Evaluating Unexpectedly Short Non-covalent Distances in X-ray Crystal Structures of Proteins with Electronic Structure Analysis

Citation: Qi, Helena W. and Heather J. Kulik. "Evaluating Unexpectedly Short Non-covalent Distances in X-ray Crystal Structures of Proteins with Electronic Structure Analysis." Journal of Chemical Information and Modeling 59, 5 (March 2019): 2199–2211 © 2019 American Chemical Society

As Published: http://dx.doi.org/10.1021/acs.jcim.9b00144

Publisher: American Chemical Society (ACS)

Persistent URL: https://hdl.handle.net/1721.1/130535

Version: Original manuscript: author's manuscript prior to formal peer review

Evaluating Unexpectedly Short Non-covalent Distances in X-ray Crystal Structures of Proteins with Electronic Structure Analysis

Helena W. Qi[1,2] and Heather J. Kulik[1,*]

[1]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

ABSTRACT: We investigate unexpectedly short non-covalent distances (< 85% of the sum of van der Waals radii) in atomically resolved X-ray crystal structures of proteins. We curate over 13,000 high quality protein crystal structures and an ultra-high resolution (1.2 Å or better) subset containing > 1,000 structures. Although our non-covalent distance criterion excludes standard hydrogen bonds known to be essential in protein stability, we observe over 82,000 close contacts in the curated protein structures. Analysis of the frequency of amino acids participating in these interactions demonstrates some expected trends (i.e., enrichment of charged Lys, Arg, Asp, and Glu) but also reveals unexpected enhancement of Tyr in such interactions. Nearly all amino acids are observed to form at least one close contact with all other amino acids, and most interactions are preserved in the much smaller ultra high-resolution subset. We quantum-mechanically characterize the interaction energetics of a subset of > 6,000 close contacts with symmetry adapted perturbation theory to enable decomposition of interactions. We observe the majority of close contacts to be favorable. The shortest favorable non-covalent distances are under 2.2 Å and are very repulsive when characterized with classical force fields. This analysis reveals stabilization by a combination of electrostatic and charge transfer effects between hydrophobic (i.e., Val, Ile, Leu) amino acids and charged Asp or Glu. We also observe a unique hydrogen bonding configuration between Tyr and Asn/Gln involving both residues acting simultaneously as hydrogen bond donors and acceptors. This work confirms the importance of first-principles simulation in explaining unexpected geometries in protein crystal structures.

1. Introduction

Protein X-ray crystal structures provide essential insight into protein structure-function relationships. Beyond providing a foundation for atomistic understanding of the behavior of biomacromolecules, crystal structures also heavily influence computational chemistry through their use in experimental tuning and validation of molecular mechanics (MM) force field models[1-4], validation of higher-level, quantum mechanical (QM) methods[5-7], and in the development of data driven models[8]. The over 100,000 protein structures in the Protein Data Bank (PDB)[9] provide a rich source of information that has been heavily mined in recent years, typically in conjunction with QM simulation, to reveal previously unknown non-covalent interactions including non-covalent carbon bonds[10-11], n to π* interactions[12-13], protein-ligand cation-π, aromatic, or other interactions[14-19], and to shed light on salt bridges[20]. Within the domain of hydrogen bonding in particular, PDB surveys have provided guidance on less well-known N-H···N[21-23], sulfur-containing[24-26], X-H π [27-28], and C-H···O[29] hydrogen bonds, among others. Despite their relative rarity, these interactions may be important for influencing relative stability of protein conformational states or for positioning substrates in catalytically competent geometries in the active sites of enzymes.

Nearly all holistic simulation of proteins is carried out with classical MM force fields, which have been noted in recent years to be in broadly improving agreement with experiment.[30] Nevertheless, for some of the most interesting challenges in biochemistry, such as intrinsically disordered proteins[31] or unexpected orientations and distances in globular proteins[32], classical force fields can fail to predict essential emergent phenomena. For example, at least a dozen unique, high-resolution (1.3-2.4 Å) crystal structures[33-39] have been solved of the enzyme catechol *O*-methyltransferase (COMT) in which the non-covalent C···O substrate distances in the

active site averaged 2.65 Å (i.e., 82% of the sum of van der Waals radii, vdW, of the two atoms). Classical MM simulations instead sample longer distances around 3.25 Å [40-41] (i.e., 100% of the sum of vdW radii of the two atoms). Large scale QM/MM simulations, only recently possible with advances in hardware and algorithms,[6, 42-50] have been shown to reproduce the experimentally observed distances[51-52]. Thus, these observations introduce unexpectedly short non-covalent distances in proteins as another class of poorly understood phenomena with which MM force fields can be expected to struggle[53].

Although for COMT, the large number of structures suggests that the unexpectedly short distances are a real physical phenomenon, shorter than van der Waals distances are often taken as a sign of poorly solved X-ray structures in conjunction with estimates based on properties of the electron density[54-55]. For instance, Molprobity[56], which is widely used to validate protein structures, assigns a clashscore defined as the number of non-bonded atoms separated by distances 0.4 Å shorter than the sum of the respective atoms' van der Waals radii. Low clashscores may be used to identify a more probable structural model. Force field refinement has been suggested to eliminate such too short non-bonded distances during crystal structure refinement.[57] Taking inspiration from COMT, we seek to identify non-covalent distances that are unexpectedly short but favorable, such that validation and force field refinement could yield structures with less physical insight. We thus employ an informatics approach to mine the PDB and combine this with careful curation of structures to eliminate cases where short distances are believed to be associated with poorly solved, low resolution structures. We down select a subset of ultra-high resolution (< 1.2 Å) structures that are expected to have on average < 0.03 Å errors in atomic positions[54, 58]. Finally, we evaluate the interactions with first-principles, symmetry

adapted perturbation theory to confirm their favorability and assess their quantum mechanical origin.

For structure mining of the PDB, we define our target to be unexpectedly short non-covalent distances in terms of sums of element-specific[59] van der Waals radii that have been commonly employed in evaluating non-bonded contact distances.[56, 60-61] Motivated by observations in COMT, we define a close contact as a non-bonded distance, d_{AB} between two heavy atoms, A and B, that is no more than 85% of the sum of vdW radii, r, of the substituent atoms:

$$d_{AB} \leq 0.85*(r_A + r_B) \quad (1)$$

To put this strict distance cutoff in context, it excludes most traditionally studied non-covalent interactions, such as conventional hydrogen bonds. For example, the van der Waals radius of an oxygen atom is 1.52 Å, giving a 2.58-Å cutoff for 85% of the sum of van der Waals distance (i.e., 3.04 Å, Figure 1). The 2.95 Å O···O distance in a water dimer hydrogen bond[62] is thus 97% of the van der Waals distance (Figure 1). A special class of low-barrier hydrogen bonds[63-64] (LBHBs) in which the barrier to hydrogen transfer is similar to the zero-point energy[65-66] is typically[67] characterized by O···O separations around 2.6 Å or lower, placing them at around the cutoff for our close contact definition (Figure 1). Other especially strong hydrogen bonds, such as charge-assisted hydrogen bonds and salt bridges can be expected to also fall within our cutoff definitions (Figure 1).[65, 68-71] What remains broadly unclear about unexpectedly short distances is how prevalent they are and whether or not they only occur between a few types of amino acids.

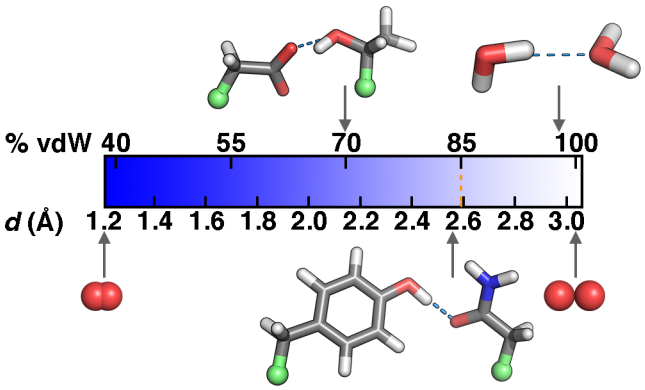


Figure 1. Examples of separation between oxygen atoms in terms of percentage of the sum of van der Waals radii (top, vdW) or distance in Å (*d*, bottom) ranging from the O_2 molecule to two O atoms separated by 100% of the sum of their vdW radii, both shown in inset. The separation in the water dimer, a hydrogen bond between a Tyr hydroxyl and Asn carbonyl oxygen, and a charge-assisted hydrogen bond between the Thr hydroxyl and Asp carboxylate are also shown. The 85% vdW cutoff used to define a close contact in this work is shown as an orange dashed line.

In this work, analysis of high quality crystal structures from the PDB reveals nearly 100,000 of such interactions, and quantum mechanical analysis of thousands of these amino acid pairs reveals the majority of them to be favorable. The rest of this manuscript is as follows. In Section 2, we describe the curation and overall characteristics of the protein data sets. In Section 3, we quantify close contacts and characterize the propensities of protein structures to form close contacts. In Sections 4 and 5, we describe the details and results of quantum mechanical simulation of over 6,000 close contact amino acid pairs. Finally, we provide our conclusions in Section 6.

2. Curation of Representative Protein Structure Sets.

Protein Data Set Curation. We selected X-ray crystal structures based on both resolution and quality criteria from a snapshot of the protein data bank (PDB)[9] obtained on October 29, 2017. From the 112.9k X-ray crystal structures in the PDB at that time, a total of 49,238 single protein entity structures of atomic resolution[54, 72] (i.e., 2.0 Å or better) were retained (Table 1). To avoid an imbalanced over counting of duplicate structures, a 90% identity cutoff filter was applied, generally excluding identical sequences or structures that differ by point mutations,

leading to an a data set we refer to as the all data (**AD**) set of 17,854 valid 2.0 Å or higher resolution X-ray crystal structures (Table 1 and Figure S1).

Table 1. Step-by-step curation of protein crystal structures from the full PDB data set. The structures column indicates the number satisfying the selection criterion on that row and all preceding criteria. *Edge cases: five structures were eliminated due to unavailability of validation reports (two cases), lack of use of standard PDB format (one case), or no standard amino acids in the structure (two cases).

Selection criteria	Structures	Set
Full PDB (on 10/29/17)	134,656	
X-ray crystal structures containing protein	112,862	
$\leq$ 2.0 Å resolution	56,107	
Single protein entity	49,238	
90% sequence identity	17,859	
Removal of edge cases*	17,854	**AD**
$R \leq 20\%$	14,418	
$R_{free} - R \leq 0.07$	14,051	
RSRZ $\leq$ 20%	13,472	**FD**
Resolution $\leq$ 1.2 Å	1,151	**HR**

To avoid excluding the unexpectedly short interactions we are aiming to study, we focused on reducing the **AD** subset by only using criteria that judge the quality of the model density fit to the diffraction pattern. The three density metrics we used were: the *R*-factor, which measures differences between calculated and experimental structure factor amplitudes, R_{free}, which is the *R*-factor computed on a set of data held out during fitting, and *RSRZ*, the *Z*-score of the real-space *R* value, which measures how well the observed and calculated densities of each residue match, scaled against the average score for that residue in other structures that have similar resolutions. We required that i) $R \leq 20\%$, a criterion well-refined structures[54] readily satisfy, ii) the R_{free}[55] differs from *R* by no more than 7%,[54] and iii) fewer than 20% of all residues are RSRZ outliers (Table 1 and Supporting Information Figures S2-S5). When automatically calculated quantities[73] were available with the PDB record, these were used rather than the self-reported values because the automatically calculated quantities usually are higher (i.e., indicating

poorer model quality), producing a more conservative data set (Supporting Information Figures S2-S4 and S6).

These three criteria eliminated 4,382 structures for a final full data (**FD**) set of 13,472 structures (Table 1). The majority of violating structures were eliminated by excluding $R > 20\%$ cases, although the criteria are not strictly independent (Table 1 and Supporting Information Figures S4-S5). We also defined a high-resolution (**HR**) 1,151 protein structure subset that satisfied all of the above criteria and additionally a high enough resolution (i.e., 1.2 Å or better) to resolve all non-hydrogen atoms[54] with anticipated[58] low positional errors (i.e., < 0.03 Å). This subset is expected to have more reliable geometric information but is only $1/10^{th}$ of the size due to the limited number of structures available at resolutions 1.2 Å or better (Supporting Information Figure S1).

In this work, we disregarded geometric criteria widely used in crystal structure validation[74] because we seek interactions associated with unexpected non-covalent distances. Reports[75-76] on structures usually include the number of clashes per 1000 atoms after adding hydrogen atoms, as defined by non-covalent distances at least 0.4 Å smaller than the sum of van der Waals radii.[56] This clashscore varies widely across our data set with only weak correlation with the resolution of the structure (Supporting Information Figure S7). Other criteria such as backbone (i.e., Ramachandran[77]) or rotameric[74] outliers were also not used to restrict the data set, though we note that in general structures are predominantly of high quality for all of these metrics (Supporting Information Figure S8).

Characteristics of the Data Sets. The 13,472 protein **FD** set reflects amino acid abundances observed in the full PDB[75-76]. Small discrepancies are observed in an overestimation in the **FD** set of Gly and Ala abundance with a corresponding underestimation of Lys or Ser

residues, but relative abundance is within 0.5% of the PDB for the remaining cases (Supporting Information Figure S9). In the **FD** set, each amino acid is present in the common secondary structure elements (Supporting Information Figure S10). As may be expected, the smaller 1,151-protein **HR** set does not match the overall PDB abundance as well as the **FD** set. Statistical tests show some overabundance of Asn, Gly, and Thr in the **HR** set and simultaneous underrepresentation of Arg, Glu, and Leu (Supporting Information Figure S9). These differences also correlate to a weak decrease in average protein size as compared to the **FD** set but overall otherwise similar distributions (Supporting Information Figures S11-S12).

Close Contact Curation. As described in the introduction, we define close contacts (CCs) as non-covalent distances between heavy atoms (i.e., C, N, O, or S) within 85% of the sum of the respective atoms' van der Waals' radii but longer than the sum of their covalent radii (Figure 1 and Supporting Information Table S1). To isolate purely non-covalent residue-residue interactions, we exclude CCs i) between covalently bound residues (e.g., residues adjacent in the primary sequence or bonded through disulfide bonds) ii) between non-standard residues, iii) within 3.5 Å of non-standard amino acids and substrates to avoid capturing indirect effects on the residue-residue interaction, and iv) with the carboxylate of the C-terminal residue of a protein. A moderate fraction (ca. 10-30%) of proteins in each set (**FD**: 1,358, **HR**: 349) have no CCs that satisfy all criteria (Supporting Information Table S2). Of the remaining proteins, these restrictions produce 82,701 CCs in 12,114 **FD** proteins and 2,326 CCs in 802 **HR** proteins, with only a small (< 5-7%) fraction corresponding to multiple CCs for the same residue-residue pair (Supporting Information Table S2). A larger number of CCs is observed for each protein in the **FD** set compared to the smaller **HR** subset (i.e., 7 vs. 3 per protein) along with a broader distribution of observed CC per structure (Supporting Information Figure S13). The difference in

protein size in the two sets can be partly accounted for by evaluating the number of CCs per residue, which narrows the difference in the observed distribution slightly but still points to a lower CC frequency in the **HR** subset (Supporting Information Figure S13). Whether the apparent resolution dependence of CC frequency is due to elimination through improved quality of the structure will be revisited in the context of quantum mechanical evaluation of a subset of these interactions in Sec. 4 (Supporting Information Figure S14).

3. Overall Analysis of Close Contacts.

Having confirmed that close contacts occur in high abundance in atomic resolution (ca. 2.0 Å or better) structures and are preserved in high-resolution (ca. 1.2 Å or better) structures, we now classify the types of residues participating in these interactions. To quantify the enrichment of a residue type in CCs, we computed the unique CC frequency relative to the abundance of an amino acid in the **FD** protein set (Figure 2, all close contacts in Supporting Information Figure S15). Enrichment of positively charged Arg or Lys and negatively charged Asp or Glu is expected, as the residues are known[78] to form salt bridges or charge-assisted hydrogen bonds that are characterized by short non-covalent distances (Figure 2 and Supporting Information Table S1). More unexpectedly, Tyr CC enrichment exceeds that of even these charged residues, and polar Ser and Thr have comparable enrichment to the charged residues (Figure 2). Breaking down CCs by the atom type participating confirms that hydroxyl oxygen participates more frequently than negatively charged carboxylate oxygen in CCs (Supporting Information Figure S16). The reverse is true for nitrogen atoms, where positively charged forms occur more frequently in close contacts than their polar, neutral counterparts (Supporting Information Figure S16). The Tyr CC enhancement can be directly attributed to its hydroxyl functional group because Phe, which differs from Tyr only by the lack of a hydroxyl group, has one of the lowest

CC frequencies (Figure 2).

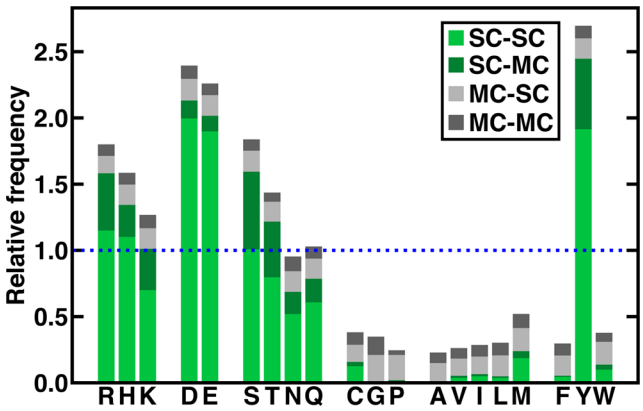


Figure 2. By-residue frequency of **FD** data set unique close contacts relative to their residue abundance. Each residue interaction is classified as being between sidechain (SC) and mainchain (MC) atoms, with the first label referring to the indicated residue and the second to its partner: SC-SC (light green) and SC-MC (dark green) are residue-specific, whereas MC-SC (light gray) and MC-MC (dark gray) are not. A relative frequency of 1 is indicated by the blue dotted horizontal bar.

Although neutral and hydrophobic residues generally have lower relative CC counts than their polar or charged counterparts, the absolute number of CCs involving these residues is significant due to their high abundance (Supporting Information Figure S9). Roughly 25% of the protein CCs in our set involve special (Cys, Gly, Pro), nonpolar (Ala, Val, Ile, Leu, Met) or non-Tyr aromatics (Phe, Trp) (a list of all CCs is provided in the Supporting Information). Overall, Met is among the highest relative frequency in this subset, followed by the also sulfur-containing Cys and then Trp (Figure 2). Trends in amino acid CC enrichment for **FD** proteins are preserved in the **HR** subset, demonstrating the relative insensitivity to the resolution at which the proteins are solved (Supporting Information Figures S17-S18).

We further break down each residue's participation in a CC by whether its main chain (MC: Cα, C, N, or O) atoms are in the close contact or if it is specific to the chemistry of that amino acid through sidechain (SC) atom participation (Figure 2). As expected, both the contribution of MC-MC interactions and MC-SC, where the labeled residue's MC atom interacts

with another residue's SC atom, are relatively constant across all amino acids (Figure 2). Conversely, significant variation in amino acid enrichment is observed in the SC-MC interactions: Ser and Thr SC atoms form the most frequent close contacts with MC atoms of another residue, followed by Tyr and the positively charged Arg or Lys (Figure 2). Negatively charged Asp or Glu and the nonpolar residues much less frequently form CCs with MC atoms of other amino acids (Figure 2). The relative distribution of interaction types is preserved in the **HR** subset, except for a reduction in contribution from SC-MC interactions particularly for Ser, Thr, and Tyr (Supporting Information Figures S17-S18). Overall, trends in relative participation in CCs for both the **HR** and **FD** sets can largely be attributed to the significant SC-SC fraction arising from chemically specific interactions. Over the **FD** set, CCs are generally distributed in all secondary structure types in roughly equal measure to the propensity in the overall proteins (Supporting Information Figures S10 and S19).

To further deduce the chemical origins of close contacts, we computed the number of SC-SC interactions for each residue pair in the **FD** set (Figure 3). In this set, CCs occur between all classes of residues, even those that are typically of the same charge (e.g., Asp to Glu or Arg to Lys) (Figure 3). Pairwise CC frequencies confirm expectations: numerous CCs are observed between Asp or Glu and Arg or Lys (Figure 3). Polar Ser or Thr, which both have –OH sidechains, form the most CCs with Asp or Glu (Figure 3). In comparison, Asn and Gln form CCs with both positively and negatively charged residues in roughly equal measure (Figure 3). Among polar residues, Ser/Thr to Asn/Gln CCs are more common than same-residue or same-type interactions, although all subsets occur with higher frequency than for the hydrophobic and aromatic amino acid types (Figure 3 and Supporting Information Figure S20).

