ACS OMEGA

Article

Cite This: ACS Omega 2017, 2, 7359-7374

http://pubs.acs.org/journal/acsodf

Characterization of Ionizable Groups' Environments in Proteins and Protein−Ligand Complexes through a Statistical Analysis of the Protein Data Bank

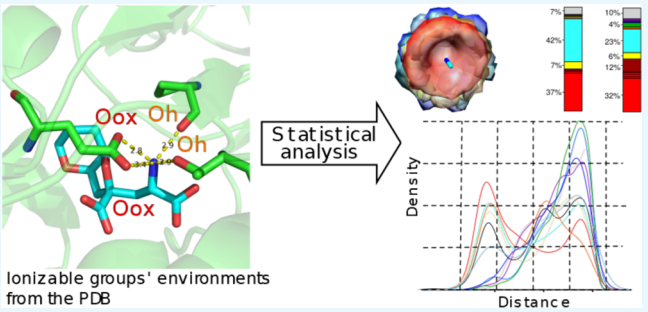
Alexandre Borrel,[†,‡,§] Anne-Claude Camproux,[†] and Henri Xhaard*[,‡]

[†]Molécules Thérapeutiques *in silico* (MT*i*), INSERM UMRS-973, University Paris Diderot, Sorbonne Paris Cité, 75205 Paris Cedex 13, France

[‡]Faculty of Pharmacy, Division of Pharmaceutical Chemistry and Technology, University of Helsinki, Viikinkaari 5E, P.O. Box 56, FI-00014 Helsinki, Finland

Supporting Information

ABSTRACT: We conduct a statistical analysis of the molecular environment of common ionizable functional groups in both protein−ligand complexes and inside proteins from the Protein Data Bank (PDB). In particular, we characterize the frequency, type, and density of the interacting atoms as well as the presence of a potential counterion. We found that for ligands, most guanidinium groups, half of primary and secondary amines, and one-fourth of imidazole neighbor a carboxylate group. Tertiary amines bind more rarely near carboxylate groups, which may be explained by a crowded neighborhood and hydrophobic character. In comparison to the environment seen by the ligands, inside proteins, an environment enriched in main-chain atoms is found, and the prevalence of direct charge neutralization by carboxylate groups is different. When the ionizable character of water molecules and phenolic or hydroxyl groups is accounted, considering a high-resolution dataset (less than 1.5 Å), charge neutralization could occur for well above 80% of the ligand functional groups considered, but for tertiary amines.



INTRODUCTION

Molecular interactions are fundamental to biochemical processes. Ionizable, basic and acidic, functional groups can form charged interactions mediated through a shared hydrogen atom, that is, salt bridges.[1] These hydrogen bonds are strong with energy of interaction estimated at 28.5−48.1 kJ/mol. They are characterized by a short distance (e.g., about 2.59−2.86 Å between the O and N atoms of a primary amine and a carboxylate group) and a ΔpK_a range of [3−11] between the acceptor and the donor.[2] Although the basic and acidic groups are often ionized at the binding sites, this is not always the case, especially considering that the local pH may differ greatly from that of the solvent.[3,4] A common way to infer ionization of a given functional group in crystallographic three-dimensional (3D) structures (which most often do not harbor hydrogen atoms) is to consider its neighborhood: if a counterion is at close range, ionization is likely.[5] If not, it is difficult to address the issue without complex quantum chemistry calculations.

In proteins, salt bridges involve a basic group such as the primary amine of a lysine side chain or the protein N-terminus, the imidazole (IMD) group of a histidine, and the guanidinium (GAI) group of an arginine and an acidic group such as the carboxylate group from an aspartate or glutamate side chain or the protein C-terminus. They play a critical role in the folding, stability, and dynamics of 3D structures at all levels, from secondary and tertiary structures to supramolecular assemblies, and have been studied for multiple aspects:

(i) their energetic contribution or electrostatic strength, especially with respect to secondary, tertiary, or quaternary structure as well as stability;[6−9] a strong correlation is observed between the secondary structure and salt bridge formation.[10] Furthermore, salt bridges form complex networks,[1,7] which are suspected to have a stabilizing effect on the protein structure, following the observed relation between the increased number of salt bridges and thermal stability;[11−13]

(ii) their geometrical characteristics; for example, salt bridges between aspartate and glutamate and histidine, arginine, or lysine display extremely well defined geometric preferences;[7]

(iii) their environment and their location (within monomers or at the interface between monomers as well as their solvent accessibility);[14] salt bridges display preferential

Received: June 6, 2017
Accepted: October 11, 2017
Published: October 30, 2017

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

formation in an environment of 30% solvent-accessible surface area;[10]

(iv) the separation of the amino acids; intrachain salt bridges are mainly separated by three or four residue salt bridges;[15]

(v) their fluctuations and nuclear magnetic resonance (NMR) conformer ensembles show that salt bridges may break and new salt bridges are formed, in good correlation with crystallographic B-factors;[16]

(vi) water molecules have important roles to play toward the stability of molecular complexes, for example, conformational stability or stabilization or mediation of ion pairs.[17,18]

A vast majority of these studies have been based on structural data extracted from the Protein Data Bank (PDB).[19] Consequently, the amount of data available to the authors has been variable, from the early work in 1995 in which Barlow and Thornton or Musafia and co-workers conducted using less than a hundred proteins[1] to 1500−2000 structures 10 years later[10,13] and up to 3644 monomers in the recent study by Donald et al. in 2011.[7] Larger datasets of course increase the robustness of the findings. The data generated for proteins in the present manuscript is the largest, that is, more than 4500 monomers, simply because of the natural growth of the PDB. The focus of the work is the environment of salt bridges and their frequencies; we include in our statistics elements such as water molecules and weakly ionizable groups that to the best of our knowledge have not been studied together so far in the literature.

In contrast to the work conducted in proteins, the environment of ionizable groups in protein−ligand complexes has received only little attention. This is probably due to the relative difficulty in identifying ionizable groups in the ligands, the absence of ready-to-use datasets, and the relative difficulty in operating cheminformatics data mining tools in the PDB. Another challenge is that until recently only limited data were available, especially considering the need to analyze enough high-resolution and diverse protein−ligand complexes. Yet, a better characterization of the interacting environment of ionizable groups would be of key interest in molecular docking simulations,[20] where such a knowledge would help to better position the bridging structural water molecules, select or optimize relevant ionization states, improve the initial placing of the ligand, and design more efficient and accurate scoring functions.[21−24]

The aim of this study is to make a quantitative and qualitative assessment of the protein molecular environments for the ligand and protein ionizable groups in the PDB. We focused on atoms forming the molecular environment in the close vicinity (3.0 and 4.0 Å) of the queried functional groups. Statistics about the density, frequency, and number of polar contacts were extracted and are discussed for both protein−ligand complexes and inside protein structures. Statistics were also extracted as to whether there is at least one contact of a given type. The scope of the study is restricted and currently excludes the long-range stabilization of basic groups either through π interactions[25] or through long-range electrostatics, although these are known to be important, for example, to protein-folding processes or to molecular recognition events.[26,27]

RESULTS

The environment of six ionizable chemical groups well-represented in the ligands is considered: primary amine (referred to as I, pK_a 7.75−10.64),[28] secondary amine (II, pK_a 9.29−11.01),[28] tertiary amine (III, pK_a 8.31−10.65),[28] IMD (pK_a 5.1−7.75),[29] GAI (pK_a 8.33−13.71),[30] and carboxylic acid (COO, pK_a 1.84−4.40)[31] (Table S6). These are referred to as query groups. The study is conducted both for ligand queries and for protein queries. Only four of these query groups are present in proteins: I (lysine side chain and N-terminus), IMD (histidine side chain), GAI (arginine side chain), and COO (aspartate and glutamate side chains and C-terminus). It is important to note that to represent the queries IMD, GAI, and COO, which contain several atoms, we used centroids (see the Experimental Section).

PDB1.5 and PDB3.0 Datasets. The work was initiated using the PDB3.0 dataset of ligand queries at 3.0 Å resolution. The study was then enriched by considering only a subset of the data at higher resolution, PDB1.5, which allowed to study more accurately the role of water molecules. Indeed, the main apparent difference between the PDB1.5 and the PDB3.0 datasets is the amount of water molecules present, that is, there are more water molecules in the PDB1.5 dataset (Figure 1).

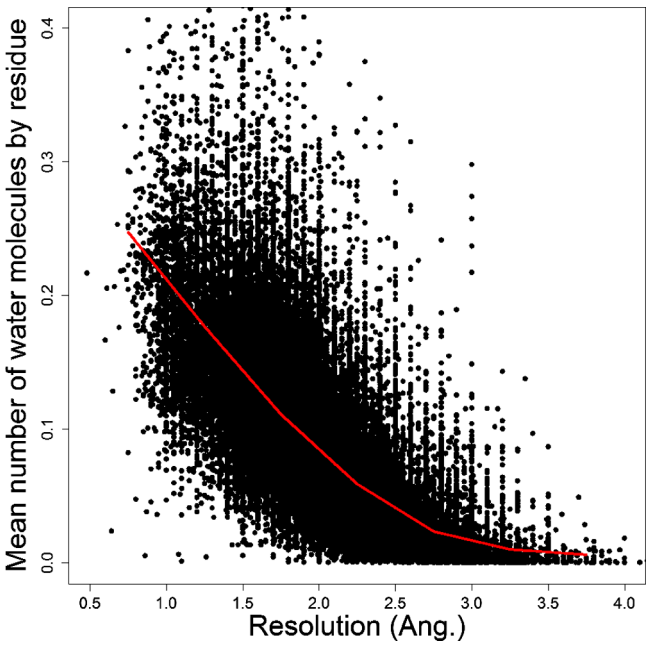


Figure 1. Mean number of water molecules by the amino acid as a function of crystallographic resolution from all proteins in the PDB. The red line represents the mean number of water molecules by the amino acid with an interval of 0.1 Å in resolution.

The study was then completed by collecting protein query interaction statistics at both resolutions. The study was also run with the PDB50 release of the PDB to eliminate potential biases due to having similar proteins in the dataset, and the results were found to be robust (see the Discussion).

The PDB1.5 dataset is composed of 387 complexes, and the PDB3.0 contains 4592 complexes (Table 1). From the dataset PDB1.5, we extracted for ligands 169 instances for the query group I, 96 for II, 70 for III, 30 for IMD, 11 for GAI, and 135 for COO. From PDB3.0, we extracted 1632 instances for the query group I, 1230 for II, 1147 for III, 264 for IMD, 146 for

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

Table 1. Content of the PDB1.5 and PDB3.0 Datasets[a]

	query groups	I	II	III	IMD	GAI	COO	any atom
PDB 1.5	number of complexes	161	91	64	26	11	96	387
	number of ligand query groups	169	96	70	30	11	135	10 314
	number of protein query groups	13 031			6227	11 380	28 146	195 913
PDB 3.0	number of complexes	1491	1113	1020	251	134	1139	4592
	number of ligand query groups	1632	1230	1147	264	146	1390	126 808
	number of protein query groups	154 979			70 474	143 529	344 848	197 306

[a]Null environments are defined from the column "any atom"

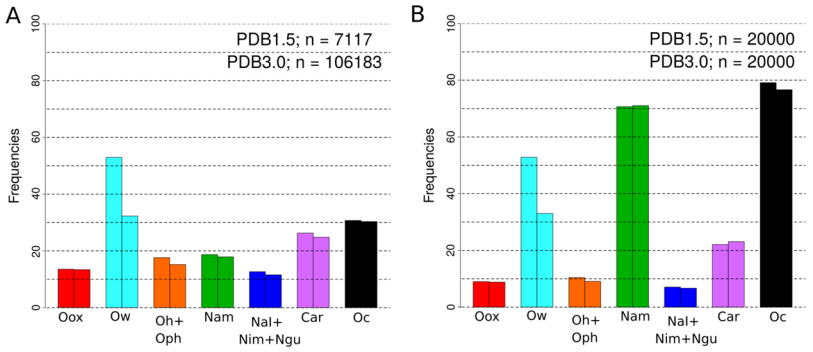


Figure 2. Null environments around (A) ligand atoms and (B) protein atoms. The graph shows the proportion of query groups with at least one Oox, Ow, Oh and Oph, Nam, NaI or Nim or Ngu, and Car atom in their neighborhood (4.0 Å). Datasets PDB1.5 (left bars) and PDB3.0 (right bars) are both shown. The following color code will be consistently used in this study: Oox (red), Oh and Oph (orange), Ow (cyan), Nam (green), Nim, Ngu, and NaI (blue), Car (purple), and Oc (black).

