, on the one hand, considering the conservation of π–π interactions in the models when comparing their interface similarity to the native complex may help. On the other hand, adding it as an additional term may also introduce noise since the number of such interactions in protein–DNA complexes is relatively small. Nevertheless it should be explored in future studies.

FUNDING INFORMATION

This work was supported by the National Science Foundation [DBI-2051491 to J.G].

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Berman HM, Bhat TN, Bourne PE, et al. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*. 2000;7:957-959.
2. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
3. Propper K, Meindl K, Sammito M, et al. Structure solution of DNA-binding proteins and complexes with ARCIMBOLDO libraries. *Acta Crystallogr D Biol Crystallogr*. 2014;70(Pt 6):1743-1757.
4. Hameduh T, Haddad Y, Adam V, Heger Z. Homology modeling in the time of collective and artificial intelligence. *Comput Struct Biotechnol J*. 2020;18:3494-3506.
5. Fan JY, Fu AL, Zhang L. Progress in molecular docking. *Quantitat Biol*. 2019;7(2):83-89.
6. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: overview, current applications, and future perspectives. *Chem Biol Drug des*. 2019;93(1):12-20.
7. Leman JK, Weitzner BD, Lewis SM, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020;17(7):665-680.
8. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303.
9. Launay G, Simonson T. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics*. 2008;9:427.
10. Chen YC. Beware of docking! *Trends Pharmacol Sci*. 2015;36(2):78-95.
11. Haddad Y, Adam V, Heger Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comput Biol*. 2020;16(4):e1007449.
12. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004;57(4):702-710.
13. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins*. 1999;Suppl 3:22-29.
14. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370-3374.
15. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins*. 2007;69(4):704-718.
16. Lensink MF, Velankar S, Wodak SJ. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins*. 2017;85(3):359-377.
17. Parisien M, Freed KF, Sosnick TR. On docking, scoring and assessing protein-DNA complexes in a rigid-body framework. *PLoS One*. 2012;7(2):e32647.
18. Basu S, Wallner B. DockQ: a quality measure for protein-protein docking models. *PLoS One*. 2016;11(8):e0161879.
19. Gao M, Skolnick J. New benchmark metrics for protein-protein docking methods. *Proteins*. 2011;79(5):1623-1634.
20. Xue LC, Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction. *Proteins*. 2014;82(2):250-267.
21. Kufareva I, Abagyan R. Methods of protein structure comparison. *Methods Mol Biol*. 2012;857:231-257.
22. Velazquez-Libera JL, Duran-Verdugo F, Valdes-Jimenez A, Nunez-Vivanco G, Caballero J. LigRMSD: a web server for automatic structure matching and RMSD calculations among identical and similar compounds in protein-ligand docking. *Bioinformatics*. 2020;36(9):2912-2914.
23. Guo F, Zou Q, Yang G, Wang D, Tang J, Xu J. Identifying protein-protein interface via a novel multi-scale local sequence and structural representation. *BMC Bioinformatics*. 2019;20(Suppl 15):483.
24. Das S, Chakrabarti S. Classification and prediction of protein-protein interaction interface using machine learning algorithm. *Sci Rep*. 2021;11(1):1761.
25. Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pac Symp Biocomput*. 2010;293-301.
26. Jain AN. Scoring functions for protein-ligand docking. *Curr Protein Pept Sci*. 2006;7(5):407-420.
27. Janin J, Henrick K, Moult J, et al. CAPRI: a critical assessment of PREdicted interactions. *Proteins*. 2003;52(1):2-9.
28. Pozzati G, Kundrotas P, Elofsson A. Scoring of protein-protein docking models utilizing predicted interface residues. *Proteins*. 2022;90(7):1493-1505.
29. Jandova Z, Vargiu AV, Bonvin A. Native or non-native protein-protein docking models? Molecular dynamics to the rescue. *J Chem Theory Comput*. 2021;17(9):5944-5954.
30. Malik FK, Guo JT. Insights into protein-DNA interactions from hydrogen bond energy-based comparative protein-ligand analyses. *Proteins*. 2022;90(6):1303-1314.
31. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. 2010;79:233-269.
32. Corona RI, Guo JT. Statistical analysis of structural determinants for protein-DNA-binding specificity. *Proteins*. 2016;84(8):1147-1161.