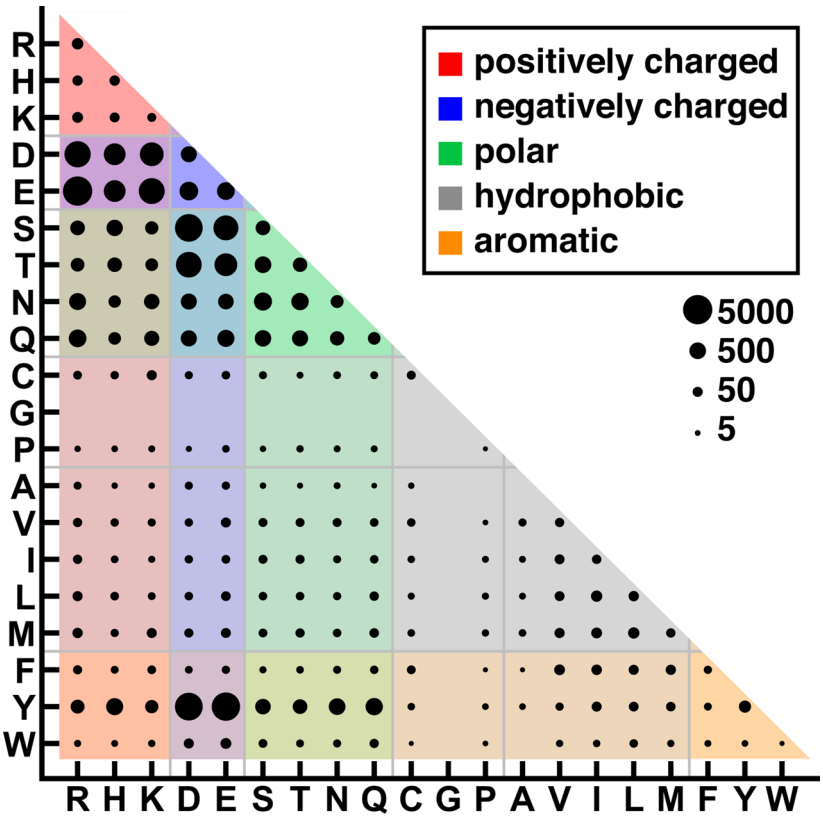


Figure 3. Matrix of absolute close contacts between residue sidechains (SC-SC) in the **FD** set between residues grouped (i.e., separated by thin gray bars) by type according to canonical charge assignment and indicated with single letter codes as indicated in inset legend: positively charged (R, H, K in red), negatively charged (D, E in blue), polar (S, T, N, Q in green), special (C, G, P in gray) or nonpolar (A, V, I, L, M in gray) indicated as hydrophobic in legend, and aromatic (F, Y, W in orange). The area of each circle represents the number of close contacts, as indicated qualitatively by the inset legend of representative circle sizes.

In the **FD** set, all residues individually form at least one CC with all other residues, with the exception of Gly which by definition has no SC atoms, and Ala that forms no CCs with itself, Trp, or Pro (Figure 3). In the smaller **HR** set, all **FD** observations are preserved, albeit with fewer and more sporadic hydrophobic-hydrophobic or hydrophobic-aromatic CCs (Supporting Information Figure S20). For example, in the **HR** subset, Cys only forms CCs with Ser, Thr, or Leu. Ala, which was noted in the **FD** set to form CCs with 17 residues, only forms CCs with Arg, His, or Asp (Supporting Information Figure S20). The **HR** subset also indicates a more pronounced preference of hydrophobic amino acids (e.g., Val or Leu) for forming CCs with

charged residues than was clear from the larger **FD** set. Without further characterization using computational chemistry (see Secs. 4-5), it would be challenging to distinguish CCs that are due to poor solution of the crystal structure (and thus are more likely in the **FD** than **HR** data set) vs. other differences in the **HR** subset due to differences in protein size and the subset size.

4. Simulation Details.

Extracting residue-residue pairs. 6,279 CC residue pairs were prepared for quantum mechanical (QM) and classical molecular mechanics (MM) modeling. This set includes all **HR** CCs as well as 4,114 CCs randomly selected from the **FD** set (see Supporting Information). Protonation and tautomeric states of ionizable residues were determined using PDB2PQR[79] at a pH of 7.0 on the relevant chain and pose of the entire protein. Cysteine residues that were predicted to be part of disulfide bonds were treated as protonated. For 95 structures, missing residues prevented automated protonation state assignment, in which case standard protonation state of the residue was chosen. Residue pairs were extracted from the relevant PDB file and protonated with the tleap module of AMBER[80] for subsequent simulation of the isolated pair. Capping hydrogen atoms were added along the protein backbone vector with a scaled bond length (1.09 Å for C-H, 1.01 Å for N-H), except for terminal residues where the PyMOL[81] *add hydrogens* feature was used due to the lack of backbone. Lists of all residue pairs and originating PDB ID, protonation states, and the Cartesian coordinates of structures are provided in the Supporting Information.

QM modeling of residue-residue pairs. Extracted residue-residue pair structures were studied with density functional theory (DFT) in TeraChem[44, 82]. Constrained geometry optimizations were carried out enforcing resolved crystal structure atom and capping H atom positions, with only the remaining H atoms relaxed. All geometry optimizations were carried out

using the L-BFGS algorithm in Cartesian coordinates, as implemented in DL-FIND[83], to default thresholds of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in self-consistent field (SCF) energy between steps. The implicit conductor-like solvent model (COSMO)[84-86] with $\varepsilon = 4$ was employed to reduce spurious proton transfer events by approximately modeling the screening of the protein environment. The H-atom-only optimizations were carried out at the B3LYP[87-89]/6-31G*[90] level of theory with empirical dispersion corrections[91]. Gas phase residue-pair QM interaction energies and decomposition (i.e., into electrostatics, exchange, induction, and dispersion) were calculated with single point energies using symmetry adapted perturbation theory (DF-SAPT0[92-94]) with the jun-cc-pVDZ basis set, as implemented in Psi4[95]. These calculations were carried out with all heavy atoms at their original crystal structure positions. Psi4 does not enforce spatial symmetry of the wavefunction during calculations of the monomer or dimer. Comparison of SAPT0 to more computationally demanding levels of symmetry adapted perturbation theory (i.e., SAPT2) showed good agreement on the systems studied (Supporting Information Figures S21-S22). We also computed B3LYP rigid interaction energies (i.e., the difference of monomer and dimer energetics) and confirmed their good agreement with SAPT0 interaction energies (Supporting Information Figure S23). Selected optimizations were also carried out enforcing only the position of select mainchain atoms (Cα, C, N) but with all other H atoms and sidechain atoms relaxed, with methodology described in the Supporting Information.

MM modeling of residue-residue pairs. To calculate MM interaction energies, MMPBSA.py[96] was used for Generalized Born (GB)[97] decomposition calculations with the AMBER ff14SB[4] force field, as implemented in AMBER. Only gas phase van der Waals and

electrostatic components were analyzed, and all solvent terms were discarded for comparison to the gas phase SAPT0 QM results.

5. Interactions in Close Contacts.

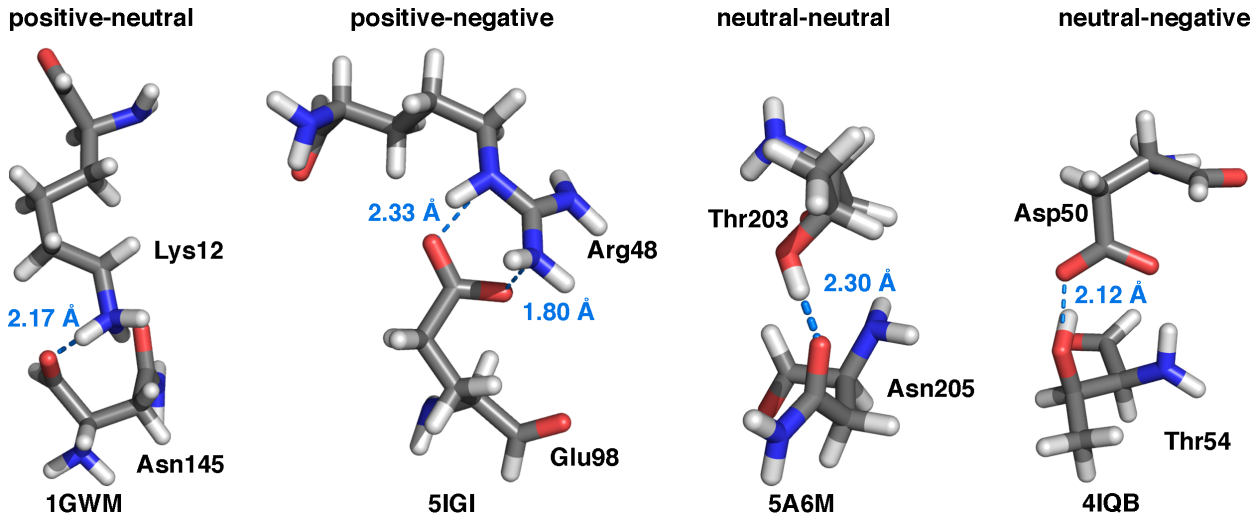
To determine the favorability of unexpectedly short non-covalent distances, we evaluated the interaction energies of over 6,000 residue-residue CC pairs extracted from their respective crystal structures (see Simulation Details). Upon placement and optimization of hydrogen atoms while keeping the crystal heavy atoms fixed, 65% (4,029 of 6,201) of the interactions are favorable at the gas phase QM level of theory (Supporting Information Figure S24). In the unscreened, gas phase QM pairwise calculations of CCs, electrostatics play a large role: as in prior work[98], none of the electrostatically-repulsive 30 positive-positive or 94 negative-negative pairs have favorable SAPT interaction energies, whereas all but one of the 786 positive-negative pairs are favorable (Supporting Information Table S3 and Figures S25-S26). The neutral-neutral CCs or the cases where a neutral residue interacts with a positive or negatively charged residue are more balanced, with 50% and 70-80% favorable interactions, respectively (Supporting Information Table S3 and Figures S25-S26). Differences in interactions by residue charge can be rationalized by the strongly favorable induction and electrostatic contributions (ca. -50 to -150 kcal/mol) to the SAPT energy that can outweigh even very large (ca. 50-100 kcal/mol) exchange repulsion contributions (Supporting Information Figure S27).

Positional errors in the full **FD** set can be anticipated to be one reason why roughly one third of the CCs were computed to be repulsive. Thus, we allowed the structures of the residue pairs to optimize sidechain positions with fixed main chain atomic positions to partially approximate the full protein environment (see Simulation Details). Focusing on the neutral set that should be least affected by the absence of a complete protein environment, we note that after

optimization the CC distances that ranged from 1.8-3.1 Å elongate significantly in almost all cases (Supporting Information Figures S28-S29). The most substantial overlap between initial CC distance and its optimized distance is observed for 2.5-3.0 Å separations, suggesting distances significantly below that threshold between neutral residue pairs are either unphysical or require the greater protein environment to enforce (Supporting Information Figures S28-S29). Often, these CCs can adjust to make the interaction favorable at the QM level of theory with only modest rearrangements of the sidechain. The median root mean squared displacement (RMSD) of all neutral-neutral CCs is only around 0.4 Å; while nearly all pairs become favorable after this optimization, they typically are no longer considered CCs by our strict geometric criteria (Supporting Information Figure S28 and Table S3). Given that we have not accounted for the full protein environment in these simulations, we thus focus on cases where the crystal structure heavy atom positions give rise to a favorable QM interaction energy.

Beyond average trends after optimization of the pairs, we determined the differences in CC distances between favorable and unfavorable residue pair QM interactions to determine whether particularly short distances were more likely to correspond to energetically unfavorable interactions. Over all CC pairs, distance distributions are comparable for favorable and unfavorable interactions, with the shortest favorable CC distance of 1.8 Å only slightly longer (vs. 1.5 Å) than the shortest distances observed in the set (Supporting Information Figure S30). A greater difference is observed for the neutral residue pairs, echoing earlier observations, as the shortest favorable CC distance of 2.3 Å is significantly longer (vs. 1.7 Å) than a number of shorter CC distances observed in the unfavorable distribution, and the favorable cases generally have a narrower CC distance distribution (Supporting Information Figure S31). Overall, four classes of residue pairs have favorable interactions: neutral, neutral-positive, neutral-negative,

and positive-negative (Figure 4). The shortest observed positive-neutral Lys-Asn (in 1GWM, N-O atom pair) CC is comparable in both distance and interaction strength to the negative-neutral Thr-Asp (in 4IQB, O-O atom pair) CC at around 2.2 Å and -14 kcal/mol (Figure 4). In both cases, strong induction and electrostatic contributions outweigh large exchange repulsion components in the SAPT energy decomposition (Figure 4 and Supporting Information Table S4). Conversely, in the MM interaction electrostatic stabilization is much weaker than the very repulsive van der Waals term, and no induction-like term is present in the force field (Supporting Information Table S4). The shortest neutral-neutral CC between the oxygen atoms of Thr and Asn (PDB: 5A6M) is somewhat longer at around 2.30 Å and only stabilized by -0.7 kcal/mol with QM due to reduced electrostatic attraction (Figure 4). Thus, it follows that oppositely charged pairs should have the shortest distances. We indeed observe a short 1.80 Å CC between the carboxylate oxygen of Glu and nitrogen of Arg (PDB: 5IGI) that is stabilized in SAPT by -33.5 kcal/mol due to strong electrostatic stabilization but extremely repulsive with MM (Figure 4 and Supporting Information Table S4). In summary, stabilizing and favorable CCs between O and N atoms should be feasible at distances as small as 70% of the sum of van der Waals radii for neutral or singly charged residue pairs and 60% for doubly charged pairs.



1GWM	
elec	-48.6
ex	66.9
ind	-19.3
disp	-12.8
total	-13.8

5IGI	
elec	-214.1
ex	215.3
ind	-43.0
disp	-27.7
total	-33.5

5A6M	
elec	-24.1
ex	42.9
ind	-11.2
disp	-8.4
total	-0.7

4IQB	
elec	-68.2
ex	115.2
ind	-43.6
disp	-15.4
total	-12.1

Figure 4. The shortest favorable close contacts (from left to right) between combinations of positive-neutral, positive-negative, neutral-neutral, and neutral-negative amino acids. Each case is annotated with its PDB ID and the relevant residue names and numbers as well as the heavy atom CC distance in Å, annotated by light blue dashed lines. For each case, components (in kcal/mol) of the SAPT0 interaction energy: electrostatic (elec), exchange repulsion (ex), induction (ind), and dispersion (disp) are shown along with the total interaction energy.

Analysis of the shortest favorable QM CC pairs has revealed significant discrepancies with MM force fields. Although a similar number of pairs are favorable when evaluated with an MM force field, the agreement between SAPT QM and MM energetics is much poorer than between hybrid DFT QM and SAPT QM (Supporting Information Figures S23-S24 and S32). MM shortcomings for some non-covalent interactions are well-established[32, 99-100], including through recent curation of residue pairs for broad analysis of method accuracy in characterizing non-covalent interactions[5]. Beyond these efforts, we briefly describe analysis of discrepancies between QM and MM pairwise interactions for unusually short non-covalent distances in close contacts that could help guide future force field development (Supporting Information Figures S33-S37). Although magnitudes of interaction energies are smaller for neutral residues and thus

in closer numerical agreement, MM interactions are generally more favorable than QM in hydrophobic-hydrophobic interactions (e.g., Leu-Val), whereas QM models form more favorable interactions with polar hydrogen bonding partners (e.g., His-Tyr/Ser/Thr) (Supporting Information Figures S33, S36, and S37). Across all CCs, including charged interactions, QM gas phase interaction energies are expectedly more favorable, with much stronger interactions involving carboxylates in QM simulations (Supporting Information Figures S34-S35). Overall, MM simulations of CCs can predict repulsive interactions that are unexpectedly favorable at the QM level, highlighting the limitations of MM force fields for characterizing these interactions (Supporting Information Figure S32).

QM simulation suggests that favorable pairwise interactions are being formed to stabilize the short distances observed in the majority of residue pairs in our large data set. Some of these interactions are anticipated to be derived from well-understood non-covalent interactions, such as salt bridges. We next characterize the specific interactions that underlie some of the most surprising and least well understood CCs observed in the overall **FD** and **HR** subsets: i) hydrophobic-negatively charged amino acid interactions, ii) interactions with ambifunctional hydrogen bonding polar amino acids, and iii) CCs that involve sulfur functional groups.

5a. Hydrophobic-Charged Interactions.

In the informatics analysis, relatively few close contacts were observed involving bulky hydrophobic amino acids, but based on the holistic QM analysis, we can anticipate these interactions, where present, to potentially be favorable if interacting with a charged residue. Thus, we examined the potential for charge-assisted interactions, including hydrogen bonds, between the bulky, hydrophobic Val, Ile, and Leu residues with negatively charged Asp and Glu residues. In the **FD** set, a small fraction (1% or 758) of CCs come from this combination of

residues, and a comparable fraction (1% or 26) is found in the **HR** subset (Supporting Information Table S5). We further focus on interactions between the sidechain carbon atoms of the hydrophobic residue with the carboxylate oxygen of Asp or Glu, which corresponds to approximately $1/10^{th}$ of those cases: 92 are present in the **FD** and 10 in the **HR** subset (Supporting Information Table S5). The QM SAPT interaction energies are favorable for half of the 14 CCs (10 **HR** plus 4 **FD**-only) in equal measure in the **HR** and **FD** sets without any apparent bias toward high-resolution structures (Supporting Information Table S6).

Comparing favorable and unfavorable cases reveals that all residues participate in both types of interactions. Broadly 2.6-2.7 Å CC distances are observed in the favorable cases, whereas all 2.2-2.4 Å CCs are repulsive for this class of CCs (Supporting Information Table S6). Although these carboxylate-hydrophobic interactions could be anticipated to consist of C-H···O hydrogen bonds, not all sidechains are oriented for productive HB angles (Figure 5). The most favorable interaction observed is a -9.8 kcal/mol stabilization between Ile and Asp (PDB ID: 3G7G) with a close contact separation of around 2.6 Å (Figure 5). In this case, a single productive HB with the most obtuse C-H···O angle in the set of 150° is observed (Figure 5 and Supporting Information Table S6). In other weakly favorable cases, less obtuse angles of around 100-130° are observed, often involving multiple O···H interactions (Figure 5 and Supporting Information Table S6). The unfavorable interactions may be grouped into two categories: i) either small C-H···O angles of 90-105° with moderate 2.4-2.6 Å C···O separation or ii) 140-150° C-H···O angles but with smaller 2.20 Å C···O separations (Figure 5).

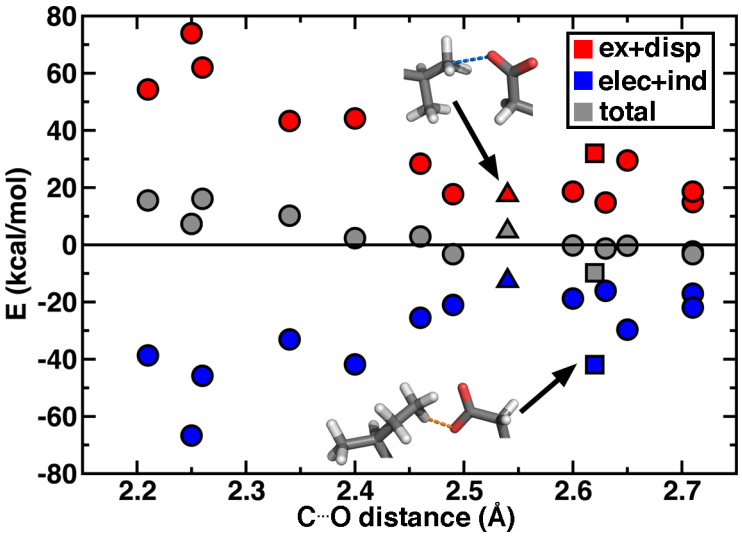


Figure 5. SAPT interaction energies (in kcal/mol) for 14 CCs between Val, Ile, or Leu and Asp or Glu as a function of C···O close contact distance (in Å). The total interaction energy is shown in gray along with the combined exchange repulsion and dispersion terms (ex+disp, in red) and electrostatic and induction terms (elec+ind, in blue). One case (square symbols) with a C-H···O hydrogen bond that stabilizes the pair is shown in inset with the H···O distance annotated by an orange dashed line, and one case is shown in inset with the methyl hydrogen atoms oriented away from the carboxylate oxygen with the C···O distance annotated by a blue dashed line (triangle symbols).