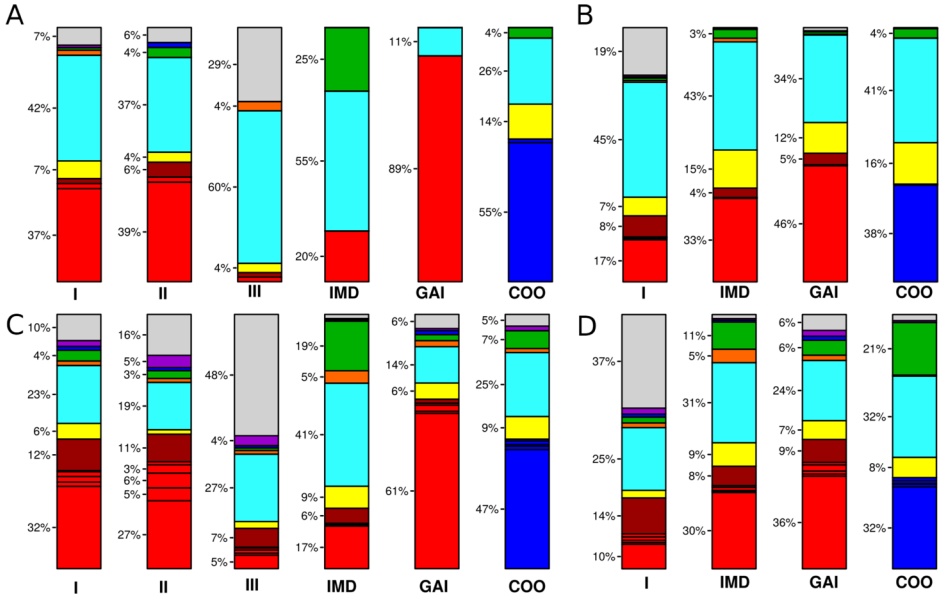


Figure 3. Neighborhoods of (A,C) ligand query groups I, II, III, IMD, GAI, and COO and (B,D) protein query groups I, IMD, GAI and COO. (A,B) is for the PDB1.5 dataset and (C,D) is for the PDB3.0 dataset. The presence of the following atom types in the neighborhood was searched and exclusively assigned to the first type found (from the bottom to the top of the bars): at least 1−4 Oox atoms within 3.0 Å; red, separators indicate the number of Oox groups from more than five (bottom) to one (top); at least one Oox atom in the 3.0−4.0 Å range (burgundy red); at least one Ow itself interacting with a Oox atom for basic query groups and interacting with a NaI, Ngu, or Nim for the acidic query group (yellow); at least one Ow (cyan); at least one Oh, Oph (orange); at least one Nam (green); at least one Ngu, Nim, or NaI (marine blue); at least one Car (purple); at least one aliphatic carbon or sulfur (gray). The color code is the same for COO but (Ngu, Nim, and NaI) are used in the place of (Oox). Note a small number of samples for IMD and GAI in panel (A).

GAI, and 1390 for COO. The numbers for ligand query data for IMD (n = 30) and GAI (n = 11) in PDB1.5 are thus too low to extract reasonable statistics. However, the results are shown because they are highly consistent with the data extracted from the PDB3.0 dataset and from the protein query data. For protein queries, the PDB1.5 dataset contains 13 031 instances of I, 6227 of IMD, 11 380 of GAI, and 28 146 of COO. In the PDB3.0 dataset, all query groups have more than 20 000 representatives.

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

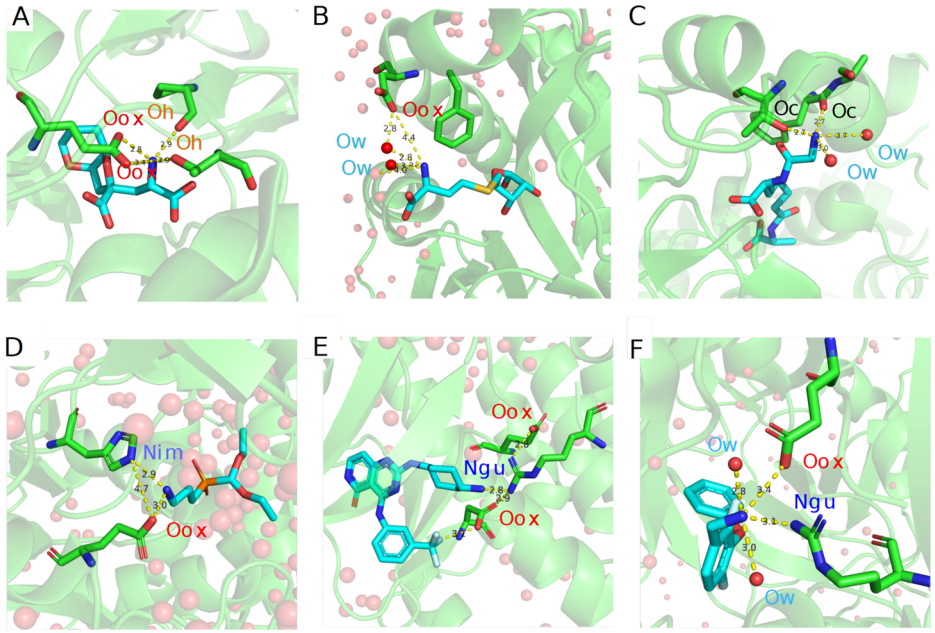


Figure 4. Examples of six different environments for query group I. (A) neutralization using a counterion (human arginase I, PDB code 3MFW); (B) neutralization using a counterion mediated by water molecules (*Helicobacter pylori* 5′-methylthioadenosine/*S*-adenosylhomocysteine nucleosidase, PDB code 4OJT); (C) only water molecules and main-chain carbonyl groups (*Streptomyces* sp. R61 DD-peptidase, PDB code 1IKI); (D) nitrogen from IMD (human GABA(B) receptor, PDB code 4MR8), (E) nitrogen from GAI (*Salmonella enterica* stationary phase survival protein, PDB code 4XJ7); and (F) nitrogen from GAI (hepatitis C virus Hcv Ns3 Protein, PDB code 4B76). Ligand carbon atoms (blue), protein carbon atoms (green), water molecules (red spheres), and protein cartoon trace (green) are shown.

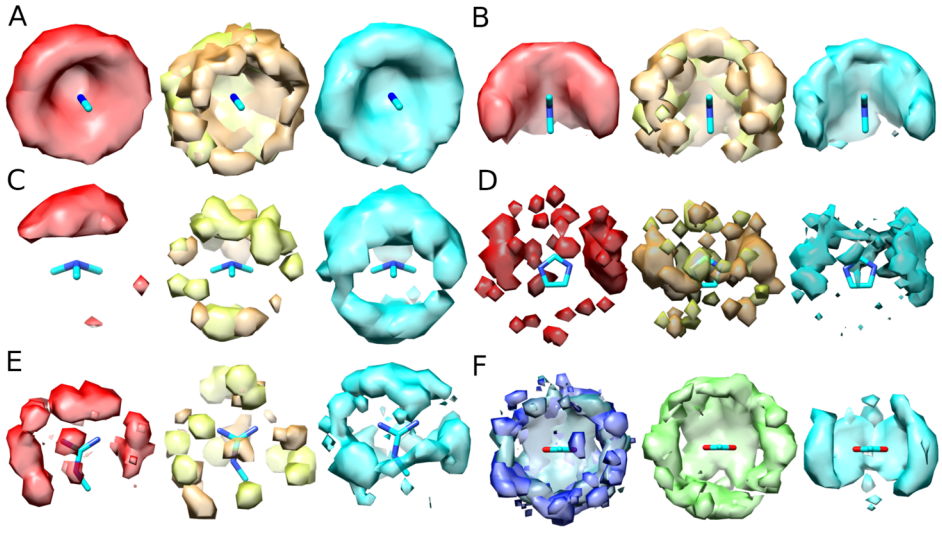


Figure 5. 3D densities of atom types around ligand queries using the dataset PDB3.0. Color code: for query group; (A) I, (B) II, (C) III, (D) IMD, and (E) GAI, Oox (red), Oh and Oph (orange, yellow), and Ow (cyan). For (F) COO, Nam (green), Nim, Ngu, and NaI (blue), and Ow (cyan).

Null Environments. A rational way to study molecular environments is to consider them in the light of the environment of any atom, that is, to a null model or the reference state. We built two null environment models, one for ligand queries and one for protein queries (Figure 2). Null environments are considered by collecting the environment of any ligand atom, that is, they are reflective of pockets binding the ligands collected in this study and a set of randomly selected protein atoms, that is, they are reflective of interactions in the protein core, especially, secondary structure elements.

Environments in the PDB1.5 and PDB3.0 datasets are very similar, save for the number of water molecules (see previous section). About 53% of any ligand atom or any protein atom has at least one water molecule (Ow) within 4.0 Å in PDB1.5, whereas these numbers drop to 32−33% in the PDB3.0 dataset.

