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 Charlotte, Charlotte, NC, 28223, USA

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

Running title: Hydrogen bonds in protein-ligand complexes

Correspondence

Jun-tao Guo, Department of Bioinformatics and Genomics, University of North Carolina Charlotte, Charlotte, NC, 28223, USA

Telephone: 704-687-7492

Email: jguo4@uncc.edu

Conflict of interest: None declared.

Funding statement: This work was supported by the National Science Foundation [DBI-2051491 to J.G]; and National Institutes of Health [R15GM132846 to J.G].

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 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 hydrogen bond energy. 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: protein-DNA, binding specificity, hydrogen bond energy, protein-ligand, minor groove

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 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 non-specific [3]. Protein-peptide (PT) interactions account for up to 40% of cellular interactions and are 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 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 hydrogen bonds 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 non-specific, intermediate complex with high energy [16–19]. While most of this recognition is expected to occur through the signature hydrogen bonding pattern of 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 side chains 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 hydrogen bonds 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 hydrogen bonds at the interface of 28 X-ray crystal structures of protein-DNA complexes. The hydrogen bonds 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 hydrogen bonds at the interface and within protein chains of 103 protein-peptide complexes. They further compared the types of hydrogen bonds 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 hydrogen bond is ~30% [32]. Zhou and Wang recently compared short hydrogen bonds, where donor-acceptor distance is less than 2.7Å, in1663 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 hydrogen bonds 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 hydrogen bonds with a single energy cutoff or a distance/angle cutoff can provide useful information about the role of hydrogen bonds in protein-ligand interaction, they have an intrinsic flaw since strong and weak hydrogen bonds are treated equally. Moreover, the distributions of interfacial hydrogen bonds 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 hydrogen bonds 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 hydrogen bonds 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.

MATERIALS AND METHODS

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] (**Error! Reference source not found.**). 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 (Supplementary 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 non-redundant 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 2Å [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/hetero dimer library, determines the oligomeric state from PDB files [40]. PHDL contains non-redundant heterodimers (Supplementary Table S1A) and homodimers (Supplementary 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, non-redundant and highly diverse datasets (Figure 1): (i) PDnrall, a 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).

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 affects 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, i.e., residues that are not on the N or C terminal of the chain were discarded. Lastly, 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.

Identification of HBs

Two widely used hydrogen bond 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 Eq. 1 [42,45]

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

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 hydrogen bonds 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 degrees or higher [43].

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 Eq. 2 [35,47]:

$$iSA = \frac{SA_P + SA_L - SA_C}{2} \qquad \textbf{\textit{Eq. 2}}$$

where SA_L and SA_P 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_C 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 side chain-side chain 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 hydrogen bonds (HBall). The SC-SC hydrogen bonds, 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.

Statistical tests:

Wilcoxon rank sum test was employed to assess if there are significant differences between samples across datasets. Chi-squared goodness of fit test was used to test the categorical distributions of types and the energy of hydrogen bonds at interface and within intrachain.

RESULTS

Hydrogen bonds 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 2 A&B). The number of HBall and HBSP in PT complexes are significantly less than those in PP complexes (Figure 2

A&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 S1 A&B). Interestingly, when the FIRST energy cutoff is set at -0.1 kcal/mol, the results are more similar to the HBPLUS data (Figure S2 A&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 hydrogen bonds, HBall and HBSP, by the interfacial surface area (iSA). Figure 2C and 2D show 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 0.043, 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 & S2C). These results are also in agreement with a previous study that PT interface has more total HBs per 100 Å^2 interface 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.

Types of hydrogen bonds 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 hydrogen bonds within proteins in all three types of complexes. BB-BB hydrogen bonds represent the largest number of overall hydrogen bonds within proteins (65-69%) followed by the Mixed (17-20%) and SC-SC (14-15%) hydrogen bonds 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 hydrogen bonds.

The distributions of the hydrogen bond 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 hydrogen bonds 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 hydrogen bonds 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 hydrogen bonds.

A previous study on protein-protein complexes indicated that the larger number of BB-BB hydrogen bonds 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 hydrogen bonds is the lowest. The percentage of interface BB-BB hydrogen bonds 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 hydrogen bonds 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 hydrogen bonds 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 an energy cutoff of –0.1 kcal/mol (Table 3, Figure S3-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-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 0.3743 and 0.6685 respectively) as well as between homodimers and heterodimers (p-values of 0.9371 and 0.9746 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 0.992 and 0.246 respectively). While there is no difference for the intrachain distributions between InterPep and LEADSPEP (p-value = 0.954), the interface distributions are different (p-value = 0.003) from FIRST HB annotations (Figure S6A). This might be a result of the relatively small LEADSPEP dataset with a small number of total hydrogen bonds (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 S6B).

Strength of hydrogen bonds at interface and within protein chain

We classified the strength of hydrogen bonds into four categories based on hydrogen bond energy from the FIRST program with different energy cutoffs used in previous studies as shown in Table 2 [17,42,48]. For intrachain hydrogen bonds within proteins, no significant differences were found among the three types of complexes (Figure 4A and Table 4). Most of the hydrogen bonds (67-70%) are strong ones with lower than -1.5 kcal/mol energy (category IV) while very few of them are of intermediate energy (less than 15% when categories II and III are combined), suggesting that the hydrogen bonds in all types of proteins have similar energy distribution with predominantly strong hydrogen bonds.

To investigate if the energy categories are related to different HB types, we compared the distributions of each type of intrachain hydrogen bonds 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 hydrogen bond energy distribution for each HB type among the PP, PT, and PD complexes. There is a higher percentage of strong BB-BB hydrogen bonds in all complexes, but relatively fewer strong ones for the Mixed HBs, suggesting that the major secondary structure types patterned by the BB-BB hydrogen bonds are optimized in terms of both distance and angle and form strong hydrogen bonds.

However, the interface hydrogen bond 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 (18.9% and 66.2%) 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.

Comparison of hydrogen bonds between HS and MS datasets

In our previous study, we demonstrated that highly specific HS protein-DNA complexes have more hydrogen bonds than the multi-specific MS protein-DNA complexes, including both total hydrogen bonds and sidechain-base hydrogen bonds [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-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 non-significant 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 hydrogen bonds and sidechain-base hydrogen bonds in the major and minor grooves. Between major and minor grooves, there is no significant difference in terms of hydrogen bond 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 hydrogen bonds (category IV) and fewer weak hydrogen bonds (category I) than those in the MS complexes in the minor groove. The MS complexes have about double the percentage of weak hydrogen bonds 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.

DISCUSSION

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

Our analyses revealed several important findings. First, for intrachain hydrogen bonds, our analysis not only corroborates several previous findings [14,25], but also provide additional information by demonstrating no significant difference in the distributions of HB energy among different complexes. Second, at the interface, the hydrogen bond distributions of PD complexes differ from both PP and PT complexes significantly in three aspects: (a) the total number of hydrogen bonds, the number of sidechain-base hydrogen bonds, 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 hydrogen bonds in protein-DNA interfaces; and (c) there is a significant difference of the minor groove hydrogen bonds 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 kcal/mol, -0.6 kcal/mol, -1.0 kcal/mol, and -1.5kcal/mol) from previous studies (Table 2) [17,42,48]. To test if similar results can be observed with different HB energy discretization, the hydrogen bonds were grouped using a larger energy range separated by -0.1 kcal/mol, -0.7 kcal/mol, -1.3 kcal/mol, and -2.0 kcal/mol (Table S4). The results of HB energy distributions, shown in Figures S9-S12 and Table S5-S6, are in agreement with conclusions (Figures 4-7, Table 4 and Table 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 hydrogen bonds is beneficial to quality assessment of protein-DNA complexes 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 hydrogen bonds 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 hydrogen bonds 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 hydrogen bond energy and strength [46]. While DNA is not a rigid molecule, the double helical nature restricts the atoms that can form optimal hydrogen bonds 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 hydrogen bonds and the sidechain-base hydrogen bonds 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 hydrogen bonds. Our results further demonstrate that the minor groove HBs play more critical roles in conferring binding specificity than previously thought.

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. Encyclopedia of Bioinformatics and Computational Biology. In: Ranganathan S, Gribskov M, Nakai K, Schönbach CBT-E of B and CB, eds. 1st ed. Academic Press; 2019:142-154. doi:https://doi.org/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 Bioinforma*. 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:https://doi.org/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, Ward SEBT-CMCIII, eds. Elsevier; 2017:154-201. doi:https://doi.org/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. Published online 2002:48-52.

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:https://doi.org/10.1006/bbrc.1999.1294

11. Hubbard RE, Kamran Haider M. Hydrogen Bonds in Proteins: Role and Strength. *eLS*. Published online February 15, 2010. doi: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:https://doi.org/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 U S A*. 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 Bioinforma*. 2009;75(2):413-429. doi:10.1002/prot.22248

24. Kota P, Ding F, Ramachandran S, Dokholyan N V. 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 A V, 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 Jr RL. 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 Bioinforma*. 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*. 1997;94(19):10172 LP - 10177. doi:10.1073/pnas.94.19.10172

47. Hubbard S, Thornton J. NACCESS: Department of Biochemistry and Molecular Biology, University College London. Published online 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*. 2003;100(22):12683 LP - 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 U S A*. 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 Bioinforma*. 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

FIGURES LEGENDS

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

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 ≤ 0.001, ** = p-value ≤ 0.01

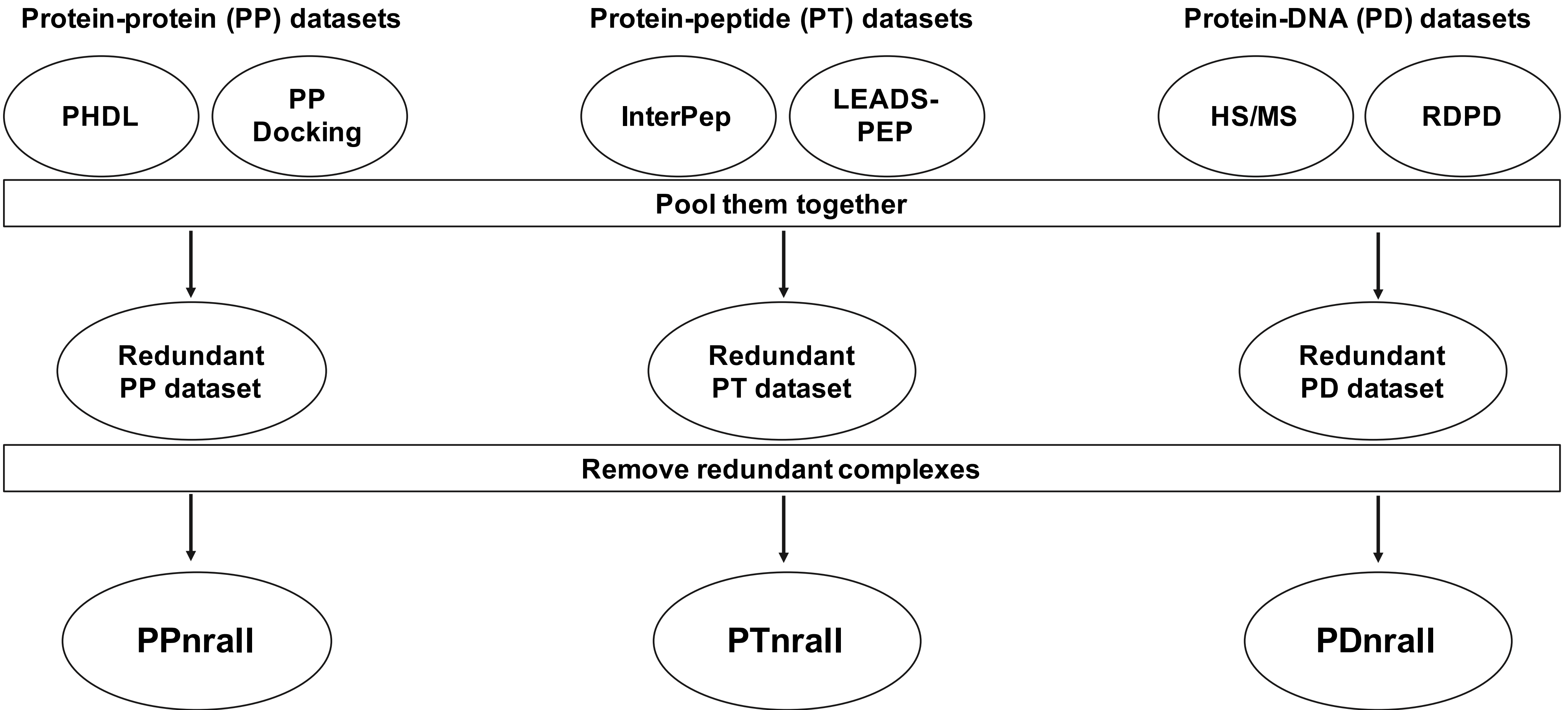
Figure 3. Comparisons of the distribution of different types of hydrogen bonds, backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and Mixed (BB-SC and SC-BB) for (**A**) intrachain within proteins and (**B**) at interface of PP, PT and PD complexes. The hydrogen bonds are annotated from the FIRST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