In all favorable cases, induction, electrostatic, and dispersion terms contribute similarly to offset of unfavorable exchange repulsion (Figure 5). The same is true of unfavorable cases, but here the short CC distance leads to very large exchange repulsion energies (Supporting Information Table S6). Evaluation of MM interaction energies for this same subset reveals all cases to be repulsive, likely owing to the lack of induction-like terms in the force field (Supporting Information Table S6). These observations are consistent with our earlier study on the effect of charge transfer between an active site Val and surrounding charged residues in the enzyme catechol *O*-methyltransferase[52, 101]. This class of interactions appears to be rare in protein crystal structures, and QM characterization is likely essential for understanding their role in protein structure and function.

5b. Ambifunctional Hydrogen Bonds.

We previously noted that tyrosine formed a disproportionate number of close contacts in our curated crystal structure data set, both in terms of its frequency relative to its modest abundance in the PDB and in absolute terms, even in very high resolution structures (see Figure 3 and Supporting Information Figures S9 and S20). The occurrence in CCs appears attributable to the hydroxyl functional group, as Tyr CC trends are more comparable to hydroxyl-containing Ser and Thr than the Phe and Trp aromatic residues. Nevertheless, the aromatic character of Tyr has the potential to make the hydroxyl oxygen more acidic and introduce effects of dispersion interactions, setting it apart from both the polar and aromatic residues. Now, we consider in more detail the involvement of the Tyr hydroxyl in stabilizing close contacts with polar residues. Specifically, the greatest number of neutral-neutral CCs for Tyr occurs with Asn and Gln: roughly 1.5% of all CCs are between these residues for a total of 1,179 in the **FD** set and 32 in the **HR** subset (Supporting Information Table S7).

Both Tyr and Asn/Gln can act as hydrogen bond donors (HBDs) and acceptors (HBAs) through changes in orientation of the hydroxyl for Tyr or through participation of the amide nitrogen vs. carbonyl in the case of Asn/Gln. After excluding a small number of CCs that involve main chain atoms, the majority of shortest distances observed in CC residue pairs (609 of 854 in **FD**, 16 of 24 in **HR**) are formed between the Tyr O and Asn/Gln O. Nevertheless, this leaves a significant number (245 of 854 in **FD**, 8 of 24 in **HR**) in which Asn/Gln N is the atom closest to the Tyr O(Supporting Information Table S7). In all cases, determination of whether these amino acid functional groups are acting as HBDs or HBAs to stabilize the interaction mandates QM characterization and geometric analysis. Over all 24 **HR** subset and 82 cases from the **FD** set, the Tyr Asn/Gln CCs are favorable at the SAPT0 level of theory in roughly 50% of the time regardless of O/O or O/N character of the CCs (Supporting Information Table S7).

Evaluation of hydrogen bond angles as well as both hydrogen atom and CC distances confirms the importance of HBs for the favorable CC cases (Figure 6 and Supporting Information Table S8). No CC shorter than 2.4 Å is favorable for either the O-type or N-type CCs, but this appears much more sensitive to the distance of the hydrogen atom and its orientation in a productive (i.e., > 120°) hydrogen bond angle (Figure 6). The strongest CCs (ca. -10 to -13 kcal/mol) are those in which Tyr OH acts as a HBD to the carbonyl O of Asn/Gln, far in excess of the -6.4 kcal/mol stabilization observed in the strongest N-type CC with Tyr O acting as an HBA to the amide N of Asn/Gln (Figure 6 and Supporting Information Table S8). This enhanced strength of the O-type CC appears to be due in part to the greater geometric flexibility in this interaction: the most favorable O-type CCs have average HB angles of 166°, which is roughly 17° higher than the average observed in the favorable N-type CCs (Supporting Information Table S9). For O-type CCs with comparably acute angles to the N-type CCs, the CC strength favorability becomes comparable (Figure 6).

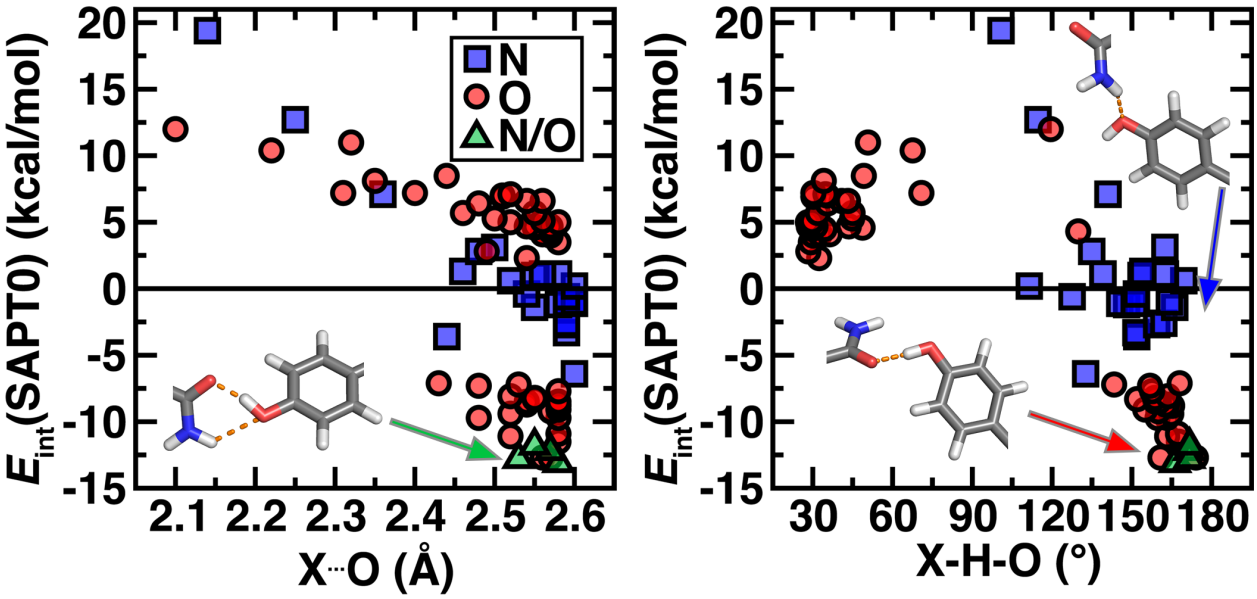


Figure 6. Plots of SAPT interaction energies (E_{int}(SAPT0)) in kcal/mol versus X···O CC distances in Å (left) or X-H-O HB angle in ° (right) for Tyr CCs with Asn or Gln. Cases are distinguished by whether the closest atom, X, is N (blue translucent squares), O (red circles), or both are proximal (N/O, green triangles). Examples of each case are shown in inset with the

X···H distance indicated by orange dashed lines and indicated with an arrow corresponding to the class of interactions the structure represents.

Returning to the characteristics that lead to the most stable O-type CCs between Tyr and Asn/Gln reveals that four of the lowest energy CCs, including the most stable case (PDB: 2W5Q, -13.1 kcal/mol), involve participation of both functional groups of the Asn/Gln in interactions with Tyr (green triangles in Figure 6). In these structures, the Tyr O-H is oriented toward the carbonyl O of Asn/Gln but the amide N-H can also form a slightly elongated HB to the Tyr O of around 2.4 Å and more acute angle of around 120-130° (Supporting Information Table S9). Comparing average energetic terms of the three categories of favorable CCs reveals that the second interaction enhances average electrostatic and induction contributions to the overall interaction over the simple O-type CCs (Supporting Information Table S9). In all cases, the order of magnitude of contributions in the SAPT energies are electrostatics > induction > exchange, rather than the more comparable magnitudes observed for charged-hydrophobic interactions in Sec. 5a. Although MM force fields predict both the double HB and standard O-type interactions to be favorable, the distinction between the two orientations is underestimated, consistent with our recent observations of more favorable double HB configurations in QM over MM simulation[102]. The weaker N-type CC energetics arise due to reduced electrostatic and induction stabilization with comparable exchange repulsion penalties (Supporting Information Table S9).

Most repulsive cases cluster at acute, unproductive hydrogen bond angles. It is possible that further sampling of hydrogen initial positions could lead to more stable CC energetics. However, we also note that there are also many-body (i.e., cooperative[103]) effects neglected in our current analysis that are known to be important in hydrogen bonding with Tyr[104]. For example, two repulsive N-type CCs from the HR subset come from the same crystal structure

(PDB ID: 4RJ2) in which Tyr27 interacts simultaneously with Gln218 and Asn222. In the present analysis, both interactions are weakly unfavorable (+1.1 kcal/mol), but the study of more residues in the environment of the greater protein could better explain the stabilizing effect of these interactions in the future (Supporting Information Table S9).

5c. Close Contacts with Sulfur.

Sulfur is the largest heavy atom in standard amino acids, with a 1.8 Å van der Waals radius, meaning that our definition of a S-containing CC is the largest at 2.82-3.06 Å (Supporting Information Table S1). This distance is still dramatically shorter than the 4.3 Å heavy atom cutoff used in a recent survey of hydrogen bonding interactions with Cys residues[24] or the average heavy atom distances observed around 3.5 Å in another survey[25]. Met and Cys are the only amino acids that contain sulfur in their sidechains. The **HR** subset contains 65 CCs with least Met or Cys, only 36 of these correspond to Met or Cys SC participation, 17 of which are CCs formed with the sidechain S atom (Supporting Information Table S10). This low frequency CC participation by S in the **HR** set is comparable in the larger **FD** set where 423 of 2,086 Met/Cys CCs involve the sidechain sulfur atom (Supporting Information Table S10). We have characterized with QM and MM 17 **HR** (55 **FD**) Cys/Met-containing CCs in which the sulfur atom forms the shortest non-bonded to a neighboring residue (Supporting Information Table S10). Very few of these interactions are favorable with QM: 1 in 17 the **HR** subset (3 of 55 for **FD**), with little apparent dependence on the CC distance (Supporting Information Table S11).

Of the three favorable cases, only one (PDB: 2WPG) corresponds to Cys stabilization (-1.9 kcal/mol) by a charged residue, which in this case is due to the Lys NH_3^+ sidechain (Figure 7). In the other two cases, polar Thr (-0.3 kcal/mol) or Gln (-4.8 kcal/mol) residues form comparably favorable interactions with Cys through their carbon or oxygen atoms (Figure 7).

Only in the Cys-Lys case is an obtuse, 142° N-H···S hydrogen bond evident with Cys acting as an HBA, and in none of the favorable cases is Cys an HBD (Figure 7). None of the CCs involving Met form favorable interactions, possibly due to the extra bulk of the thioether group (Figure 7 and Supporting Information Table S11). The lowest energy Met interaction is also with a Thr oxygen atom (+0.5 kcal/mol) but with much higher exchange repulsion (Figure 7). In both favorable and unfavorable sulfur close contacts, relatively weak electrostatic stabilization in comparison to the O···N/O interactions described in Sec. 5b is offset by comparable exchange repulsion (Supporting Information Table S11). Overall, these observations suggests that i) sulfur may not form close contacts as readily as lighter elements, ii) sulfur is not abundant enough in **HR** protein structures for examination of non-covalent interactions, or iii) a larger van der Waals cutoff should be used to examine favorable sulfur-containing non-covalent interactions in future work.

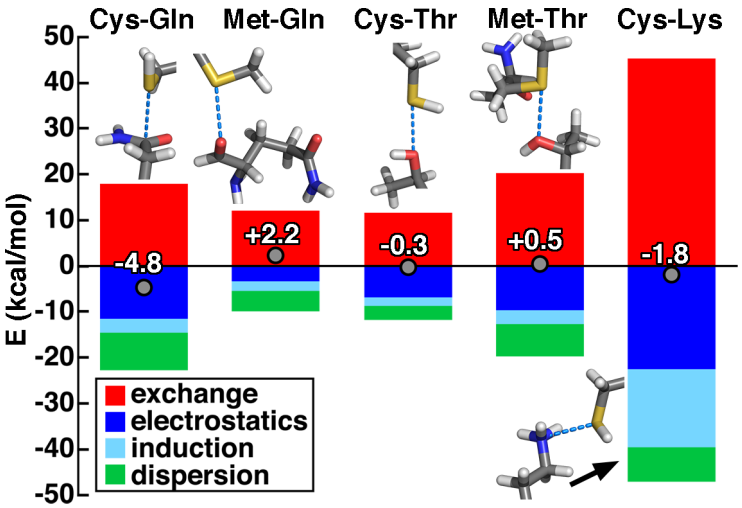


Figure 7. SAPT interaction energies (in kcal/mol) for (left to right) Cys-Gln, Met-Gln, Cys-Thr, Met-Thr, and Cys-Lys CC pairs. The total interaction energy (in kcal/mol) is shown as a gray circle, and the components are shown as a stacked bar graph: exchange repulsion (in red), electrostatics (in dark blue), induction (in light blue), and dispersion (in green). The close contact in each case is indicated by a light blue dashed line in the inset structures.

6. Conclusions

We have investigated the existence of unusually short non-covalent distances in atomically resolved and ultra-high resolution (i.e., 1.2 Å or better) X-ray crystal structures of proteins. We first defined the scope of our search in terms of non-covalent distances that were shorter than 85% of the sum of the van der Waals radii of the substituent heavy atoms, excluding standard hydrogen bonds. Over a survey of over 13,000 high quality protein structures in the PDB (i.e., the **FD** set) and a very high-resolution > 1,000 protein **HR** subset, we observed thousands of such interactions to occur. Quantifying the relative frequency of these close contacts with respect to an individual amino acid's abundance revealed the frequent occurrence of Tyr in such interactions. In addition to expected interactions, such as short HBs between oppositely charged residues, nearly every amino acid was observed to form at least one close contact with another residue.

We extracted these close contact pairs of amino acids in order to characterize them with quantum-mechanical symmetry adapted perturbation theory that enables decomposition of interaction energies into terms arising from sterics (i.e., exchange repulsion and dispersion) as well as electronic properties (i.e., electrostatic attraction and induction). This energetic analysis revealed that the majority of close contacts extracted from crystal structures were favorable due to pairwise interactions between the relevant amino acids. Searching for the shortest favorable close contacts revealed exceptionally short (ca. 2.2 Å) heavy atom distances could be favorable, although restricting this analysis to neutral-neutral amino acid interactions generally indicated that close contacts with 2.4-2.8 Å separations were most favorable. Evaluation of the same close contacts with MM force fields instead yielded repulsive interaction energies. Evaluation of the limited number of sulfur-containing close contacts revealed few to be favorable, owing to a diminished contribution from electrostatic stabilization and enhanced exchange repulsion.

Analysis of interactions between hydrophobic residues, such as Val, Ile, and Leu, and carboxylates from Asp or Glu, revealed the presence of C-H···O hydrogen bonds, and identified the relative importance of electrostatic and induction terms that stabilized these interactions. The relative strength of hydrogen bonds in which a Tyr hydroxyl acted as a hydrogen bond donor or acceptor to Asn or Gln sidechains was examined, and it was revealed the most stable interactions occurred when both Tyr and Asn/Gln acted simultaneously as both HB acceptors and donors. These observations motivate next steps beyond the pairwise analysis of close contacts to also investigate patterns in cooperative hydrogen bonds in protein active sites. Future work will focus on identifying the role that these close contacts play in local protein stability or in affecting substrate positioning in enzyme active sites.

ASSOCIATED CONTENT

Supporting Information. Protein set curation details; resolution of proteins in data set; *R*-value R_{free}, and *RSRZ* characteristics of data sets; geometric criteria of data sets including clashscore and Ramachandran/rotamer outliers; residue abundance in HR and FD sets relative to PDB and colored by secondary structure; distribution of protein size in HR vs. FD data sets; close contact distance cutoffs; close contact statistics in HR and FD subsets including by resolution; frequency of all and unique close contacts in the FD and HR sets; CCs by atom type or secondary structure element in FD set; matrix of HR set residue-residue sidechain-sidechain close contacts; comparison of SAPT2 vs. SAPT0, B3LYP vs. SAPT0, and MM vs. SAPT0 interaction energies for crystal heavy atom structures with optimized H atoms; overall interaction energy statistics for CCs evaluated with SAPT0 and MM as well as divisions by CC type; and SAPT2 interaction energies; comparison of B3LYP and SAPT0 interaction energies; distribution of interaction energy components in SAPT0 calculations; rearrangement of CC distances and RMSD with

sidechain optimization; CC distance distributions for favorable/unfavorable interaction energies for all cases and the neutral subset; decision tree analysis of the determinants for interactions being more favorable with QM vs. MM; charged-hydrophobic CC summary statistics, energetics, and geometric properties; ambifunctional Tyr-Asn/Gln CC summary statistics, energetics, and geometric properties; sulfur CC summary statistics, energetics, and geometric properties. (PDF)

Characteristics of all CCs obtained from the PDB; coordinates of 6,279 crystal, H-optimized and sidechain-optimized CCs; lists of all interaction energies from SAPT0 and MM of 6,279 characterized CCs. (ZIP)

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*email: hjkulik@mit.edu phone: 617-253-4584

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported in part by an NEC Corporation Grant from the MIT Research Support Committee. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, which supported the work. H.W.Q. was supported in part by a Department of Energy Computational Science Graduate Fellowship (DOE-CSGF). Support for this research was provided by a core center grant P30-ES002109 from the National Institute of Environmental

Health Sciences, National Institutes of Health. The authors acknowledge Adam H. Steeves for providing a critical reading of the manuscript.

References

1. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D., Optimization of the Additive Charmm All-Atom Protein Force Field Targeting Improved Sampling of the Backbone Φ, Ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012,** *8*, 3257-3273.
2. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved Side-Chain Torsion Potentials for the Amber Ff99sb Protein Force Field. *Proteins: Struct., Funct., Bioinf.* **2010,** *78*, 1950-1958.
3. Wickstrom, L.; Okur, A.; Simmerling, C., Evaluating the Performance of the Ff99sb Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009,** *97*, 853-856.
4. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., Ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J. Chem. Theory Comput.* **2015,** *11*, 3696-3713.
5. Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D., The Biofragment Database (Bfdb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017,** *147*, 161727.
6. Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J., Ab Initio Quantum Chemistry for Protein Structure. *J. Phys. Chem. B* **2012,** *116*, 12501-12509.
7. Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P., Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010,** *110*, 5023-5063.
8. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein Structure Prediction Using Rosetta. In *Methods in Enzymology*, Elsevier: 2004; Vol. 383, pp 66-93.
9. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.
10. Bauzá, A.; Frontera, A., Rch3···O Interactions in Biological Systems: Are They Trifurcated H-Bonds or Noncovalent Carbon Bonds? *Crystals* **2016,** *6*, 26.
11. Mundlapati, V. R.; Sahoo, D. K.; Bhaumik, S.; Jena, S.; Chandrakar, A.; Biswal, H. S., Noncovalent Carbon-Bonding Interactions in Proteins. *Angew. Chem., Int. Ed.* **2018,** *0*.
12. Newberry, R. W.; Raines, R. T., The N→π* Interaction. *Acc. Chem. Res.* **2017,** *50*, 1838-1846.
13. Bartlett, G. J.; Woolfson, D. N., On the Satisfaction of Backbone-Carbonyl Lone Pairs of Electrons in Protein Structures. *Protein Sci.* **2016,** *25*, 887-897.
14. An, Y.; Bloom, J. W.; Wheeler, S. E., Quantifying the π-Stacking Interactions in Nitroarene Binding Sites of Proteins. *J. Phys. Chem. B* **2015,** *119*, 14441-14450.
15. Hudson, K. L.; Bartlett, G. J.; Diehl, R. C.; Agirre, J.; Gallagher, T.; Kiessling, L. L.; Woolfson, D. N., Carbohydrate–Aromatic Interactions in Proteins. *J. Am. Chem. Soc.* **2015,** *137*, 15152-15160.

16. Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S., Cation–π Interactions in Protein–Ligand Binding: Theory and Data-Mining Reveal Different Roles for Lysine and Arginine. *Chem. Sci.* **2018,** *9*, 2655-2665.
17. Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R., The Good, the Bad and the Twisted: A Survey of Ligand Geometry in Protein Crystal Structures. *J. Comput.-Aided Mol. Des.* **2012,** *26*, 169-183.
18. Salonen, L. M.; Ellermann, M.; Diederich, F., Aromatic Rings in Chemical and Biological Recognition: Energetics and Structures. *Angew. Chem., Int. Ed.* **2011,** *50*, 4808-4842.
19. Gallivan, J. P.; Dougherty, D. A., Cation-π Interactions in Structural Biology. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** *96*, 9459-9464.
20. Kurczab, R.; Śliwa, P.; Rataj, K.; Kafel, R.; Bojarski, A. J., Salt Bridge in Ligand–Protein Complexes—Systematic Theoretical and Statistical Investigations. *J. Chem. Inf. Model.* **2018,** *58*, 2224-2238.
21. Iyer, A. H.; Krishna Deepak, R. N. V.; Sankararamakrishnan, R., Imidazole Nitrogens of Two Histidine Residues Participating in N–H···N Hydrogen Bonds in Protein Structures: Structural Bioinformatics Approach Combined with Quantum Chemical Calculations. *J. Phys. Chem. B* **2018,** *122*, 1205-1212.
22. Holcomb, M.; Adhikary, R.; Zimmermann, J.; Romesberg, F. E., Topological Evidence of Previously Overlooked Ni+1–H···Ni H-Bonds and Their Contribution to Protein Structure and Stability. *J. Phys. Chem. A* **2018,** *122*, 446-450.
23. Deepak, R. N. V. K.; Sankararamakrishnan, R., Unconventional N-H…N Hydrogen Bonds Involving Proline Backbone Nitrogen in Protein Structures. *Biophys. J.* **2016,** *110*, 1967-1979.
24. Mazmanian, K.; Sargsyan, K.; Grauffel, C.; Dudev, T.; Lim, C., Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions. *J. Phys. Chem. B* **2016,** *120*, 10288-10296.
25. Zhou, P.; Tian, F.; Lv, F.; Shang, Z., Geometric Characteristics of Hydrogen Bonds Involving Sulfur Atoms in Proteins. *Proteins: Struct., Funct., Bioinf.* **2009,** *76*, 151-163.
26. Mundlapati, V. R.; Gautam, S.; Sahoo, D. K.; Ghosh, A.; Biswal, H. S., Thioamide, a Hydrogen Bond Acceptor in Proteins and Nucleic Acids. *J. Phys. Chem. Lett.* **2017,** *8*, 4573-4579.
27. Nishio, M.; Umezawa, Y.; Fantini, J.; Weiss, M. S.; Chakrabarti, P., CH–π Hydrogen Bonds in Biological Macromolecules. *Phys. Chem. Chem. Phys.* **2014,** *16*, 12648-12683.
28. Steiner, T.; Koellner, G., Hydrogen Bonds with π-Acceptors in Proteins: Frequencies and Role in Stabilizing Local 3D Structures11edited by R. Huber. *J. Mol. Biol.* **2001,** *305*, 535-557.
29. Yesselman, J. D.; Horowitz, S.; Brooks, C. L.; Trievel, R. C., Frequent Side Chain Methyl Carbon-Oxygen Hydrogen Bonding in Proteins Revealed by Computational and Stereochemical Analysis of Neutron Structures. *Proteins: Struct., Funct., Bioinf.* **2014,** *83*, 403-410.
30. Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S., Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* **2012,** *8*, 1409-1414.
31. Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H., Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015,** *11*, 5513-5524.

32. Jiang, Z.; Biczysko, M.; Moriarty, N. W., Accurate Geometries for "Mountain Pass" Regions of the Ramachandran Plot Using Quantum Chemical Calculations. *Proteins: Struct., Funct., Bioinf.* **2018,** *86*, 273-278.
33. Bonifácio, M. J.; Archer, M.; Rodrigues, M. L.; Matias, P. M.; Learmonth, D. A.; Carrondo, M. A.; Soares-da-Silva, P. c., Kinetics and Crystal Structure of Catechol-O-Methyltransferase Complex with Co-Substrate and a Novel Inhibitor with Potential Therapeutic Application. *Mol. Pharmacol.* **2002,** *62*, 795-805.
34. Vidgren, J.; Svensson, L. A.; Liljas, A., Crystal Structure of Catechol O-Methyltransferase. *Nature* **1994,** *368*, 354-358.
35. Palma, P. N.; Rodrigues, M. L.; Archer, M.; Bonifácio, M. J.; Loureiro, A. I.; Learmonth, D. A.; Carrondo, M. A.; Soares-da-Silva, P., Comparative Study of Ortho- and Meta-Nitrated Inhibitors of Catechol-O-Methyltransferase: Interactions with the Active Site and Regioselectivity of O-Methylation. *Mol. Pharmacol.* **2006,** *70*, 143-153.
36. Tsuji, E.; Okazaki, K.; Takeda, K., Crystal Structures of Rat Catechol-O-Methyltransferase Complexed with Coumarine-Based Inhibitor. *Biochem. Biophys. Res. Commun.* **2009,** *378*, 494-497.
37. Rutherford, K.; Le Trong, I.; Stenkamp, R. E.; Parson, W. W., Crystal Structures of Human 108v and 108m Catechol O-Methyltransferase. *J. Mol. Biol.* **2008,** *380*, 120-30.
38. Ellermann, M.; Lerner, C.; Burgy, G.; Ehler, A.; Bissantz, C.; Jakob-Roetne, R.; Paulini, R.; Allemann, O.; Tissot, H.; Grunstein, D.; Stihle, M.; Diederich, F.; Rudolph, M. G., Catechol-O-Methyltransferase in Complex with Substituted 3'-Deoxyribose Bisubstrate Inhibitors. *Acta Crystallogr., Sect. D* **2012,** *68*, 253-260.
39. Harrison, S. T.; Poslusney, M. S.; Mulhearn, J. J.; Zhao, Z.; Kett, N. R.; Schubert, J. W.; Melamed, J. Y.; Allison, T. J.; Patel, S. B.; Sanders, J. M.; Sharma, S.; Smith, R. F.; Hall, D. L.; Robinson, R. G.; Sachs, N. A.; Hutson, P. H.; Wolkenberg, S. E.; Barrow, J. C., Synthesis and Evaluation of Heterocyclic Catechol Mimics as Inhibitors of Catechol-O-Methyltransferase (COMT). *ACS Medicinal Chemistry Letters* **2015,** *6*, 318-323.
40. Patra, N.; Ioannidis, E. I.; Kulik, H. J., Computational Investigation of the Interplay of Substrate Positioning and Reactivity in Catechol O-Methyltransferase. *PLoS ONE* **2016,** *11*, e0161868.
41. Lau, E. Y.; Bruice, T. C., Importance of Correlated Motions in Forming Highly Reactive near Attack Conformations in Catechol O-Methyltransferase. *J. Am. Chem. Soc.* **1998,** *120*, 12387-12394.
42. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **2008,** *4*, 222-231.
43. Ufimtsev, I. S.; Martinez, T. J., Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *J. Chem. Theory Comput.* **2009,** *5*, 1004-1015.
44. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009,** *5*, 2619-2628.
45. Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J., Excited-State Electronic Structure with Configuration Interaction Singles and Tamm-Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2011,** *7*, 1814-1823.
46. Ufimtsev, I. S.; Luehr, N.; Martínez, T. J., Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011,** *2*, 1789-1793.

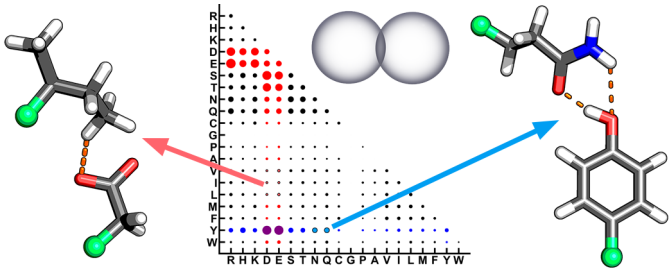
47. Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S., Linear-Scaling Methods in Quantum Chemistry. *Rev. Comput. Chem.* **2007,** *23*, 1.
48. Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R., Auxiliary Basis Sets for Main Row Atoms and Transition Metals and Their Use to Approximate Coulomb Potentials. *Theor. Chem. Acc.* **1997,** *97*, 119-124.
49. Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R., Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995,** *240*, 283-290.
50. Karelina, M.; Kulik, H. J., Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **2017,** *13*, 563-576.
51. Zhang, J.; Kulik, H. J.; Martinez, T. J.; Klinman, J. P., Mediation of Donor–Acceptor Distance in an Enzymatic Methyl Transfer Reaction. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112*, 7954-7959.
52. Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J., How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016,** *120*, 11381-11394.
53. Li, P.; Soudackov, A. V.; Hammes-Schiffer, S., Fundamental Insights into Proton-Coupled Electron Transfer in Soybean Lipoxygenase from Quantum Mechanical/Molecular Mechanical Free Energy Simulations. *J. Am. Chem. Soc.* **2018,** *140*, 3068-3076.
54. Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M., Protein Crystallography for Non-Crystallographers, or How to Get the Best (but Not More) from Published Macromolecular Structures. *FEBS J.* **2008,** *275*, 1-21.
55. Brunger, A. T., Free R Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures. *Nature* **1992,** *355*, 472-475.
56. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010,** *66*, 12-21.
57. Bell, J. A.; Ho, K. L.; Farid, R., Significant Reduction in Errors Associated with Nonbonded Contacts in Protein Crystal Structures: Automated All-Atom Refinement with Primex. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012,** *68*, 935-952.
58. Who Checks the Checkers? Four Validation Tools Applied to Eight Atomic Resolution Structures11edited by I. A. Wilson. *J. Mol. Biol.* **1998,** *276*, 417-436.
59. Bondi, A., Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964,** *68*, 441-451.
60. Rowland, R. S.; Taylor, R., Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from Van Der Waals Radii. *J. Phys. Chem.* **1996,** *100*, 7384-7391.
61. McConkey, B. J.; Sobolev, V.; Edelman, M., Discrimination of Native Protein Structures Using Atom–Atom Contact Scoring. *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100*, 3215-3220.
62. Odutola, J.; Dyke, T., Partially Deuterated Water Dimers: Microwave Spectra and Structure. *J. Chem. Phys.* **1980,** *72*, 5062-5070.
63. Cleland, W.; Kreevoy, M., Low-Barrier Hydrogen Bonds and Enzymic Catalysis. *Science* **1994,** *264*, 1887-1890.
64. Frey, P.; Whitt, S.; Tobin, J., A Low-Barrier Hydrogen Bond in the Catalytic Triad of Serine Proteases. *Science* **1994,** *264*, 1927-1930.
65. Perrin, C. L.; Nielson, J. B., "Strong" Hydrogen Bonds in Chemistry and Biology. *Annu. Rev. Phys. Chem.* **1997,** *48*, 511-544.

66. Wang, L.; Fried, S. D.; Boxer, S. G.; Markland, T. E., Quantum Delocalization of Protons in the Hydrogen-Bond Network of an Enzyme Active Site. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111*, 18454-18459.
67. Ishikita, H.; Saito, K., Proton Transfer Reactions and Hydrogen-Bond Networks in Protein Environments. *J. R. Soc., Interface* **2014,** *11*.
68. Desiraju, G. R., A Bond by Any Other Name. *Angew. Chem., Int. Ed.* **2011,** *50*, 52-59.
69. Gilli, P.; Pretto, L.; Bertolasi, V.; Gilli, G., Predicting Hydrogen-Bond Strengths from Acid−Base Molecular Properties. The pKa Slide Rule: Toward the Solution of a Long-Lasting Problem. *Acc. Chem. Res.* **2009,** *42*, 33-44.
70. Gilli, P.; Gilli, G., Hydrogen Bond Models and Theories: The Dual Hydrogen Bond Model and Its Consequences. *J. Mol. Struct.* **2010,** *972*, 2-10.
71. Gilli, P.; Pretto, L.; Gilli, G., Pa/pKa Equalization and the Prediction of the Hydrogen-Bond Strength: A Synergism of Classical Thermodynamics and Structural Crystallography. *J. Mol. Struct.* **2007,** *844*, 328-339.
72. Ilari, A.; Savino, C., Protein Structure Determination by X-Ray Crystallography. In *Bioinformatics: Data, Sequence Analysis and Evolution*, Keith, J. M., Ed. Humana Press: Totowa, NJ, 2008; pp 63-87.
73. Yang, H.; Peisach, E.; Westbrook, J. D.; Young, J.; Berman, H. M.; Burley, S. K., Dcc: A Swiss Army Knife for Structure Factor Analysis and Validation. *J. Appl. Crystallogr.* **2016,** *49*, 1081-1084.
74. Read, Randy J.; Adams, Paul D.; Arendall, W. B.; Brunger, Axel T.; Emsley, P.; Joosten, Robbie P.; Kleywegt, Gerard J.; Krissinel, Eugene B.; Lütteke, T.; Otwinowski, Z.; Perrakis, A.; Richardson, Jane S.; Sheffler, William H.; Smith, Janet L.; Tickle, Ian J.; Vriend, G.; Zwart, Peter H., A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **2011,** *19*, 1395-1412.
75. PDBMetrics Http://Www.Cbi.Cnptia.Embrapa.Br/Sms/Pdb_Metrics/Frequency.Html. (accessed 11/02/17).
76. Fileto, R.; Kuser, P. R.; Yamagishi, M. E.; Ribeiro, A. A.; Quinalia, T. G.; Franco, E. H.; Mancini, A. L.; Higa, R. H.; Oliveira, S. R.; Santos, E. H.; Vieira, F. D.; Mazoni, I.; Cruz, S. A.; Neshich, G., Pdb-Metrics: A Web Tool for Exploring the Pdb Contents. *GMR, Genet. Mol. Res.* **2006,** *5*, 333-41.
77. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V., Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963,** *7*, 95-99.
78. Marqusee, S.; Baldwin, R. L., Helix Stabilization by Glu-... Lys+ Salt Bridges in Short Peptides of De Novo Design. *Proc. Natl. Acad. Sci. U. S. A.* **1987,** *84*, 8898-8902.
79. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., Pdb2pqr: An Automated Pipeline for the Setup of Poisson–Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004,** *32*, W665-W667.
80. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman Amber 2015, University of California, San Francisco. 2015.
81. Schrodinger, L. L. C., The PyMOL Molecular Graphics System, Version 1.7.4.3. 2010.
82. Petachem. http://www.petachem.com. (accessed 10/31/18).

83. Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P., DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations. *J. Phys. Chem. A* **2009,** *113*, 11856-11865.
84. Klamt, A.; Schuurmann, G., Cosmo: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993,** *2*, 799-805.
85. Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J., Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.* **2015,** *11*, 3131-3144.
86. Liu, F.; Sanchez, D. M.; Kulik, H. J.; Martinez, T. J., Exploiting Graphical Processing Units to Enable Quantum Chemistry Calculation of Large Solvated Molecules with Conductor-Like Polarizable Continuum Models. *Int. J. Quantum Chem.* **2019,** *119*, e25760.
87. Becke, A. D., Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.
88. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988,** *37*, 785--789.
89. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994,** *98*, 11623-11627.
90. Harihara, P. C.; Pople, J. A., Influence of Polarization Functions on Molecular-Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973,** *28*, 213-222.
91. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010,** *132*, 154104.
92. Jeziorski, B.; Moszynski, R.; Szalewicz, K., Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of Van Der Waals Complexes. *Chem. Rev.* **1994,** *94*, 1887-1930.
93. Hohenstein, E. G.; Sherrill, C. D., Density Fitting and Cholesky Decomposition Approximations in Symmetry-Adapted Perturbation Theory: Implementation and Application to Probe the Nature of π-π Interactions in Linear Acenes. *J. Chem. Phys.* **2010,** *132*, 184111.
94. Hohenstein, E. G.; Parrish, R. M.; Sherrill, C. D.; Turney, J. M.; Schaefer, H. F., Large-Scale Symmetry-Adapted Perturbation Theory Computations via Density Fitting and Laplace Transformation Techniques: Investigating the Fundamental Forces of DNA-Intercalator Interactions. *J. Chem. Phys.* **2011,** *135*, 174107.
95. Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D., Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**.
96. Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E., Mmpbsa.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012,** *8*, 3314-3321.
97. Tsui, V.; Case, D. A., Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations. *Biopolymers* **2000,** *56*, 275-291.

98. Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrášek, J., Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins. *J. Chem. Theory Comput.* **2010,** *6*, 2191-2203.
99. Paton, R. S.; Goodman, J. M., Hydrogen Bonding and π-Stacking: How Reliable Are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *J. Chem. Inf. Model.* **2009,** *49*, 944-955.
100. Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J., Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated Ab Initio Quantum Chemical and Empirical Potential Procedures. *J. Chem. Theory Comput.* **2009,** *5*, 982-992.
101. Qi, H. W.; Karelina, M.; Kulik, H. J., Quantifying Electronic Effects in QM and QM/MM Biomolecular Modeling with the Fukui Function. *Acta Phys.-Chim. Sin.* **2018,** *34*, 81-91.
102. Mehmood, R.; Qi, H. W.; Steeves, A. H.; Kulik, H. J., The Protein's Role in Substrate Positioning and Reactivity for Biosynthetic Enzyme Complexes: The Case of SyrB2/Syrb1. *chemrxiv: 10.26434/chemrxiv.7234058.v1* **2018**.
103. Li, J.; Wang, Y.; Chen, J.; Liu, Z.; Bax, A.; Yao, L., Observation of A-Helical Hydrogen-Bond Cooperativity in an Intact Protein. *J. Am. Chem. Soc.* **2016,** *138*, 1824-1827.
104. Pinney, M. M.; Natarajan, A.; Yabukarski, F.; Sanchez, D. M.; Liu, F.; Liang, R.; Doukov, T.; Schwans, J. P.; Martinez, T. J.; Herschlag, D., Structural Coupling Throughout the Active Site Hydrogen Bond Networks of Ketosteroid Isomerase and Photoactive Yellow Protein. *J. Am. Chem. Soc.* **2018,** *140*, 9827-9843.

for Table of Contents use only



"Evaluating Unexpectedly Short Non-covalent Distances in X-ray Crystal Structures of Proteins with Electronic Structure Analysis" by Helena W. Qi and Heather J. Kulik

Supporting Information for

Evaluating Unexpectedly Short Non-covalent Distances in X-ray Crystal Structures of Proteins with Electronic Structure Analysis

Helena W. Qi[1,2] and Heather J. Kulik[1,*]

[1]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

Contents

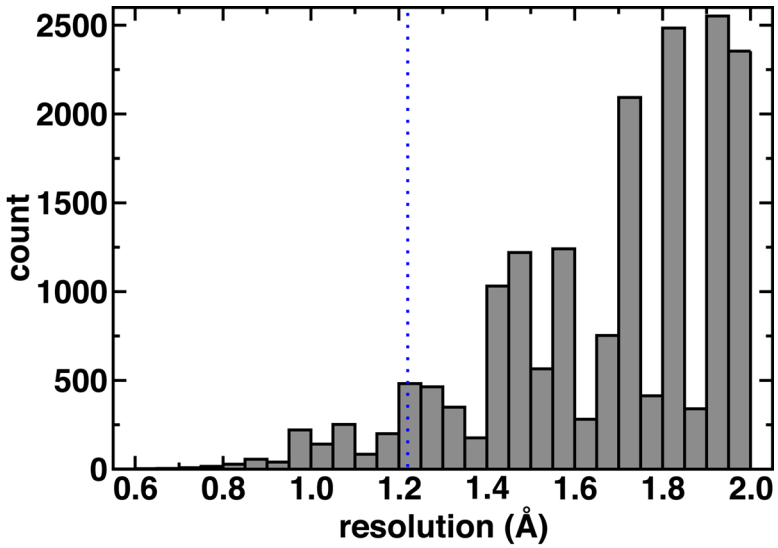


Figure S1. Histogram of reported X-ray diffraction resolution of all 17,854 proteins in the AD data set without normalization. Resolutions are binned by 0.05 Å increments, and the resolution cutoff for the HR data set is indicated as a vertical dotted blue line.

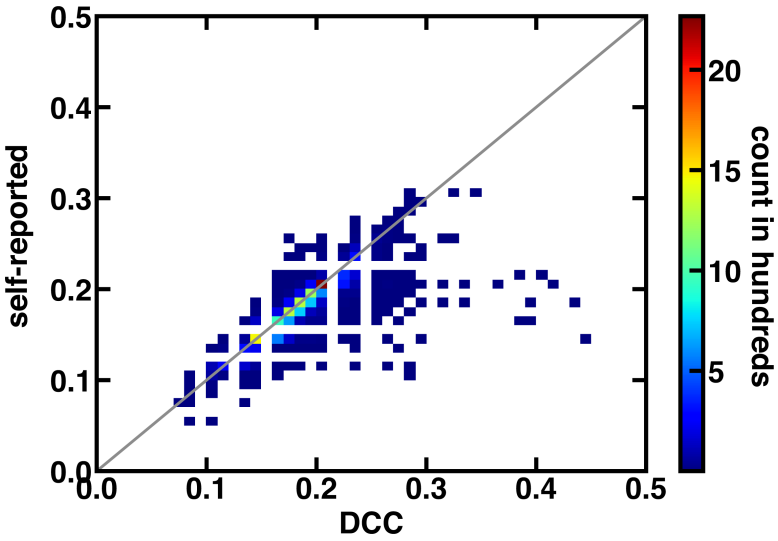


Figure S2. 2D histogram comparison of *R*-values for the **AD** data set from the *DCC*[1] program and self-reported values in the PDB file with points colored as indicated in color bar at right. A gray parity line is shown, and the two values typically agree, with less than 1% of values differing by more than 0.05. *DCC* values are used when available because they are reported as they are typically higher than the self-reported values and are therefore more conservative.