Comparing the environments of ligand and protein atoms uncovers a major difference. The environment of protein atoms is significantly enriched in amide groups (Nam) [18% (any ligand atom) against 71% (any protein atom)] as well as in carbonyl groups [(Oc) 30% (any ligand atom) against 77% (any protein atom)] (values are from the PDB3.0 dataset; very similar values are obtained from the PDB1.5 dataset). This can be explained by the contact formed by secondary structure elements in proteins and by the lower exposition of the main-chain atoms to the ligand-binding sites. The environment of ligand atoms is slightly enriched in charged and polar amino:

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

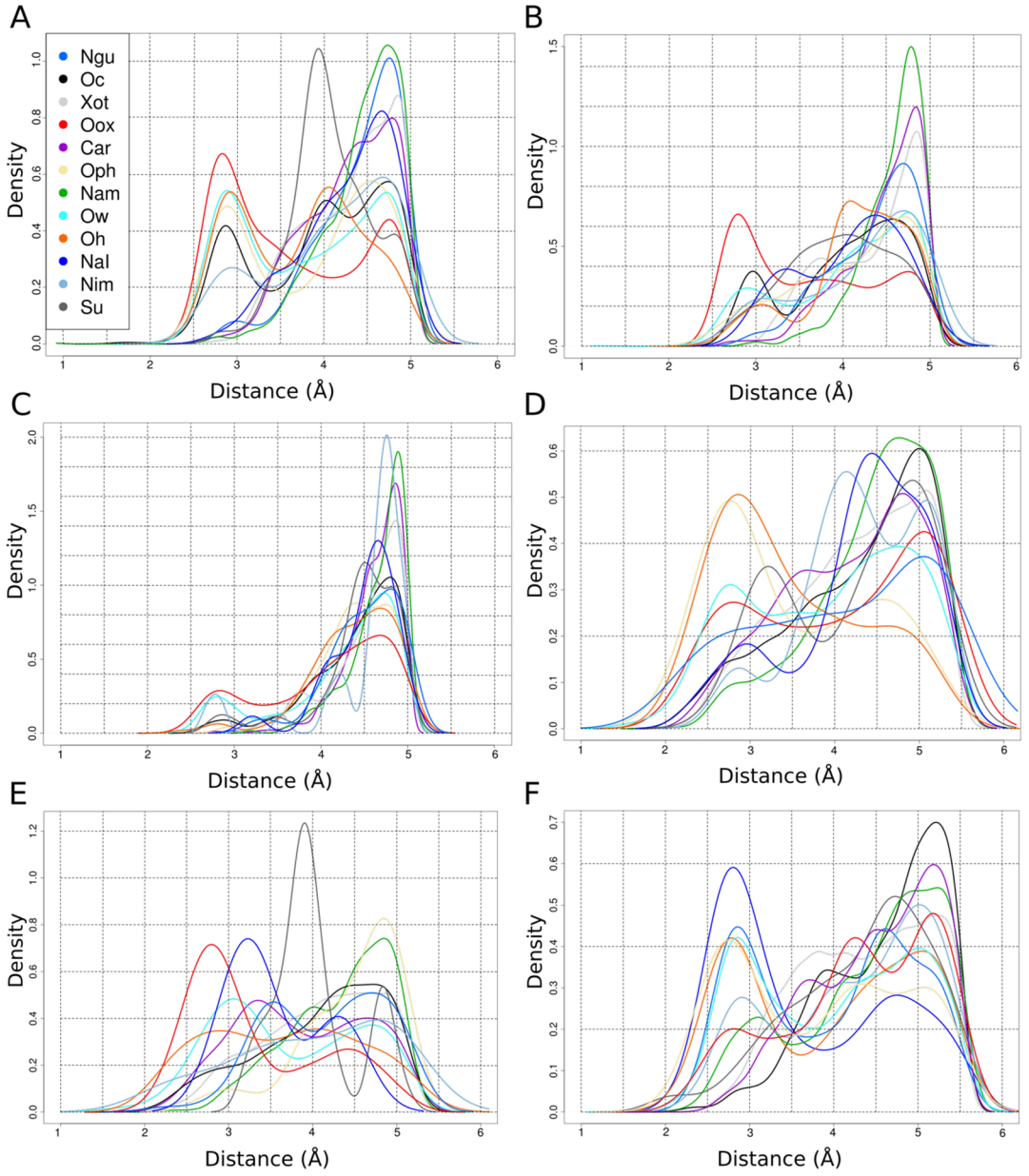


Figure 6. Density of presence for selected protein atoms in the neighborhood of ligand queries. The *Y* axis represents the relative density value for all atoms collected within 6.0 Å distance from the query group. I (A), II (B), III (C), IMD (D), GAI (E), and COO (F) using the dataset PDB3.0. Density curves are colored as follows: Oox (red), Oh (orange), Oph (light orange), Oc (black), Ow (cyan), Nam (green), Ngu (light blue), NaI (blue), Car (purple), and Xot (gray).

carboxylic acid (Oox; 13 vs 9%), phenolic and hydroxyl (Oh and Oph; 17 vs 9%), and positively charged groups (NaI, Nim, and Ngu; 12% vs 7%). Car appears equally in ligand and protein null environments (23−25%).

Neutralization at the Level of the Functional Group. We start the Results section by presenting an overview of the neutralization of the charge at the level of a query group (Figure 3) and subsequently present details about the different environments and in particular their composition. These different types of environments are illustrated in Figure 4, taking the case of a primary amine. Classical environments are salt bridge interaction with a carboxylate group (Figure 4A), interaction with a carboxylate group mediated by a water molecule (Figure 4B), and environment formed by water molecules and carbonyl groups (Figure 4C). Less classical environments for primary amines are, for example, interaction with an IMD group (Figure 4D) or with a GAI group (Figure 4E,F). The interacting atoms were analyzed by placing the ligand query fragments in the same referential (Figure 5; data available in .pdb format in the Supporting Information). This was done by computing the rotation/translation matrices using an in-house implementation of the Kabsch's algorithm.[32,33] For III and to a lesser extent II, interactions occur predominantly in the axial position from the tetrahedron formed by nitrogen on the top and to a lesser extent below the three connected carbons (Figure 5B,C). Note that the superimposition of I functional groups is fuzzy because of the rotational freedom around the C···N bond.

Strong contacts (short interaction distances) were found between the six functional groups studied and the atoms Oox, Oc, Oh, Oph, and Ow and to a lower extent Nim. For the five basic queries, we sequentially cumulatively looked at

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

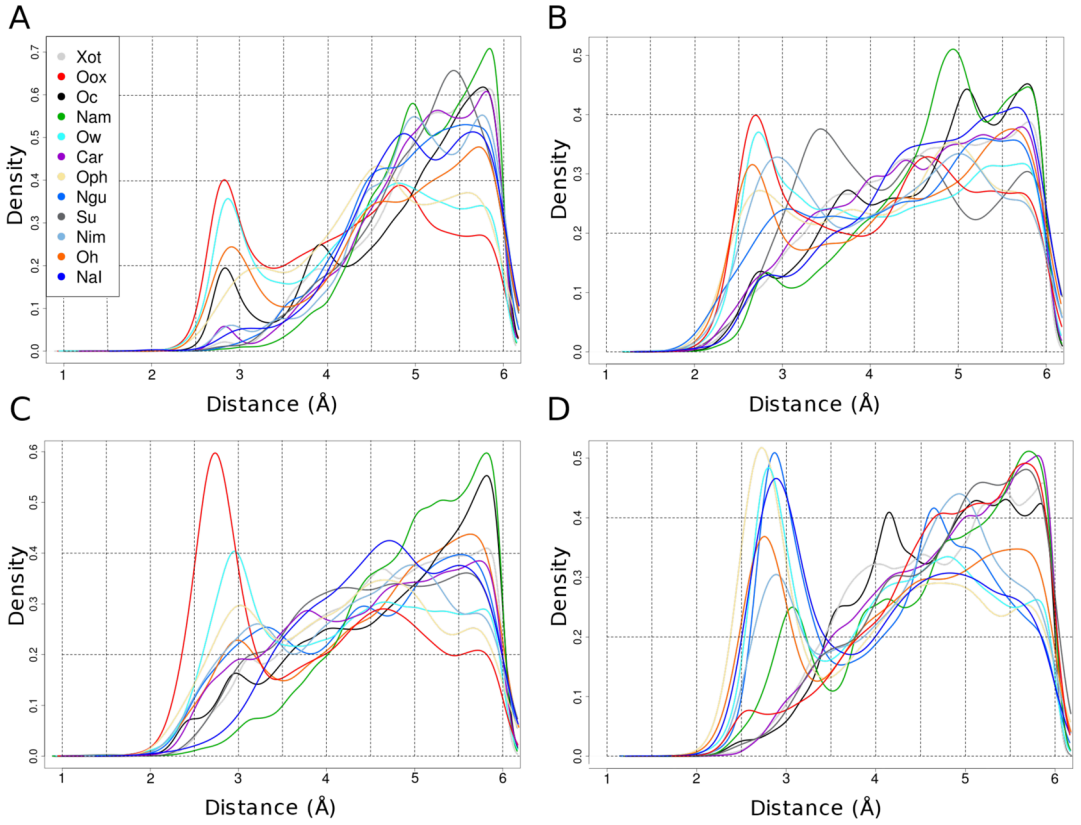


Figure 7. Density of presence for selected protein atoms in the neighborhood of protein queries. The *Y* axis represents the relative density value all atoms collected within 6.0 Å distance from the query group: I (A), IMD (B), GAI (C), and COO (D) using the dataset PDB1.5. Density curves are colored as follows: Oox (red), Oh (orange), Oph (light orange), Oc (black), Ow (cyan), Nam (green), Ngu (light blue), NaI (blue), Car (purple), and Xot (gray).

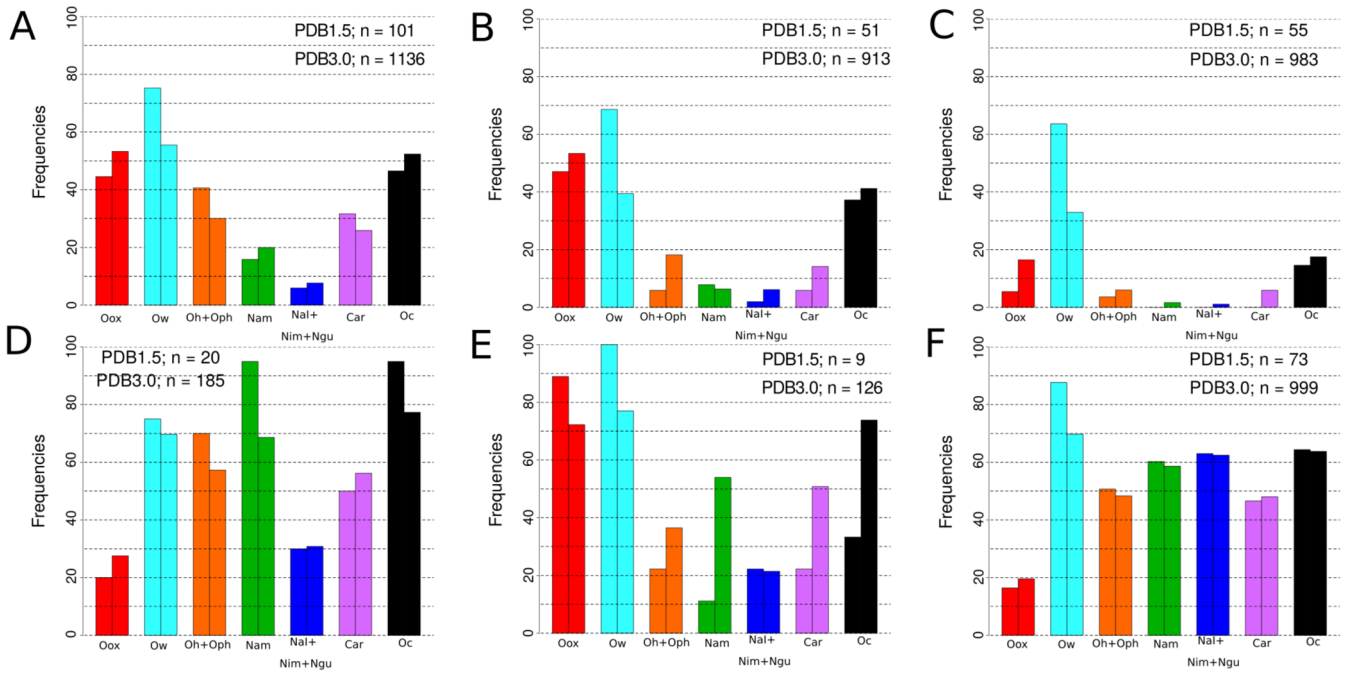


Figure 8. Proportion of ligand query group I (A), II (B), III (C), IMD (D), GAI (E), and COO (F) with at least one type of neighbor atom type at a distance of 4.0 Å. For each atom type, proportions are represented using the datasets PDB1.5 (left bars) and PDB3.0 (right bars). Color code is the same as above: Oox (red), Ow (cyan), Oh and Oph (orange), Nam (green), Nim, Ngu and NaI (blue), Car (purple), and Oc (black).

