Received: 22 October 2021 | Revised: 17 January 2022 | Accepted: 31 January 2022

DOI: 10.1002/prot.26313

RESEARCH ARTICLE

PROTEINS WILEY

Insights into protein–DNA interactions from hydrogen bond energy-based comparative protein–ligand analyses

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

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

[2]Research Center of Modeling and Simulation, National University of Science and Technology, Islamabad, 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
This work was supported by the National Science Foundation (DBI-2051491 to Jun-tao Guo) and the National Institutes of Health (R15GM132846 to Jun-tao Guo).

Abstract

Hydrogen bonds play important roles in protein folding and protein–ligand interactions, particularly in specific protein–DNA recognition. However, the distributions of hydrogen bonds, especially hydrogen bond energy (HBE) in different types of protein–ligand complexes, is unknown. Here we performed a comparative analysis of hydrogen bonds among three non-redundant datasets of protein–protein, protein–peptide, and protein–DNA complexes. Besides comparing the number of hydrogen bonds in terms of types and locations, we investigated the distributions of HBE. Our results indicate that while there is no significant difference of hydrogen bonds within protein chains among the three types of complexes, interfacial hydrogen bonds are significantly more prevalent in protein–DNA complexes. More importantly, the interfacial hydrogen bonds in protein–DNA complexes displayed a unique energy distribution of strong and weak hydrogen bonds whereas majority of the interfacial hydrogen bonds in protein–protein and protein–peptide complexes are of predominantly high strength with low energy. Moreover, there is a significant difference in the energy distributions of minor groove hydrogen bonds between protein–DNA complexes with different binding specificity. Highly specific protein–DNA complexes contain more strong hydrogen bonds in the minor groove than multi-specific complexes, suggesting important role of minor groove in specific protein–DNA recognition. These results can help better understand protein–DNA interactions and have important implications in improving quality assessments of protein–DNA complex models.

KEYWORDS
binding specificity, hydrogen bond energy, minor groove, protein–DNA, protein–ligand

1 | INTRODUCTION

Proteins interact with DNA, peptides, and other proteins to form macromolecular assemblies that carry out fundamental and essential biological functions.[1] Protein–DNA (PD) complexes, for example, play critical roles in the regulation of gene expression, histone packaging, DNA replication, repair, modification, and recombination.[2] The interactions between protein and DNA display different degrees of specificity that ranges from highly specific to nonspecific.[3] Protein–peptide (PT) interactions account for up to 40% of cellular interactions and are

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Proteins: Structure, Function, and Bioinformatics* published by Wiley Periodicals LLC.

involved in mediating signal transduction, regulating apoptotic pathways, and immune responses.[4–6] Protein–protein (PP) interactions form essential complexes like hormone–receptor, antibody–antigen, and protease–inhibitor, which control cell signaling, electron transport, signal transduction, and cell metabolism.[7] Disruptions in these interactions can cause serious medical conditions such as cancer, cardiovascular, and neurodegenerative disorders.[7–10] Knowledge of detailed interactions among these complexes at atomic resolution is therefore essential to understanding the underlying mechanisms that govern biochemical processes. It also has important implications in biomedical applications such as protein–ligand docking, *in silico* design of inhibitors and interfaces, and virtual screening of drugs library in the pharmaceutical industry.

Hydrogen bonds (HBs) play key roles in conferring binding specificity of macromolecular complexes.[11–14] An HB is generally considered as a weak, electrostatic interaction between a polar acceptor atom that carries a lone pair of electrons and a hydrogen atom that is covalently linked to a polar atom, oriented toward each other at an equilibrium distance. This orientation- and distance-dependent nature of HBs is vital in providing the shape and chemical complementarity for selective recognition and binding of complexes.[12] In PD complexes, for example, HBs play a key role in DNA base readout by proteins and act as the major contributor to binding specificity that is vital for the biomolecular function of protein–DNA complexes.[15] The recognition of DNA by proteins is guided by an innate hydrogen-bonding pattern that generates an initial unstable nonspecific, intermediate complex with high energy.[16–19] While most of this recognition is expected to occur through the signature hydrogen-bonding pattern in the major groove, many DNA binding proteins also bind to the minor groove through hydrogen bonding and shape readout.[15,20] Later, this complex transitions to a stable and highly specific low energy state through reversible structural deformations that are also guided by a specific HB pattern.[12] In PP complexes, HBs influence stability as well as binding specificity at the interface.[14] Interfacial hydrophilic sidechains of a PP complex have a high charge density that is stabilized primarily through hydrogen bonding. Buried polar atoms at the interface not involved in hydrogen bonding may destabilize the complex.[21–24] Peptide binding, on the other hand, utilizes HBs to improve interface packing density as well as minimize the entropic cost of transitioning from a highly flexible, unstructured peptide to a well-defined rigid structure in a complex with protein.[25] On average, PT interface contains more HBs per 100 Å^2 interface area when compared to PP interface and PT interface HBs generally are more linearly oriented.[25] In addition to binding, HBs are the primary driving force in folding of protein chains into core secondary structures such as alpha helices and beta sheets and base pairing in nucleic acids.[11] HBs also bring flexibility to the structure, which is central to the dynamic nature of proteins and plays a key role in allosteric, catalytic, and binding activities.[11,26]

The role of HBs in binding and folding of complexes has previously been studied as individual cases as well as a group of cases.[18,27–30] Mandel-Gutfreund et al. studied different types of HBs at the interface of 28 X-ray crystal structures of protein–DNA complexes. The HBs were classified according to the types of donor and acceptor atoms, such as backbone, sidechain, or base edges.[13] Xu et al. performed a similar analysis on 319 protein–protein complexes.[14] London et al. compared the types of HBs at the interface and within protein chains of 103 protein–peptide complexes. They further compared the types of HBs in protein–peptide complexes to those in protein–protein complexes.[25] Rawat and Biswas in 2011 performed a comparison of HBs along with several other structural features to investigate the role of flexibility in protein–DNA, protein–RNA, and protein–protein complexes.[31] Jiang et al. demonstrated that in protein–protein complexes, the average energy contribution of a HB is ∼30%.[32] Zhou and Wang recently compared short HBs, where donor–acceptor distance is less than 2.7 Å, in 1663 high-quality protein, protein–ligand, and protein–nucleic acid structures.[33] Itoh et al. showed that the interaction energy of even the weaker N^+—C—H···O HBs is comparable to other protein–ligand interactions such as π–π interactions suggesting the importance of considering HB energy in drug design.[34]

While analyses based on the number of HBs with a single energy cutoff or a distance/angle cutoff can provide useful information about the role of HBs in protein–ligand interaction, they have an intrinsic flaw since strong and weak HBs are treated equally. Moreover, the distributions of interfacial HBs in terms of HB strength or HB energy in protein–ligand complexes, and more importantly, the distributions of interfacial HB energy among different types of protein-ligand complexes remain unknown. To address these issues, in this study, we performed a holistic statistical comparative analysis of HBs across interfaces and within protein chains (intrachain) among PP, PT, and PD complexes to get an insight into their roles in each type of complexes. In addition to comparing the types and locations of HBs in each type of complexes, we investigated the HB energy distributions and found significant differences among these three types of complexes, especially a unique pattern in protein–DNA complexes. To the best of our knowledge, an HB energy-based large-scale comparison of macromolecular complexes has never been explored before.

2 | MATERIALS AND METHODS

2.1 | Datasets

Seven previously published and widely used datasets of protein–DNA, protein–peptide, and protein–protein complexes were selected, including three datasets of protein–DNA complexes: highly specific (HS), multi-specific (MS),[3] and rigid docking protein–DNA (RDPD) complexes[35]; two protein–peptide complex datasets: LEADS–PEP,[36] and InterPep[37]; and two datasets for protein–protein complexes: an updated M-TASSER dimer library[38] and the protein–protein Docking benchmark (RDPP, version 5)[39] (Table 1). Since the M-TASSER dimer library was published over 10 years ago, we generated an updated dataset, called Protein Homo/Heterodimer Library (PHDL) using some of the guidelines described in the original paper (Table S1).