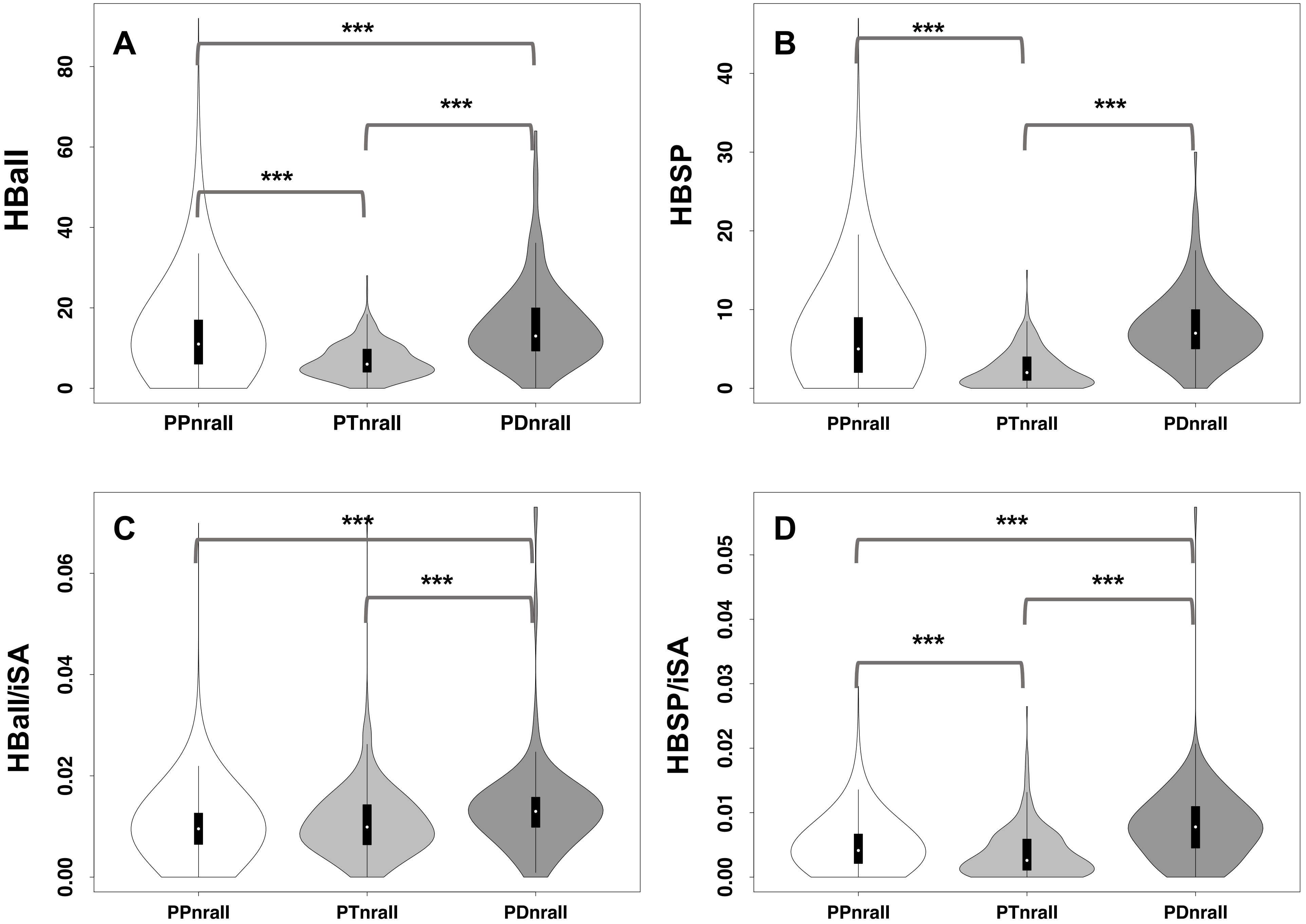
Figure 4. Comparisons of the distributions of hydrogen bond energy for (**A**) intrachain and (**B**) at interface.

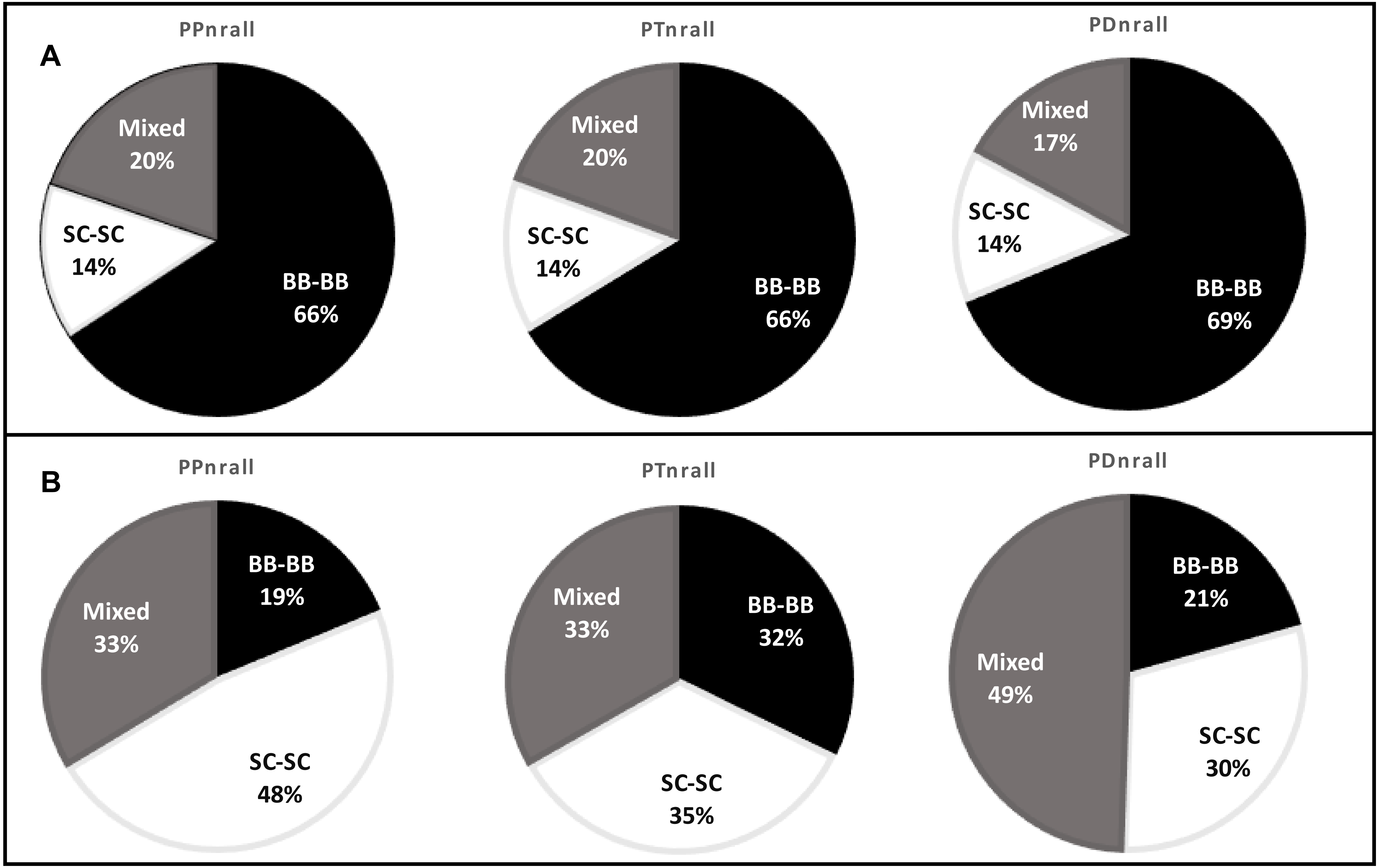
Figure 5. Comparison of (**A**) intrachain hydrogen bond energy and (**B**) interface hydrogen bond energy in different hydrogen bond types.

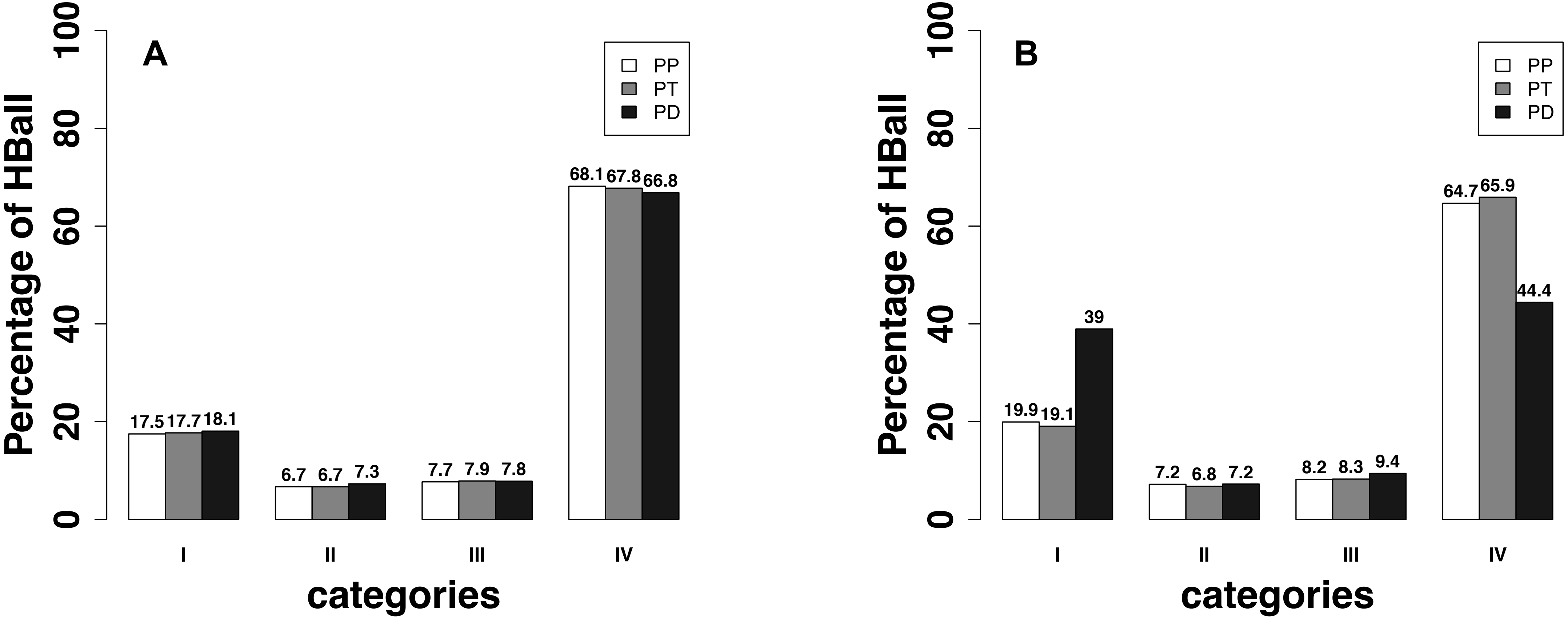
Figure 6. Comparison of major groove for (**A**) HBall and (**B**) HBSP energy distributions between HS and MS complexes.