possibilities of charge neutralization not only by carboxylate groups (Oox) but also by acidic groups that provide opportunities for hydrogen bonds with a charge-transfer component (Oh, Oph, and Ow). When we account for the

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

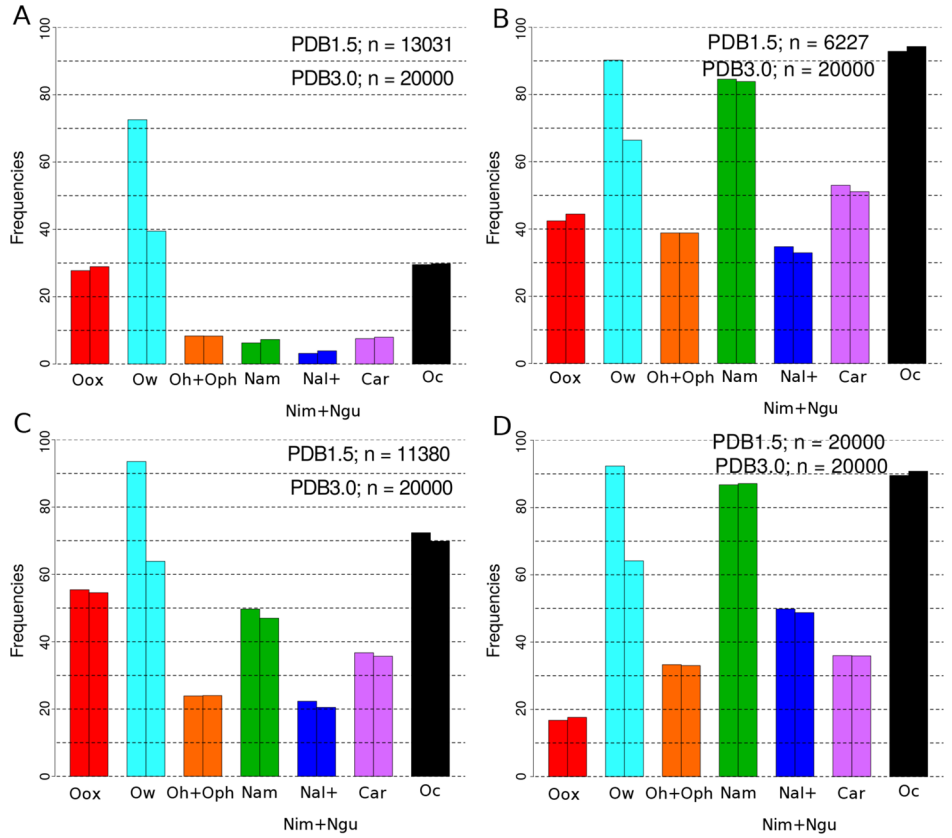


Figure 9. Proportion of protein query group I (A), IMD (B), GAI (C), and COO (D) with at least one type of neighbor atom type at a distance of 4.0 Å. For each atom type, proportions are represented using the datasets PDB1.5 (left bars) and PDB3.0 (right bars). Colors are as follows: Oox (red), Ow (cyan), Oh and Oph (orange), Nam (green), Car (purple), Oc (black), and Nim, Ngu, and NaI (blue).

functional groups of ionizable character in the neighborhood, considering only the well-solvated highest resolution dataset (PDB1.5), we assess that direct counterions are present within 4.0 Å for ligand queries I in 93% of cases, for II in 88%, for III in 71%, for IMD in 85%, for GAI nearly all, and for COO in 96% of the cases; for protein queries, these numbers are 81% for I, 97% for IMD, 98% for GAI, and 96% for COO. These numbers are much higher than those obtained by considering only direct carboxylate counterion neutralization.

We refined the analysis to consider separately the cases where water molecules mediate ionic contacts (yellow in Figure 3).[34] Water molecules were defined to mediate an ionic interaction if the water molecule itself is within 3.0 Å of a potential counterion (Oox for I, II, III, IMD, and GAI; NaI, Nim, or Ngu for COO); a corrective number was used to calibrate distances in the case of centroids (see the Experimental Section). As a result, water molecules were found to mediate ionic contacts for 7% of I, 4% of II, 4% of III, and 14% of COO in ligand queries and 7% of I, 15% of IMD, 12% of GAI, and 16% of COO for protein queries. For all queries, there are slightly but consistently more intervening water molecules detected in the PDB1.5 dataset, supporting a better refinement of the structures.

Similarly, the fraction of carboxylate counterions in the 3.0–4.0 Å distance range from the basic queries—that indicates ionic interactions but not charge-reinforced hydrogen bonds—is for all functional groups considered lower in the higher resolution dataset (compare the burgundy red on Figure 3A,C and B,D): for example, 2% against 12% for primary amines or 6% against 11% for secondary amines (ligand queries). This phenomena is accompanied by an increase in the close range interaction with Oox in the higher resolution dataset. This could reflect a nonoptimal refinement in the lower resolution crystal structures, a suggestion well in line with the recent work about halogen bonds.[35] It is interesting that the phenomena of poor refinement could be observed for classical functional groups that are expected to be well-represented by current force fields, as opposed to halogen atoms.

Carboxylate Contacts. Carboxylate oxygens (Oox) are often involved in charge-reinforced hydrogen bonds (Figures 4A,B and 5A–E, left-hand densities).[36] The distribution of Oox around the functional groups I, II, III, IMD, and GAI shows a strong density peak at 2.8 Å, seen especially for I and II (Figures 6, 7, and S3–S5) as well as for GAI. For III and IMD, a weak peak of density is also found at 2.8 Å. Similarly, for COO, the peak of Ngu, NaI, and Nim is also found at 2.8 Å. This value of 2.8 Å is typical of salt bridges, as reported elsewhere.[2]

The high propensity of the query bases to form salt bridges with Oox atoms is corroborated by their frequent close contacts (Figures 8 and 9): ligand GAI (72–89% combining both datasets), primary and secondary amines (45–54%), and IMD (20–28%) often neighbor a carboxylate group in their binding sites. Tertiary amines bind less near carboxylate groups (5–16%), which may be explained by a more crowded neighborhood and a more hydrophobic character (see the Discussion). In proteins, the prevalence of direct charge neutralization by carboxylate groups is different: GAI (54–55%), IMD (42–44%), and primary amine (28–29%). Ligand

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359–7374

Table 2. *p*-Values and Significance [Represented by the Number of (*)] of Tests of Comparison of the Environments, i.e., between Contingency Tables of Atom Type Composition by Query Group and the Null Environments[a]

Type and dataset	Queries	Oox	Ow	Oh+Oph	Nam	NaI+Nim +Ngu	Car	Oc
	I	0 ***	5.7e^{-1} -	2.1e^{-3} *	5.3e^{-11} ***	1.1e^{-20} ***	2.5e^{-19} ***	4.9e^{-1} -
	II	0 ***	9.5e^{-1} -	5.8e^{-1} -	3.9e^{-26} ***	2.5e^{-11} ***	9e^{-22} ***	4.1e^{-2} -
	III	0 ***	0 ***	7e^{-4} *	0 ***	2.2e^{-16} ***	1.1e^{-15} ***	6.8e^{-1} -
	IMD	4.2e^{-3} -	5.3e^{-2} -	1.9e^{-3} *	1.4e^{-4} **	6.5e^{-1} -	1.4e^{-2} -	1 -
	GAI	1.1e^{-10} ***	2.3e^{-1} -	4.7e^{-1} -	2.7e^{-1} -	4e^{-1} -	1.9e^{-2} -	3.5e^{-1} -
	COO	1.2e^{-25} ***	8.5e^{-5} **	1.4e^{-06} ***	3.4e^{-12} ***	1.8e^{-93} ***	3.9e^{-12} ***	1.4e^{-07} ***
	I	0 ***	0 ***	1.1e^{-14} ***	0 ***	1.4e^{-09} ***	1.1e^{-77} ***	0 ***
	IMD	1.5e^{-53} ***	7e^{-05} **	1.6e^{-55} ***	2.2e^{-64} ***	8e^{-72} ***	6.6e^{-19} ***	8.2e^{-49} ***
	GAI	0 ***	4e^{-89} ***	1.8e^{-73} ***	0 ***	0 ***	1.1e^{-11} ***	0 ***
	COO	2.2e^{-1} -	4.8e^{-76} ***	0 ***	0 ***	0 ***	1.2e^{-5} ***	0 ***

[a]The PDB3.0 is preferred over the PDB1.5 dataset for ligand queries because of lack of data in the latter. Three significance levels are defined: not significant if corrected *p*-value is more than 0.1; (*) if corrected *p*-value is less than 0.1; (**) if corrected *p*-value is less than 0.05; and (***) if corrected *p*-value is less than 0.01. Boxes are colored when the *p*-value is significant: in red when query neighborhoods are enriched and in blue when query neighborhoods are depleted.

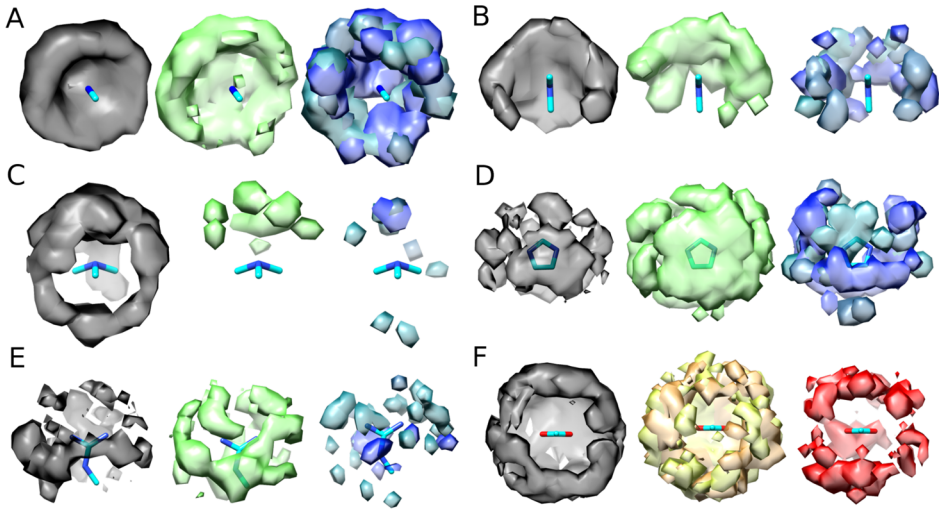


Figure 10. 3D densities of contact atoms using the dataset PDB3.0 for ligand queries (A) I, (B) II, (C) III, (D) IMD, (E) GAI, and (F) COO. Color code: Oc (black), Nam (green), Nim, Ngu, and NaI (blue), Oph and Oh (yellow and orange), and Oox (red).

and protein carboxylate groups are similarly neutralized (49–63%).