Each of the three datasets for PD represents a specific category of protein–DNA complexes. The HS dataset comprises 29 PD complexes with high binding specificity between protein and DNA whereas the MS dataset comprises 104 cases, in which proteins can bind to multiple conserved DNA sequences.[3] The RDPD dataset consists of 38 highly diverse nonredundant TF-DNA complexes that cover 11 structural folds, 15 super-families, and 28 families.[35]

The two PT complex datasets differ mainly in the peptide chain lengths. InterPep comprises protein complexes with peptides ranging from 5 to 25 amino acids whereas peptides in LEADS-PEP are 3–12 amino acids long.[36,37] InterPep is a larger dataset with 502 X-ray and NMR structures, which was originally developed for testing a peptide-binding site prediction pipeline.[37] LEADS-PEP, on the other hand, is a much smaller dataset with 53 carefully curated and widely used complexes designed specifically for peptide-based therapeutics and peptide docking. It contains only X-ray crystal structures with a resolution better than 2 Å.[36]

The complexes in the PP datasets differ mainly in size and definition of interaction unit. The protein–protein docking benchmark (RDPP) has 230 complex structures that were experimentally solved with corresponding unbound components available.[39] The structures in the RDPP dataset represent a diverse combination of antigen–antibody, enzyme–substrate, enzyme–regulatory complex, GPCR proteins, and several other classes of proteins. The docking benchmark defines a true interaction as one that has functional significance as identified in the literature and agreed upon by the scientific community. The second PP dataset PHDL, a protein homo/heterodimer library, determines the oligomeric state from PDB files.[40] PHDL contains nonredundant heterodimers (Table S1A) and homodimers (Table S1B), where no two chains share more than 30% sequence identity with each other and each interacting partner has at least 40 amino acids.

In addition to these individual datasets, we pooled the datasets of the same type of complexes together and generated three larger, nonredundant, and highly diverse datasets (Figure 1): (i) PDnrall, a

TABLE 1 The protein–DNA, protein–peptide, and protein–protein datasets

Types	Datasets	Number of complexes	Experimental method and selection criteria	Ligand	Average interface area
Protein–DNA	Highly specific	28	X-ray (≤3 Å) R-factor <0.3	Double-stranded DNA	~1100 Å^2
	Multi-specific	105	X-ray (≤3 Å) R-factor <0.3	Double-stranded DNA	~700 Å^2
	Rigid docking	38	X-ray (≤3 Å)	Double-stranded DNA	~1100 Å^2
Protein–peptide	InterPep	502	X-ray (≤3 Å) or NMR	5–25 residues	~665 Å^2
	LEADS-PEP	53	X-ray (<2 Å) R-factor < 0.3	3–12 residues	~512 Å^2
Protein–protein	Protein homo/ heterodimer library	2608	X-ray (≤3 Å)	>40 residues per protein chain	~1374 Å^2
	Docking Benchmark V5	230	X-ray (≤3.25 Å)	≥30 residues per protein chain	~1847 Å^2

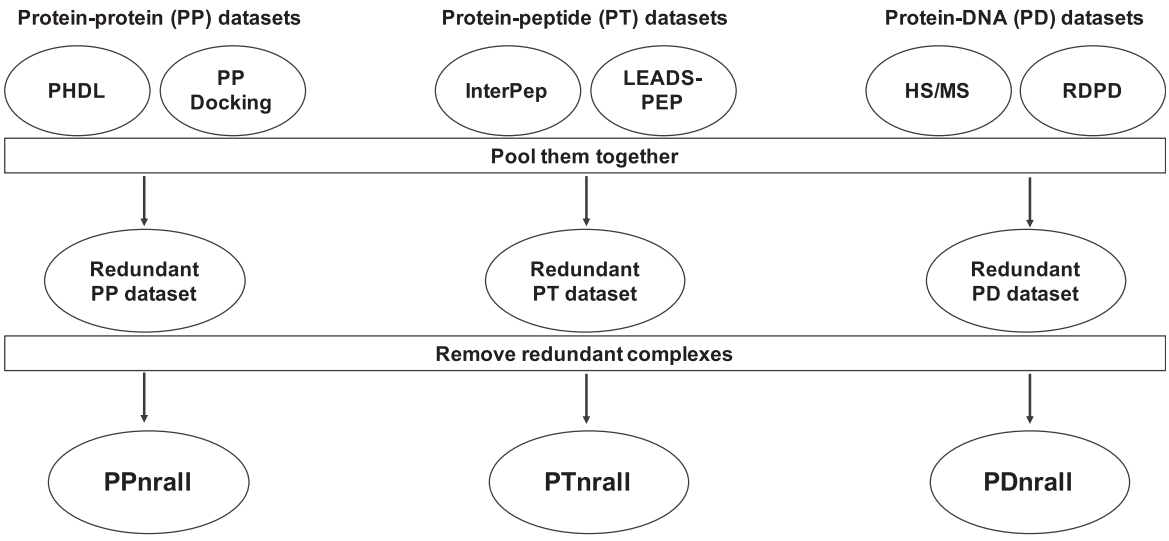

FIGURE 1 A flow chart for generating non-redundant datasets of protein–protein, protein–peptide, and protein–DNA complexes

protein–DNA dataset comprising HS, MS, and RDPD; (ii) PTnrall, a protein–peptide dataset comprising LEADS-PEP and InterPep; and (iii) PPnrall, a protein–protein dataset comprising PHDL and RDPP. The redundancy after combining the respective datasets was removed with PISCES using a sequence identity cutoff of 30%,[41] which resulted in 2724 non-redundant protein–protein complexes (PPnrall), 346 non-redundant protein–peptide complexes (PTnrall), and 126 non-redundant protein–DNA complexes (PDnrall).

2.2 | Dataset processing

The datasets were filtered rigorously for accurate analysis. In case of multiple models for one native structure as in the NMR entries, only the first model was selected. All the heteroatoms, including water molecules, were removed since we do not consider solvation effects for the sake of simplicity and fair comparison. Proteins that have residues with insertion codes were renumbered accordingly. Since considering the alternate locations of a residue in an experimentally solved crystal structure may result in over counting the number of HBs, only the state with the highest occupancy for a given residue was included for analysis. The complexes with internal missing residues, that is, residues that are not on the N or C terminal of the chain were discarded. Finally, interactions between proteins and ligands were calculated based on interaction units for complexes composed of multiple chains of proteins and ligands. For example, 4FQI protein unit has two chains H, L and the ligand unit has six chains A, B, C, D, E, and F. For such cases, we only considered the inter-unit interaction between protein and ligand. In the case of 4FQI, H and L were identified as one unit while ABCDEF as another unit.