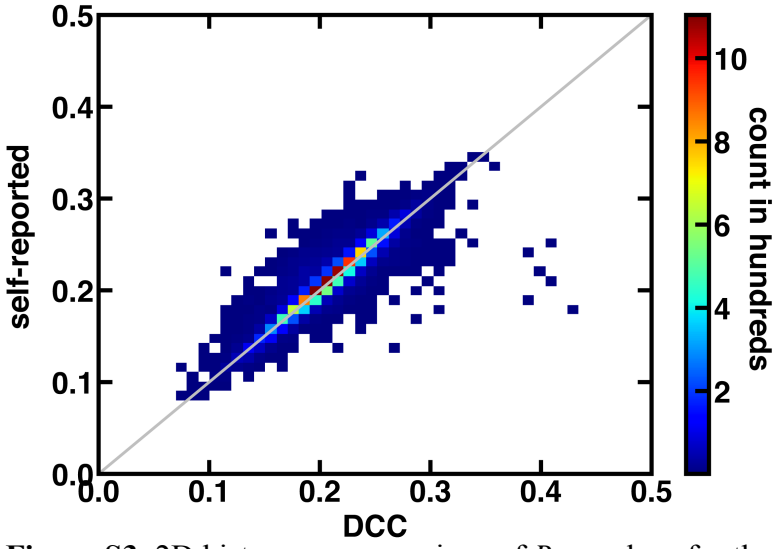


Figure S3. 2D histogram comparison of R_{free}-values for the **AD** data set from the *DCC*[1] program and self-reported values in the PDB file with points colored as indicated in color bar at right. A gray parity line is shown, and the two values typically agree, with less than 1% of values differing by more than 0.05. *DCC* values are used when available because they are reported as they are typically higher than the self-reported values and are therefore more conservative.

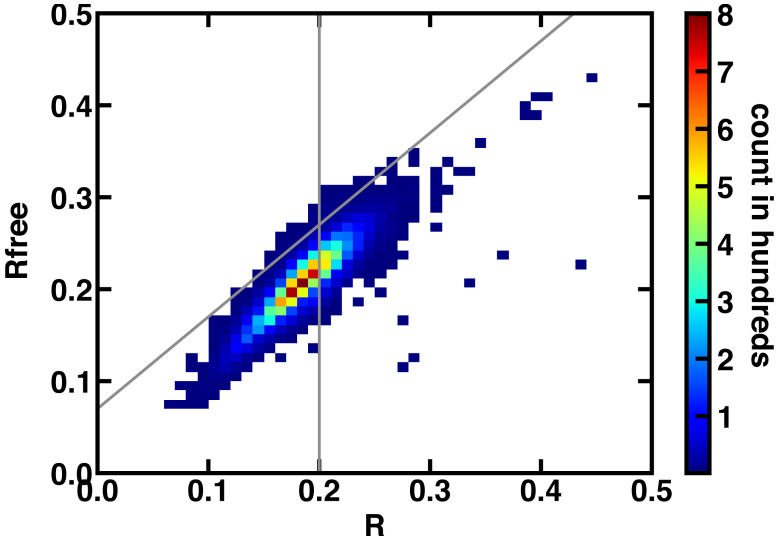


Figure S4. 2D histogram comparison of R vs. R_{free} for the **AD** data set, where values are from the DCC program unless unavailable in which case they are obtained directly from the PDB with points colored as indicated in the color bar at right. A total of 17,543 of 17,854 protein structures are shown. The $R \leq 0.20$ cutoff is shown as a vertical gray line along with an $R_{free} - R = 0.07$ cutoff line also shown in gray. The included structures in the FD set must sit in the lower left triangle of the graph bounded by these two lines.

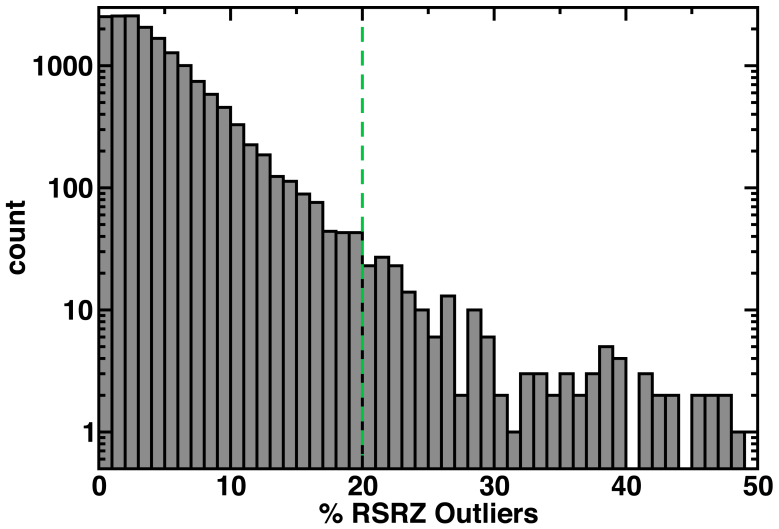


Figure S5. The percentage of RSRZ outliers for the **AD** dataset, where reported (16,912 of 17,854 cases) shown on a log scale without normalization with the 20% cutoff used in generating the **FD** data set indicated as a green vertical dashed line. One case with 100% RSRZ outliers has been truncated to make visualization of the distribution clearer.

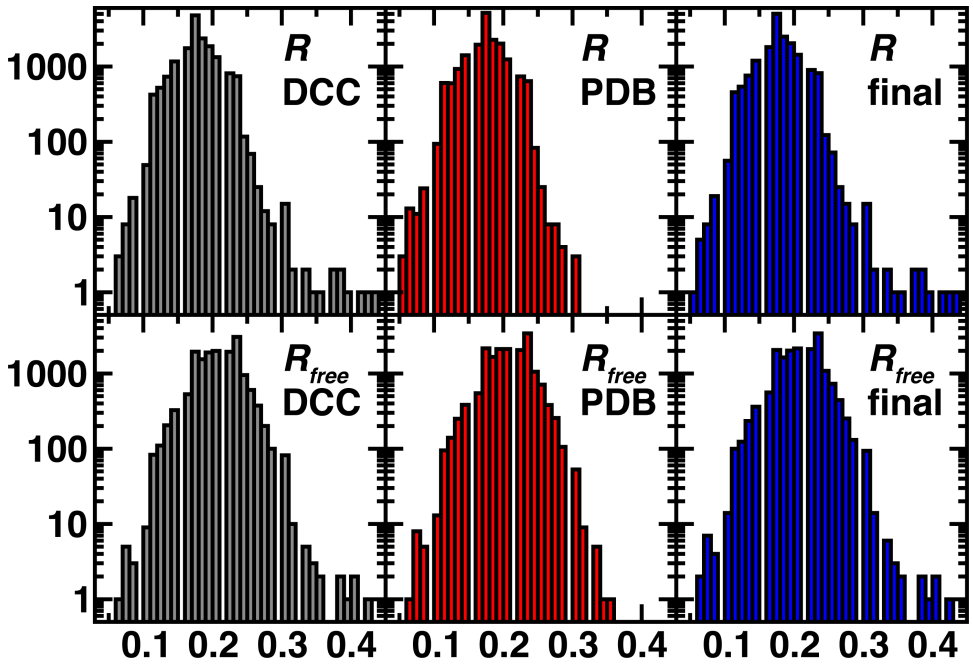


Figure S6. Unnormalized histograms with counts shown on a log plot for the R (top) and R_{free} (bottom) values obtained from the DCC program (left), directly from the PDB file (middle), and the final value chosen (right) in this work, which is the DCC value where available. Partial truncation of the x-axis has excluded a handful of points. Of 17,854 structures, the following values are available: DCC R: 16,846; DCC R_{free}: 16,008; PDB R: 17,839; PDB R_{free}: 17,522; final R: 17,844; final R_{free}: 17,544.

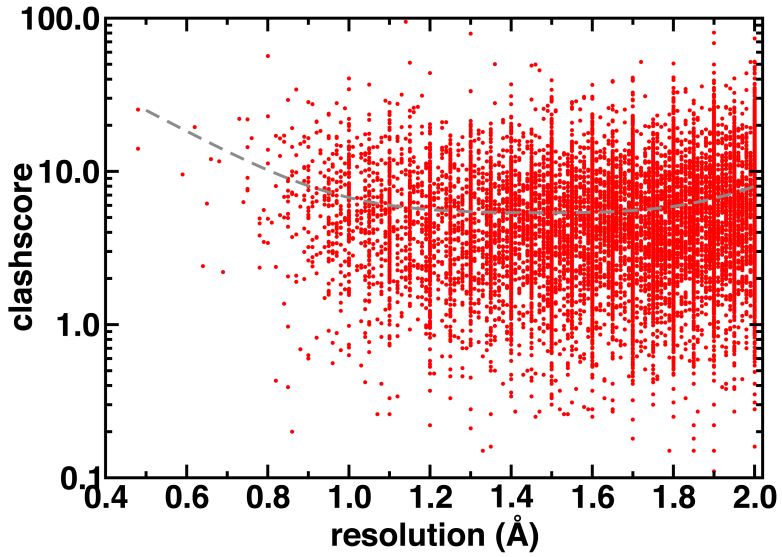


Figure S7. Reported clashscore (number of clashes found per 1000 atoms including hydrogen atoms where resolved) for structures vs. resolution (in Å) of structures shown as red dots for the complete **AD** data set. The clashscore is shown on a log scale. A fifth order regression of the data is shown as a gray dashed line. Relatively weak clashscore dependence is observed with resolution of the structures.

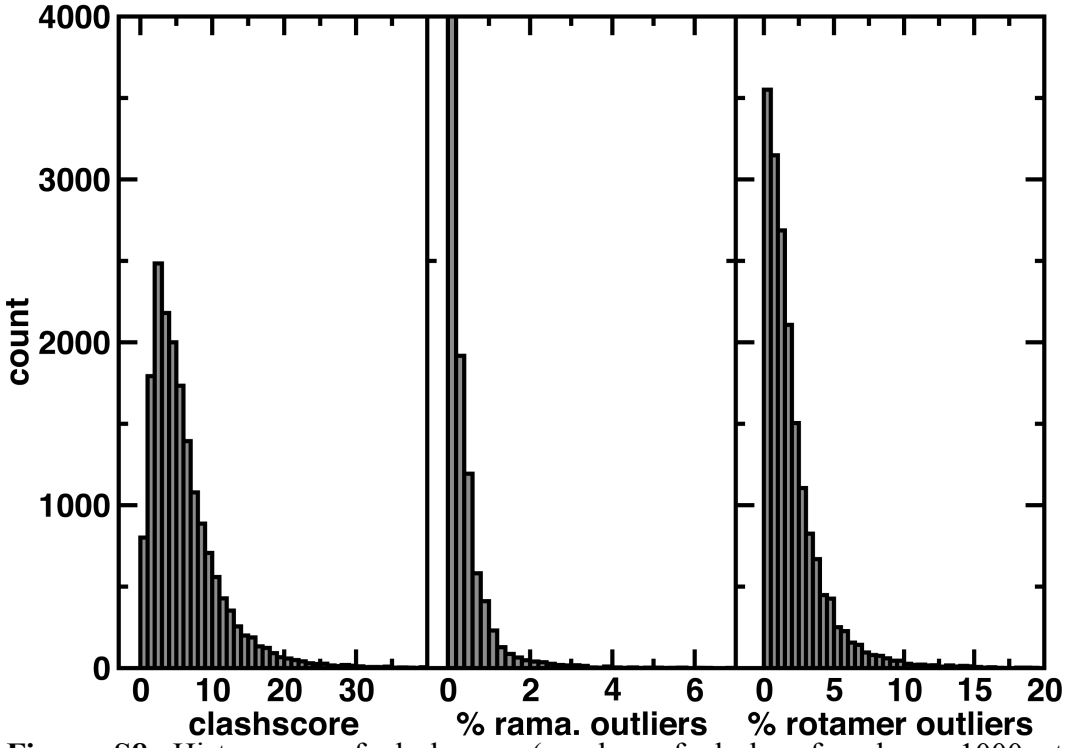


Figure S8. Histograms of clashscore (number of clashes found per 1000 atoms including hydrogen atoms where resolved, left), % Ramachandran outliers (middle), and % rotamer outliers (right) for all values reported (17,854 for clashscore, 17,849 for Ramachandran, 17,850 for rotamer) in the **AD** data set without normalization. The middle plot peak is truncated to resolve the features of the histograms, and some x-axis outliers are omitted. Bin sizes are: 1 for clashscore, 0.2% for Ramachandran outliers, and 0.5% for rotamer outliers.

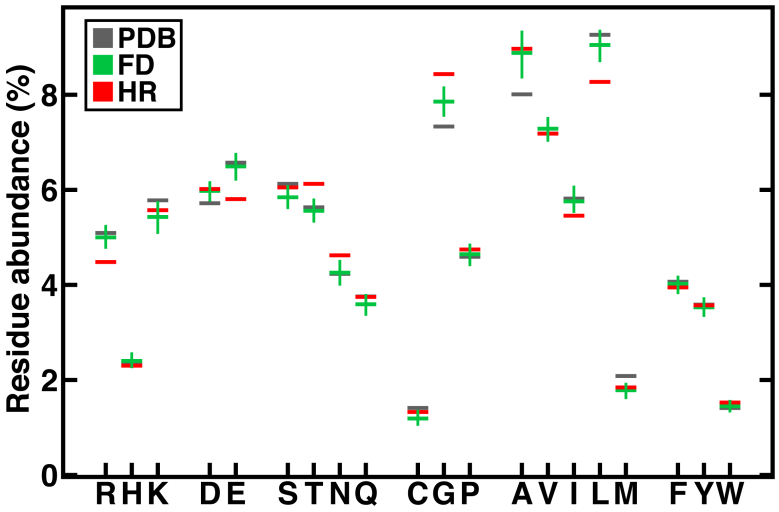


Figure S9. Abundance (%) of residues, as indicated by single letter codes and grouped by residue type (i.e., charged: R, H, K, D, E; polar: S, T, N, Q; special: C, G, P; nonpolar: A, V, I, L, M; and aromatic: F, Y, W), in the **FD** (green), **HR** (red), and the full Protein Data Bank (PDB, gray). The range on **FD** values indicates possible residue abundances obtained by randomly drawing 5000 samples of 1,151 proteins from the **FD** dataset to account for the smaller size of the **HR** vs. **FD** data set.

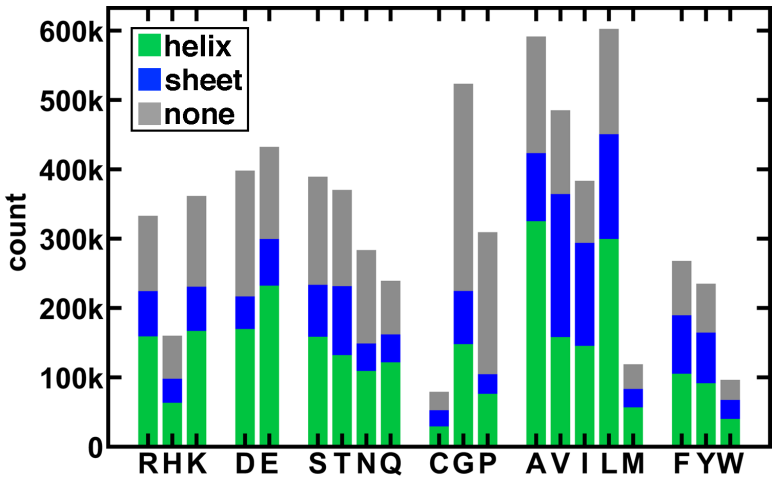


Figure S10. Total residue count in the **FD** data set colored by the secondary structure element that the residue is in, as reported by the PDB: helix (green), sheet (blue), or none (gray).

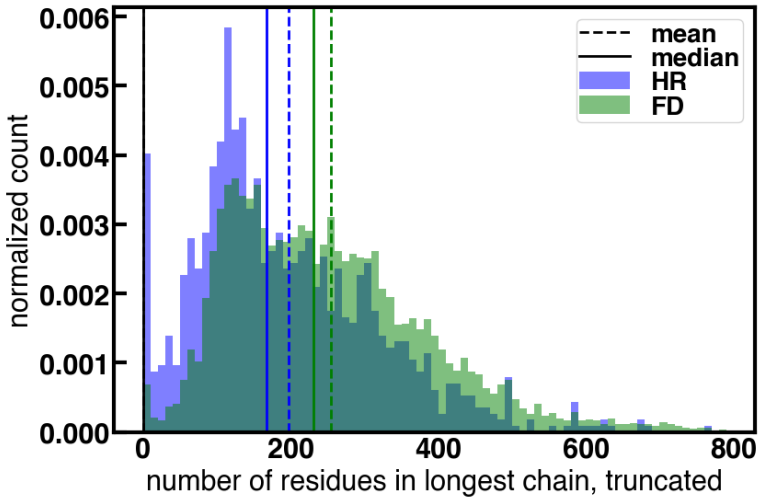


Figure S11. Comparison of the normalized histogram for chain length of proteins in the **HR** (blue) and **FD** (green) data sets. The x-axis has been truncated to 800 residues, excluding only a small number of outliers to visualize the majority of the distribution. The bin size is 10 residues for both distributions. The median is shown as a vertical solid line and the median is shown as a vertical dashed line colored blue for HR statistics and green for FD statistics.

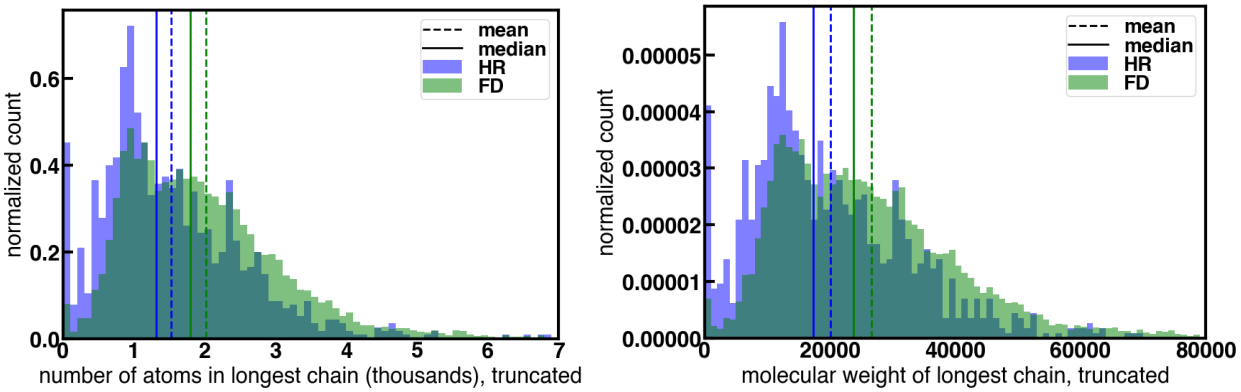


Figure S12. Comparison of the normalized histogram for number of heavy atoms (left) and molecular weight in au based on heavy atoms (right) for proteins in the **HR** (blue) and **FD** (green) data sets. The x-axis for the number of atoms (molecular weight) has been truncated to 7,000 (80,000) to exclude a small number of outliers and visualize the majority of the distribution. The bin size is 100 atoms (1000 au) for both distributions for number of atoms (molecular weight). The median is shown as a vertical solid line and the median is shown as a vertical dashed line colored blue for **HR** statistics and green for **FD** statistics.