The null environments can be used to evaluate the significance of the query to Oox interactions. The preference for Oox is significantly higher for four out of five basic functional groups considered (Table 2). Preference for Oox by the ligand and protein queries is clearly seen for I, II, and GAI that have at least one Oox in their neighborhood in 44–89% of cases compared to 14% for null environments. The number of Oox (or other polar atoms, water molecules excepted) interacting with III is surprisingly low, much lower than what would be expected from the null environment (see the Discussion). As should be expected, the COO to Oox is significantly lower than for the null environment (Table 2). Even if occurring less, carboxyl–carboxylate interactions, which require both carboxylic acid oxygens to be in the neutral form, are strong, as discussed elsewhere.[37]

Hydroxyl and Phenol Contacts. For the hydroxyl (Oh) and phenol groups (Oph), the interaction distance peaks at 2.8 Å seen for Oox are also found (Figures 6, 7, and S3–S5). For IMD and COO, an equivalent peak found at a distance of 2.8 Å suggests that strong hydrogen bonds with a charge-transfer component, comparable to salt bridges, are formed. For GAI, the Oh and Oph interaction is shifted toward 3.0 Å for both types of queries. This indicates weaker hydrogen bonds and may relate to the charge of GAI groups being most often already neutralized by a carboxylate group (in 72–89% of the cases, see Figure 3).

At least one hydroxyl or phenol group is found in the vicinity of a ligand IMD query in 70% (PDB1.5) and 58% (PDB3.0) of the cases (Figure 8); these numbers are lower for protein

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359–7374

queries, about 23−24% (Figure 9). This could point to specific recognition motifs at the binding sites toward the IMD query group. Favorable Oh and Oph interaction for the IMD and COO queries may be linked with the delocalized nature of the electrons on the IMD ring and carboxylate. Contacts between ligand IMD and Oph in the absence of carboxylate or water molecules in the vicinity are found for both ligand and protein queries (orange on Figure 3C,A in the PDB3.0 dataset but not the PDB1.5 dataset. This may reflect an incomplete refinement of the PDB3.0 dataset (importantly, the protein data are of significant size), or simply the fact that some water molecules are not seen in lower resolution structures.

For most query groups (ligand I and IMD and all protein queries), hydroxyl and phenol groups interact significantly more than in the null environment. Query III shows significantly less Oh or Oph contacts than in the null environment, in accordance with its specific environment (see the Discussion). For environments showing interactions between hydroxyl or phenol groups and GAI, statistical significance could not be demonstrated for ligand queries. This probably indicates lack of data (only 134 neighborhoods considered for GAI using the PDB3.0 dataset).

Water Molecules and Charge Neutralization. In terms of contact density, water molecules exhibit a peak at 2.8 Å for all considered queries, closely resembling those of Oox (Figures 6, 7, and S3−S5). In proteins where there are plenty of data, this peak in the density at 2.8 Å is visible for I and IMD (Figures 7 and S5). For GAI, the peak of water molecule density is shifted to longer distances, as was observed for Oh and Oph. This may again be explained because the GAI query groups are almost always neutralized by a salt bridge with a carboxylate. Similar to hydroxyl and phenolic groups, water molecules can form hydrogen bonds that have a proton-transfer component and therefore may act as counterions (Figure 4C). Water molecules also have an amphoteric character and therefore can act both as a counterion of basic groups (I, II, III, IMD, and GAI) and the acidic group (COO).

Water molecules (Ow) were found in the close vicinity of all query groups for I, II, III, IMD, GAI, and COO, whereby at least 60% of the query groups considered have at least one water molecule within 4.0 Å in the PDB1.5 dataset (Figures 4 panels B, C and F, 8, and 10). Water molecules are over-represented in comparison to the null environment of ligand queries IMD and COO. The large number of water molecules interacting with III to some extent compensates the lower amount of interacting protein atoms, as can be seen in Figure 3A (see also Figure 8D).

IMD to Base Close Contacts and Other Base−Base Interaction. The data collected highlight the interaction of IMD (either as a query IMD or as a target atom Nim) with, surprisingly, bases (for a complete composition of the neighborhoods at 4.0 Å, see Tables S1 and S2). The nature of the contact between, for example, a primary amine and an IMD group is exemplified for I with Nim (Figure 4D). This contact has not been described in the literature but may take the form of hydrogen bonding with a proton being shared between the uncharged IMD and the protonated amine. A strong interaction is corroborated by a density peak at a distance of 2.8 Å for both IMD ligand and protein queries (Figures 6, 7, and S3−S5). The atom types Nim, NaI, and Ngu are within 4.0 Å of 30−34% of the IMD queries in both datasets (Figures 8 and 9). Altogether, there are sufficient number of occurrences of IMD-Nim in the PDB1.5 dataset for protein queries (1459 occurrences, 3588 in PDB3.0) to rule out refinement errors. These numbers are also consequent for protein queries for IMD−NaI (296 occurrences in PDB1.5 and 1015 occurrences in PDB3.0) and IMD−Ngu (1865 occurrences in PDB1.5 and 4809 occurrences in PDB3.0). In terms of significance, IMD to NaI, Nim, and Ngu is not significant because of lack of data (n = 185) for ligand queries, but it is significantly above background in protein queries (Table 2).

The case of the other basic groups I, II, and III is different (Figure 4E,F). These groups carry a positive charge under physiological conditions and are likely to repel each other, although there is evidence for cation−cation interactions in ionic liquids.[38] An unlikely interaction is seen in the density proportion with the absence of NaI and Ngu peaks at 2.8 Å for I, II, and III. These groups are very rarely positioned near (<3.0 Å) the basic queries in terms of raw numbers, for NaI, six occurrences in PDB3.0 and for Ngu, 16 occurrences in PDB3.0 (Tables S4 and S5). Accordingly, the environment of I, II, and III in terms of NaI, Ngu, and Nim is significantly below the null environment (Table 2). There are however density peaks near 3.4 Å (Figures 6, 7, and S3−S5). This reflects another aspect of the interaction formed by basic groups, that is, network of charges and secondary contacts (Figure 4D−F).

Amide and Carbonyl Contacts. Carbonyl oxygen (Oc) forms a suitable environment for basic groups as a hydrogen bond acceptor (Figure 4C). In proteins, carbonyls belong exclusively to main-chain and side-chain amide functional groups. In proteins, the main-chain carbonyl groups carry a permanent partial charge and very often benefit from aligned dipoles; thus, they make strong hydrogen bonds. Oc densities show a strong peak in the distribution for interaction with I and II at 2.9 and 3.0 Å (Figures 6, 7, and S3−S5), slightly longer than for hydrogen bonds that involves basic queries and Oox, Oh, and Oph. This is fully in line with the other work.[2,39] In terms of representation, Oc is present near the queries I, II, IMD, GAI, and COO: for ligands, from 33 to 77% (PDB3.0 dataset, where there are enough samples for all query groups) (Figure 8) and for protein queries, from 29 to 94%, similar in both datasets (Figure 9). For query III, Oc is present in only 15−17% of the neighborhoods. For COO in ligands, Oc is surprisingly significantly more represented than in the null environment (Table 2). Instead, for protein queries, Oc is always less represented in the neighborhood than in the null environments.

Main-chain and side-chain amide groups (Nam) are almost never found in the 3.0 Å vicinity of I, II, or III (n = 19 for protein−ligand interactions in PDB3.0) (Tables S4 and S5). For the IMD query, Nam is located above and below the plane of the IMD ring (Figure 10). Amide (Nam) shows density peaks close (3.0 Å) to IMD and COO. In terms of significance, Nam is significantly less represented than the environment for ligand and protein queries of II, III, and GAI (Table 2). For I and IMD, the over-representation is found in both systems. It may reflect a favorable arrangement of atoms without the hydrogen bond, but a fraction is to represent IMD to main-chain nitrogen interactions.[40,41]

Distance Threshold to Define Polar Contacts. When considering data within a sphere of 3.0 Å radius, the number of neighboring atoms is lower for simple groups (1.6 ± 1.3 for I; 1.0 ± 1.0 for II; and 0.3 ± 0.5 for III) compared to larger functional groups defined using a centroid (6.5 ± 2.8 neighboring atoms for IMD; 5.6 ± 3.4 for GAI; and 5.6 ±

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

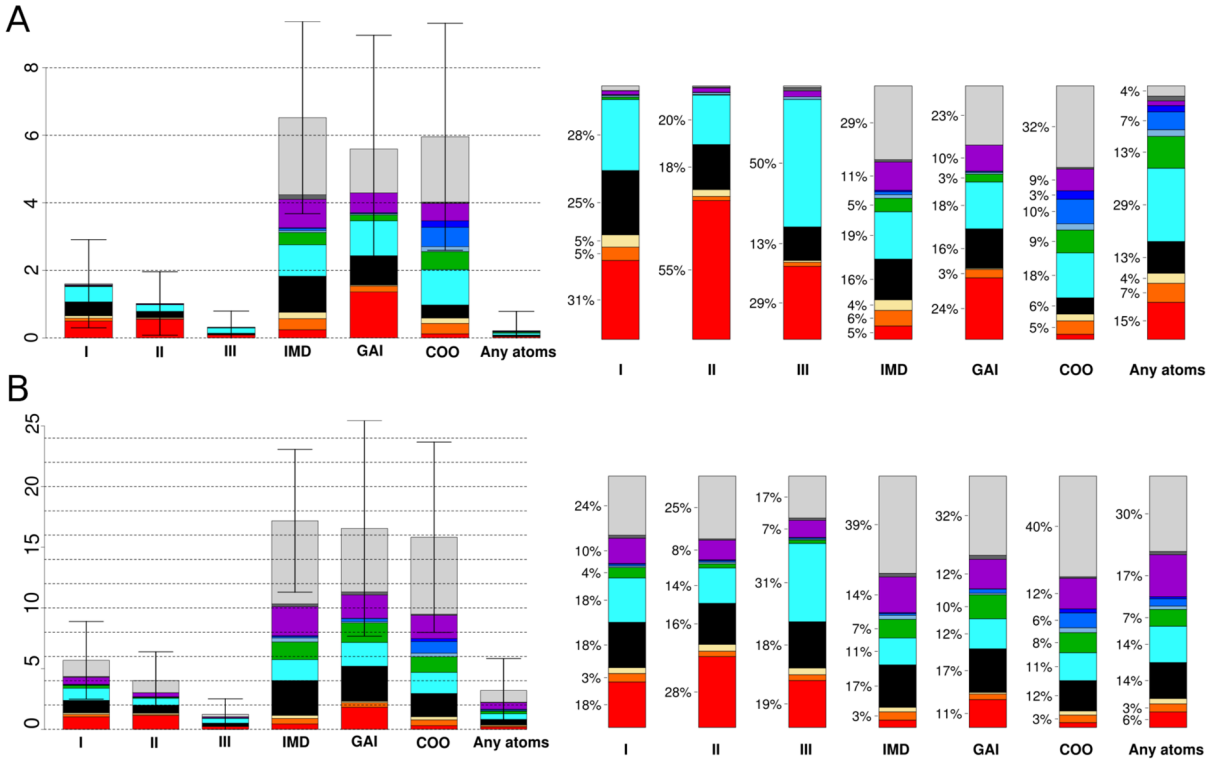


Figure 11. Influence of the distance threshold on the number of atoms (left panels) and atom type frequency (right panels) for ligand queries using the dataset PDB3.0. Neighborhood defined (A) using a data collection distance of 3.0 Å and (B) using a distance of 4.0 Å. Atom types are colored as follows: Oox (red), Oh (orange), Oph (light orange), Oc (black), Ow (cyan), Nam (green), Ngu (light blue), NaI (blue), Car (purple), Su (dark gray), and Xot (gray). Note the different y-axis scales for the left-hand panels.