2.3 | Identification of HBs

Two widely used HB annotation programs, FIRST (Floppy Inclusion and Rigid Substructure Topography) and HBPLUS, were used to identify HBs with default parameters.[42,43] Reduce was used to add hydrogen atoms to pdb files for FIRST HB calculations while HBPLUS calculates the hydrogen atom positions within the program.[44] FIRST employs an energy-based approach and the HB energy is calculated as in Equation (1).[42,45]

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

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.[46] The angle term $F(\theta,\phi,\varphi)$ is calculated based on the hybridization state of the acceptor and donor atoms, where θ is the donor–hydrogen–acceptor angle, ϕ is the hydrogen–acceptor–base angle, and φ is the angle between the normals of the planes defined by the six atoms attached to the sp^2 center as described by Dahiyat et al.[45] The FIRST program was used for both the number of HBs annotations using a widely used HB energy cutoff of −0.6 kcal/mol as well as for HB energy-based analysis. HBPLUS identifies HB with a distance–angle approach and defines the optimal distance between the donor and acceptor as 2.5 Å or smaller and the optimal angle as 90° or higher.[43]

2.4 | Interface analysis and comparison

Since the interface sizes are different among different types of complexes (Table 1), in order to accurately assess the roles of HB at the interface of PP, PT, and PD complexes, the numbers of HBs were compared with respect to the interfacial surface area. The interfacial surface area (iSA) of a complex was calculated using NACCESS v2.1.1 with default parameters as shown in Equation (2):

$$\text{iSA} = \frac{SA_P + SA_L - SA_C}{2} \quad (2)$$

where SA_p and SA_L represent the surface area of protein and ligand, respectively, and SA_C is the surface area of the protein–ligand complex. For multichain components, SA_P is the surface area of the protein unit while SA_L is the surface area of the ligand unit.

The HB distributions were compared at three different aspects: HB types, HB locations, and HB energy ranges. The types of HB were grouped depending on the types of atoms involved in hydrogen bonding, sidechain (or base in DNA), or backbone. HB types include SC–SC (representing sidechain–sidechain in PP and PT or sidechain–base in PD), BB–BB (for backbone–backbone), and Mixed type (for SC–BB or BB–SC). A union of all three types encompasses all HBs (HBall). The SC–SC HBs, also termed here as HBSP, are generally considered more specific in molecular recognition and binding as the backbone atoms are the same for each type of molecules, protein, or DNA. There are two different HB location types, interface (between proteins and ligands), and intrachain (within proteins).

We divided hydrogen bond energy (HBE) from the FIRST program into four categories based on different energy cutoffs used in previous studies[17,42,48] and personal communication with the FIRST program developer as shown in Table 2.

2.5 | Statistical tests

Wilcoxon rank-sum test was employed to assess if there are significant differences between samples across datasets. Chi-squared

TABLE 2 Hydrogen bond energy (HBE) categories based on energy ranges

Category	HBE range (kcal/mol)
I	−0.6 ≤ HBE < −0.1
II	−1.0 ≤ HBE < −0.6
III	−1.5 ≤ HBE < −1.0
IV	HBE < −1.5

goodness of fit test was used to test the categorical distributions of types and the energy of HBs at interface and within intrachain.

3 | RESULTS

3.1 | HBs at the interface of complexes

We first compared the number of HBall and HBSP in PDnrall, PPnrall, and PTnrall datasets. Based on HB annotations from FIRST with the widely used energy cutoff of −0.6 kcal/mol,[48] we found that the number of interface HBall and the number of interface HBSP in PD complexes are significantly higher than those in the PP and PT complexes (Figure 2A,B). The number of HBall and HBSP in PT complexes are significantly less than those in PP complexes (Figure 2A,B). Results from HBPLUS are consistent with the data from FIRST except that the number of HBSP in PP complexes is larger than that in PD complexes with HBPLUS (Figure S1A,B). Interestingly, when the FIRST energy cutoff is set at −0.1 kcal/mol, the results are more similar to the HBPLUS data (Figure S2A,B).

Since the interface areas among the three types of complexes are different with PP complexes having the largest average interfacial area and PT complexes having the smallest average interfacial area (Table 1), comparing the raw number of interface HBs might be biased towards the complexes with a larger contact surface. Therefore, we normalized the number of interface HBs, HBall, and HBSP, by the interfacial surface area (iSA). Figure 2C,D shows that both HBall/iSA and HBSP/iSA ratios of PD complexes are significantly higher than those in the PP complexes and PT complexes. There is a clear pattern for the iSA normalized HBSP, PD > PP > PT. When the analyses were carried out with HBPLUS, the results are consistent with the results from FIRST (Figure S1). Even though no significant difference of the ratio HBall/iSA from FIRST is found between PP and PT complexes for a two-tailed test (Figure 2C), one-tailed test with a null hypothesis that HBall/iSA in PP is not smaller than HBall/iSA in PT results a *p*-value of .043, which is in line with the result from HBPLUS as well as that from FIRST with an energy cutoff at −0.1 kcal/mol: the ratio of HBall/iSA in PT complexes is significantly higher than PP complexes (Figure S1C and S2C). These results are also in agreement with a previous study that PT interface has more total HBs per 100 Å^2 interface

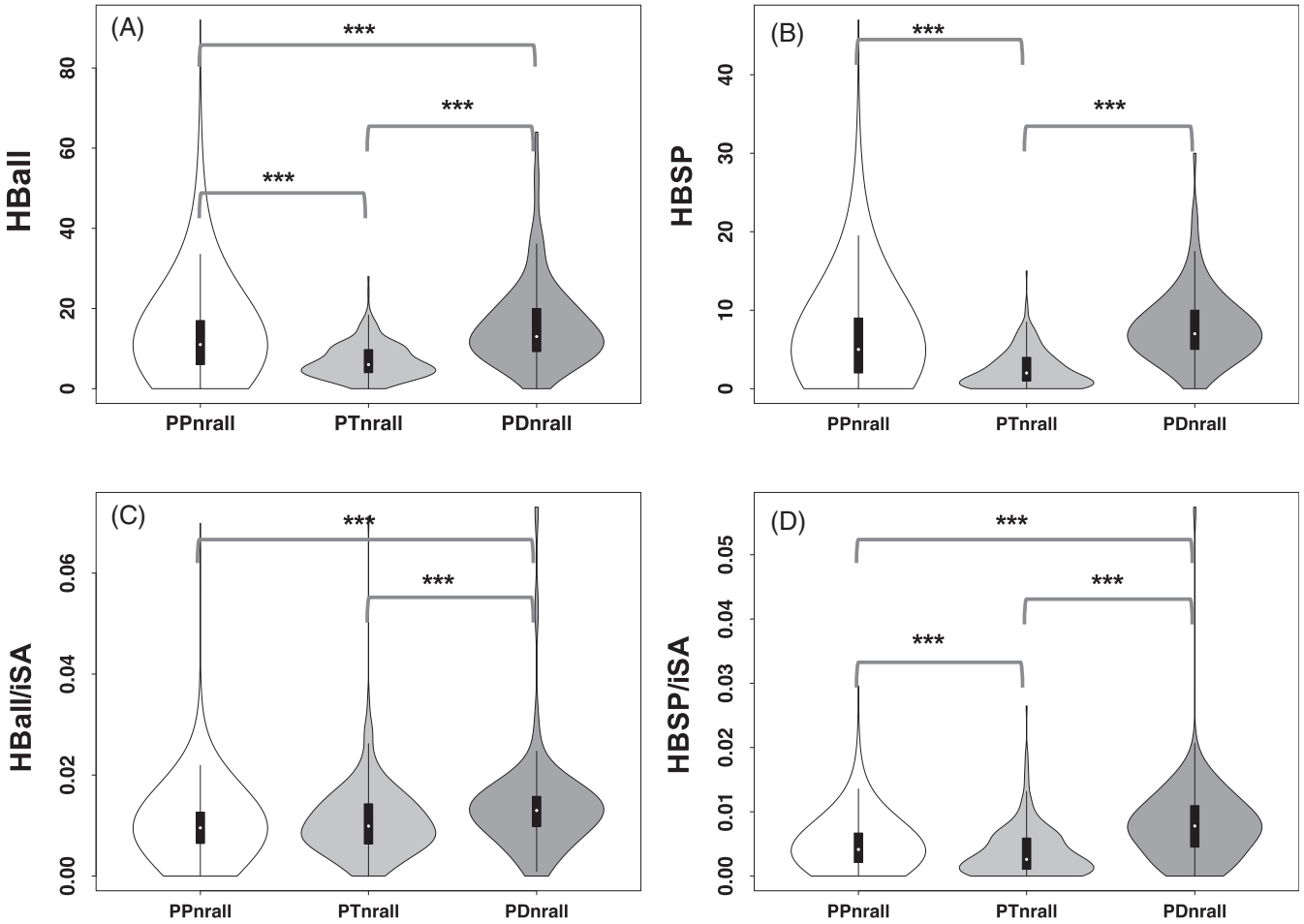

FIGURE 2 Comparison of interfacial hydrogen bonds based on FIRST with an energy cutoff of −0.6 kcal/mol: (A) the number of total hydrogen bonds (HBall); (B) the number of SC–SC or SC–Base hydrogen bonds (HBSP); (C) the ratio of HBall to interfacial surface area (iSA); and (D) the ratio of HBSP to iSA. ***p-value ≤ .001, **p-value ≤ .01

area than that in PP.[25] However, the HBSP/iSA ratio is the opposite, suggesting relatively fewer interface HBSP in PT complexes when the interface area is taken into consideration.