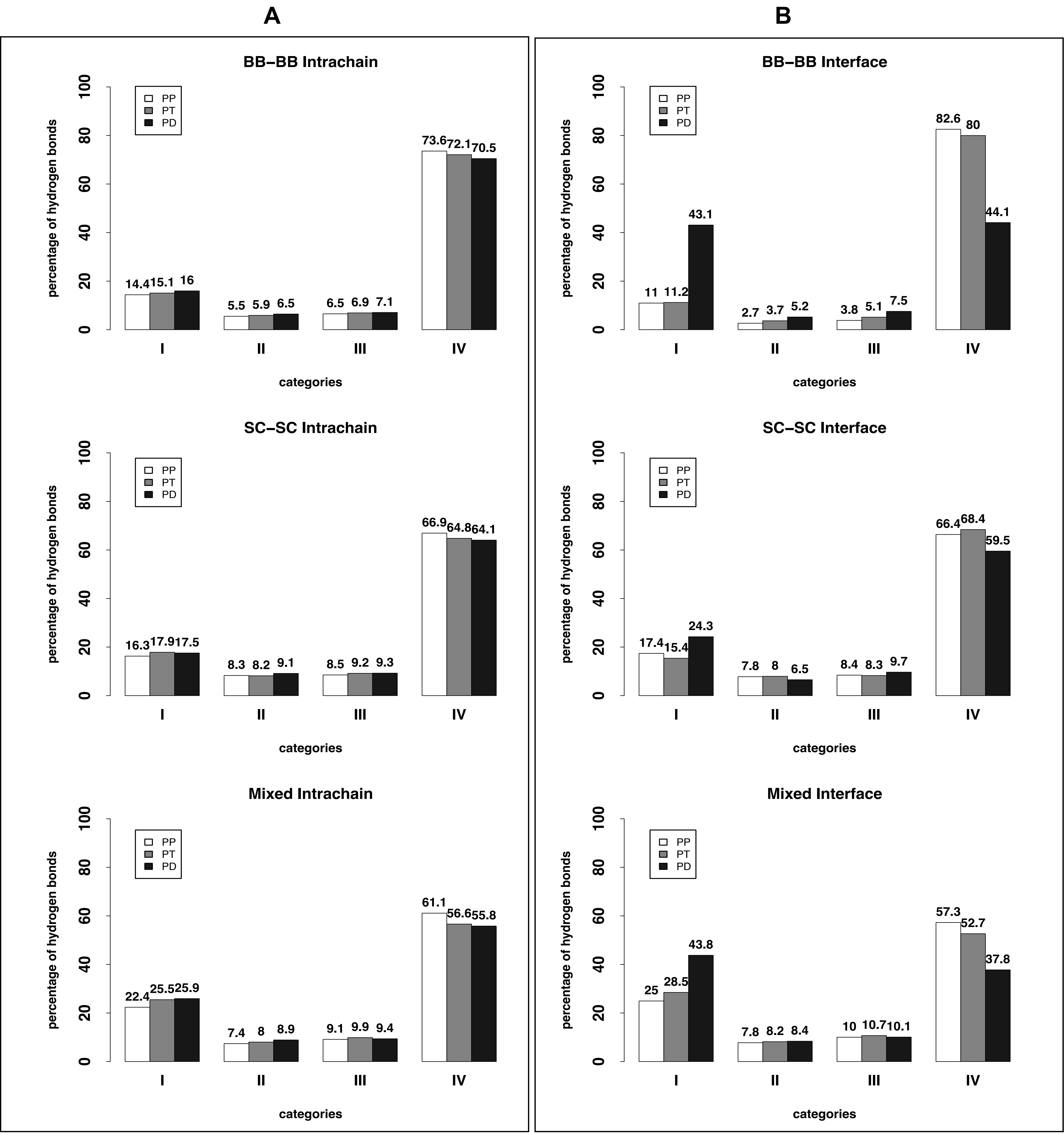
Figure 7. Comparison of minor groove for (**A**) HBall and (**B**) HBSP energy distributions between HS and MS complexes.

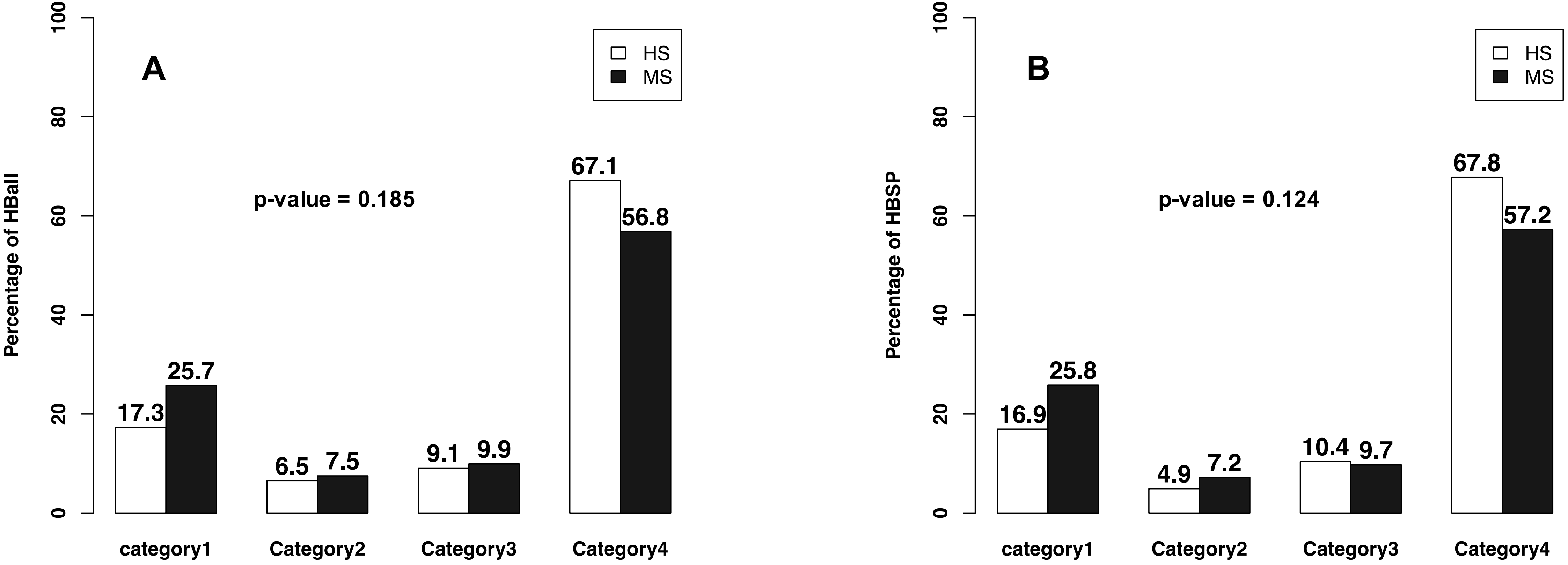

Protein-protein (PP) datasets
Protein-peptide (PT) datasets
Protein-DNA (PD) datasets
PHDL
PP Docking
InterPep
LEADS-PEP
HS/MS
RDPD
Pool them together
Redundant PP dataset
Redundant PT dataset
Redundant PD dataset
Remove redundant complexes
PPnrall
PTnrall
PDnrall


A
B
C
D
HBall
HBSP
HBall/iSA
HBSP/iSA
***
PPnrall
PTnrall
PDnrall
0
20
40
60
80
10
30
0.00
0.01
0.02
0.03
0.04
0.05
0.06


A
PPnrall
Mixed 20%
SC-SC 14%
BB-BB 66%
PTnrall
Mixed 20%
SC-SC 14%
BB-BB 66%
PDnrall
Mixed 17%
SC-SC 14%
BB-BB 69%
B
PPnrall
BB-BB 19%
Mixed 33%
SC-SC 48%
PTnrall
Mixed 33%
BB-BB 32%
SC-SC 35%
PDnrall
BB-BB 21%
Mixed 49%
SC-SC 30%


A
PP
PT
PD
Percentage of HBall
0
20
40
60
80
100
17.5
17.7
18.1
6.7
6.7
7.3
7.7
7.9
7.8
68.1
67.8
66.8
I
II
III
IV
categories
B
PP
PT
PD
Percentage of HBall
0
20
40
60
80
100
19.9
19.1
39
7.2
6.8
7.2
8.2
8.3
9.4
64.7
65.9
44.4
I
II
III
IV
categories


A
B
BB–BB Intrachain
percentage of hydrogen bonds
□ PP
▨ PT
■ PD
14.4 15.1 16
5.5 5.9 6.5
6.5 6.9 7.1
73.6 72.1 70.5
I
II
III
IV
categories
BB–BB Interface
11 11.2 43.1
2.7 3.7 5.2
3.8 5.1 7.5
82.6 80 44.1
SC–SC Intrachain
16.3 17.9 17.5
8.3 8.2 9.1
8.5 9.2 9.3
66.9 64.8 64.1
SC–SC Interface
17.4 15.4 24.3
7.8 8 6.5
8.4 8.3 9.7
66.4 68.4 59.5
Mixed Intrachain
22.4 25.5 25.9
7.4 8 8.9
9.1 9.9 9.4
61.1 56.6 55.8
Mixed Interface
25 28.5 43.8
7.8 8.2 8.4
10 10.7 10.1
57.3 52.7 37.8
0
20
40
60
80
100


A
B
□ HS
■ MS
Percentage of HBall
Percentage of HBSP
0
20
40
60
80
100
p-value = 0.185
p-value = 0.124
17.3
25.7
6.5
7.5
9.1
9.9
67.1
56.8
16.9
25.8
4.9
7.2
10.4
9.7
67.8
57.2
category1
Category2
Category3
Category4

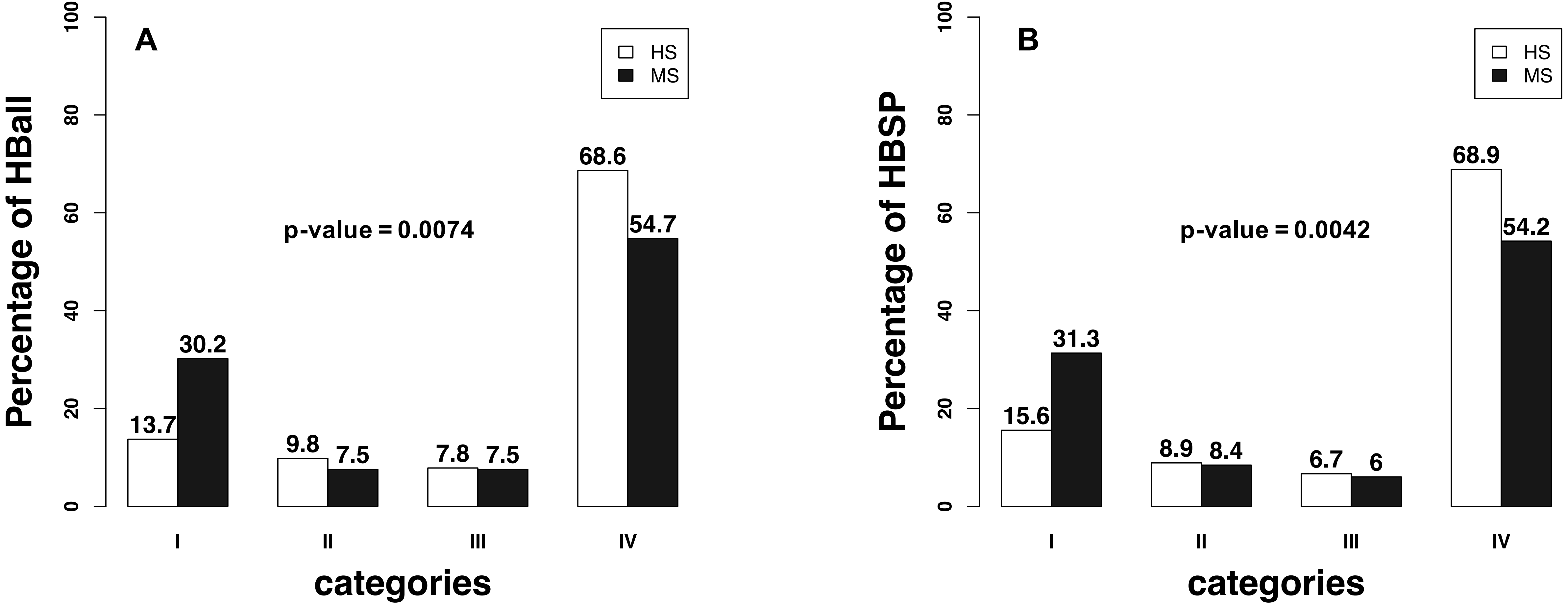

A
Percentage of HBall
0
20
40
60
80
100
□ HS
■ MS
13.7
30.2
9.8
7.5
7.8
7.5
68.6
54.7
p-value = 0.0074
I
II
III
IV
categories
B
Percentage of HBSP
0
20
40
60
80
100
□ HS
■ MS
15.6
31.3
8.9
8.4
6.7
6
68.9
54.2
p-value = 0.0042
I
II
III
IV
categories

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 < 2Å, R-factor < 0.3	3-12 residues	~512 Å^2
Protein-Protein	Protein homo/hetero dimer 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

Table 1. 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

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

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

Supplementary data

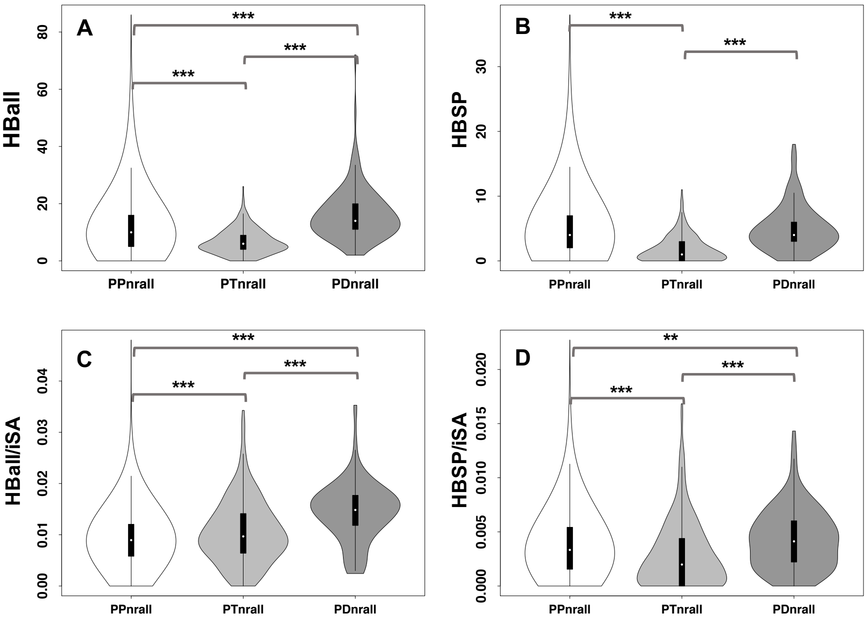


Figure S1. Comparison of interfacial hydrogen bonds based on HBPLUS with default parameters: (**A**) the number of total hydrogen bonds (HBall); (**B**) the number of SC-SC or SC-base hydrogen bonds (HBSP); (**C**) the ratio HBall to interfacial surface area (iSA); and (**D**) the ratio of HBSP to iSA.