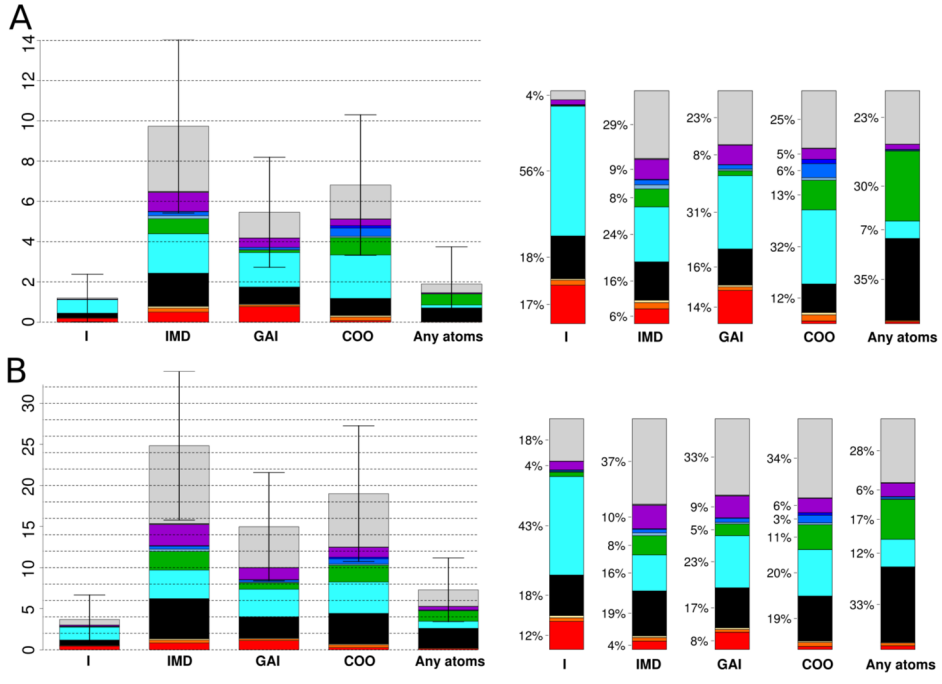


Figure 12. Influence of the distance threshold on the number of atoms (left panels) and atom type frequency (right panels) for protein queries using the dataset PDB3.0. Neighborhood is defined (A) using a threshold distance of 3.0 Å and (B) using a threshold distance of 4.0 Å. Atom types are colored as follows: Oox (red), Oh (orange), Oph (light orange), Oc (black), Ow (cyan), Nam (green), Ngu (light blue), NaI (blue), Car (purple), Su (dark gray), and Xot (gray). Note the different y-axis scales for the left-hand panels.

3.3 for COO) (Figures 11 and 12). This is easily explained because complex functional groups contain several atoms. The interaction shell collected within 3.0 Å of the query groups is composed mostly of polar atoms (Figures 11, 12, and Tables S3−S5). Indeed, query groups I, II, and III have 72%, 74%, and 92% of polar neighbors (Oox, Oh, Oph, Ow, Oc, Nam, Nim,

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

Ngu, and NaI) against 57% for any atoms in the null environment (data from PDB1.5, Tables S2 and S4). The proportion of neighboring oxygen and nitrogen polar atoms is in contrast lower for IMD (60%), GAI (50%), and COO (51%), which may reflect favorable interactions with carbon atoms, for example, COO to ring edge anion−π contacts.[42] Additionally, it could be a difference introduced by the data collection method, either a sphere centered on a point charge or a centroid; the latter may lead to contacts farther away to be included.

When using a longer radius for selecting neighbors (Figures 11 and 12), 4.0 Å compared to 3.0 Å, the number of neighboring atoms increase by 2−3 fold: 5.7 ± 3.2 for I; 4.0 ± 2.4 for II; 1.2 ± 1.2 for III; 17.2 ± 5.9 for IMD; 16.6 ± 8.9 for GAI; and 15.8 ± 7.8 for COO. Interestingly, III keeps a small number of atoms in its neighborhood even at a distance of 4.0 Å. The relative proportion of polar interacting atoms (Oox, Oh, Oph, Ow, Oc, Nam, Nim, Ngu, and NaI) decreases, which reflects the inclusion in the statistics of hydrophobic contacts as well as carbons connected to polar atoms, such as the central carbon atom belonging to carboxylate groups.

Generally, increasing the radius of the collection sphere brings the distribution of neighbors toward that observed for our null environment (tested up to 6.0 Å, data not shown). For the null environments, the number of atoms included in the neighborhood is much lower compared to the other query groups, that is, 0.2 ± 0.6 at 3.0 Å for the ligand query. This is explained by the fact that "any atom" in a ligand is usually carbon connected to two or three atoms, and that the 3.0 Å sphere represents strong polar contacts. Similar results were observed for protein queries.

DISCUSSION

Robustness of the Study toward a Potential Bias in the Dataset. In this manuscript, we present diverse statistics extracted from the PDB, which may be sensitive to biases in the dataset because of too many close homologues. We thus decided to run the study a second time using the PDB50 release, that is, a release that contains no two sequences sharing over 50% identity (statistics about the number of groups extracted are found in the Supporting Information Table S1). For protein queries, in which a subset of query groups are randomly extracted, we already control that the sample taken is robust over five different random extractions (see the Experimental Section). Not surprisingly, the statistics derived are more or less unaffected by using PDB50 (Supporting Information Figures S1 and S2). For ligand queries, we remove biases by keeping only one structure for each unique ligand (see the Experimental Section). The data obtained from PDB50 thus follow closely the statistics obtained from the complete PDB, especially for the groups having enough data (100 or more queries). The positive effect of using PDB50 on eliminating possible biases originating from the presence of several close homologues is nonetheless counterbalanced by a severe depletion in the data available. The resulting low number of ligand queries, especially for IMD and GAI in PDB3.0 and for almost all query groups in PDB1.5, leads to discrepancies between PDB50 and PDB100. Altogether, the study on the nonredundant PDB50 nonetheless confirms all trends observed with PDB100.

Interaction Environments of III are Clearly Different than I and II. One of the surprising findings of this study is that III forms salt bridges less frequently with carboxylate groups in comparison to I and II (see the Results). This is especially unexpected because pK_a of III is about the same as pK_a of I in the 8−10 range.[28] As elaborated in the Results section, water molecules can function as counterions and are frequently found near III (64% in the PDB1.5), especially in the absence of a carboxylate counterion. A reason for III to favor water molecules over protein counterions is the limited space available around the query (Figure 5). This limited space is corroborated by the low number of interacting atoms (Figure 11). Furthermore, the density curves for III are low at a close range (Figures 6C and S3). Taken together, this suggests that the distinct interacting environment of III is a consequence of its low accessible volume. Accessibility has been known for long in chemistry to relate to chemical reactivity. This is the first instance to show the importance of space available affecting the ability to form molecular interactions. Query III is furthermore stabilized by hydrophobic contacts. This is not seen in this study because the sphere of 4.0 Å radius used for data collection around III does not capture hydrophobic contacts made by the attached carbon atoms. Indeed, less than 6−8% of III has at least one aromatic carbon (Car) within 6.0 Å, in comparison to 25−26% for the null environment (Figures 2 and 8C).

Charged Groups without Neutralization by a Counterion. This manuscript is centered on the neutralization of charges, but what happens to the remaining complexes is of interest. First, the majority species is not always the ionized one, especially for IMD that has a pK_a range of 5.1−7.75 (Table S6). In addition, long-range contacts where charges are not directly neutralized by salt bridges are not accounted here. In particular, cation−π interactions are not studied in detail. Their number is nonetheless bounded by the number of aromatic carbons seen in the vicinity of the queries. For the respective ligand and protein queries, using the PDB3.0 dataset, there is at least one Car near I in 26 and 8% of the cases, for II in 14% of the cases, for III in 5% of the cases, for IMD in 37 and 36% of the cases, for GAI in 51 and 20% of the cases, and for COO in 44 and 30% of the cases. The cation−π or anion−π contacts are not the focus of this study because more complex geometric parameters as well as longer distances should be used to study them in more detail.[25,43] More generally, for ligand queries, we filtered out metals in the vicinity of the ligands as well as nonbonded ligand contacts, eliminating potential unexpected counterions.

Multiple Atom Interactions from Functional Groups. The peaks of densities collected at distances longer than about 3.5 Å need to be carefully interpreted because they often relate to atoms that do not directly interact with the query groups but are constrained by the chemistry of proteins. These can be connected atoms, for example, the carboxyl carbon and the second oxygen of a carboxylate group. This is seen in Figure 7A where the peak for I is followed by a weaker peak starting at 3.4 Å that corresponds to the second carboxylate oxygen (see also Figure 13). Another typical example of secondary contacts is the oxygen carbonyl Oc or the amide Nam in proteins. Secondary structure elements explain very well the shape of Nam with marked peaks at 5.0 Å on density proportion (Figure 7).

Another type of secondary molecular contact occurs when networks of hydrogen bonds of ionic side chains are in place (Figure 4E,F). Generally, arginine amino acid serves as a branching unit and therefore a key node in salt bridge networks.[1] In our dataset, considering protein-only contacts and only the salt bridge, about one-third of GAI and half of I,

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

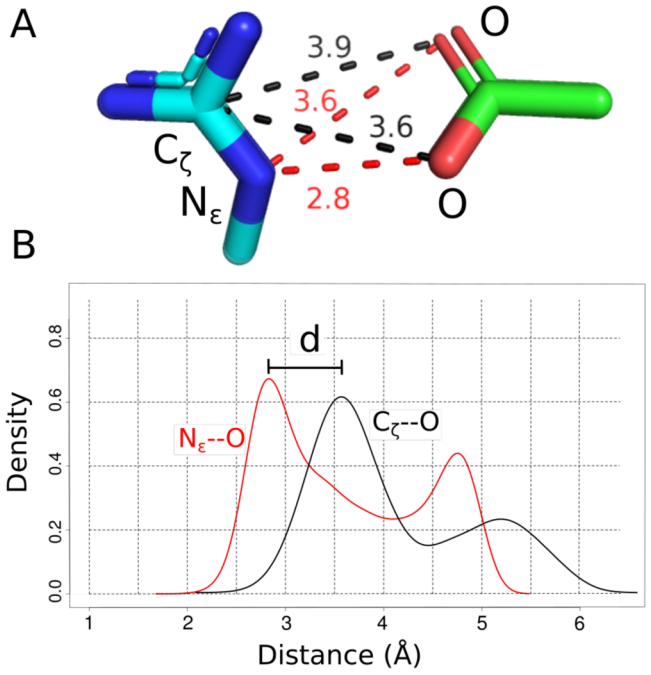


Figure 13. Empirical correction of the data collection sphere radius for complex functional groups, exemplified by the Nε−Oox distance. (A) Actual Nε−Oox hydrogen-bonding distances and Cζ−Oox distances presented in this manuscript. (B) Densities of Oox atom distribution used to define the corrective factor *d*. The peak of strong interaction is found at 2.8 Å for Nε−Oox and calibrated at this value for Cζ−Oox by subtracting *d*.

IMD, and COO are part of a complex network (Table 3). Very interestingly, the numbers we obtain are similar for ligands and proteins, with the notable exceptions of III and GAI (Table 3). We found that two-thirds of the tertiary amine salt bridges are actually ionic networks, and for GAI in the ligand, seldom a salt bridge network is present. This is likely to reflect the characteristic of the binding sites that accommodated these ligands.

The numbers we obtained for intraprotein salt bridges agree well with the study of Musafia et al., who reported one-third of all residues participating in salt bridges to be part of complex salt bridges.[1] In a different study, Donald and co-workers reported instead that most (over 95%) of the salt bridges are local and not involved in complex networks[7] in contrast to ours and Musafia's study and suggested that this was due to a methodological difference, that is, a focus on intra-subunit salt bridges.