3.2 | Types of HBs at interface and within intrachain

We compared the distributions of the HB types at complex interface or within protein (intrachain) in PP, PT, and PD complexes and between individual complexes of the same type of complexes. Figure 3A and Table 3 show that there is no significant difference among the types of HBs within proteins in all three types of complexes. BB–BB HBs represent the largest number of overall HBs within proteins (66%–69%) followed by the Mixed (17%–20%), and SC–SC (14%) HBs, respectively (Figure 3A). This is not surprising because the two major secondary structure types of the core protein structure, α-helices, and β-sheets, are stabilized by backbone–backbone HBs.

The distributions of the HB types at interface, however, are significantly different from the intrachain and among the three types of complexes (Figure 3B and Table 3). The percentages of SC–SC HBs at interface increase dramatically when compared with those within proteins while the BB–BB is the least type in all three complexes. The proportions of BB–BB HBs at the interface are approximately one-third of those from intrachain in PP and PD complexes and approximately half of that in PT complexes (Figure 3). The proportions of interface SC–SC HBs are at least twice more than those in intrachain in all three types of complexes. There is an increase of the Mixed HB type at interface when compared with intrachain. In PD complexes, the Mixed HB type consists of about half of all interfacial HBs.

A previous study on protein–protein complexes indicated that the larger number of BB–BB HBs within protein chains as compared to the interface is likely due to the differences in the degrees of freedom available to the corresponding atoms.[14] On both PP and PT interfaces, the highest proportion of HB types is SC–SC between interacting components while the percentage of BB–BB HBs is the lowest. The percentage of interface BB–BB HBs in PT complexes is higher than those in the PP and PD complexes. It has been suggested that a higher number of interface BB–BB HBs in PT complexes is a result of bridging beta strands at the interface between interacting peptides and protein molecules.[25] Once the interfacial beta-sheet containing complexes are removed from the dataset, BB–BB HBs are comparable between PP and PT complexes.[25] Similar results were observed for the comparison of HB types annotated by HBPLUS and by FIRST with

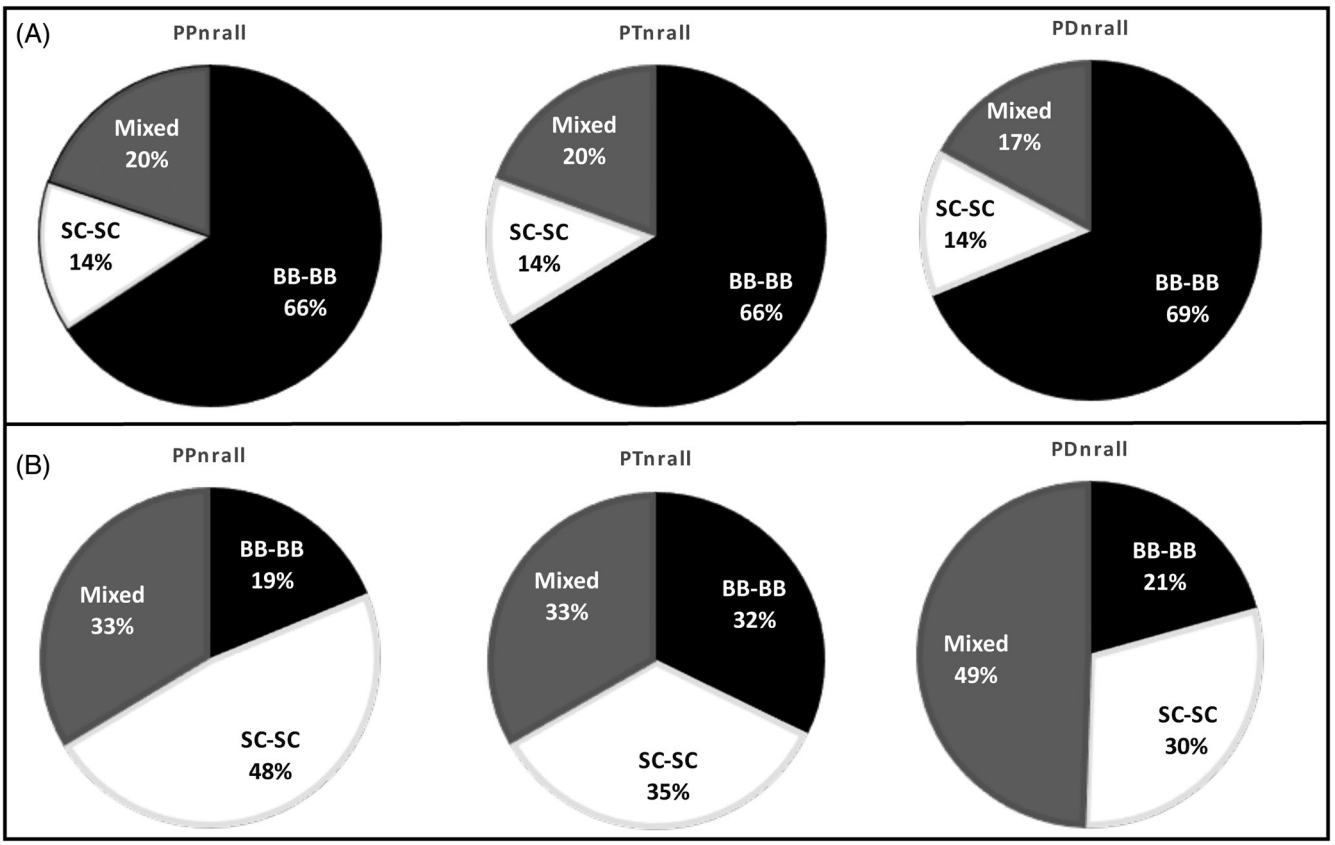

FIGURE 3 Comparisons of the distribution of different types of hydrogen bonds, backbone–backbone (BB–BB), sidechain–sidechain (SC–SC), and Mixed (BB–SC and SC–BB) for (A) intrachain within proteins and (B) at interface of PP, PT, and PD complexes. The hydrogen bonds are annotated from the FIRST program with an energy cutoff of −0.6 kcal/mol

an energy cutoff of −0.1 kcal/mol (Table 3, Figures S3 and S4 and Table S2).

Besides comparisons among the three different types of non-redundant complexes, we also compared the distributions between individual datasets for each type of complexes (Figures S5 and S6). For example, PHDL is composed of homodimers and heterodimers and the PD dataset has HS and MS complexes with different binding specificity. We found that there is no significant difference in the distribution of HB types for both intrachain and at interface between HS and MS (p values of .3743 and .6685, respectively) as well as between homodimers and heterodimers (p values of .9371 and .9746, respectively) from FIRST (Figure S5A). There is also no significant difference of HB type distributions for intrachain and at interface between PHDL and RDPP (p values of .992 and .246, respectively). While there is no difference for the intrachain distributions between InterPep and LEADS-PEP (p-value = .954), the interface distributions are different (p-value = .003) from FIRST HB annotations (Figure S6A). This might be a result of the relatively small LEADS-PEP dataset with a small number of total HBs (Figure 2). Similarly, no significant differences were found between any two of the above datasets of the same types of protein–ligand complexes based on HBPLUS annotations (Figures S5B and S6).