*** = p-value ≤ 0.001; ** = p-value ≤ 0.01

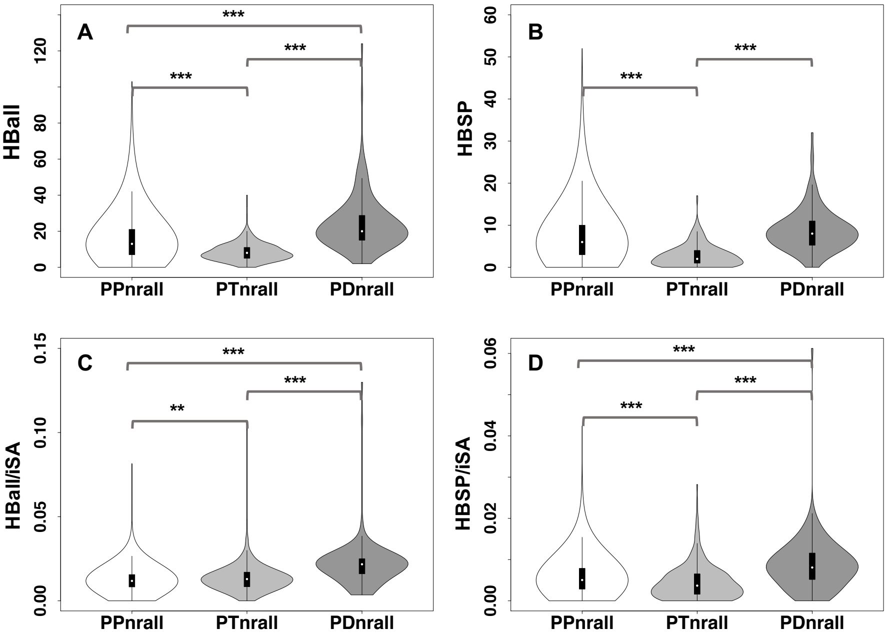


Figure S2. Comparison of interfacial hydrogen bonds based on FIRST with an energy cutoff of -0.1 kcal/mol: (**A**) the number of total hydrogen bonds (HBall); (**B**) the number of SC-SC or SC-Base hydrogen bonds (HBSP); (**C**) the ratio of HBall to interfacial surface area (iSA); and (**D**) the ratio of HBSP to iSA.

*** = p-value ≤ 0.001, ** = p-value ≤ 0.01

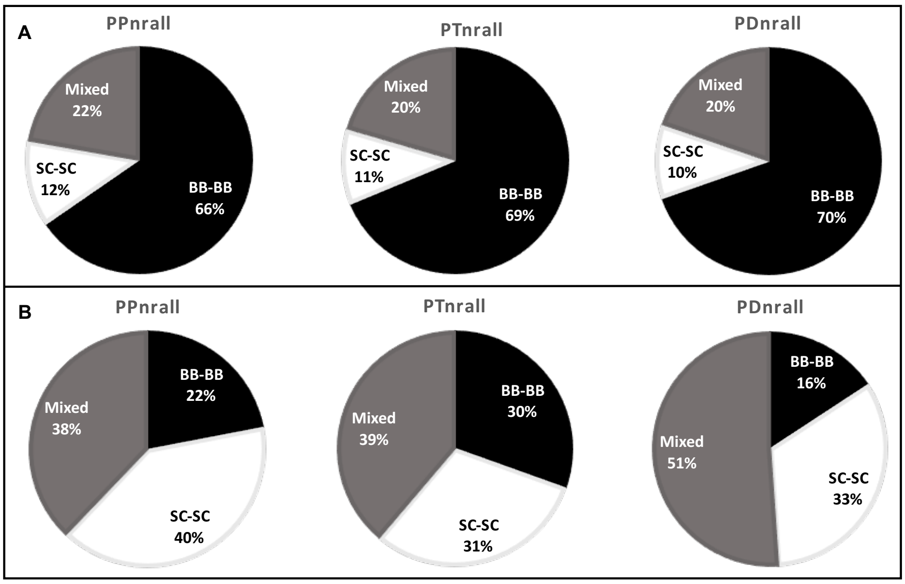


Figure S3. Comparison of the distributions of hydrogen bond types with HBPLUS: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) at (**A**) intrachain and (**B**) interface of PP, PT and PD complexes. (See p-values in Table 3)

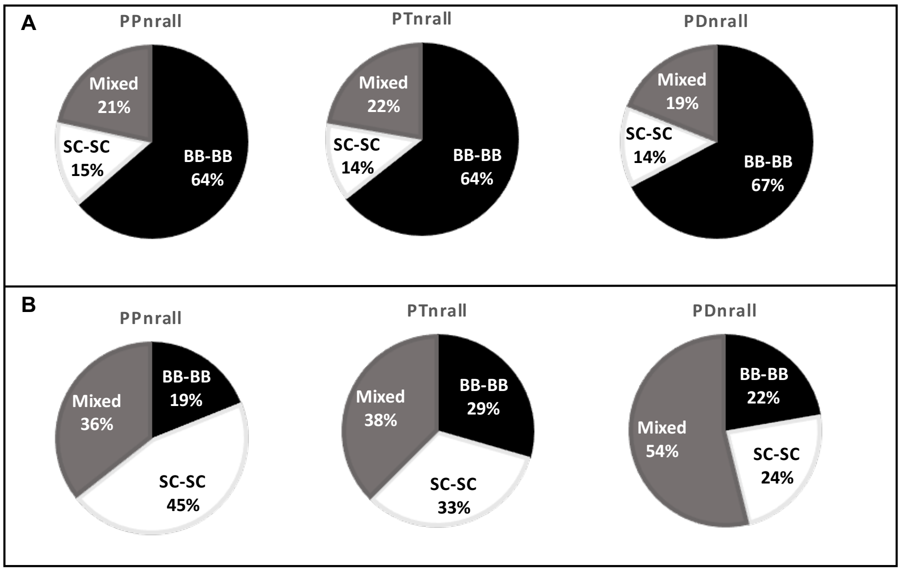


Figure S4. Comparisons of the distribution of different types of hydrogen bonds, backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and Mixed (BB-SC and SC-BB) for (**A**) intrachain within proteins and (**B**) at interface of PP, PT and PD complexes. The hydrogen bonds are annotated from the FIRST program with an energy cutoff of -0.1 kcal/mol. (See p-values in Table S2)

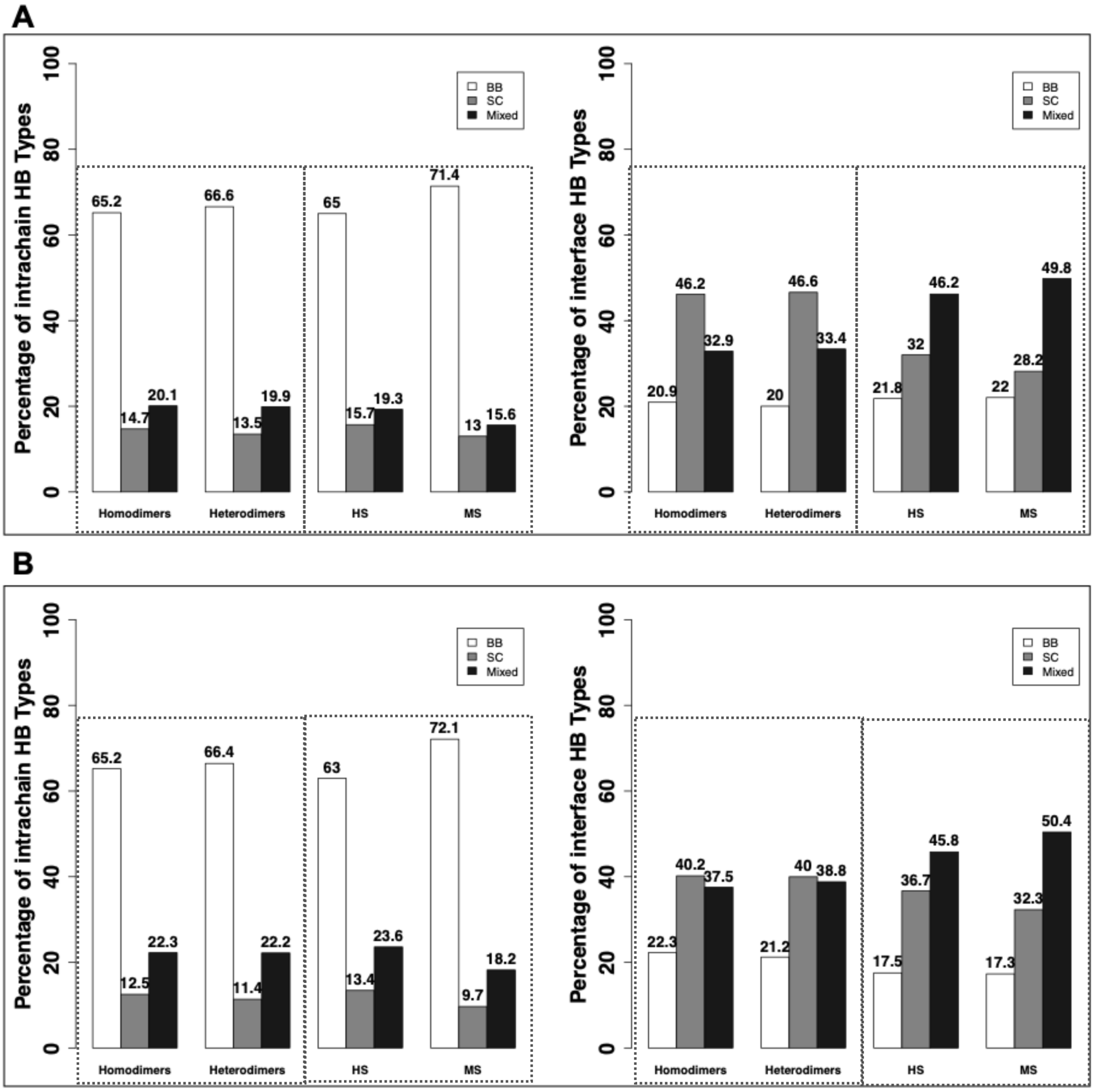


Figure S5. Comparison of the percentages of HB types: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) in intrachain and interface of homodimers, heterodimers, highly specific and multi-specific protein-DNA complexes. (**A**)The hydrogen bonds are annotated by FIRST with an energy cutoff of -0.6 kcal/mol. (**B**) The hydrogen bonds are annotated by HBPLUS.