CONCLUSIONS

This manuscript presents for the first time a characterization of the molecular environments of ionizable groups in protein−ligand complexes, and the data are placed in the light of intra- and inter-subunit interactions in protein structures. We include in our statistics elements such as water molecules and weakly ionizable groups, which together with the increased amount of data resulting from the natural growth of the PDB, make all aspects of this work novel. The findings in this manuscript can be summarized by a few principles. Taken together or individually they have a broad application toward the initial placement of docking poses, scoring the quality of protein structure or protein−ligand complexes and positioning water molecules in binding sites.

(1) The data collected, protein−ligand interaction of both at 1.5 and 3.0 Å resolution and intraprotein interaction at 1.5 Å resolution, show a consistent picture about the type and frequency of the interacting atoms. A notable difference in the environment is the over-representation of Oc and Nam in protein structures. This means that conclusions can be inferred from proteins about ligand−protein complexes and reciprocally, but also highlights that caution should be taken when deriving statistical interaction data.

(2) A sphere of 3.0 Å radius from point charges carries the majority of information about polar contacts. The strong polar contacts can be selectively captured by such a method. This avoids considering potentially noninteracting groups, as can be seen, for example, from the densities for I and Nam or Ngu (Figures 7 and 8). Getting a longer threshold to consider molecular interactions, as is often done in the literature by considering a 4.0 Å threshold,[6,7,44,45] probably shadows the strong charged-reinforced hydrogen-bonding data.

(3) Acidic and basic groups interact within 4.0 Å with a counterion in 45−89% of cases for I, II, GAI, and COO. When functional groups of ionizable character (Oh, Oph, and Ow) are accounted, this number increases to above 80% but for IMD and tertiary amine, it increases above 70%. Formation of net−neutral pairs has been indeed demonstrated for arginine−tyrosine pairs in aprotic environments using a combination of experimental and

Table 3. Frequency and Number of Ionizable Side Chains within 4.0 Å of the Query Groups, Indicative of Ionic Networks[a]

	frequency						raw numbers				
number of ionizables side chains within 4 Å	SB	none	one	two	three	four and more	none	one	two	three	four and more
I (ligand)	0.36	0.44	0.19	0.24	0.10	0.03	501	220	272	110	3
I (protein)	0.54	0.70	0.16	0.11	0.02	0.01	13 865	3332	2213	436	154
II (ligand)	0.36	0.45	0.20	0.19	0.07	0.10	413	181	169	61	89
III (ligand)	0.64	0.83	0.11	0.06	0	0	816	108	55	3	1
IMD (ligand)	0.50	0.64	0.18	0.14	0.02	0.02	118	34	26	4	3
IMD (protein)	0.38	0.45	0.21	0.18	0.08	0.08	8918	4192	3653	1670	1567
GAI (ligand)	0.09	0.23	0.07	0.36	0.11	0.21	29	9	48	14	26
GAI (protein)	0.28	0.41	0.17	0.24	0.09	0.09	8117	3373	4916	1809	1785
COO (ligand)	0.33	0.38	0.21	0.17	0.13	0.11	375	208	173	129	114
COO (protein)	0.45	0.50	0.22	0.16	0.07	0.04	10 138	4457	3140	1444	821

[a]"SB" refers to the frequency of ionic networks when only queries involved in at least one salt bridge are considered.

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

Table 4. Protein Atom Types Used in This Study

atoms	atoms in PDB format with the corresponding amino acid	atom type abbreviation
oxygen in carboxylate	Glu (OE1, OE2), aspartic acid (OD1, OD2)	Oox
oxygen in water molecule	HOH (O)	Ow
oxygen in hydroxyl or phenol	threonine (OG1), serine (OG)	Oh
oxygen in phenol	tyrosine (OH)	Oph
oxygen in carbonyl	protein main chain (O), asparagine (OE1), glutamine(OD1)	Oc
nitrogen in amide	asparagine (ND2), glutamine(NE2), protein main-chain (N)	Nam
nitrogen in IMD side-chain	histidine (NE2, ND1)	Nim
nitrogen in GAI side-chain	arginine (NH1, NH2, NHE, CZ)	Ngu
nitrogen in lysine side-chain	lysine (NZ)	NaI
carbon sp2 and nitrogen sp2 in an aromatic ring	phenylalanine (CG, CD1, CE2, CZ, CE1, CD2), tyrosine (CG, CD1, CD2, CE1, CE2, CZ), tryptophan (CG, CD1, CD2, NE1, CE2, CE3, CZ3, CH2, CZ2)	Car
sulfur atoms	cysteine (SG), methionine (SD)	Su
carbon atoms	carbons not included in the above-mentioned groups	Xot

computational methods.[46] A parsimonious way to have a protonated (basic) group at a binding site or in a protein is to have a proton-donating (acidic) group directly interacting with it. This could be taken advantage of, for example, in enumerating protonation states in docking simulations. This study does not characterize what happens in the remaining cases: interactions with other acidic groups, interactions not seen, for example, due to crystal packing, or the group may not be ionized. In particular, phosphate groups are widely present in endogenous ligands[47] and do form charged interactions with the protein.

(4) Tertiary amines have a specific interaction shell: they form much less salt bridges, for example, than primary amines (5−16% against 45−54%), although they have roughly the same pK_a. They form less observed polar contact with the protein than "any atom" from the ligand. By contrast, water molecules appear to be the most prevalent strong polar contact made by tertiary amines. There is a strong hydrophobic component in their binding subsite, which can be inferred from the prevalence of carbon atom neighbors (Figures 5 and 6), although it has not been directly studied here. This highlights the role of accessibility in forming molecular contacts. Contact accessibility is not taken into account by current scoring functions and would deserve further study.

(5) Water molecules play a key role in the stabilization of polar groups, especially in the absence of salt bridges. Water molecules are prevalent at binding sites. Their contribution to binding is critical but difficult to measure, especially in terms of enthalpic or entropic contributions. This study highlights an interesting new possible role for water molecules, that is, to act as a counterion to neutralize ionizable groups through hydrogen bonds that have a charge-transfer character. This role may also be taken by phenolic or hydroxyl groups. Quantum chemical calculation is necessary to study this phenomenon in more detail.

EXPERIMENTAL SECTION

Computational Tools. All scripts developed for this study, developed in Python 2.7, are provided as is from the platform GitHub (https://github.com/ABorrel/saltbridges). All plots and statistical analyses were conducted using the R package (version 3.2.2).[48] Proteins were visualized using Pymol (version 1.4.1),[49] and 3D densities were created using Chimera (version 1.10).[50]

Data Extraction. Crystallographic complexes were extracted from the PDB,[19] October 2015 release, 112 968 structures. Structures elucidated by NMR or including DNA or RNA were not selected. Two global criteria of quality were used for filtering, a resolution less than either 1.5 or 3.0 Å and an R-free[51] value less than 0.25. These values are standards for the analysis of proteins or protein−ligand complexes.[7,10,34] Two datasets, named PDB1.5 or PDB3.0 depending on the resolution range considered, were thus built, where the PDB1.5 dataset is a subset of the PDB3.0 dataset. To control the robustness of the statistics obtained for ligand queries, in particular toward biases that may arise from the presence of close homologues in the dataset, the complete study was run a second time on the PDB50 release of the PDB, which features no pairs of structure with a percentage of sequence identity above 50%.

All ligands present in the PDB, about 14 000 ligands, were first queried. Query groups were identified using in-house scripts. Briefly, to avoid errors due to incomplete data, the connectivity matrix of each ligand was rebuilt by defining bonds when the distance between two atoms is less than 1.42 Å. Tertiary amine groups were defined as such when not planar, that is, when the distance between the N atom and the plane formed by the three carbon atoms is less than 1.00 Å. These values were empirically defined at the start of the study based on their distribution in the PDB (Figure S6). Queries with no protein interaction (no protein atoms within 4.0 Å) were removed. Ligand query groups returning a nonbonded interaction (upper limit 4.0 Å) with an ion or any ligand atom were also removed. To eliminate a source of redundancy, when a ligand (based on the PDB ligand identifier) was present in several, not necessarily homologous PDB, structures, the structure with the best resolution was selected. In cases where several ligands bearing a query group were included in one structure, the first ligand occurrence in the PDB file was selected. Note that a single ligand may contain several query groups.

Query groups and their environments were also retrieved from protein-only structural data (both intrachain and interchain contacts for a given PDB file). In that setup, protein query groups were deduced directly from the atom names in the PDB file. Because there are plenty of data, to limit the computational workload, protein-only contacts were limited to

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

20 000 randomly extracted samples for each query group. The extraction process for each query group was repeated five times with different random seeds, and nearly identical results were obtained.

Definition of Molecular Environment. The molecular environment of the query groups was defined by all atoms present within the sphere(s) centered on either a point charge atom for queries I, II, and III or a single point (a centroid) representing the functional group for queries IMD, GAI, and COO. A centroid is used for these latter groups to avoid combining the interacting environment of individual atoms. For IMD, the centroid was defined by the center of mass of the side-chain aromatic nitrogen atoms, for GAI by the Cζ carbon, and for COO by the center of mass of the side-chain carboxylate oxygen atoms.

Twelve protein atom types, deduced using the PDB files annotation, were used to describe the environments. Oox, carboxyl oxygen atoms; Oh, hydroxyl oxygen atoms; Oph, phenol oxygen atoms; Ow, water molecule oxygen atoms; Oc, side-chain or main-chain carbonyl oxygen atoms; Nam, side-chain or main-chain amide nitrogen atoms; Nim, IMD nitrogen atoms; Ngu, GAI nitrogen atoms; NaI, primary amine nitrogen atoms; Car, aromatic carbons; Su, sulfur atoms; and Xot, remaining carbon atoms (see Table 4 for a complete description).

Three types of analyses were conducted using both PDB1.5 and PDB3.0 datasets for both ligand queries and protein queries: (i) for each atom type, we measured if at least one representative was found near the query groups I, II, III, IMD, and GAI as well as COO; (ii) we collected the relative densities of the presence of a given atom type within a sphere collection radius, up to 6.0 Å; and (iii) we investigated the composition of the neighborhood in terms of atom type frequency at 3.0 and 4.0 Å. These values were chosen because it is common practice to use a 4.0 Å sphere when studying salt bridges,[44,45] and a sphere of 3.0 Å radius allows to focus on stronger (shorter) hydrogen-bonded interactions. For centroids, the radii of the spheres used for data collection were corrected by subtracting an empirically defined distance d that cancels the offset introduced by the use of centroids (Figure 13). Distance d takes the values +1.0 Å for IMD, +1.1 Å for GAI, and +0.8 Å for COO.

Null Environments. The so-called "null environments" were defined as references and used to compare the environment seen by each ligand and protein query group with the environment seen by (i) any ligand atom and (ii) any protein atom. For ligand queries, the environments of all ligand atoms, a total of 126 808 atoms at 3.0 Å of resolution and 10 314 atoms at 1.5 Å resolution, were extracted. For proteins, the null environments were defined using a set of 200 000 random protein atoms. The comparison of null environments against query group environments (global counts of occurrence by atom types, grouped together by the query group) was conducted using contingency table comparison statistical tests. In the case of large effectives (more than one thousand data points), a Pearson's chi-square test was realized. In the case of smaller effectives, the exact goodness-of-fit test was preferred. In the case of multinomial tests from a contingency table containing more than 2 × 2 entries, a Bonferroni correction was applied on p-value thresholds of significance. For further information about the statistical methods used, see ref 52.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00739.