3.3 | Strength of HBs at interface and within protein chain

We classified the strength of HBs into four categories based on HB energy from the FIRST program with different energy cutoffs used in previous studies as shown in Table 2.[17,42,48] For intrachain HBs within proteins, no significant differences were found among the three types of complexes (Figure 4A and Table 4). Most of the HBs (66–69%) are strong ones with lower than −1.5 kcal/mol energy (category IV) while very few of them are of intermediate energy (<16% when categories II and III are combined), suggesting that the HBs in all types of proteins have similar energy distribution with predominantly strong HBs.

To investigate if the energy categories are related to different HB types, we compared the distributions of each type of intrachain HBs in each energy category (Figure 5A and Table S3). Similar trends for BB–BB, SC–SC, and Mixed types were observed among the three types of complexes and there is no significant difference of intrachain HB energy distribution for each HB type among the PP, PT, and PD complexes. There is a higher percentage of strong BB–BB HBs in all complexes, but relatively fewer strong ones for the Mixed HBs, suggesting that the major secondary structure types patterned by the BB–BB HBs are optimized in terms of both distance and angle and form strong HBs.

TABLE 3 p values of chi-square tests between HB types from FIRST (−0.6 kcal/mol cutoff) and HBPLUS at interface and intrachain

	Intrachain		Interface		Interface/Intrachain		
Dataset1/Dataset2	FIRST	HBPLUS	FIRST	HBPLUS	Dataset	FIRST	HBPLUS
PPnrall, PDnrall	0.720	0.647	2.2e−16	0.025	PDnrall	<2.2e−16	<2.2e−16
PTnrall, PDnrall	0.874	0.945	0.002	0.0005	PPnrall	<2.2e−16	<2.2e−16
PTnrall, PPnrall	0.972	0.774	2.2e−16	<2.2e−16	PTnrall	8.904e−14	<2.2e−16

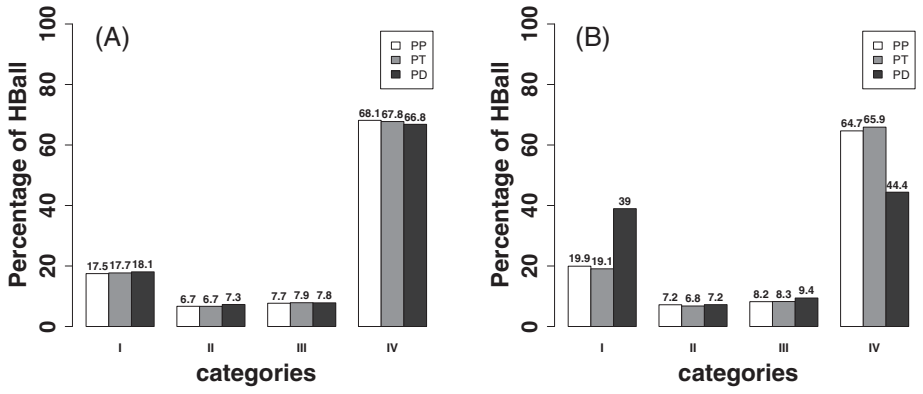

FIGURE 4 Comparisons of the distributions of hydrogen bond energy for (A) intrachain and (B) at interface

TABLE 4 p values of chi-square tests between HBE categories at interface and within intrachain

Dataset1/Dataset2	Intrachain	Interface	Dataset	Interface/intrachain
PPnrall, PDnrall	0.919	2.2e−16	PDnrall	5.3e−07
PTnrall, PDnrall	0.994	3.73e−06	PPnrall	0.871
PTnrall, PPnrall	0.995	0.5247	PTnrall	0.979

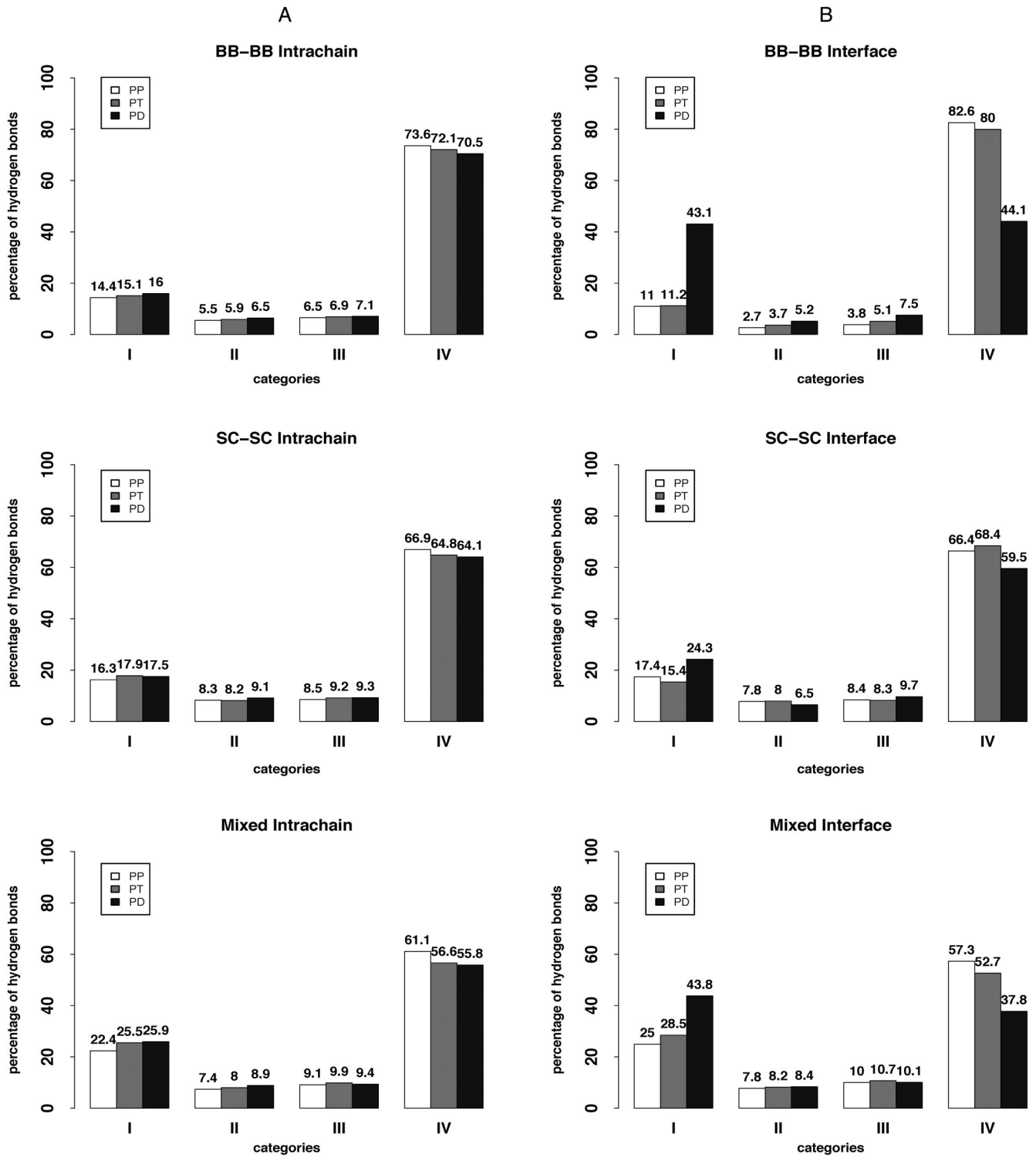

FIGURE 5 Comparison of (A) intrachain hydrogen bond energy and (B) interface hydrogen bond energy in different hydrogen bond types

However, the interface HB energy distributions among different types of complexes are significantly different and exhibit a unique pattern for the PD complexes (Figure 4B and Table 4). There is a higher percentage of weak HB (category I) at PD complex interface when compared to those in PP and PT complexes as well as the intrachain HB energy in PD complexes. PD has the smallest percentage of strong HBs (category IV) among the three types of complexes. The difference between category I and IV HB percentage is much smaller in PD complexes (39% and 44.4%) than those in PP (19.9% and 64.7%), and PT (19.1% and 65.9%) complexes (Figure 4B). PP and PT complexes have similar distributions of interface HB energy categories. In addition, the interface and intrachain

HB energy distributions in both PP and PT complexes are also similar (Table 4).