Table S1. Definition of covalent[2] radii and van der Waals radii for individual and combinations of C, N, O, and S atoms, all in Å. In select cases where ordinary and charge-assisted (CA) hydrogen bond (HB) distances are available[3], they are also indicated. Close contacts that were within the sum of their atoms' covalent bond radii were also removed from the dataset, along with any close contact between the two residues involved to be consistent with the no covalently bound residues criteria.

	covalent			vdW				HB	
1···2	1	2	sum	1	2	sum	85%	ordinary[3]	CA[3]
C···C	0.76	0.76	1.52	1.70	1.70	3.40	2.89	--	--
N···N	0.71	0.71	1.42	1.55	1.55	3.10	2.64	3.05	2.59-2.68
O···O	0.66	0.66	1.32	1.52	1.52	3.04	2.58	2.70	2.36-2.40
S···S	1.05	1.05	2.10	1.80	1.80	3.60	3.06	4.00	3.45
C···N	0.76	0.71	1.47	1.70	1.55	3.25	2.76	--	--
C···O	0.76	0.66	1.42	1.70	1.52	3.22	2.74	--	--
C···S	0.76	1.05	1.81	1.70	1.80	3.50	2.98	--	--
N···O	0.71	0.66	1.37	1.55	1.52	3.07	2.61	2.87	2.51
N···S	0.71	1.05	1.76	1.55	1.80	3.35	2.85	--	--
O···S	0.66	1.05	1.71	1.52	1.80	3.32	2.82	--	--

Table S2. CCs of standard amino acids obtained from residues that are not covalently bound (through both sequence ordering and disulfide bonds or linkages) or when there's an observation of two atoms being within the sum of covalent radii (see Table S1). This also excludes CCs between CD of proline and CA or C of adjacent residues. Those CCs that included the carboxylate of a terminal residue were also excluded, leading to the following statistics, as indictaed. Some steps involved removal at the full **AD** set, whereas others are only relevant for **FD** and **HR**, leading to some columns indicated by '--'. Only first listed pose is chosen in the same chain. Multiple chains indicate duplicate copies of the same protein, given that single entity filtering was completed. *Close contacts for residues within 3.5 Å of ligands/small molecules (excluding water) or non-standard amino acids were also excluded at this step.

	AD	**FD**	**HR**
Total number of proteins satisfying criteria	17,854	13,472	1,151
Number of proteins with no UCCs	1,618	1,358	349
Number of proteins with UCCs	16,236	12,114	802
Initial close contacts	123,435	--	--
After OXT removal	123,199	--	--
Any two residues within covalent bond distance but not covalently bonded	122,884	85,644	--
Overall close contacts preserved (CC)*	119,043	82,701	2,326
Unique close contacts (UCC)	112,265	78,422	2,165
HB candidates (any X-H···X)	--	81,212	2,276
Salt bridge candidates (RK to DE)	--	10,579	328

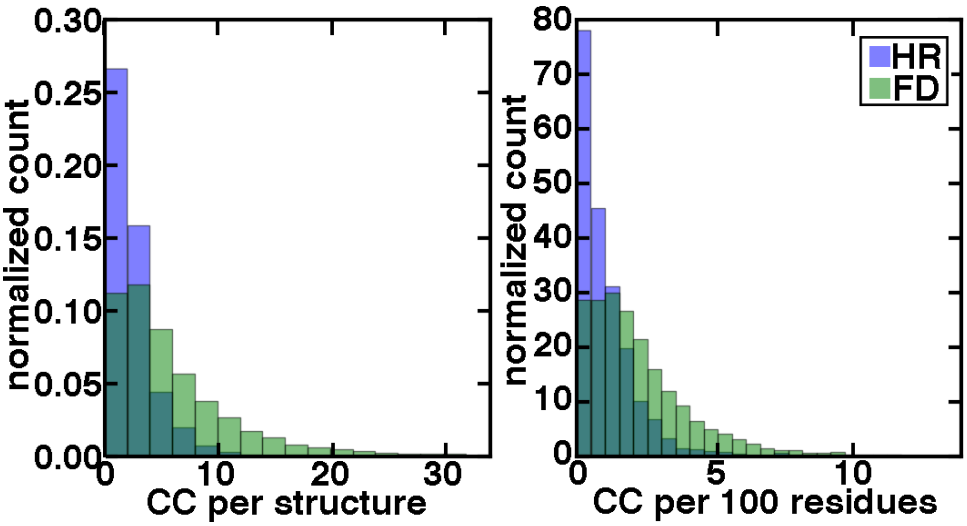


Figure S13. Normalized histograms of unique close contact (CC) frequency in the **HR** (blue) and **FD** (green) data sets: CCs per structure (left) and CCs per 100 residues (right). The median size of **HR** proteins is reduced with respect to **FD** but that only partly accounts for the increased frequency of CCs in the **FD** set over the **HR** set as seen at right. Bin sizes are 2 CC per structure at left and 0.5 CC per 100 residues at right.

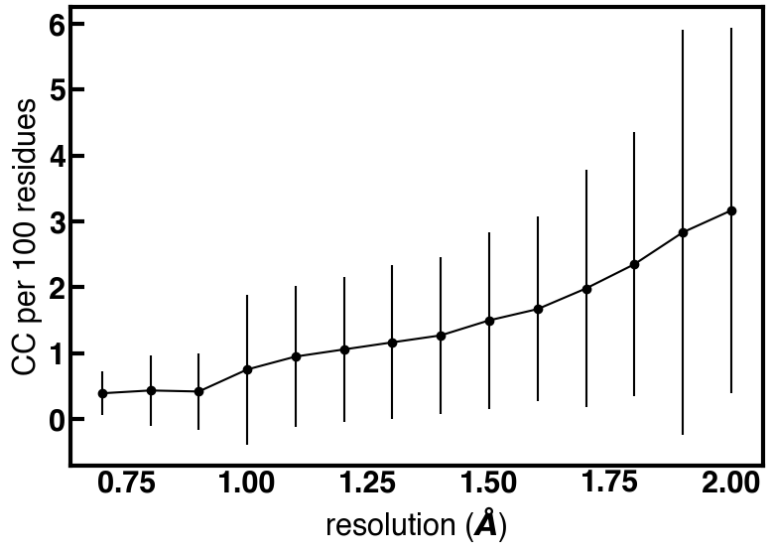


Figure S14. Average number of close contacts (CC) per 100 residues grouped by resolution of the crystal structure. The standard deviation of this quantity is shown as a vertical range for each resolution. One outlier at 1.00 Å has been removed to avoid accentuating the standard deviation.

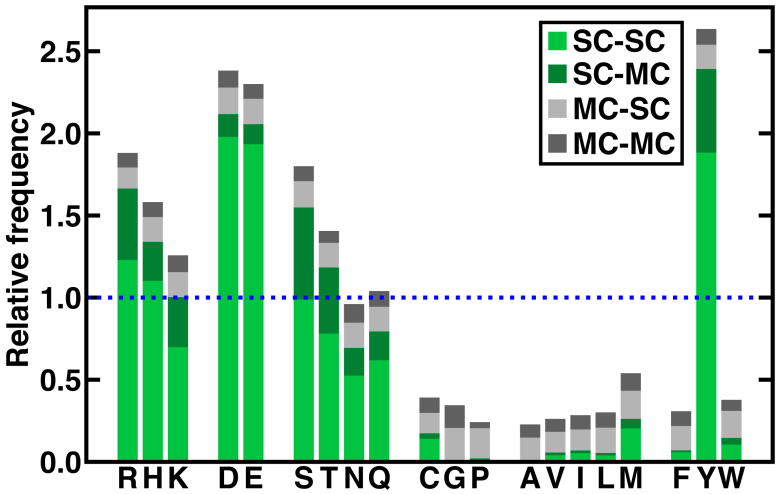


Figure S15. By-residue frequency of all close contacts in the **FD** data set relative to residue abundance in the data set classified by interactions between sidechain (SC) and mainchain (MC) atoms. Here, SC-SC (light green) and SC-MC (dark green) refer to sidechain participation by the labeled residue, whereas MC-SC (light gray) and MC-MC (dark gray) refer to main chain atoms of the labeled residue participating in a close contact. A relative frequency of 1 is indicated by the blue dotted horizontal bar.

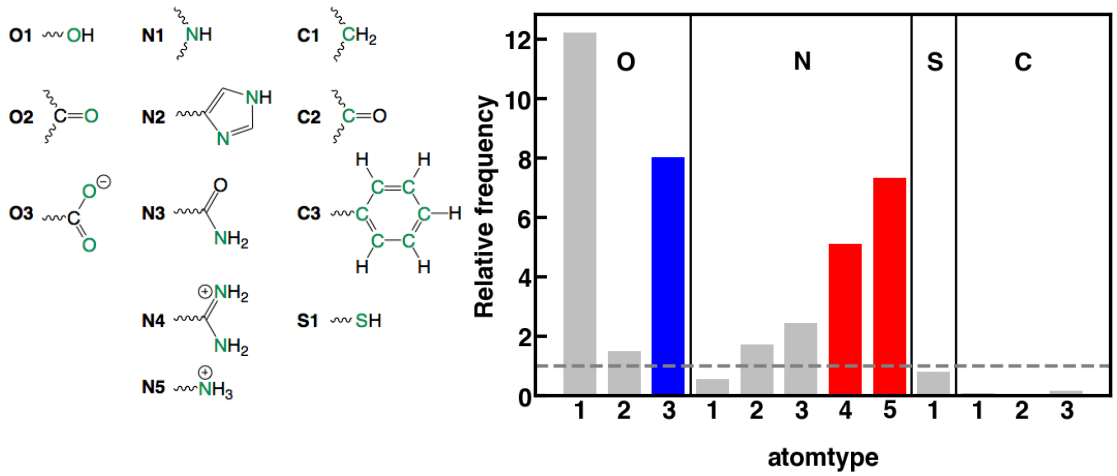


Figure S16. (Left) Representative structures of atom type definitions in skeleton representation with relevant atoms colored in green. For O: hydroxyl O (1), carbonyl O (2), and negatively charged carboxylate O (3). For N: neutral amine N (1), aromatic N (2), N from Asn or Gln (3), positively charged Arg N (4), and positively charged Lys N (5). For S: a single type (1) is defined shown as –SH in Cys but used for both Cys and Met. For C: sp^3 C (1), sp^2 C (2), or aromatic C (3). (Right) Relative frequency of CCs by atom type in **FD** data set. A relative frequency of 1 is indicated by the gray dashed horizontal bar. Absolute frequencies are nonzero for all atom types but due to the high fraction of amino acid atoms that are carbon atoms, the relative frequency of C1 and C2 close contacts appears to approach zero.

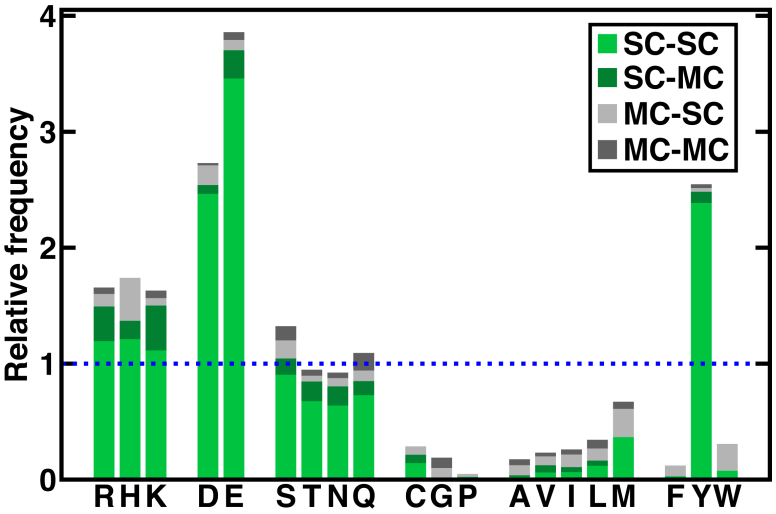


Figure S17. By-residue frequency of all close contacts in the **HR** data set relative to residue abundance in the data set classified by interactions between sidechain (SC) and mainchain (MC) atoms. SC-SC (light green) and SC-MC (dark green) refer to sidechain participation by the labeled residue, whereas MC-SC (light gray) and MC-MC (dark gray) refer to main chain atoms of the labeled residue participating in a close contact. A relative frequency of 1 is indicated by the blue dotted horizontal bar.

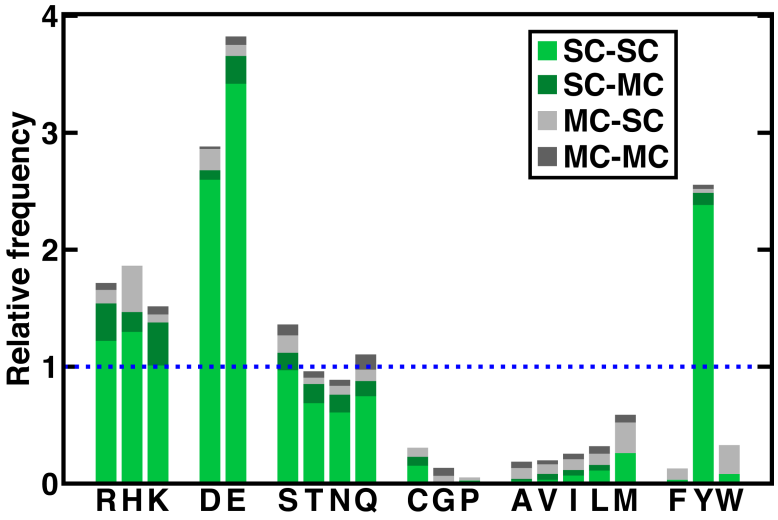


Figure S18. By-residue frequency of unique close contacts in the **HR** data set relative to residue abundance in the data set classified by interactions between sidechain (SC) and mainchain (MC) atoms. Here, unique close contacts refers to only counting a single contact between two residues, even if multiple are formed. SC-SC (light green) and SC-MC (dark green) refer to sidechain participation by the labeled residue, whereas MC-SC (light gray) and MC-MC (dark gray) refer to main chain atoms of the labeled residue participating in a close contact. A relative frequency of 1 is indicated by the blue dotted horizontal bar.

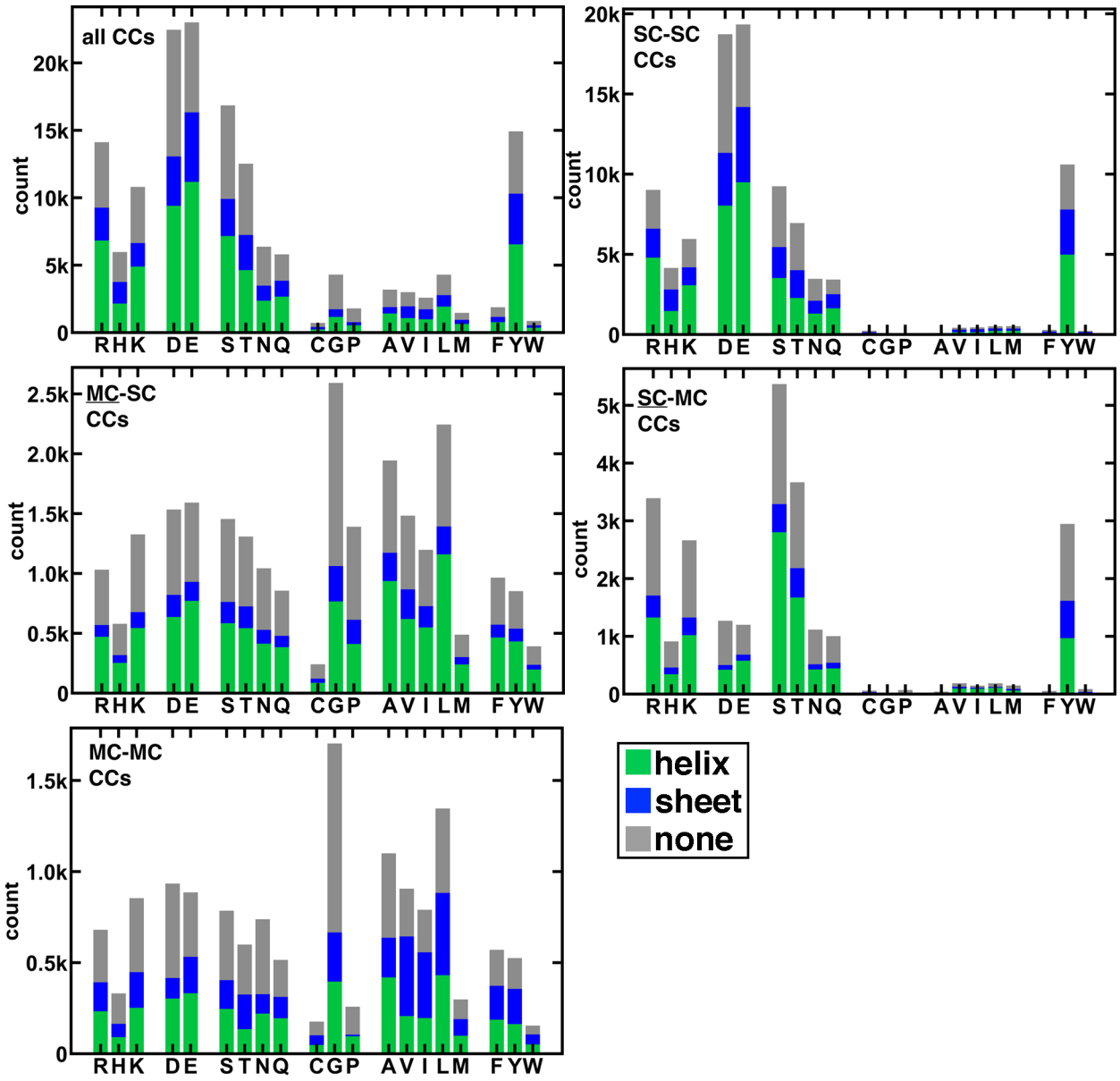


Figure S19. Residue count (absolute) only for unique close contacts in the **FD** data set colored by the secondary structure element that the residue is in, helix (green), sheet (blue), or none (gray), for all CCs (top, left), SC-SC CCs (top, right), MC-SC CCs, where the labeled residue's MC is involved with the SC of the unlabeled residue (middle, left), SC-MC CCs, where the labeled residue's SC is involved with the MC of the unlabeled residue (middle, right), and MC-MC CCs (bottom, left). The legend for all graphs is shown at bottom right.

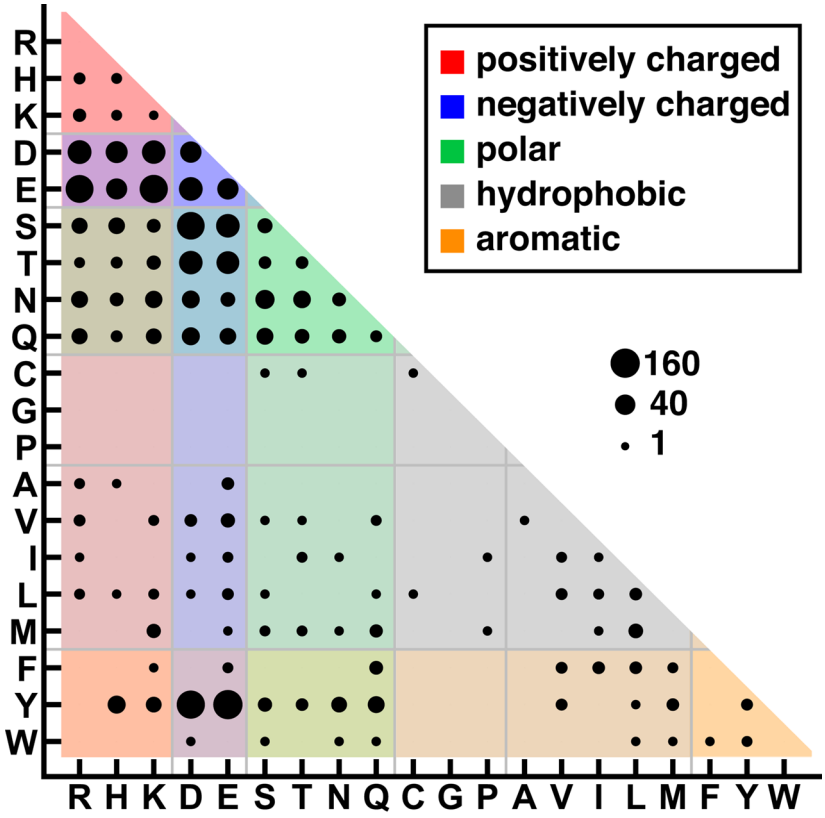


Figure S20. Matrix of absolute close contact frequency between residue sidechains (SC-SC) in the **HR** data set between residues grouped (i.e., separated by thin gray bars) by type according to canonical charge assignment and indicated with single letter codes: positively charged (R, H, K in red), negatively charged (D, E in blue), polar (S, T, N, Q in green), special (C, G, P in gray) or nonpolar (A, V, I, L, M in gray), and aromatic (F, Y, W in orange). Each region is colored translucently, and resulting combinations lead to blended colors between different residue types, as indicated also in the inset legend. The area of each circle represents the number of close contacts, as indicated qualitatively by the inset legend of representative circle sizes. Only the non-redundant lower triangle of the matrix is shown for clarity.