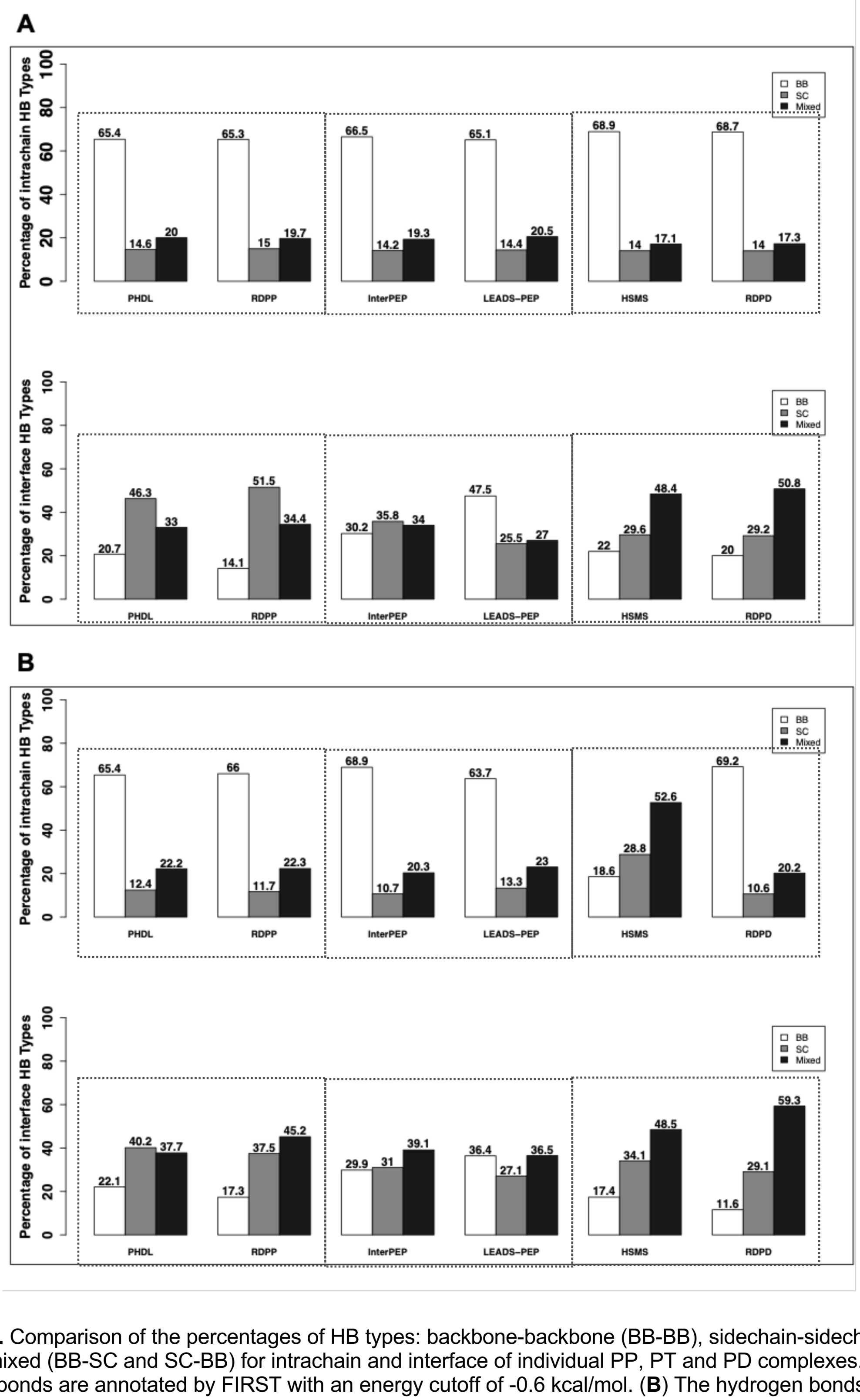


Figure S6. Comparison of the percentages of HB types: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) for intrachain and interface of individual PP, PT and PD complexes. (**A**) The hydrogen bonds are annotated by FIRST with an energy cutoff of -0.6 kcal/mol. (**B**) The hydrogen bonds are annotated by HBPLUS.

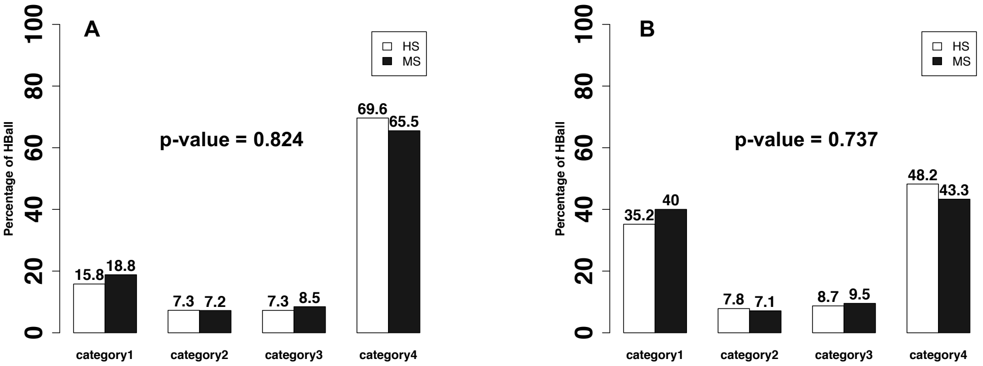


Figure S7. Comparison of the categories of hydrogen bond energy (based on Table 2) between HS and MS complexes. (**A**) intrachain; (**B**) interface.

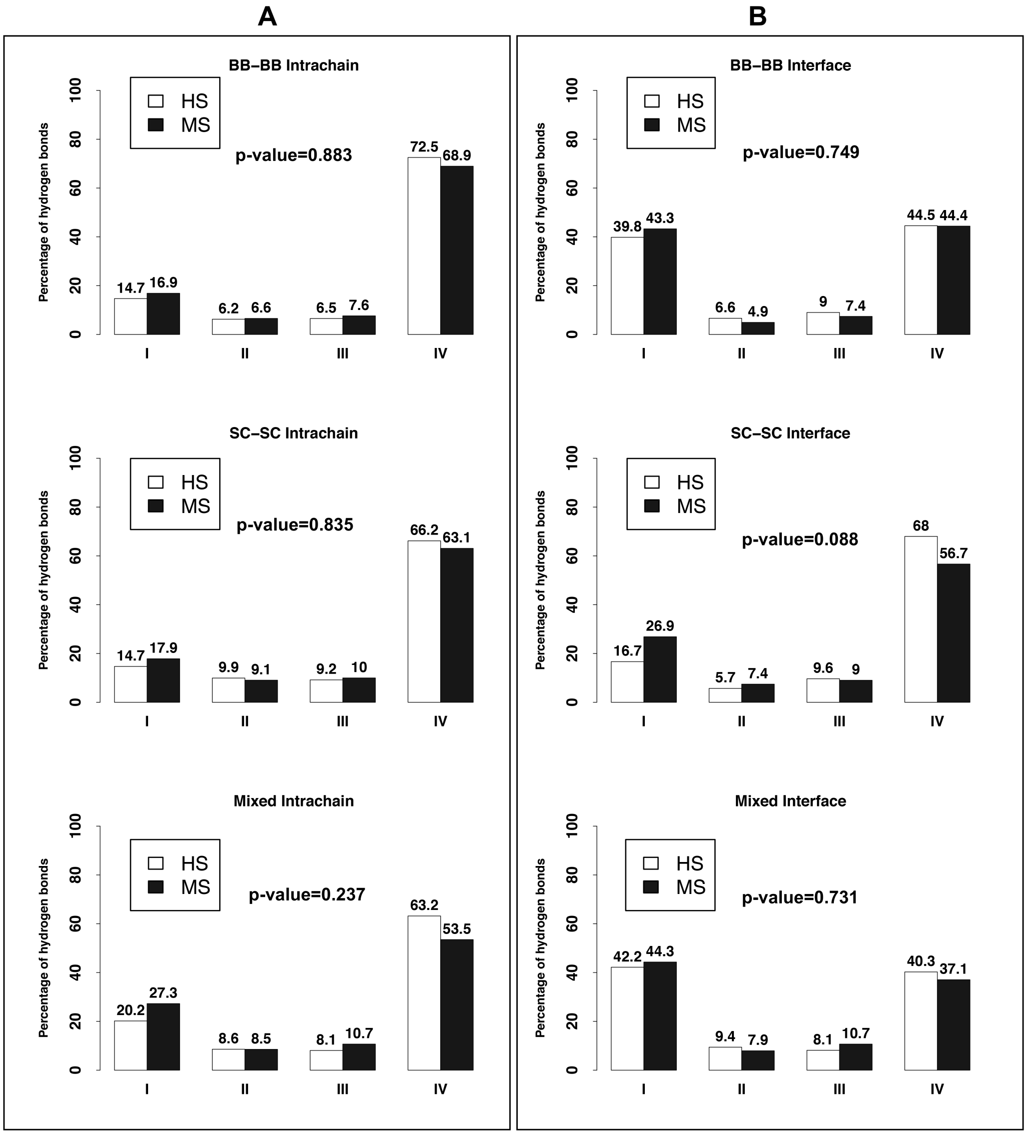


Figure S8. Comparison of hydrogen bond energy categories (based on Table 2) in different hydrogen bond types between HS and MS complexes. (**A**) intrachain; (**B**) interface.

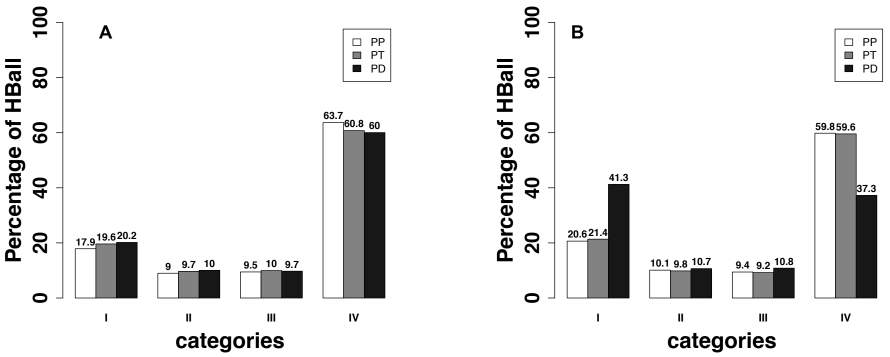


Figure S9. Comparisons of the distributions of hydrogen bond energy based on the discretization in Table S4 for (**A**) intrachain and (**B**) at interface. (See Table S5 for p-values).

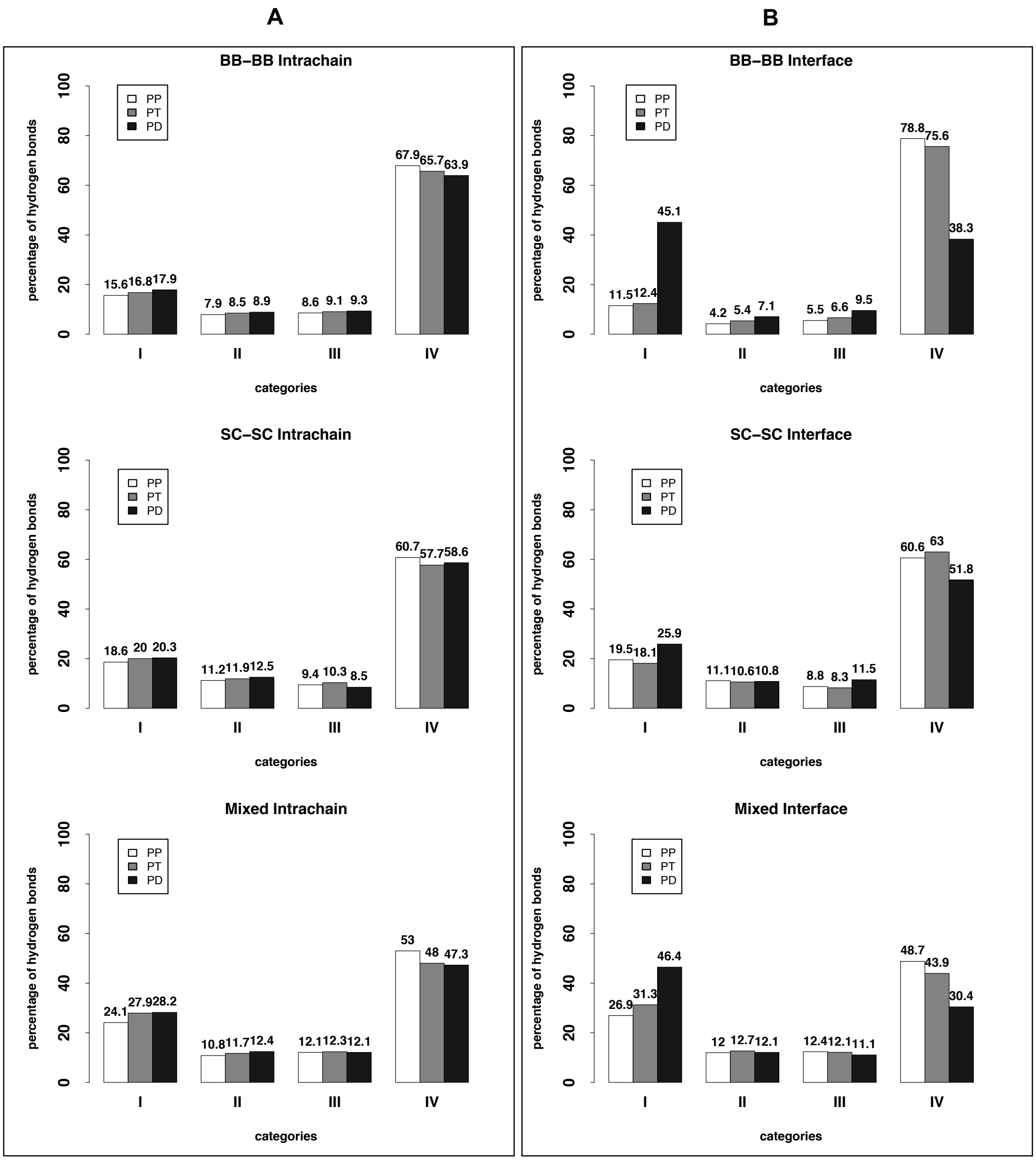


Figure S10. Comparison of (**A**) intrachain hydrogen bond energy and (**B***)* interface hydrogen bond energy (based on the discretization in Table S4) in different hydrogen bond types (See Table S6 for p-values).