We also compared the energy distributions of each HB type across interfaces (Figure 5B). Similar to the pattern observed for all HBs in PD, energy distributions of different types of interface HB in PD complexes also differ significantly from PP and PT complexes while there is no significant difference between PP and PT complexes (Table S3). Interestingly, SC–SC HBs in PD complexes have a much larger percentage of strong, category IV HBs (59.5%) while the BB–BB and Mixed types in PD complexes have more weak, category I HBs (43.1% and 43.8%, respectively) than the SC–SC HBs (24.3%), suggesting important functional applications of HBs in specific protein–DNA interactions.

3.4 | Comparison of HBs between HS and MS datasets

In our previous study, we demonstrated that highly specific HS protein–DNA complexes have more HBs than the multi-specific MS protein–DNA complexes, including both total HBs and sidechain–base HBs.[3] It is intriguing to see whether there is any relationship between the HB strength and protein–DNA binding specificity. We first compared the HB types and energy categories within proteins as well as at the interface of HS and MS complexes. No significant differences between HS and MS complexes were found in terms of energy categories (Figures S7 and S8) while there are significant differences between the intrachain and interface for both HS (p-value: 9.673e−07) and MS complexes (p-value: 6.413e−07). We did observe some statistically nonsignificant small differences. For example, the number of SC–SC interface HBs in HS (32%) is slightly higher than that in MS (28.2%; Figure S5A). Both HS and MS complexes show similar interface HB energy distributions with an overall balance of strong and weak HBs, but HS complexes have a slightly higher percentage of HBs in category IV (Figure S7).

Since both major and minor grooves are known to play important roles in the base and shape readout mechanisms in specific protein–DNA recognition,[3,15,20,49] we compared the energy distributions of total HBs and sidechain–base HBs in the major and minor grooves. Between major and minor grooves, there is no significant difference in terms of HB energy distributions within each type of PD complexes, PDnrall, HS, and MS with high p values (data not shown). For major groove HBs, while we observed more strong and fewer weak major groove HBs in HS complexes than those in the MS complexes, the differences in the energy distributions of HBall and HBSP in the major groove between HS and MS complexes are not statistically significant (Figure 6). However, we observed a significant difference in the energy distributions in the minor groove for both HBall and HBSP between HS and MS complexes (Figure 7). In general, HS complexes have more strong HBs (category IV) and fewer weak HBs (category I) than those in the MS complexes in the minor groove. The MS complexes have about doubled the percentage of weak HBs in category I than that in HS complexes. These results suggest a clear and important role of HB energy of the minor groove in specific protein–DNA interaction.

4 | DISCUSSION

Despite the generally known importance of HBs in protein–ligand interactions, the relative contribution of different types of HBs, especially their energy in different types of complexes, is unknown. Previous studies mainly focused on analyses of the number of HBs. Here, we performed a systematic comparative analysis of HBs and their energy at the interface and within protein chains among three nonredundant protein–ligand complexes, PP, PT, and PD. To the best of our knowledge, this is the first study that compares the energy of HBs in different types of complexes. In addition, our use of large nonredundant datasets not only maximizes the diversity of the complexes but also avoids potential biases. Results between HBPLUS and FIRST

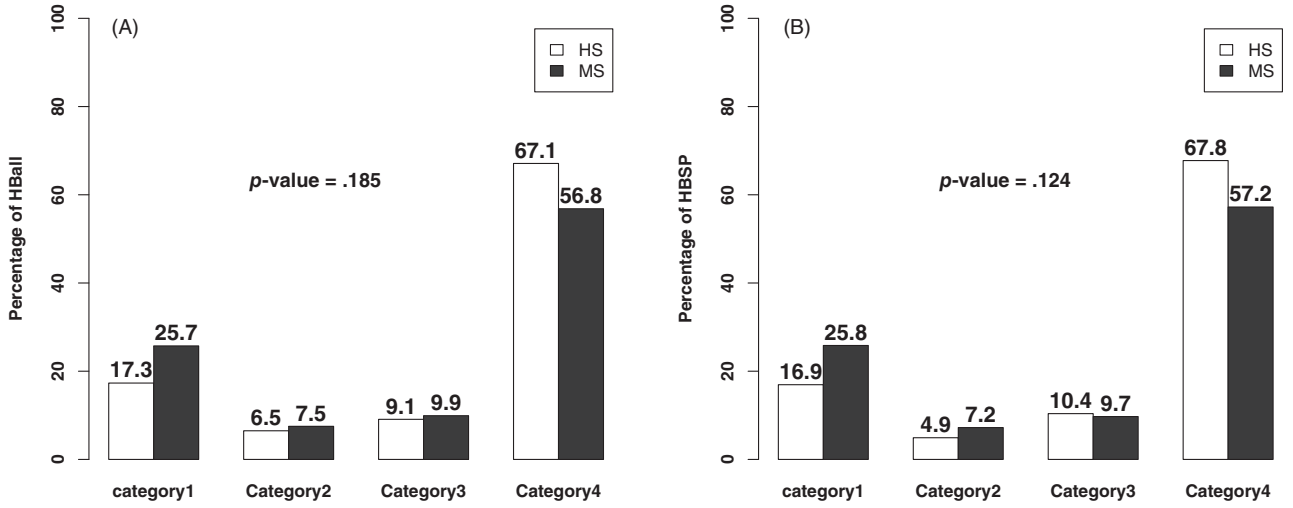

FIGURE 6 Comparison of major groove for (A) HBall and (B) HBSP energy distributions between HS and MS complexes

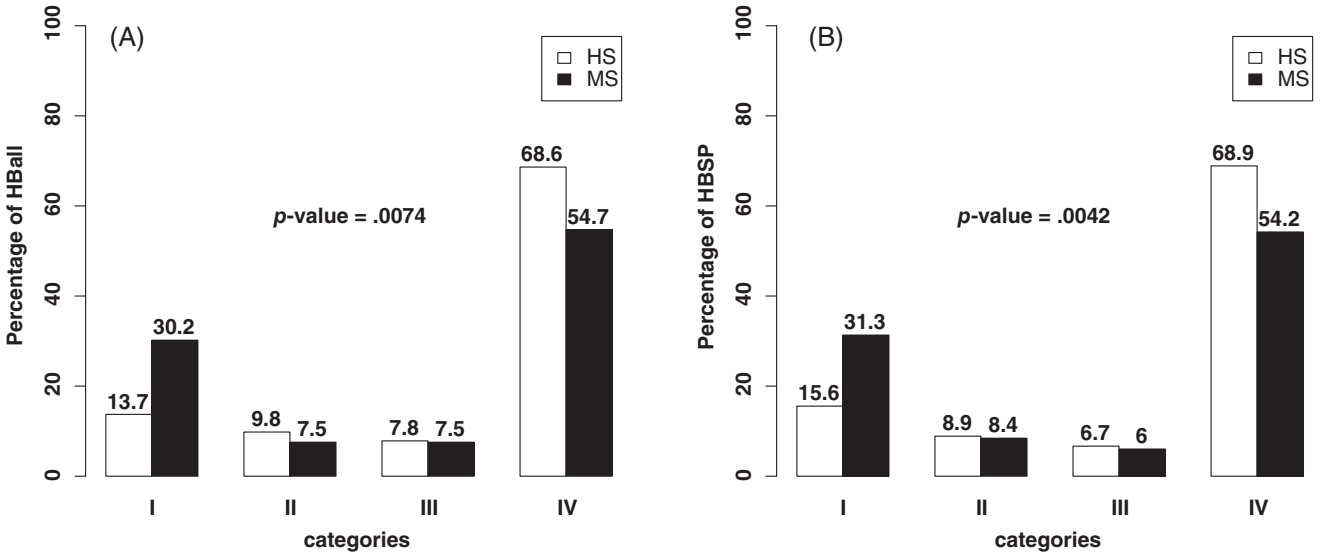

FIGURE 7 Comparison of minor groove for (A) HBall and (B) HBSP energy distributions between HS and MS complexes