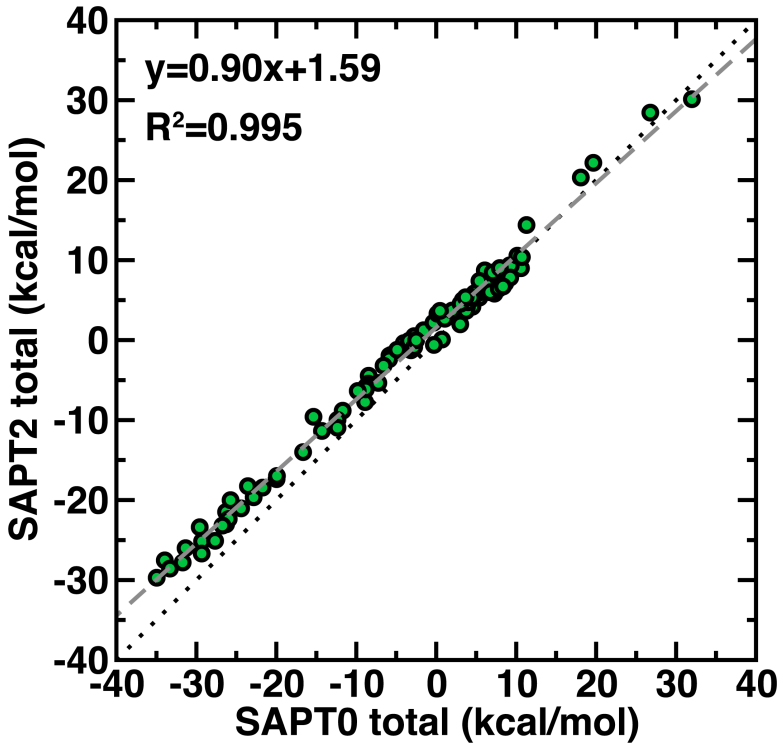


Figure S21. Total SAPT0 vs. SAPT2 interaction energy for 88 residue-residue pairs with a parity line shown in dotted black as well as the best fit line shown in dashed gray.

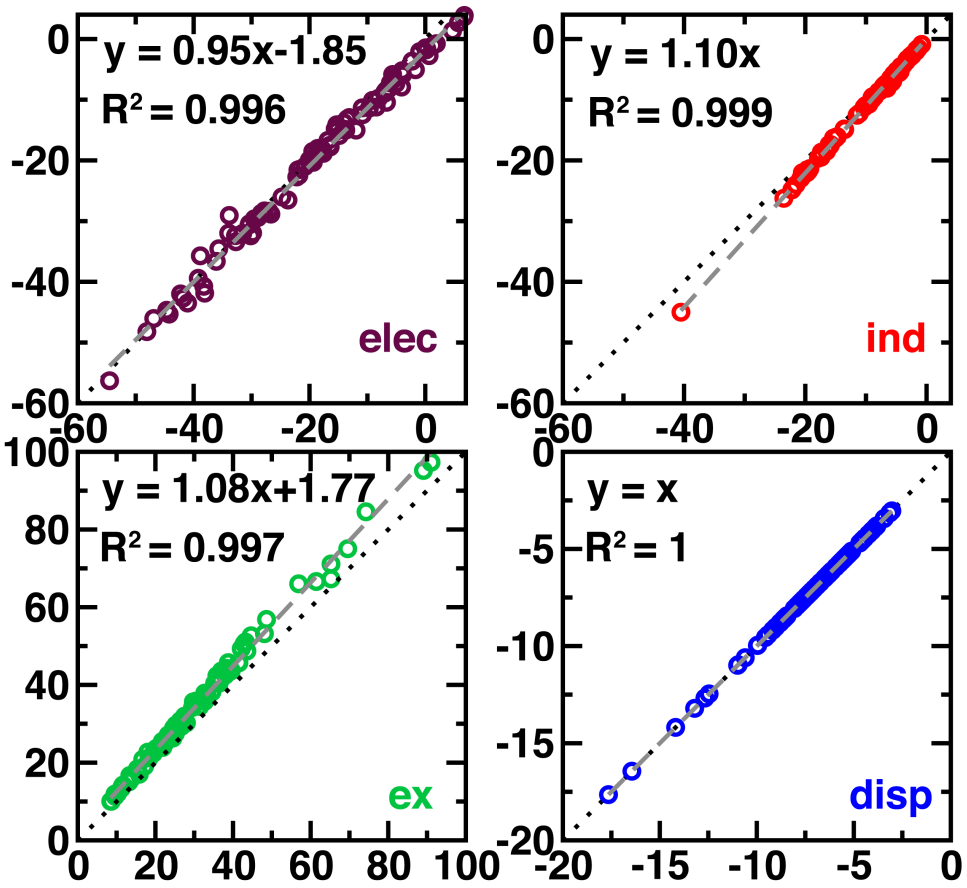


Figure S22. Components of SAPT0 vs. SAPT2 interaction energy (top, left: electrostatic; top, right: induction; bottom, left: exchange; bottom, right: dispersion) in kcal/mol for 88 residue-residue pairs. A dotted black parity line is shown along with a best fit line (dashed gray) in inset with correlation coefficient. Dispersion terms are unchanged in SAPT0 and SAPT2 and therefore the relationship is trivial.

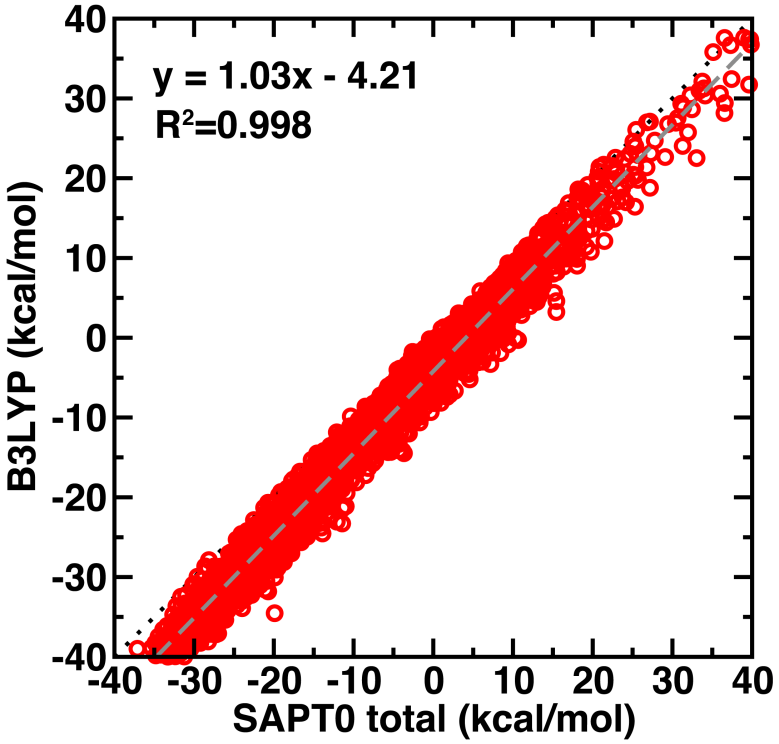


Figure S23. Total gas phase B3LYP vs. SAPT0 interaction energy for 6,201 residue-residue pairs that do not exhibit proton transfer. The -40 to +40 kcal/mol range shows roughly 50% of the data to keep the range consistent with the SAPT0/SAPT2 comparison and to focus on weaker interactions. The parity line is shown in dotted black as well as the best fit line shown in dashed gray, with equation and correlation coefficient shown in inset.

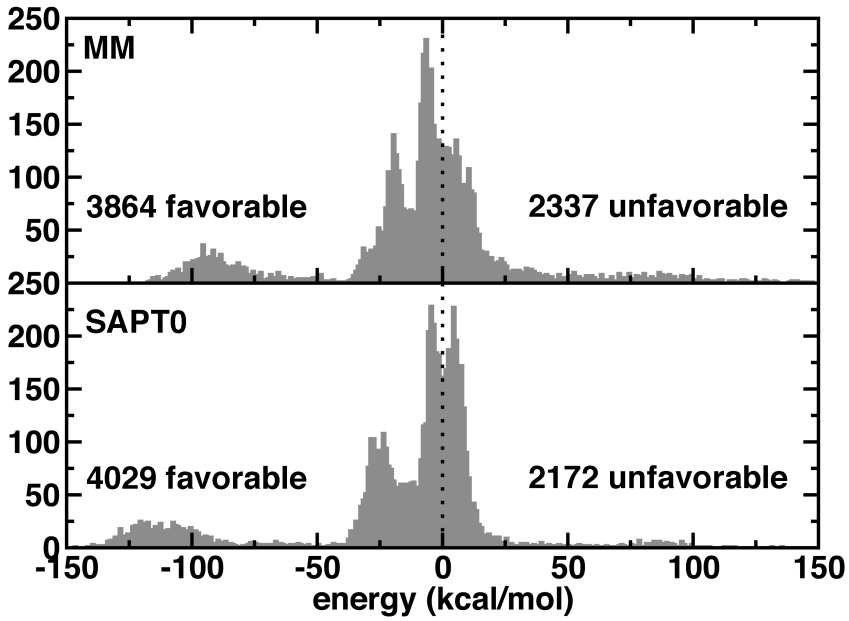


Figure S24. Histogram of total interaction energies evaluated with SAPT0 (bottom) and MM (gas phase electrostatics and vdW terms, top) over the -150 to +150 kcal/mol range with 1 kcal/mol bins for 6,201 CCs. The 0 kcal/mol interaction cutoff is shown as a vertical dotted black line.

Table S3. Summary statistics of the type of close contacts studied starting from a pool of 6,279 candidates. Canonical protonation states are assumed for the initial count (His, Arg, Lys all positively charged, Asp or Glu negatively charged). Excluded due to proton transfer (PT) means that during optimization of coordinates, protonation state changed in ε=4 solvation and so the structure was excluded from further analysis. Non-canonical indicates the number of cases that enter a category after a non-standard protonation state is assigned (i.e., neutral Glu, Asp, Lys, or neutral His). +/- MM energetics are not tabulated due to some cases leading to poor assignment of properties due to strong sharing of protons or partial transfer. Most are strongly unfavorable.

	+/-	++	--	nn	n+	n-	total
initial count	984	54	263	2,320	753	1,905	6,279
excluded due to PT	61	0	16	0	1	0	78
non-canonical change	-137	-24	-153	+426	-225	+113	--
FD total	786	30	94	2,746	527	2,018	6,201
Crystal H-opt energetics							
SAPT0 favorable	786	0	0	1,365	415	1,463	4,029
SAPT0 unfavorable	0	30	94	1,381	112	555	2,172
MM favorable	746	0	0	1,412	347	1,359	3,864
MM unfavorable	40	30	94	1,334	180	659	2,337
Shortest SAPT0-favorable interaction	5igi: Glu-Arg O-N 1.80 Å -33.5 kcal/mol	N/A	N/A	5a6m: Thr-Asn O-O 2.30 Å -0.73 kcal/mol	1gwm: Lys-Asn N–O 2.17 Å -13.8 kcal/mol	4iqb: Thr-Asp O-O 2.12 Å -12.1 kcal/mol	N/A
SC-SC opt energetics							
SAPT0 favorable	785	0	0	2,693	523	2,013	6,014
SAPT0 unfavorable	1	30	94	53	4	5	187
MM favorable	--	0	0	2,644	516	1,799	--
MM unfavorable	--	30	94	102	11	219	--
HR total	414	10	61	507	273	849	2,114
Crystal H-opt energetics							
SAPT0 favorable	414	0	0	202	217	581	1,414
SAPT0 unfavorable	0	10	61	305	56	268	700
MM favorable	391	0	0	181	176	527	1,275
MM unfavorable	23	10	61	326	97	322	839
SC-SC opt energetics							
SAPT0 favorable	80	0	0	498	270	796	1,644
SAPT0 unfavorable	0	10	61	9	3	5	88
MM favorable	79	0	0	489	267	692	1,527
MM unfavorable	1	10	61	18	6	109	205

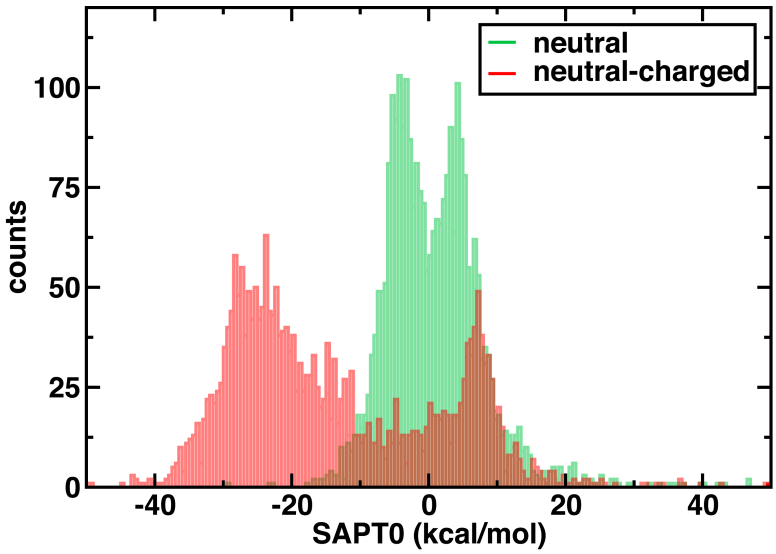


Figure S25. Unnormalized histograms of SAPT0 total interaction energies (in kcal/mol) for 2,746 neutral-neutral and 2,545 neutral-charged (2,018 neutral-negative, 527 neutral-positive) close contacts. Bins correspond to 0.5 kcal/mol increments.

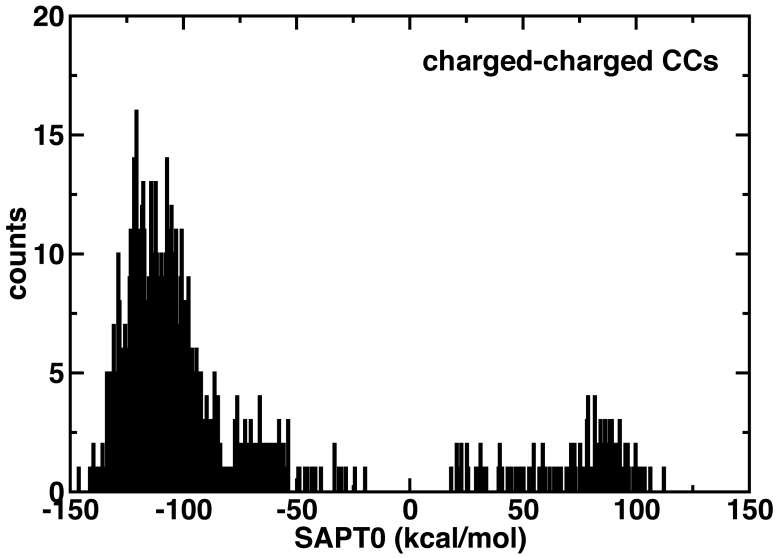


Figure S26. Unnormalized histogram of SAPT0 total interaction energies (in kcal/mol) for 910 charged-charged (786 positive-negative, 30 positive-positive, and 94 negative-negative) close contacts: repulsive SAPT0 energies correspond predominantly to same charge pairs. Bins correspond to 1.0 kcal/mol increments.

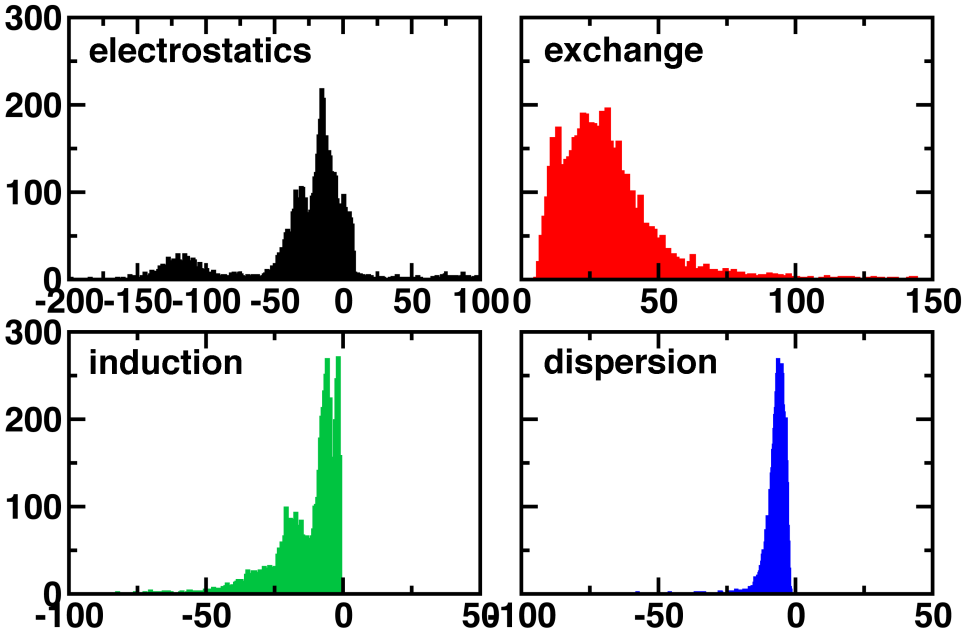


Figure S27. Unnormalized histograms of SAPT0 total energy components: electrostatic (top, left), exchange repulsion (top, right), induction (bottom, left) and dispersion (bottom, right) components of SAPT0 total interaction energies for both charged and neutral CCs. The y-axis is the number of counts for each binned interaction energy (up to 300), which is the same for all graphs. Bins for the graphs are: 1 kcal/mol for electrostatics, exchange, 0.5 kcal/mol for induction, and 0.25 kcal/mol for dispersion. The x-axis range for each graph (in kcal/mol) is: -200 to +100 kcal/mol for electrostatics, +0 to +150 kcal/mol for exchange, -100 to +50 kcal/mol for induction or dispersion (i.e., the range of the electrostatics graph is double the other three, which are the same).

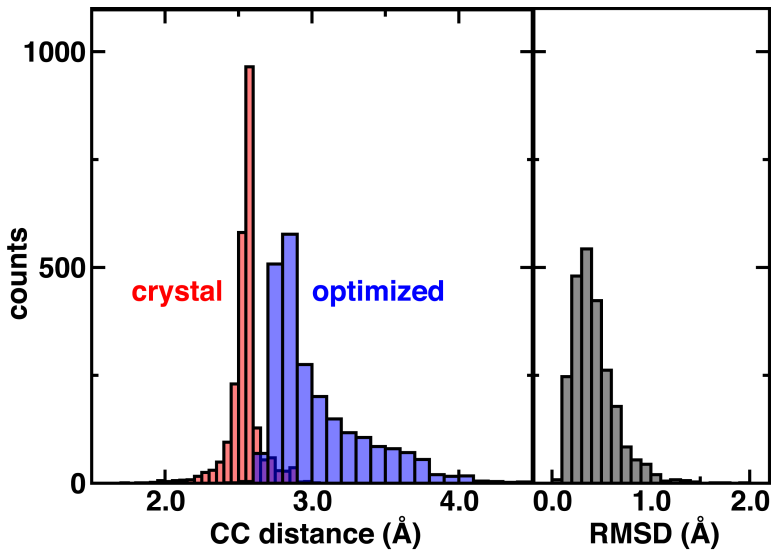


Figure S28. (left) Unnormalized histogram of 2,746 neutral close contact initial distances (in Å) from the crystal structure (crystal, red histogram, 0.05 Å bins) and after sidechain optimization with ε=4 (optimized, blue histogram, 0.10 Å bins). (right) Root mean square deviation (RMSD,

in Å) of the optimized pairs with respect to initial crystal structures. Sidechain optimization was carried out in the gas phase with the long-range corrected ωPBEh functional.