Neighborhoods of ligand query groups I, II, III, IMD, GAI, and COO using the nonredundant PDB50 release of the PDB with a resolution below 1.5 Å (A) and below 3.0 Å (B); neighborhoods of protein query groups I, IMD, GAI, and COO using the nonredundant PDB50 release of the PDB with a resolution below 1.5 Å (A) and below 3.0 Å (B); density of presence for selected protein atoms in the neighborhood of ligand queries for I, II, III, and IMD; density of presence for the selected protein atoms in the neighborhood of ligand queries for GAI and COO; density of presence for the selected protein atoms in the neighborhood of protein queries for I, IMD, GAI, and COO; ligand bond length of (A) CC, (B) CO, (C) CN, and (D) minimal distance between N and the plane formed by the three C connected to N for any tertiary amine found in the PDB not filtered; content of the dataset collected using the nonredundant PDB50 release of the PDB with resolution below 1.5 Å and below 3.0 Å; null environments defined from the column "any atoms"; composition of the protein−ligand neighborhood at 4.0 Å from ligand queries using the PDB1.5 dataset; composition of the protein−ligand neighborhood at 4.0 Å from ligand queries using the PDB3.0 dataset; composition of protein−ligand neighborhood at 3.0 Å from ligand queries using the PDB1.5 dataset; composition of the protein−ligand neighborhood at 3.0 Å from the ligand queries using the PDB3.0 dataset; and pK_a considered for each query group (PDF)

Coordinates (PDB format) of the superimposed ligand query groups and interacting atoms (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: henri.xhaard@helsinki.fi. Phone +358-2941-59190 (H.X.).

ORCID

Alexandre Borrel: 0000-0001-6499-4540

Present Address

§Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA (A.B.).

Author Contributions

A.B. developed the PDB data mining. A.B. and H.X. conducted the data analysis and wrote the manuscript. All authors have given approval to the final version of the manuscript.

Funding

This study was funded by a grant from the French research ministry (A.B.). The Magnus Ehrnrooth foundation and the KAKSIN program from the French embassy in Finland are thanked for additional resources.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Drug Discovery and Chemical Biology—Biocenter Finland network and the Center for Scientific Computing (CSC-IT) for organizing computational resources

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

and local infrastructure. We thank the Integrative Life Science—Informational and Structural Biology Doctoral Program for organizing graduate studies.

REFERENCES

(1) Musafia, B.; Buchner, V.; Arad, D. Complex salt bridges in proteins: statistical analysis of structure and function. *J. Mol. Biol.* **1995**, *254*, 761−770.
(2) Gilli, P.; Pretto, L.; Bertolasi, V.; Gilli, G. Predicting hydrogen-bond strengths from acid−base molecular properties. The pK_a slide rule: toward the solution of a long-lasting problem. *Acc. Chem. Res.* **2009**, *42*, 33−44.
(3) Onufriev, A. V.; Alexov, E. Protonation and pK changes in protein−ligand binding. *Q. Rev. Biophys.* **2013**, *46*, 181−209.
(4) Kim, M. O.; McCammon, J. A. Computation of pH-dependent binding free energies. *Biopolymers* **2016**, *105*, 43−49.
(5) Forrest, L. R.; Honig, B. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 296−309.
(6) Kumar, S.; Nussinov, R. Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. *Biophys. J.* **2002**, *83*, 1595−1612.
(7) Donald, J. E.; Kulp, D. W.; DeGrado, W. F. Salt bridges: Geometrically specific, designable interactions. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 898−915.
(8) Meuzelaar, H.; Tros, M.; Huerta-Viga, A.; van Dijk, C. N.; Vreede, J.; Woutersen, S. Solvent-Exposed Salt Bridges Influence the Kinetics of α-Helix Folding and Unfolding. *J. Phys. Chem. Lett.* **2014**, *5*, 900−904.
(9) Marqusee, S.; Sauer, R. T. Contributions of a hydrogen bond/salt bridge network to the stability of secondary and tertiary structure in λ repressor. *Protein Sci.* **1994**, *3*, 2217−2225.
(10) Sarakatsannis, J. N.; Duan, Y. Statistical characterization of salt bridges in proteins. *Proteins* **2005**, *60*, 732−739.
(11) Yip, K. S. P.; Britton, K. L.; Stillman, T. J.; Lebbink, J.; de Vos, W. M.; Robb, F. T.; et al. Insights into the molecular basis of thermal stability from the analysis of ion-pair networks in the glutamate dehydrogenase family. *Eur. J. Biochem.* **1998**, *255*, 336−346.
(12) Olson, C. A.; Spek, E. J.; Shi, Z.; Vologodskii, A.; Kallenbach, N. R. Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 123−132.
(13) Gvritishvili, A. G.; Gribenko, A. V.; Makhatadze, G. I. Cooperativity of complex salt bridges. *Protein Sci.* **2008**, *17*, 1285−1290.
(14) Lyu, P. C.; Gans, P. J.; Kallenbach, N. R. Energetic contribution of solvent-exposed ion pairs to alpha-helix structure. *J. Mol. Biol.* **1992**, *223*, 343−350.
(15) Kumar, S.; Nussinov, R. Salt bridge stability in monomeric proteins. *J. Mol. Biol.* **1999**, *293*, 1241−1255.
(16) Kumar, S.; Nussinov, R. Fluctuations in ion pairs and their stabilities in proteins. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 433−454.
(17) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817−2831.
(18) Ranjani, C. V.; Rangarajan, S.; Michael, D.; Roy, S.; Sekar, K. Role of water molecules and ion pairs in Dps and related ferritin-like structures. *Int. J. Biol. Macromol.* **2008**, *43*, 333−338.
(19) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(20) Ben-Shimon, A.; Shalev, D. E.; Niv, M. Y. Protonation States in Molecular Dynamics Simulations of Peptide Folding and Binding. *Curr. Pharm. Des.* **2013**, *19*, 4173−4181.
(21) Kroemer, R. T. Structure-Based Drug Design: Docking and Scoring. *Curr. Protein Pept. Sci.* **2007**, *8*, 312−328.
(22) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213−228.
(23) Huang, S.-Y.; Zou, X. Advances and challenges in Protein-ligand docking. *Int. J. Mol. Sci.* **2010**, *11*, 3016−3034.
(24) Petukh, M.; Stefl, S.; Alexov, E. The Role of Protonation States in Ligand-Receptor Recognition and Binding. *Curr. Pharm. Des.* **2013**, *19*, 4182−4190.
(25) Dougherty, D. A. The cation−π interaction. *Acc. Chem. Res.* **2013**, *46*, 885−893.
(26) Niggemann, M.; Steipe, B. Exploring local and non-local interactions for protein stability by structural motif engineering. *J. Mol. Biol.* **2000**, *296*, 181−195.
(27) Gromiha, M. M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.
(28) Hall, H. K. Correlation of the Base Strengths of Amines. *J. Am. Chem. Soc.* **1957**, *79*, 5441−5444.
(29) Bruice, T. C.; Schmir, G. L. Imidazole catalysis. II. The reaction of substituted imidazoles with phenyl acetates in aqueous solution. *J. Am. Chem. Soc.* **1958**, *80*, 148−156.
(30) Albert, A.; Goldacre, R.; Phillips, J. The Strength of Heterocyclic Bases. *J. Chem. Soc.* **1948**, 2240−2249.
(31) Wood, E. Data for Biochemical Research (third edition). *Biochem. Educ.* **1987**, *15*, 97.
(32) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.
(33) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 827−828.
(34) Sabarinathan, R.; Aishwarya, K.; Sarani, R.; Vaishnavi, M. K.; Sekar, K. Water-mediated ionic interactions in protein structures. *J. Biosci.* **2011**, *36*, 253−263.
(35) Zhang, Q.; Xu, Z.; Zhu, W. The Underestimated Halogen Bonds Forming with Protein Side Chains in Drug Discovery and Design. *J. Chem. Inf. Model.* **2017**, *57*, 22−26.
(36) Williams, M. A.; Ladbury, J. E. Hydrogen Bonds in Protein-Ligand Complexes. In *Protein-Ligand Interact: From Molecular Recognition to Drug Design*; Böhm, H.-S., Schneider, G., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA, 2003; pp 137−161.
(37) Sawyer, L.; James, M. N. G. Carboxyl−carboxylate interactions in proteins. *Nature* **1982**, *295*, 79−80.
(38) Knorr, A.; Ludwig, R. Cation-cation clusters in ionic liquids: Cooperative hydrogen bonding overcomes like-charge repulsion. *Sci. Rep.* **2015**, *5*, 17505.
(39) Gilli, G.; Gilli, P. Towards an unified hydrogen-bond theory. *J. Mol. Struct.* **2000**, *552*, 1−15.
(40) Deepak, R. N. V. K.; Sankararamakrishnan, R. N−H···N Hydrogen Bonds Involving Histidine Imidazole Nitrogen Atoms: A New Structural Role for Histidine Residues in Proteins. *Biochemistry* **2016**, *55*, 3774−3783.
(41) Preimesberger, M. R.; Majumdar, A.; Rice, S. L.; Que, L.; Lecomte, J. T. J. Helix-Capping Histidines: Diversity of N−H···N Hydrogen Bond Strength Revealed by 2h J NN Scalar Couplings. *Biochemistry* **2015**, *54*, 6896−6908.
(42) Jackson, M. R.; Beahm, R.; Duvvuru, S.; Narasimhan, C.; Wu, J.; Wang, H.-N.; et al. A preference for edgewise interactions between aromatic rings and carboxylate anions: the biological relevance of anion−quadrupole interactions. *J. Phys. Chem. B* **2007**, *111*, 8242−8249.
(43) Schwans, J. P.; Sunden, F.; Lassila, J. K.; Gonzalez, A.; Tsai, Y.; Herschlag, D. Use of anion-aromatic interactions to position the general base in the ketosteroid isomerase active site. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 11308−11313.
(44) Gupta, P. S. S.; Nayek, A.; Banerjee, S.; Seth, P.; Das, S.; Sur, V. P.; et al. SBION2: Analyses of Salt Bridges from Multiple Structure Files, Version 2. *Bioinformation* **2015**, *11*, 039−042.
(45) Gupta, P. S. S. P.; Mondal, S.; Mondal, B.; Islam, R. N. U.; Banerjee, S.; Bandyopadhyay, A. K. SBION: A Program for Analyses of

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374

Salt-Bridges from Multiple Structure Files. *Bioinformation* **2014**, *10*, 164−166.
(46) Banyikwa, A. T.; Goos, A.; Kiemle, D. J.; Foulkes, M. A. C.; Braiman, M. S. Experimental and Computational Modeling of H-Bonded Arginine−Tyrosine Groupings in Aprotic Environments. *ACS Omega* **2017**, *2*, 5641−5659.
(47) Zhang, Y.; Borrel, A.; Ghemtio, L.; Regad, L.; Boije af Gennäs, G.; Camproux, A.-C.; et al. Structural Isosteres of Phosphate Groups in the Protein Data Bank. *J. Chem. Inf. Model.* **2017**, *57*, 499−516.
(48) Team R Core (R Foundation for Statistical Computing). *R: A Language and Environment for Statistical Computing*, 2015.
(49) DeLano, W. L. *The PyMOL Molecular Graphics System*. http://www.pymol.org/, 2002.
(50) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; et al. UCSF Chimera−a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(51) Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992**, *355*, 472−475.
(52) McDonald, J. H. *Handbook of Biological Statistics*, 2nd ed.; Spartky House Publishing: Baltimore, MD, U.S.A., 2009.

DOI: 10.1021/acsomega.7b00739
ACS Omega 2017, 2, 7359−7374