are in high agreement even though they use different algorithms for identifying HBs. We also showed similar results between individual datasets for each type of complexes suggesting the results are robust regardless of the datasets and the tools used for HBs annotations.

Our analyses revealed several important findings. First, for intrachain HBs, our analysis not only corroborates several previous findings,[14,25] but also provides additional information by demonstrating no significant difference in the distributions of HB energy among different complexes. Second, at the interface, the HB distributions of PD complexes differ from both PP and PT complexes significantly in three aspects: (a) the total number of HBs, the number of sidechain–base HBs, and the normalized numbers by interface area in PD complexes are significantly higher than those in both PP and PT complexes; (b) more importantly, PD complexes have significantly different distributions of HB types and energy than those of either PP or PT complexes. There is a unique balance between strong and weak HBs in protein–DNA interfaces; and (c) there is a significant difference of the minor groove HBs between HS and MS complexes with HS having more low energy strong HBs.

Our comparative analyses on energy categories are based on HB energy cutoffs (−0.1, −0.6, −1.0, and −1.5 kcal/mol) from previous studies (Table 2).[17,42,48] To test if similar results can be observed with different HB energy discretization, the HBs were grouped using a larger energy range separated by −0.1, −0.7, −1.3, and −2.0 kcal/mol (Table S4). The results of HB energy distributions, shown in Figures S9–S12 and Tables S5 and S6, are in agreement with conclusions (Figures 4–7, Tables 4 and S3) with energy ranges in Table 2, suggesting our key findings are not affected by different discretization of HB energy.

The above findings have important functional and practical implications. While omitting HB information in assessing predicted PP and PT complex models may have minimal effect, our results suggest consideration of HBs is beneficial to quality assessment of protein–DNA complex models since both the raw number and the normalized number of interface HBs in PD complexes are much higher than those in PP and PT complexes. The use of conserved numbers of native HBs in models was suggested to evaluate the quality of protein–peptide models.[50] We found that using the number of HBs can improve quality assessment of protein–DNA complex models.[51] However, due to the unique pattern of interface HB energy distributions in PD complexes and the dynamic nature of macromolecules, it could help model evaluations by considering the HB energy instead of using the raw number of HBs. We demonstrated in our previous study that the accuracy of structure-based prediction of transcription factor binding sites could be improved by adding an HB energy term.[52,53]

Our data also provide an insight into the mechanism of binding specificity between protein and DNA. We observed an approximate balance of high and low energy interface HBs in PD complexes, but not in the other two types of complexes (Figures 4B and 5B). One possibility of such difference lies in the geometry of interacting components as geometry is one of the key factors affecting HBE and strength.[46] While DNA is not a rigid molecule, the double-helical nature restricts the atoms that can form optimal HBs with protein sidechains while the peptide and protein surfaces have a relatively higher flexibility to position atoms for stronger HBs. Other than the unique structure of DNA double-helix that contributes to the pattern of energy distribution, it may also reflect the kinetics of protein–DNA recognition and binding, and the functions of many DNA binding proteins. For example, most of the DNA binding proteins are transcription factors, which bind to conserved DNA binding sequences while allowing variations at certain sites to regulate gene expression. Recent structural and dynamic analyses have shown that transcription factors typically bind to a preferred strand of the DNA double helix.[19,54] A fine balance of strong and weak HBs helps transcription factors bind to conserved yet different sequences by allowing easier association and disengagement. This is further supported by the comparison

between protein–DNA complexes of different binding specificity. Highly specific DNA binding proteins have more strong HBs than the MS group comprising transcription factors (Figures 6 and 7).[3]

The most interesting finding is from the DNA minor groove HB analysis. Both the energy of all HBs and the sidechain–base HBs of highly specific protein–DNA complexes are significantly different from that of multi-specific protein–DNA complexes (Figure 7). While it is generally thought that minor groove contacts play little role in conferring specific protein–DNA interactions, more studies have shown that this might not be the case. It has been reported that local sequence-dependent minor groove shape plays an important role in specific recognition between protein and DNA.[15,20,55–57] The number of contacts in minor grooves of HS complexes is more than that in MS complexes and the HS complexes contain wider minor grooves than MS,[3] thus making it possible for optimal orientation of atoms to form stronger HBs. Our results further demonstrate that the minor groove HBs play more critical roles in conferring binding specificity than previously thought.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary file and references cited in the paper. The PDB files are from https://www.rcsb.org.