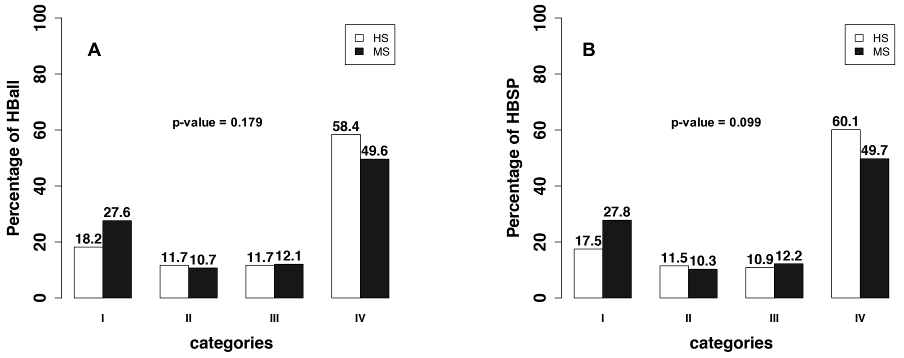


Figure S11. Comparison of major groove for (**A**) HBall and (**B**) HBSP energy distributions (based on the discretization in Table S4) between HS and MS complexes.

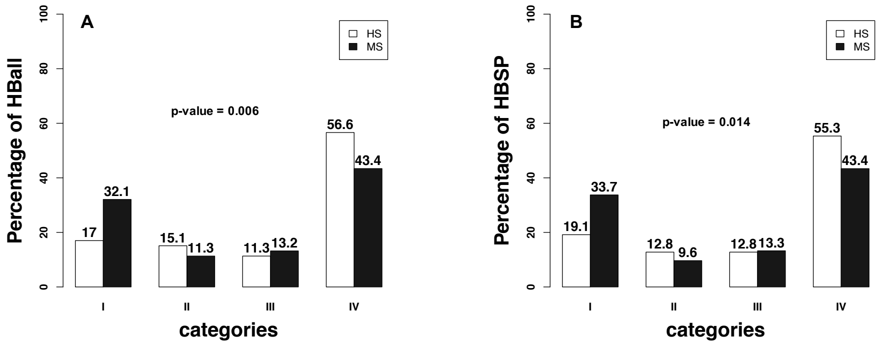


Figure S12. Comparison of minor groove for (**A**) HBall and (**B**) HBSP energy distributions (based on the discretization in Table S4) between HS and MS complexes.

Table S1. PDB ids in the protein homo/heterodimer library (PHDL)

(**A**) PDB ids of the heterodimers in PHDL

1AY7	1BDJ	1BH9	1BVN	1CXZ	1D0D	1D4T	1DJ7	1DOW	1DS6	1E44	1E96	1EUV	1F3V	1FM2
1FXW	1GK9	1J2J	1JIW	1JQL	1KTP	1M2T	1MK2	1MTP	1MZW	1NME	1NPE	1OO0	1ORY	1PDK
1QGE	1QTX	1R0R	1R8S	1SPP	1SVD	1T0P	1TA3	1TMQ	1U0S	1UGH	1V5I	1V74	1W98	1WMH
1WQJ	1WRD	1WYW	1XG2	1XTG	1Y43	1Z0J	1Z3E	1Z5Y	2A5D	2A9K	2AQ2	2BCG	2BKR	2C1M
2C7M	2D5R	2DVW	2EHB	2F4M	2FCW	2FHZ	2FTX	2GSK	2H7Z	2H9A	2HRK	2HTH	2IE4	2O3B
2OOB	2OZN	2P45	2P8Q	2PA8	2PTT	2QWO	2R25	2UUY	2V3B	2V8S	2V9T	2VPB	2WBW	2WWX
2WY8	2XJY	2XN6	2XPP	2YGG	2Z30	3A2F	3A8G	3AA7	3AON	3AQF	3B0C	3BH7	3BS5	3BY4
3CF4	3CKI	3CNQ	3D3B	3D6N	3DAW	3DGP	3EGV	3F1P	3F75	3FJU	3FMO	3FPU	3FXE	3GJ3
3GOV	3K1R	3KNB	3L51	3LQC	3MCB	3ME0	3MKR	3MXN	3N1M	3NCE	3NVN	3NY7	3O2P	3O3O
3ONA	3OQ3	3P71	3P73	3PLV	3PNL	3PT8	3QDR	3QHY	3QQ8	3SDE	3SHG	3TBI	3TJ5	3UB5
3V61	3VF0	3VRD	3VYR	3VZ9	3WDG	3X37	3YGS	3ZG9	4APX	4BVX	4C2A	4C4P	4CMM	4CRW
4CSR	4DBG	4DHI	4DRI	4F7G	4FBJ	4G01	4G6T	4GN4	4H5S	4H6J	4HST	4HT3	4IU3	4IUM
4J38	4JE3	4JS0	4K12	4K5A	4KAX	4L2I	4LJO	4LLD	4LZX	4M0W	4MRT	4NBX	4NTQ	4NUT
4OB0	4PAS	4PZ5	4QJF	4QLP	4QO1	4R1D	4RCA	4RHZ	4RLJ	4U9H	4UAF	4UHZ	4UN2	4UQZ
4UYQ	4UZZ	4W8P	4WKS	4X86	4X8K	4XAX	4XYD	4YH8	4YI0	4YYP	4ZGM	4ZHY	4ZQU	5AQV
5B64	5B78	5BY8	5BZ0	5C50	5CEC	5CHL	5D6J	5DYN	5EU0	5EUI	5F22	5FOY	5G1X	5GNA
5GXW	5GZT	5H3J	5HE9	5HKQ	5HKY	5I4H	5INB	5IVA	5JCA	5JP1	5JW9	5KYC	5L0R	5L0V
5L3D	5L9Z	5LSI	5LXR	5M0Y	5M2O	5M72	5MAW	5ML9	5MS2	5MU7	5NCW	5NRM	5O33	5OOV
5OW0	5OXZ	5OYL	5SVH	5T51	5T86	5TUU	5TVQ	5TZP	5UIW	5UN7	5UNI	5UUK	5V7P	5VGB
5VKO	5VMO	5WUJ	5WXK	5XA5	5XEC	5XLU	5Y27	5Y38	5YCA	5YR0	5YWR	5Z51	5ZNG	5ZWL
5ZZA	6APP	6AU8	6BN1	6BSC	6BW9	6DLM	6DRE	6DXZ	6EH4	6EM7	6ES1	6F2G	6F6R	6FDK
6FFA	6FUD	6GHO	6GR8	6H02	6H9U	6HM3	6HUL	6IUA	6J4P	6JLE	6JXH	6K06	6K3B	6KGC
6KHS	6KMJ	6KXD	6L4P	6L8G	6LBX	6LKI	6LPH	6M0J	6MBB	6MGN	6MIB	6MS4	6NE2	6NVX
6ODD	6OP8	6OQ7	6OVM	6OX6	6Q00	6QBA	6QUP	6R6M	6RCX	6RM9	6RTW	6S07	6S3F	6S8Q
6SWT	6U3B	6U54	6UUI	6V7M	6VE5	6VJJ	6W0V	6W9S	6WCW	6WG4	6WH1	6WJC	6WUD	6XRU
6XZU	6YX5	6YZ5	6ZXW	7A48	7BQV	7BZK	7C96	7CE4	7CN7	7CQ3	7EDP	7JTU	7MC5	