33. Lin M, Guo JT. New insights into protein-DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res*. 2019;47(21):11103-11113.
34. Farrel A, Murphy J, Guo JT. Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics*. 2016;32(12):i306-i313.
35. Takeda T, Corona RI, Guo JT. A knowledge-based orientation potential for transcription factor-DNA docking. *Bioinformatics*. 2013;29(3):322-330.
36. Corona RI, Sudarshan S, Aluru S, Guo JT. An SVM-based method for assessment of transcription factor-DNA complex models. *BMC Bioinformatics*. 2018;19(Suppl 20):506.
37. Farrel A, Guo JT. An efficient algorithm for improving structure-based prediction of transcription factor binding sites. *BMC Bioinformatics*. 2017;18(1):342.
38. Marcu O, Dodson EJ, Alam N, et al. FlexPepDock lessons from CAPRI peptide-protein rounds and suggested new criteria for assessment of model quality and utility. *Proteins*. 2017;85(3):445-462.
39. Hubbard RE, Kamran HM. Hydrogen Bonds in Proteins: Role and Strength. *Encyclopedia of Life Sciences*. John Wiley & Sons; 2010.
40. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A*. 1997;94(19):10172-10177.
41. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. *Proteins*. 2001;44(2):150-165.
42. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*. 1994;238(5):777-793.
43. van Dijk M, Bonvin AM. A protein-DNA docking benchmark. *Nucleic Acids Res*. 2008;36(14):e88.
44. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302-2309.
45. Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics*. 2003;19(18):2500-2501.
46. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics*. 2016;54:5 6 1-5 6 37.
47. Honorato RV, Roel-Touris J, Bonvin A. MARTINI-based protein-DNA coarse-grained HADDOCKing. *Front Mol Biosci*. 2019;6:102.
48. Vangone A, Rodrigues JP, Xue LC, et al. Sense and simplicity in HADDOCK scoring: lessons from CASP-CAPRI round 1. *Proteins*. 2017;85(3):417-423.
49. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*. 1999;285(4):1735-1747.
50. Dixit SB, Arora N, Jayaram B. How do hydrogen bonds contribute to protein-DNA recognition? *J Biomol Struct Dyn*. 2000;17(Suppl 1):109-112.
51. Sheu SY, Yang DY, Selzle HL, Schlag EW. Energetics of hydrogen bonds in peptides. *Proc Natl Acad Sci U S A*. 2003;100(22):12683-12687.
52. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. 2010;26(7):889-895.
53. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*. 2003;53(Suppl 6):334-339.
54. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins*. 1995;23(3):ii-v.
55. Zahiri J, Emamjomeh A, Bagheri S, et al. Protein complex prediction: a survey. *Genomics*. 2020;112(1):174-183.
56. Fuxreiter M, Simon I, Bondos S. Dynamic protein-DNA recognition: beyond what can be seen. *Trends Biochem Sci*. 2011;36(8):415-423.
57. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. *Nature*. 2007;450(7172):964-972.
58. Song WY, Guo J-T. Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. *J Biomol Struct Dyn*. 2015;33(10):2083-2093.
59. Wilson KA, Kellie JL, Wetmore SD. DNA-protein pi-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res*. 2014;42(10):6726-6741.
60. Wilson KA, Wells RA, Abendong MN, Anderson CB, Kung RW, Wetmore SD. Landscape of pi-pi and sugar-pi contacts in DNA-protein interactions. *J Biomol Struct Dyn*. 2016;34(1):184-200.
61. Gallivan JP, Dougherty DA. Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A*. 1999;96(17):9459-9464.
62. Baker CM, Grant GH. Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers*. 2007;85(5–6):456-470.
63. Brylinski M. Aromatic interactions at the ligand-protein interface: implications for the development of docking scoring functions. *Chem Biol Drug des*. 2018;91(2):380-390.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Malik FK, Guo J. ComparePD: Improving protein–DNA complex model comparison with hydrogen bond energy-based metrics. *Proteins*. 2023;91(8):1077-1088. doi:10.1002/prot.26493