REFERENCES

1. Du X, Li Y, Xia Y-L, et al. Insights into protein-ligand interactions: mechanisms, models, and methods. *Int J Mol Sci*. 2016;17(2):144. doi:10.3390/ijms17020144
2. Pandey P, Hasnain S, Ahmad S. ABC of Bioinformatics. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, eds. *Encyclopedia of Bioinformatics and Computational Biology*. 1st ed. Academic Press; 2019:142-154. doi:10.1016/B978-0-12-809633-8.20217-3
3. Corona RI, Guo J. Statistical analysis of structural determinants for protein–DNA-binding specificity. *Proteins Struct Funct Bioinform*. 2016;84(8):1147-1161. doi:10.1002/prot.25061
4. Stanfield RL, Wilson IA. Protein-peptide interactions. *Curr Opin Struct Biol*. 1995;5(1):103-113. doi:10.1016/0959-440X(95)80015-S
5. Mendoza F, Espino P, Cann K, Bristow N, McCrea K, Los M. Anti-tumor chemotherapy utilizing peptide-based approaches—apoptotic pathways, kinases and proteasome as targets. *Arch Immunol Ther Exp (Warsz)*. 2005;53:47-60.
6. Trellet M, Melquiond ASJ, Bonvin AMJJ. A unified conformational selection and induced fit approach to protein-peptide docking. *PLoS One*. 2013;8(3):e58769.
7. Hardcastle IR. 5.06—protein–protein interaction inhibitors in cancer. In: Chackalamannil S, Rotella D, eds. *Ward SEBT-CMCIII*. Elsevier; 2017:154-201. doi:10.1016/B978-0-12-409547-2.12392-3
8. Filippova GN, Qi C, Ulmer JE, et al. Advances in brief tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity 1. *Cancer Res*. 2002;62:48-52.
9. Göhler T, Jäger S, Warnecke G, Yasuda H, Kim E, Deppert W. Mutant p53 proteins bind DNA in a DNA structure-selective mode. *Nucleic Acids Res*. 2005;33(3):1087-1100. doi:10.1093/nar/gki252
10. Chène P. Mutations at position 277 modify the DNA-binding specificity of human p53 in vitro. *Biochem Biophys Res Commun*. 1999;263(1):1-5. doi:10.1006/bbrc.1999.1294
11. Hubbard RE, Kamran HM. Hydrogen bonds in proteins: role and strength. *eLS*. 2010. doi:10.1002/9780470015902.a0003011.pub2
12. Coulocheri SA, Pigis DG, Papavassiliou KA, Papavassiliou AG. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie*. 2007;89(11):1291-1303. doi:10.1016/j.biochi.2007.07.020
13. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol Biol*. 1995;253(2):370-382. doi:10.1006/JMBI.1995.0559
14. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*. 1997;10(9):999-1012. doi:10.1093/protein/10.9.999
15. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. 2010;79(1):233-269. doi:10.1146/annurev-biochem-060408-091030
16. Luscombe NM, Thornton JM. Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol*. 2002;320(5):991-1009. doi:10.1016/S0022-2836(02)00571-5
17. Dixit SB, Arora N, Jayaram B. How do hydrogen bonds contribute to protein-DNA recognition? *J Biomol Struct Dyn*. 2000;17(sup1):109-112. doi:10.1080/07391102.2000.10506610
18. Mukherjee S, Majumdar S, Bhattacharyya D. Role of hydrogen bonds in protein−DNA recognition: effect of nonplanar amino groups. *J Phys Chem B*. 2005;109(20):10484-10492. doi:10.1021/jp0446231
19. Dai L, Xu Y, Du Z, Su XD, Yu J. Revealing atomic-scale molecular diffusion of a plant-transcription factor WRKY domain protein along DNA. *Proc Natl Acad Sci USA*. 2021;118(23):1-10. doi:10.1073/pnas.2102621118
20. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. *Nature*. 2009;461(7268):1248-1253. doi:10.1038/nature08473
21. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. *Biophys J*. 2003;84(3):1895-1901. doi:10.1016/S0006-3495(03)74997-2
22. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol*. 2010;403(4):660-670. doi:10.1016/J.JMB.2010.09.028
23. Worth CL, Blundell TL. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. *Proteins Struct Funct Bioinform*. 2009;75(2):413-429. doi:10.1002/prot.22248
24. Kota P, Ding F, Ramachandran S, Dokholyan NV. Gaia: automated quality assessment of protein structure models. *Bioinformatics*. 2011;27(16):2209-2215. doi:10.1093/bioinformatics/btr374
25. London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. *Structure*. 2010;18(2):188-199. doi:10.1016/J.STR.2009.11.012
26. Song W, Guo J-T. Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. *J Biomol Struct Dyn*. 2015;33(10):2083-2093. doi:10.1080/07391102.2014.997797
27. Laederach A, Reilly PJ. Specific empirical free energy function for automated docking of carbohydrates to proteins. *J Comput Chem*. 2003;24(14):1748-1757. doi:10.1002/jcc.10288
28. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res*. 2005;33(18):5781-5798. doi:10.1093/nar/gki875
29. Eildal JNN, Hultqvist G, Balle T, et al. Probing the role of backbone hydrogen bonds in protein–peptide interactions by amide-to-ester mutations. *J Am Chem Soc*. 2013;135(35):12998-13007. doi:10.1021/ja402875h

30. Stranges PB, Kuhlman B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci*. 2013;22(1):74-82. doi:10.1002/pro.2187
31. Rawat N, Biswas P. Shape, flexibility and packing of proteins and nucleic acids in complexes. *Phys Chem Chem Phys*. 2011;13(20):9632-9643. doi:10.1039/C1CP00027F
32. Jiang L, Lai L. CH···O hydrogen bonds at protein-protein interfaces. *J Biol Chem*. 2002;277(40):37732-37740. doi:10.1074/jbc.M204514200
33. Zhou S, Wang L. Unraveling the structural and chemical features of biological short hydrogen bonds. *Chem Sci*. 2019;10(33):7734-7745. doi:10.1039/c9sc01496a
34. Itoh Y, Nakashima Y, Tsukamoto S, et al. N(+)-C-H···O Hydrogen bonds in protein-ligand complexes. *Sci Rep*. 2019;9(1):767. doi:10.1038/s41598-018-36987-9
35. Kim R, Corona RI, Hong B, Guo J. Benchmarks for flexible and rigid transcription factor-DNA docking. *BMC Struct Biol*. 2011;11:45. doi:10.1186/1472-6807-11-45
36. Hauser AS, Windshügel B. LEADS-PEP: a benchmark data set for assessment of peptide docking performance. *J Chem Inf Model*. 2016;56(1):188-200. doi:10.1021/acs.jcim.5b00234
37. Johansson-Åkhe I, Mirabello C, Wallner B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Sci Rep*. 2019;9(1):4267. doi:10.1038/s41598-019-38498-7
38. Chen H, Skolnick J. M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J*. 2008;94(3):918-928. doi:10.1529/biophysj.107.114280
39. Vreven T, Moal IH, Vangone A, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J Mol Biol*. 2015;427(19):3031-3041. doi:10.1016/J.JMB.2015.07.016
40. Berman HM, Battistuz T, Bhat TN, et al. The Protein Data Bank. *Acta Crystallogr Sect D*. 2002;58(6 Part 1):899-907. doi:10.1107/S0907444902003451
41. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003;19(12):1589-1591. doi:10.1093/bioinformatics/btg224
42. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. *Proteins Struct Funct Bioinform*. 2001;44(2):150-165. doi:10.1002/prot.1081
43. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*. 1994;238(5):777-793. doi:10.1006/JMBI.1994.1334
44. 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. doi:10.1006/jmbi.1998.2401
45. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. *Protein Sci*. 1997;6(6):1333-1337. doi:10.1002/pro.5560060622
46. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA*. 1997;94(19):10172-10177. doi:10.1073/pnas.94.19.10172
47. Hubbard S, Thornton J. *NACCESS: Department of Biochemistry and Molecular Biology*, University College London. 1993. http://www.bioinf.manchester.ac.uk/naccess/.
48. Sheu S-Y, Yang D-Y, Selzle HL, Schlag EW. Energetics of hydrogen bonds in peptides. *Proc Natl Acad Sci USA*. 2003;100(22):12683-12687. doi:10.1073/pnas.2133366100
49. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA*. 1976;73(3):804-808. doi:10.1073/pnas.73.3.804
50. Marcu O, Dodson E-J, Alam N, et al. FlexPepDock lessons from CAPRI peptide–protein rounds and suggested new criteria for assessment of model quality and utility. *Proteins Struct Funct Bioinform*. 2017;85(3):445-462. doi:10.1002/prot.25230
51. Corona RI, Sudarshan S, Aluru S, Guo J. An SVM-based method for assessment of transcription factor-DNA complex models. *BMC Bioinformatics*. 2018;19(20):506. doi:10.1186/s12859-018-2538-y
52. Farrel A, Murphy J, Guo J. Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics*. 2016;32(12):i306-i313. doi:10.1093/bioinformatics/btw264
53. Farrel A, Guo J. An efficient algorithm for improving structure-based prediction of transcription factor binding sites. *BMC Bioinformatics*. 2017;18(1):342. doi:10.1186/s12859-017-1755-0
54. 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. doi:10.1093/nar/gkz963
55. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci*. 2014;39(9):381-399. doi:10.1016/j.tibs.2014.07.002
56. Chiu T-P, Rao S, Mann RS, Honig B, Rohs R. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res*. 2017;45(21):12565-12576. doi:10.1093/nar/gkx915
57. Dantas Machado AC, Cooper BH, Lei X, Di Felice R, Chen L, Rohs R. Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout. *Nucleic Acids Res*. 2020;48(15):8529-8544. doi:10.1093/nar/gkaa642

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: Malik FK, Guo J. Insights into protein–DNA interactions from hydrogen bond energy-based comparative protein–ligand analyses. *Proteins*. 2022;90(6):1303-1314. doi:10.1002/prot.26313