(**B**) PDB ids of the homodimers in PHDL

1A8U	1AA7	1AOC	1B2P	1B43	1B5E	1B6Z	1BD9	1BDY	1BO4	1BYF	1C77	1C8U	1CI9	1CKM
1CKU	1CQX	1D0C	1D0Q	1D1G	1D2C	1D2O	1D7F	1D9C	1DEB	1DJ0	1DL5	1DOK	1DPG	1DQE
1DQP	1DQZ	1DU5	1DYS	1E0B	1E19	1E5L	1E7L	1E8U	1EAJ	1EBF	1EE8	1EEJ	1EF0	1EJF
1ELU	1EVX	1EXT	1EYQ	1EZG	1F08	1F0K	1F46	1F5V	1F86	1FBQ	1FBT	1FLM	1FN9	1G29
1G64	1G8E	1G8L	1G8M	1GDE	1GE7	1GQI	1GT1	1GU7	1GVJ	1GXJ	1GYO	1GYX	1H18	1HKQ
1HQS	1HRU	1HYO	1HZ5	1I07	1I0R	1I4S	1I4U	1I6W	1I78	1I7N	1IAZ	1IG0	1IGU	1II7
1IPS	1IQ8	1IRQ	1ITU	1ITV	1IUJ	1IX2	1IX9	1IYB	1IZY	1J0H	1J3M	1J49	1JAD	1JAY
1JK6	1JLY	1JNP	1JR8	1JYA	1JZT	1K3Y	1K4Z	1K66	1KAE	1KDG	1KFI	1KJN	1KKO	1KNQ
1KPT	1KQL	1KQP	1KTJ	1KV0	1KZQ	1LGQ	1LJM	1LN0	1LQ9	1LQA	1M2D	1M4J	1M76	1MBY
1MK4	1MKF	1MKK	1MKZ	1MO9	1MXR	1MY7	1MZG	1N1E	1N2Z	1NBC	1ND4	1NKI	1NNW	1NO5
1NS5	1NSZ	1NU4	1NV7	1NWP	1NWW	1NXM	1NXU	1O5X	1O63	1O6A	1OC2	1OC9	1OCK	1OFZ
1OH0	1OI6	1OKI	1ON2	1OOE	1OR4	1ORD	1ORU	1OSY	1OTK	1OVN	1OX8	1P1C	1P4O	1P5T
1P65	1P6O	1PC6	1PIW	1PIX	1PKV	1PL5	1PPV	1PSR	1Q6O	1QAH	1QC5	1QFH	1QH5	1QI9
1QKS	1QL0	1QLW	1QMH	1QO8	1QQ5	1QSD	1QUP	1QVE	1QVZ	1QXR	1R11	1R12	1R1D	1R61
1R7A	1RDO	1REG	1RFY	1RKT	1RKU	1RW0	1S0P	1S4K	1S9R	1SBY	1SD4	1SEI	1SFN	1SGM
1SJ1	1SMO	1SNN	1SQS	1SU2	1SXH	1SXR	1SZQ	1T06	1T1V	1T3C	1T6S	1T6T	1T7S	1T92
1TBX	1TE2	1TE5	1TEJ	1TJ7	1TLJ	1TU1	1TV8	1TVN	1TXG	1U07	1U5U	1U6R	1U6Z	1UCR
1UDV	1UIX	1UKK	1USC	1USO	1UWK	1UZ3	1V4E	1V58	1V5V	1V5X	1V6P	1V6Z	1V7L	1V7O
1V8H	1VB5	1VC4	1VH5	1VHD	1VHQ	1VHZ	1VJH	1VJQ	1VL7	1VSC	1W5R	1W9C	1WKV	1WLG
1WMX	1WPN	1WR8	1WRA	1WTJ	1WWA	1WWP	1WWZ	1WY2	1WY5	1WZ3	1WZD	1X2I	1X7D	1X9I
1X9Z	1XEQ	1XGS	1XHK	1XJ4	1XNF	1XNG	1XRK	1XRU	1XSV	1XTA	1XVI	1XVS	1Y0H	1Y0U
1Y2O	1Y7M	1Y7R	1Y89	1Y9B	1Y9W	1YDY	1YGA	1YLM	1YLQ	1YLR	1YLX	1YOC	1YRB	1Z41
1Z4E	1Z5B	1ZBO	1ZBR	1ZC6	1ZK8	1ZKI	1ZO2	1ZQ9	1ZUO	1ZV1	1ZVF	1ZZG	2A0U	2A2J
2A4N	2A9U	2AIB	2AKZ	2ANX	2AQ6	2AQP	2AQX	2ARC	2ASK	2AUW	2AXW	2AYT	2B4H	2B6C
2B9D	2BDR	2BJI	2C0D	2C1L	2C2I	2C49	2C5A	2CAR	2CB5	2CC0	2CDU	2CH7	2CMG	2CO3
2CO5	2CTZ	2CU6	2CUN	2CVI	2CWK	2CXN	2D4G	2D73	2D7V	2D8D	2DBS	2DC0	2DC1	2DC3
2DC4	2DCT	2DFJ	2DJ5	2DKJ	2DLB	2DM9	2DOU	2DQL	2DR1	2DS5	2DSJ	2DSK	2DST	2DTC
2DXQ	2E2N	2E2X	2E5F	2E5Y	2E85	2EBE	2ECS	2ECU	2EG4	2EGD	2EGV	2EIX	2EJN	2EK0
2ERB	2ESR	2ETX	2EV1	2F02	2F07	2F1F	2F22	2F2E	2F48	2F5G	2F62	2F96	2F9H	2FA1
2FAE	2FBN	2FCA	2FFG	2FG0	2FHQ	2FIU	2FJR	2FM6	2FNU	2FP1	2FRE	2FSW	2FTR	2FUR
2FXV	2FZF	2FZT	2G3W	2G84	2GA1	2GAN	2GAX	2GBO	2GEC	2GEX	2GFF	2GIY	2GJ3	2GJA
2GKM	2GLZ	2GOM	2GSV	2GU9	2GUD	2GV8	2H1T	2H28	2H2N	2H2R	2H8G	2H98	2HA8	2HBV
2HDW	2HHJ	2HIN	2HIQ	2HO1	2HQ7	2HQY	2HS1	2HXR	2HZG	2I2O	2I5E	2I5G	2I7R	2I8D
2I9U	2IAB	2IB0	2IG3	2IGI	2IPR	2IUT	2IYC	2J05	2J85	2J8W	2J98	2JD3	2JDJ	2JHF
2NLV	2NNH	2NOG	2NQL	2NQT	2NS9	2NTT	2NX9	2NXV	2NYS	2NZ5	2NZ7	2O4C	2O6P	2O7M
2OB3	2OD0	2OD4	2ODA	2OEM	2OFC	2OGB	2OGI	2OHC	2OKU	2OKX	2OM6	2OMD	2OND	2ONF
2OPI	2OPL	2OQB	2OR2	2ORI	2OTA	2OU3	2OU5	2OXL	2OY9	2P08	2P0M	2P12	2P1A	2P1J
2P23	2P2S	2P3Y	2P4P	2P4R	2P62	2P64	2P8U	2P97	2P9H	2PA7	2PEB	2PFW	2PH0	2PIH
2PJS	2PL7	2PO3	2PR8	2PRV	2PS1	2PS5	2PUZ	2Q03	2Q0X	2Q24	2Q3V	2Q5C	2Q6Q	2Q7A
2Q8O	2Q8V	2QBU	2QE8	2QFR	2QGY	2QH9	2QHQ	2QIW	2QJD	2QJF	2QL8	2QLX	2QMX	2QND
2QQZ	2QRR	2QSI	2QU7	2QV0	2QVH	2QXY	2QYC	2QZZ	2R15	2R1F	2R1I	2R5O	2R74	2R8Q
2R8W	2RAS	2RB7	2RBB	2RBG	2RC8	2RCZ	2RDC	2RDE	2RGM	2RK0	2UUZ	2UW1	2V27	2V6K
2V9B	2VD8	2VGX	2VH3	2VKJ	2VOK	2VQ3	2VSW	2VVW	2W1T	2W1V	2W2K	2W31	2W3G	2W43
2W6A	2W8X	2WCR	2WCU	2WD6	2WK4	2WLV	2WMM	2WNS	2WNW	2WU9	2WUF	2WVF	2WW4	2WZV
2X2W	2X65	2X7X	2XDG	2XFN	2XFV	2XGG	2XHF	2XJ3	2XMJ	2XOL	2XR4	2XT2	2XUA	2XW7

```
2XWL  2XZ8  2XZ9  2Y27  2Y43  2YA8  2YEQ  2YFA  2YIO  2YMA  2YMQ  2YMY  2YR2  2YVE  2YVS
2YW2  2YWL  2YWW  2YXH  2YYB  2YYV  2YYY  2YZI  2Z0U  2Z5E  2Z6R  2Z73  2Z76  2Z8R  2ZB9
2ZDP  2ZEW  2ZGL  2ZGY  2ZO9  2ZOG  2ZVX  2ZVY  2ZW2  2ZW5  2ZX2  2ZYQ  2ZZV  3A1D  3A3D
3A8R  3AAB  3ABH  3AIA  3AJ6  3ALY  3AMI  3AOW  3ATJ  3B0F  3B42  3B4U  3B73  3B8X  3BA3
3BBD  3BBZ  3BCW  3BED  3BGA  3BHQ  3BJE  3BKX  3BL4  3BMZ  3BNK  3BOS  3BRC  3BRU  3BS9
3BWS  3BYP  3BZY  3C1Q  3C3Y  3C8C  3CCD  3CGU  3CJL  3CJP  3CKA  3CNK  3CP7  3CQR  3CRN
3CRY  3CSX  3CT6  3CTP  3CU2  3CW9  3CZ1  3CZ6  3CZZ  3D0F  3D34  3D3I  3D5P  3D7A  3DA5
3DFU  3DMC  3DME  3DN7  3DNF  3DP7  3DS2  3DSB  3DUP  3DUW  3DXO  3E1W  3E2C  3E2D  3E48
3E4V  3E7Q  3E8O  3E96  3EDE  3EDN  3EFY  3EGO  3EIK  3EKG  3ENT  3EO6  3EOF  3EOQ  3EPY
3EQZ  3ER7  3ERX  3ES4  3EUU  3EVI  3EWW  3EY8  3EZH  3F08  3F1L  3F3S  3F5H  3F6G  3F6O
3F6T  3F7E  3F84  3F9S  3F9T  3F9U  3FA5  3FCH  3FD4  3FD7  3FF9  3FGV  3FGY  3FH3  3FHU
3FIL  3FJ4  3FK9  3FKR  3FLD  3FOU  3FPF  3FPK  3FQM  3FR7  3FRQ  3FV6  3FVV  3FX7  3FYB
3G0T  3G16  3G1P  3G3Q  3G3S  3G3Z  3G46  3G4E  3G67  3G8K  3G8R  3GAE  3GAZ  3GB3  3GBY
3GDW  3GE6  3GFA  3GKX  3GLV  3GMG  3GMX  3GO6  3GOC  3GPV  3GR3  3GRD  3GRN  3GRO  3GU3
3GVE  3GW4  3GWK  3GWL  3GWN  3GWO  3GWR  3GZR  3H2B  3H3N  3H6R  3H8L  3H8U  3HA2  3HCN
3HDO  3HEB  3HG9  3HIM  3HIN  3HJ9  3HJG  3HL4  3HLU  3HLX  3HM4  3HMT  3HN0  3HNW  3HO7
3HOA  3HPE  3HPF  3HR0  3HS3  3HU5  3HUP  3HV2  3I0Z  3I2Z  3I3W  3I5Q  3I9F  3IA1  3IAV
3IBS  3IBW  3ICY  3IGR  3IJM  3IKK  3ILW  3IN6  3IPO  3ITF  3IUO  3IUP  3IUW  3IX1  3IX3
3IX7  3JSL  3JU7  3JX9  3JXO  3K0Z  3K2N  3K67  3K86  3K8R  3K9U  3K9V  3KBY  3KD4  3KD6
3KE7  3KEA  3KF3  3KGZ  3KHF  3KIZ  3KKB  3KKZ  3KMA  3KPH  3KUV  3KUZ  3KWR  3KWS  3KZP
3KZT  3L0Q  3L32  3L46  3L5Z  3L6I  3L6U  3LAG  3LAS  3LF5  3LF6  3LFI  3LGD  3LHN  3LHR
3LIA  3LID  3LJD  3LM2  3LMB  3LQ6  3LQS  3LR2  3LRT  3LS9  3LV4  3LVC  3LYD  3LYN  3LYY
3LZX  3LZZ  3M33  3M8J  3MAB  3MAD  3MBK  3MC1  3MCW  3MCZ  3MEX  3MGD  3MGG  3MGJ  3MGK
3MIL  3MIZ  3MJQ  3MMH  3MOZ  3MQM  3MQQ  3MTR  3MUJ  3MUQ  3MUX  3MVE  3MVG  3MWJ  3MZ2
3N08  3N10  3N1E  3N8B  3NAU  3NAW  3NDO  3NEK  3NI0  3NI6  3NJ2  3NO7  3NOI  3NPF  3NPI
3NPP  3NQB  3NQW  3NRL  3NS6  3NTL  3NTV  3NUF  3NVA  3NX3  3NYD  3O0L  3O4W  3O5Y  3O6V
3O7O  3OAJ  3OCP  3OFG  3OHE  3OMT  3OMY  3ONX  3OOO  3OPC  3OQ2  3OQP  3OT2  3OTN  3OVP
3OY2  3OZI  3OZY  3P1X  3P2C  3P6B  3P6K  3P7J  3P8T  3P9V  3PA8  3PC7  3PDY  3PET  3PFO
3PIJ  3PJT  3PJV  3PJY  3PMC  3PMR  3PN3  3PPB  3PPL  3PPM  3PSM  3PU9  3PUB  3PUH  3PX2
3Q18  3Q20  3Q31  3Q4N  3QBM  3QGU  3QHA  3QKC  3QTA  3QWU  3QYF  3R3P  3R41  3R5G  3R89
3RA5  3RAU  3RBY  3RKC  3ROT  3RQ9  3RQB  3RRI  3RRS  3S06  3S18  3S84  3SBU  3SG8  3SK2
3SLZ  3SON  3SY6  3T2Z  3T6S  3T7Y  3T8K  3TAK  3TB6  3TC9  3TDQ  3TE8  3TFJ  3THF  3TJ8
3TP9  3TRI  3TY2  3TYY  3U1Y  3U4Z  3U6G  3U7R  3U96  3UB6  3UBU  3UEJ  3UEP  3UF6  3UFE
3UHA  3UMO  3UMZ  3UPL  3URY  3USS  3UT4  3UUN  3UV0  3UV1  3UX3  3V1E  3V4K  3V4M  3V67
3V6G  3VAY  3VB8  3VCC  3VEJ  3VK5  3VM9  3VRC  3VTX  3VW9  3VZX  3W08  3W0E  3W1O  3W36
3W77  3WAE  3WGT  3WHA  3WJE  3WRB  3WSC  3WV8  3WX7  3X3Y  3ZFI  3ZIG  3ZIT  3ZJL  3ZRP
3ZRX  3ZTB  3ZTH  3ZX4  3ZYL  3ZYY  4A7U  4AB5  4AE4  4AG0  4AG7  4AML  4AUU  4AVR  4AXO
4AYN  4B0N  4B0Z  4B54  4BE3  4BE9  4BF5  4BG7  4BG8  4BI3  4BK0  4BLG  4BND  4BOL  4BRC
4BWO  4BWV  4BX2  4C0R  4C86  4CHI  4CI8  4CJN  4CL3  4COB  4CWC  4D3D  4D3Q  4DCZ  4DJN
4DMG  4DNN  4DNX  4DO2  4DT5  4DZZ  4E0A  4E0U  4EBG  4EF0  4EGU  4EHS  4EHU  4EI0  4EIB
4EIR  4EJR  4EP4  4EPU  4EQ7  4EQQ  4EQS  4ESW  4ETK  4EU9  4EVX  4EW5  4EZG  4FBM  4FDI
4FKB  4FKZ  4FRY  4FU3  4FVF  4FYP  4FZL  4G06  4G3V  4G5A  4GEK  4GHO  4GI2  4GIT  4GKM
4GOF  4GP7  4GR6  4GXO  4GYT  4H5B  4H7L  4H8A  4HAH  4HBE  4HBQ  4HCE  4HCF  4HEH  4HEI
4HEQ  4HFQ  4HFS  4HHV  4HI7  4HIA  4HL2  4HMS  4HU7  4HW5  4HWV  4HYL  4I1Q  4I4K  4I4O
4I6R  4I6Y  4IBG  4IC3  4ICS  4ID0  4IGU  4IHU  4IJ5  4IJ7  4IJZ  4IKB  4IP5  4IQD  4IQI
4ITB  4IV9  4IX3  4IXN  4IY4  4IYJ  4J0N  4J3Y  4J42  4J5R  4J6C  4J7R  4J8C  4J8E  4J8Z
4JAW  4JEM  4JG9  4JGP  4JLE  4JN9  4JOQ  4JTM  4JXR  4JYS  4K0U  4K26  4K28  4K6H  4KEM
4KR5  4KTP  4KTW  4KV2  4L1J  4L3K  4L3R  4L57  4L7A  4L9C  4LAN  4LIR  4LJ3  4LJI  4LJL
4LM4  4LMY  4LS9  4LSM  4LTB  4LXQ  4M0Q  4M0S  4M73  4M7Y  4MAC  4MAE  4MAK  4MAM  4MDU
4MEB  4MGE  4MIS  4MJD  4MN7  4MPM  4MT8  4MUV  4MYP  4N04  4N06  4N0R  4N0V  4N6J  4N7W
4N8O  4NAD  4NC7  4NDS  4NEX  4NK2  4NLH  4NOG  4NPR  4NQ8  4NQF  4NRN  4NSV  4NTC  4NU3
4O6I  4O6Y  4O7J  4O9K  4OH9  4OHJ  4OK4  4OKE  4OKI  4OM8  4OO4  4OPM  4OQQ  4OS3  4OTN
4OYU  4OZ0  4P33  4P5N  4P7C  4P93  4P94  4PAG  4PE0  4PHJ  4PIC  4PRS  4PUH  4PVC  4PXE
4PYQ  4PZK  4Q04  4Q1V  4Q25  4Q51  4Q69  4Q6Z  4Q7O  4Q9A  4Q9V  4QBN  4QE0  4QGB  4QGE
4QGX  4QHJ  4QI3  4QJY  4QNC  4QR8  4QUS  4R16  4R27  4R3N  4R60  4R8D  4R8O  4R8Z  4R9X
4RAY  4RBR  4RDZ  4RE5  4RGB  4RGD  4RGP  4RLZ  4RO3  4RP3  4RPT  4RRQ  4RSW  4RT5  4RUN
4RVS  4RZ3  4RZB  4S1H  4S23  4S26  4S3I  4S3P  4TLJ  4TMT  4TN5  4TPV  4TQJ  4TR6  4TRH
4TRT  4TSD  4TT0  4TTY  4TVI  4TWL  4TX5  4U13  4U5G  4U9N  4UAB  4UAI  4UC2  4UEJ  4UG1
4UIQ  4UNU  4UOP  4UP3  4UR6  4USK  4UTU  4UU3  4UUL  4UX7  4UXU  4UZ8  4V15  4V17  4V29
4W7Y  4W9R  4WBP  4WF5  4WH5  4WJT  4WPM  4WWF  4WX0  4WZN  4X08  4X3L  4X51  4X6X  4X8Y
4XFW  4XIN  4XO6  4XQ4  4XQC  4XVV  4XWT  4XZZ  4Y1R  4Y7D  4YEA  4YEP  4YMG  4YNX  4YPO
4YSL  4YT2  4YTD  4YTO  4YX1  4YY5  4YZG  4YZZ  4Z24  4Z27  4Z39  4Z4A  4ZBD  4ZBW  4ZCE
4ZDS  4ZFV  4ZKY  4ZO2  4ZSI  4ZUR  4ZV5  4ZVA  4ZVC  5A3V  5A48  5A9D  5ACS  5AIF  5AL7
5AMT  5AQ0  5AVN  5AWI  5AXG  5AYV  5AZW  5B08  5B0H  5B0P  5B1Q  5B4N  5B5I  5B7G  5BIR
5BJX  5BNC  5BR4  5BTU  5BU6  5BWI  5C04  5C1F  5C40  5C5Z  5C7Q  5C8Z  5CES  5CL2  5CQG
5CR4  5CRB  5CRH  5CUO  5CX8  5CXO  5CYJ  5D1P  5D1R  5D1V  5D3A  5DCL  5DY1  5E2C  5ECC
5EDX  5EIU  5EK5  5EQ2  5ER9  5EUV  5F29  5F2K  5F46  5F5N  5F6R  5FAV  5FCN  5FFP  5FFQ
5FI3  5FID  5FIS  5FLH  5FVJ  5FXD  5FZP  5G4I  5GGY  5GPK  5GSM  5GT5  5GUK  5GVY  5GX8
5GXE  5GXX  5GY7  5H1N  5H34  5H3Z  5H78  5HB6  5HCB  5HDM  5HEE  5HHJ  5HI8  5HIF  5HJL
5HOP  5HRA  5HS7  5HTL  5HWV  5HX0  5I0Y  5I5M  5I90  5I96  5IDB  5IN1  5IOJ  5IPY  5IRB
5IT3  5ITJ  5IW9  5IXV  5IZ3  5J0A  5J41  5J4I  5J7M  5J90  5JAZ  5JBR  5JE6  5JEL  5JHX
5JIP  5JKJ  5JNP  5JNU  5JSI  5JTD  5JWC  5K3X  5K4W  5KAY  5KEF  5KHD  5KO4  5KX4  5L0L
5L73  5LLJ  5LTL  5LVS  5LWK  5LZK  5M7C  5M97  5M99  5MOZ  5MQ8  5MUY  5MWX  5N6X  5NCK
5NCR  5NEG  5NL6  5NLZ  5NO5  5NZO  5O10  5O2Z  5OI7  5OLY  5OO7  5ORG  5OVY  5SY4  5T3E
5T3U  5TD6  5TFP  5TJJ  5TO5  5TTA  5TXC  5U35  5U4H  5U5N  5U85  5UCT  5UE1  5UE7  5UEJ
5UF5  5UFN  5UH7  5UI9  5UJD  5UKV  5UQS  5UUO  5UZX  5V01  5V4A  5V4P  5V4R  5V5U  5V6I
5VAZ  5VDN  5VHT  5VJ4  5VM2  5VSJ  5VT2  5VX1  5W4Z  5W8Q  5WEC  5WFX  5WHX  5WI2  5WPP
5WUT  5WWD  5X03  5X56  5X9I  5XAQ  5XGT  5XKT  5XNA  5XNE  5XOM  5XP0  5XPV  5XSP  5XUN
5XVJ  5XXA  5Y78  5Y8L  5Y9Q  5Y9Z  5YA6  5YAD  5YAT  5YDD  5YET  5YGE  5YGH  5YHR  5YJC
```

5YKR	5YKZ	5YN4	5YNX	5YRH	5YZ1	5Z11	5Z16	5Z28	5Z2G	5Z2H	5Z2V	5Z49	5Z50	5Z8O
5ZFK	5ZI1	5ZI2	5ZKT	5ZQJ	5ZUM	5ZVV	5ZXN	6A51	6A55	6A5F	6A6F	6A71	6A80	6AE9
6AEF	6AEP	6ALL	6AMG	6AQE	6AR4	6AT3	6AWL	6AWR	6B7C	6B9F	6BHY	6BIE	6BND	6BSU
6BSY	6C0G	6C3C	6C5B	6C6N	6C8R	6CDB	6CKK	6CMK	6COF	6CPB	6CPD	6CQP	6CS9	6CW0
6CWW	6D2W	6D3V	6D41	6DAO	6DB1	6DBP	6DEB	6DGK	6DGM	6DJC	6DKK	6DQP	6DT3	6DVR
6E28	6EDQ	6EID	6EJT	6EL2	6ENI	6EP6	6ES9	6EW7	6EWM	6EY5	6F1J	6F43	6F5C	6FDC
6FF2	6FHG	6FIY	6FP5	6FU3	6G6U	6G96	6GDJ	6GF6	6GFB	6GHU	6GU1	6GYG	6GZA	6H1W
6H31	6H59	6H6O	6H86	6H8F	6HAT	6HAZ	6HBV	6HIU	6HJO	6HK8	6HNM	6HPQ	6HQ2	6HQZ
6HTJ	6HZY	6I1A	6I5B	6I6S	6IAU	6IFQ	6ILS	6IME	6IOW	6IPT	6IRP	6J1O	6J25	6J3E
6J4K	6J5X	6J66	6J6A	6J6L	6J8L	6J94	6JDH	6JHV	6JIE	6JNJ	6JQW	6JSX	6K2F	6K2Y
6K62	6K7C	6K8V	6KEW	6KFM	6KGJ	6KHL	6KI2	6KLK	6KNL	6L2U	6L3Q	6L5H	6L6G	6L85
6LAC	6LCQ	6LEB	6LF1	6LGI	6LH6	6LIY	6LPN	6LZH	6M2O	6M31	6M4B	6M9G	6MB8	6MRV
6MTW	6MX1	6MXV	6N7O	6N91	6N9Q	6NAL	6NDI	6NIM	6NJC	6NK3	6NL2	6NNH	6NNR	6NNW
6NP6	6NQY	6NRX	6O0B	6O14	6O5K	6O6Y	6O8N	6OH8	6OIB	6OJF	6OMP	6ON4	6OVP	6OZU
6P1E	6P2I	6P58	6P73	6PCE	6PNR	6PT4	6PT8	6Q2C	6QJ6	6QLA	6QSI	6QUW	6QWO	6R5J
6R6U	6RCH	6RIV	6RJB	6RK0	6RK1	6RS4	6RWD	6RYK	6S2R	6S33	6S6F	6S7X	6S95	6SAN
6SCB	6SCQ	6SEK	6SF4	6SFH	6SI6	6SIZ	6SJ8	6SRB	6SSG	6SU3	6SW4	6T7O	6TCB	6TEK
6TJ8	6TJR	6TL7	6TVV	6TY0	6TY2	6TYK	6U2U	6U60	6UBL	6UBO	6UD6	6UH8	6UN8	6URE
6USS	6UXU	6V1B	6V3Z	6V42	6VD8	6VH6	6VJC	6VJU	6VPE	6VTV	6VUD	6VZ0	6W40	6W6X
6WE8	6WJA	6WN2	6WU7	6WXW	6XB6	6XNO	6XPH	6Y04	6Y1W	6Y1Y	6Y7F	6YF6	6YIZ	6YJ9
6YKB	6Z68	6ZA0	6ZII	6ZK8	6ZMB	6ZN7	6ZT4	7A1F	7A5C	7AED	7AG6	7AO3	7APP	7ASV
7B5J	7B67	7BB3	7BIO	7BJN	7BM8	7BR1	7BRA	7BU2	7C02	7C23	7C38	7C4A	7C5Y	7C64
7C8G	7C8P	7CBI	7CCB	7CDV	7CIK	7CJ3	7CJ7	7CKH	7CMA	7CSV	7CWQ	7EV1	7JJV	7JKV
7JW2	7KB9	7KL8	7KPZ	7KQA	7KSB	7KWD	7LZG	7MBK	7NBI	7NET	7NUU	7O39	12AS	

Table S2. p-values of chi-square tests between hydrogen bond types from FIRST with an energy cutoff of -0.1 kcal/mol.

	Intrachain	**Interface**	**Interface/Intrachain**	
Dataset1/ Dataset2	**p-values**	**p-values**	**Dataset**	**p-values**
PPnrall, PDnrall	0.858	0.0047	PDnrall	6.941e-10
PTnrall, PDnrall	0.845	0.0043	PPnrall	3.831e-10
PTnrall, PPnrall	0.963	0.0137	PTnrall	3.369e-06

Table S3. p-values of chi squared tests comparing proportions of different types of HB energy categories based on Table 2.

Intrachain		**Interface**	
Dataset1/Dataset2	**p-values**	**Dataset1/Dataset2**	**p-values**
BB-BB: PDnrall, PPnrall	0.924	*BB-BB: PDnrall, PPnrall	2.2e-16
BB-BB: PDnrall, PTnrall	0.986	*BB-BB: PDnrall, PTnrall	2.2e-16
BB-BB: PPnrall, PTnrall	0.991	*BB-BB: PPnrall, PTnrall	0.703
SC-SC: PDnrall, PPnrall	0.948	SC-SC: PDnrall, PPnrall	4.031e-06
SC-SC: PDnrall, PTnrall	0.989	SC-SC: PDnrall, PTnrall	1.036e-10
SC-SC: PPnrall, PTnrall	0.994	SC-SC: PPnrall, PTnrall	0.399
Mixed: PDnrall, PPnrall	0.741	Mixed: PDnrall, PPnrall	2.2e-16
Mixed: PDnrall, PTnrall	0.987	Mixed: PDnrall, PTnrall	2.2e-16
Mixed: PPnrall, PTnrall	0.839	Mixed: PPnrall, PTnrall	0.816

* Since the numbers of HBs of the interface BB-BB types for category II and III are small, the chi-square statistical analysis was performed by combining the numbers in category II and III.

Table S4. HB energy (HBE) categories based on different energy ranges

Category	HBE range (kcal/mol)
I	-0.7 ≤ HBE < -0.1
II	-1.3 ≤ HBE < -0.7
III	-2.0 ≤ HBE < -1.3
IV	HBE < -2.0

Table S5. p-values of chi-square tests between hydrogen bond energy categories (based on the discretization in Table S4) at interface and within intrachain.

Dataset1/Dataset2	p-values (intrachain)	p-values (interface)	Dataset	p-values (interface/intrachain)
PPnrall, PDnrall	0.959	0.007	PDnrall	0.005
PTnrall, PDnrall	0.999	0.009	PPnrall	0.944
PTnrall, PPnrall	0.980	0.999	PTnrall	0.99

Table S6. p-values of chi squared tests comparing proportions of different types of HB energy categories based on the discretization in Table S4.

Intrachain		Interface	
Dataset1/Dataset2	p-values	Dataset1/Dataset2	p-values
BB-BB: PDnrall, PPnrall	0.859	*BB-BB: PDnrall, PPnrall	2.2e-16
BB-BB: PDnrall, PTnrall	0.986	BB-BB: PDnrall, PTnrall	2.2e-16
BB-BB: PPnrall, PTnrall	0.973	*BB-BB: PPnrall, PTnrall	0.726
SC-SC: PDnrall, PPnrall	0.926	SC-SC: PDnrall, PPnrall	6.935e-15
SC-SC: PDnrall, PTnrall	0.948	SC-SC: PDnrall, PTnrall	4.83e-12
SC-SC: PPnrall, PTnrall	0.946	SC-SC: PPnrall, PTnrall	0.968
Mixed: PDnrall, PPnrall	0.926	Mixed: PDnrall, PPnrall	9.802e-05
Mixed: PDnrall, PTnrall	0.948	Mixed: PDnrall, PTnrall	0.009
Mixed: PPnrall, PTnrall	0.946	Mixed: PPnrall, PTnrall	0.756

* Since the numbers of HBs of the interface BB-BB types for category II and III are small, the chi-square statistical analysis was performed by combining the numbers in category II and III.