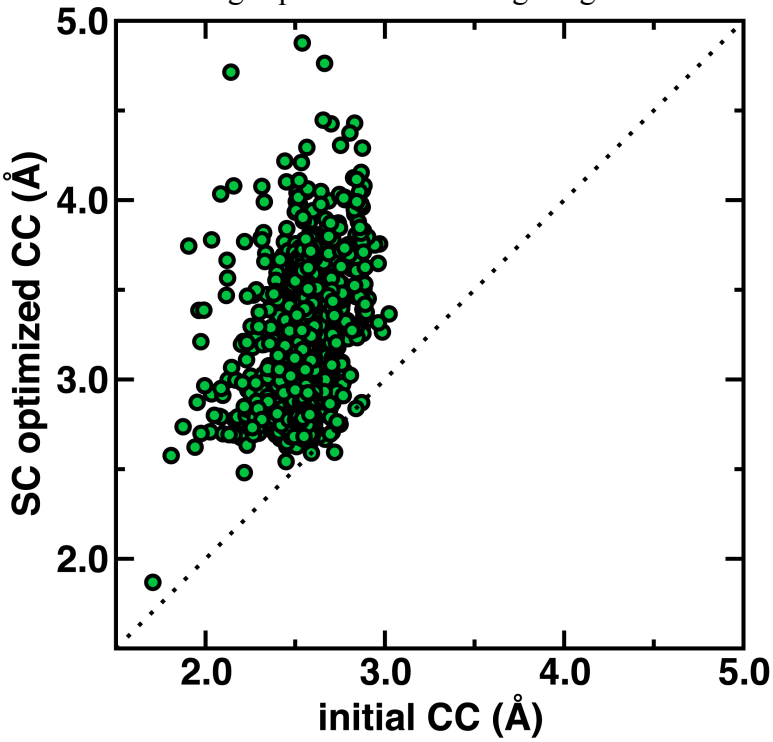


Figure S29. Scatter plot of initial (crystal) and final CC distances after optimization of sidechains for 2,746 neutral close contacts. A dotted black parity line is shown. Sidechain optimization was carried out in the gas phase with the long-range corrected ωPBEh functional.

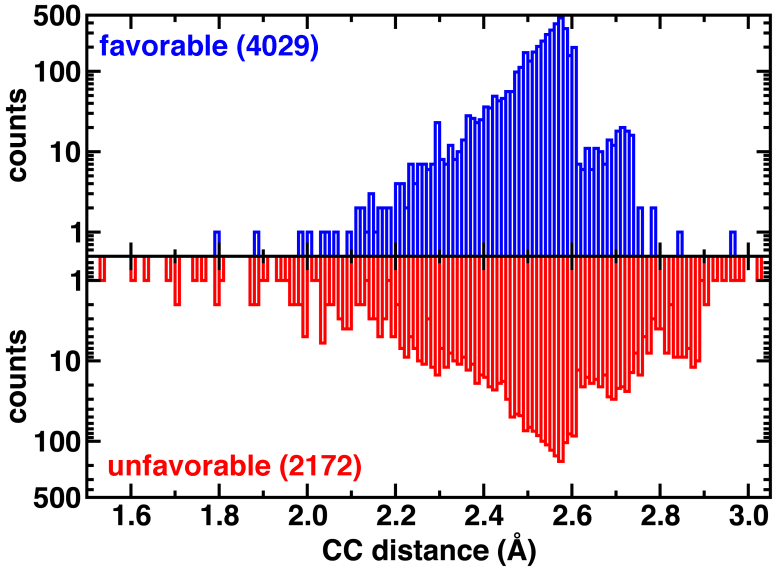


Figure S30. Unnormalized histogram of counts of favorable (4,029 total blue, top) and unfavorable (red, bottom, 2,172) close contacts by CC distance (in Å) binned to 0.01 Å. The plots are shown on a logarithmic scale with the unfavorable distribution reflected. Favorable is defined as < 0 kcal/mol SAPT0 interaction energies with optimized hydrogen atoms and unfavorable is ≥ 0 kcal/mol.

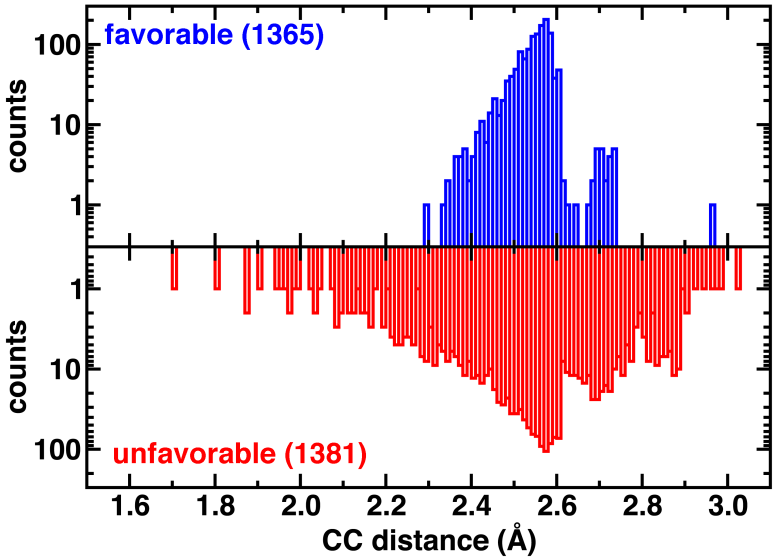


Figure S31. Unnormalized histogram of counts of favorable (1,365 total blue, top) and unfavorable (red, bottom, 1,381) neutral-neutral (2,746 total) close contacts by CC distance (in Å) binned to 0.01 Å. The plots are shown on a logarithmic scale with the unfavorable distribution reflected. Favorable is defined as < 0 kcal/mol SAPT0 interaction energies with optimized hydrogen atoms and unfavorable is ≥ 0 kcal/mol.

Table S4. Details of the shortest favorable SAPT0 interactions by each class of residue type as well as the MM interaction energies for the same set of residues.

	n+	+/-	nn	n-
Shortest favorable SAPT interaction characteristics	1gwm: Lys-Asn N–O 2.17 Å -13.8 kcal/mol	5igi: Glu-Arg O-N 1.80 Å -33.5 kcal/mol	5a6m: Thr-Asn O-O 2.30 Å -0.73 kcal/mol	4iqb: Thr-Asp O-O 2.12 Å -12.1 kcal/mol
SAPT electrostatics	-48.6	-214.1	-24.1	-68.2
SAPT exchange	66.9	215.3	42.9	115.2
SAPT induction	-19.3	-43.0	-11.2	-43.6
SAPT dispersion	-12.8	-27.70	-8.4	-15.4
SAPT total	-13.8	-33.5	-0.7	-12.1
MM vdw	52.6	617.7	14.93	48.9
MM eel	-30.7	-110.0	-11.49	-38.4
MM total	21.9	507.7	3.43	10.5

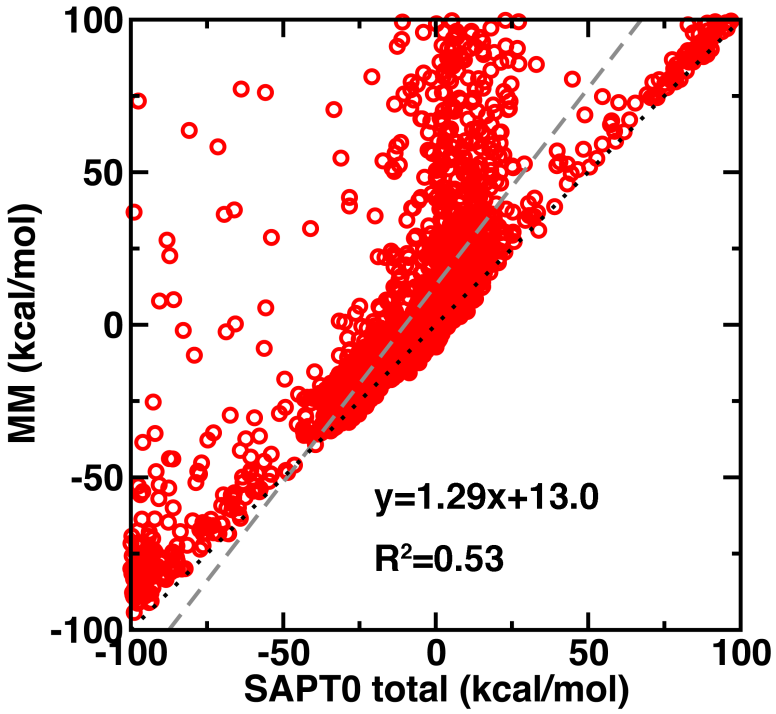


Figure S32. Total gas phase MM (electrostatic and vdW terms) vs. SAPT0 interaction energy for 6,201 residue-residue pairs that do not exhibit proton transfer. The -100 to +100 kcal/mol range excludes major outliers in the MM data, captures most physical interactions, and is used to restrict the fit of the gray dashed best fit line (correlation coefficient and fitting line shown in inset). The parity line is shown in dotted black.

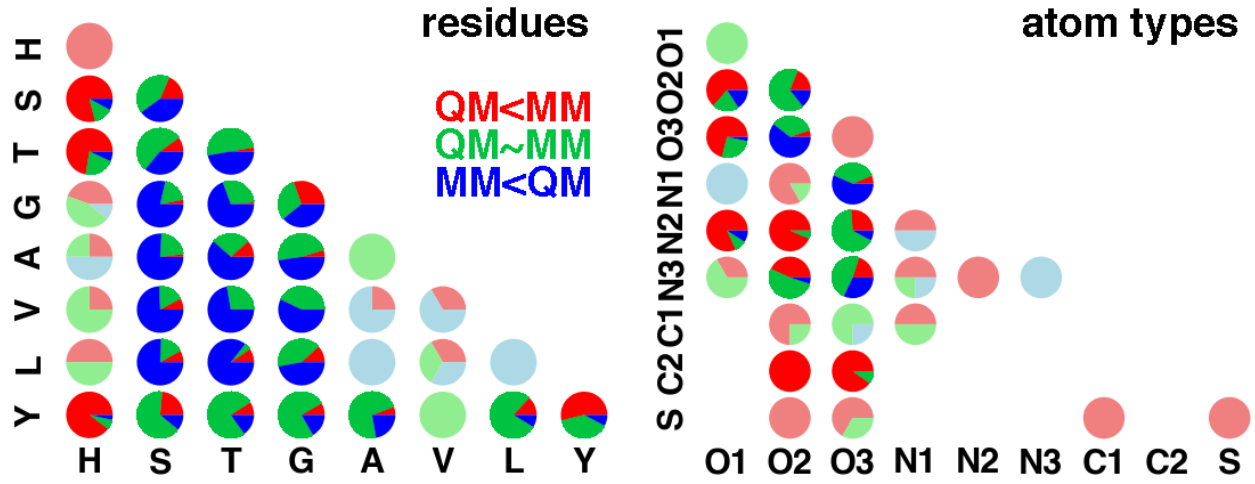


Figure S33. Classification of an 2,745 CC subset of neutral residue-residue interactions: His, Ser, Thr, Gly, Ala, Val, Leu, Tyr across the full CC subset grouped by residue (left) or by the atom types involved in the CC (right). The atom type definitions are the same as in Figure S16. Interactions are designated as QM<MM (QM is more favorable than MM, red), QM~MM (QM is within 1.5 kcal/mol of MM, green), and MM<QM (MM is more favorable than QM, blue). If there is no data, no circle is shown. If there are fewer than 10 data points, the circle is shown in lighter colors.

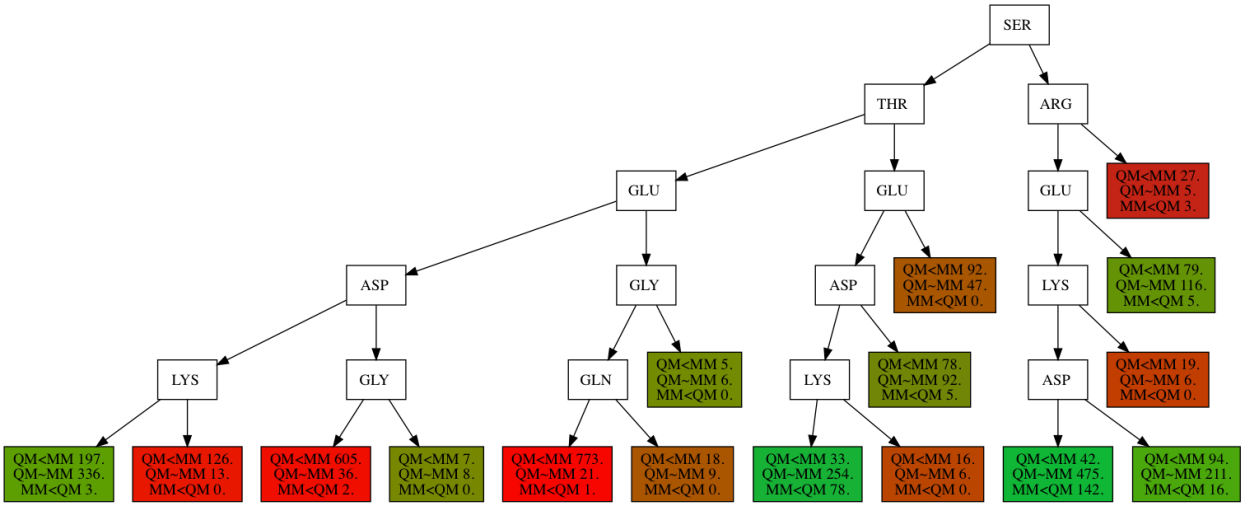


Figure S34. Decision tree based on residue type of 4,107 CCs of any charge for which both SAPT and MM crystal structure, H-optimized interaction energies were available and the SAPT0 interaction energy was no more than 0 kcal/mol (i.e., was favorable). Each case is classified by MM<QM, QM>MM, or QM~MM, where QM~MM corresponds to cases within 3 kcal/mol of each other. The decision tree was created with a maximum depth of 5 and a minimum number of 10 samples in each leaf. The possible divisions were by any residue identity. The decision tree should be read from top to bottom with the right arrow corresponding to true and left corresponding to false. Each leaf lists the number of samples in each category and is colored according to MM-lean (blue), mixed (green), or QM lean (red).

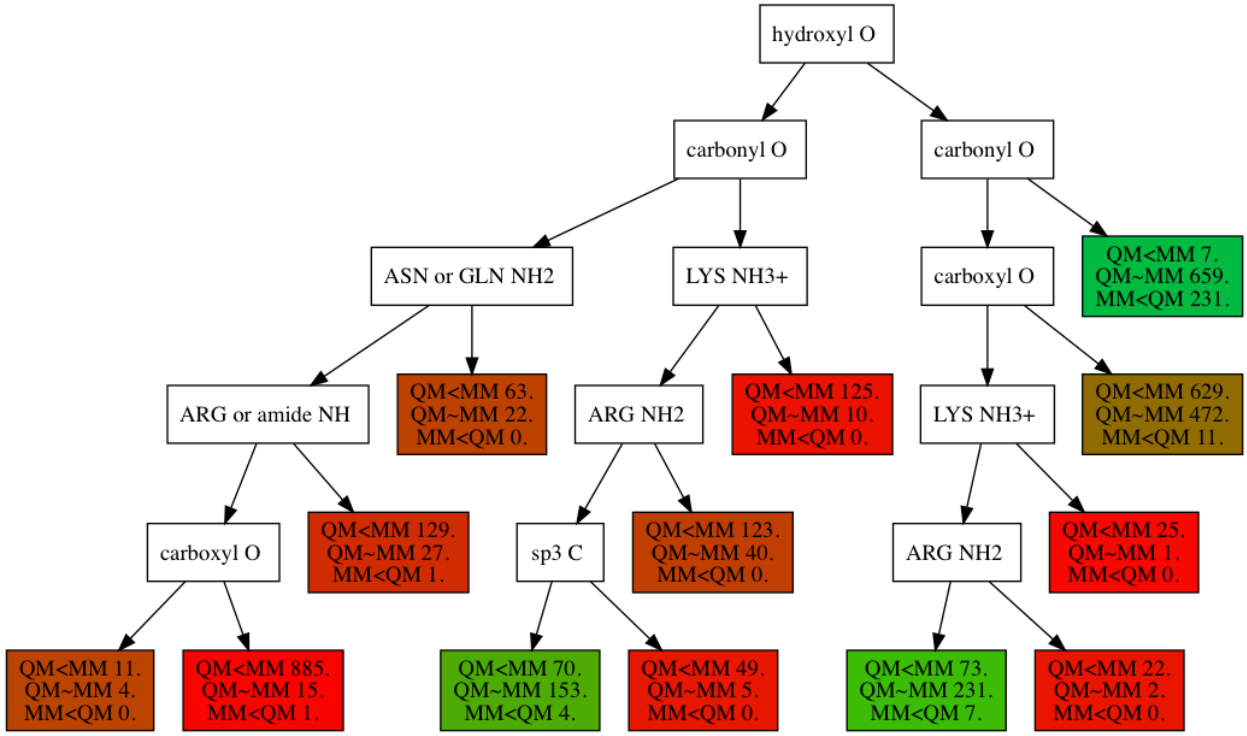


Figure S35. Decision tree based on atom type of 4,107 CCs of any charge for which both SAPT and MM crystal structure, H-optimized interaction energies were available and the SAPT0 interaction energy was no more than 0 kcal/mol (i.e., was favorable). Each case is classified by

MM<QM, QM>MM, or QM~MM, where QM~MM corresponds to cases within 3 kcal/mol of each other. The decision tree was created with a maximum depth of 5 and a minimum number of 10 samples in each leaf. The possible divisions were by atom types for S, O (carboxyl, hydroxyl, carbonyl), N (amide NH or Arg NH, Asn or Gln NH, Asn or Gln NH2, ring N, Arg NH2, Lys NH3+), and C (aromatic, sp2, or sp3). These types correspond roughly to those in the earlier atom type figure, Figure S16. The decision tree should be read from top to bottom with the right arrow corresponding to true and left corresponding to false. Each leaf lists the number of samples in each category and is colored according to MM-lean (blue), mixed (green), or QM lean (red).

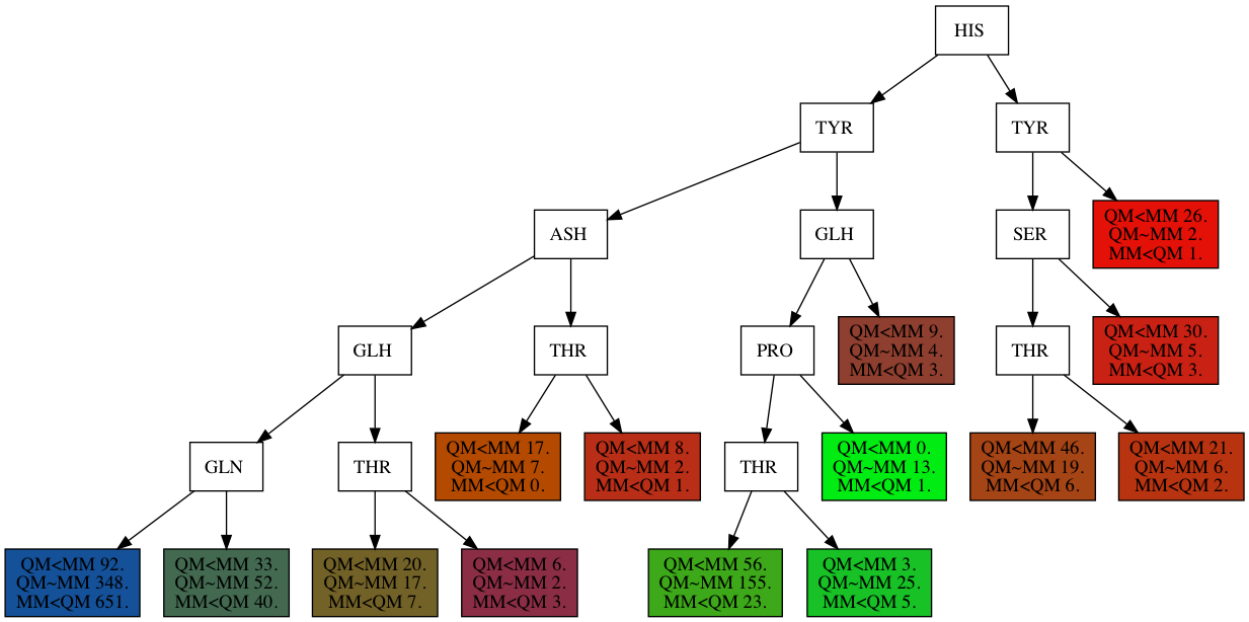


Figure S36. Decision tree based on residue type of 1,770 neutral residue CCs for which both SAPT and MM crystal structure, H-optimized interaction energies were available and the SAPT0 interaction energy was no more than +3 kcal/mol. Each case is classified by MM<QM, QM>MM, or QM~MM, where QM~MM corresponds to cases within 1.5 kcal/mol of each other. The decision tree was created with a maximum depth of 5 and a minimum number of 10 samples in each leaf. The possible divisions were by residues for only neutral cases: Gly, Pro, Ala, Val, Ile, Leu, Met, Phe, Trp,. Ser, Thr, Asn, Gln, Cys, Tyr, and His as well as non-canonical protonated Glu or Asp and neutral Lys. The decision tree should be read from top to bottom with the right arrow corresponding to true and left corresponding to false Each leaf lists the number of samples in each category and is colored according to MM-lean (blue), mixed (green), or QM lean (red).

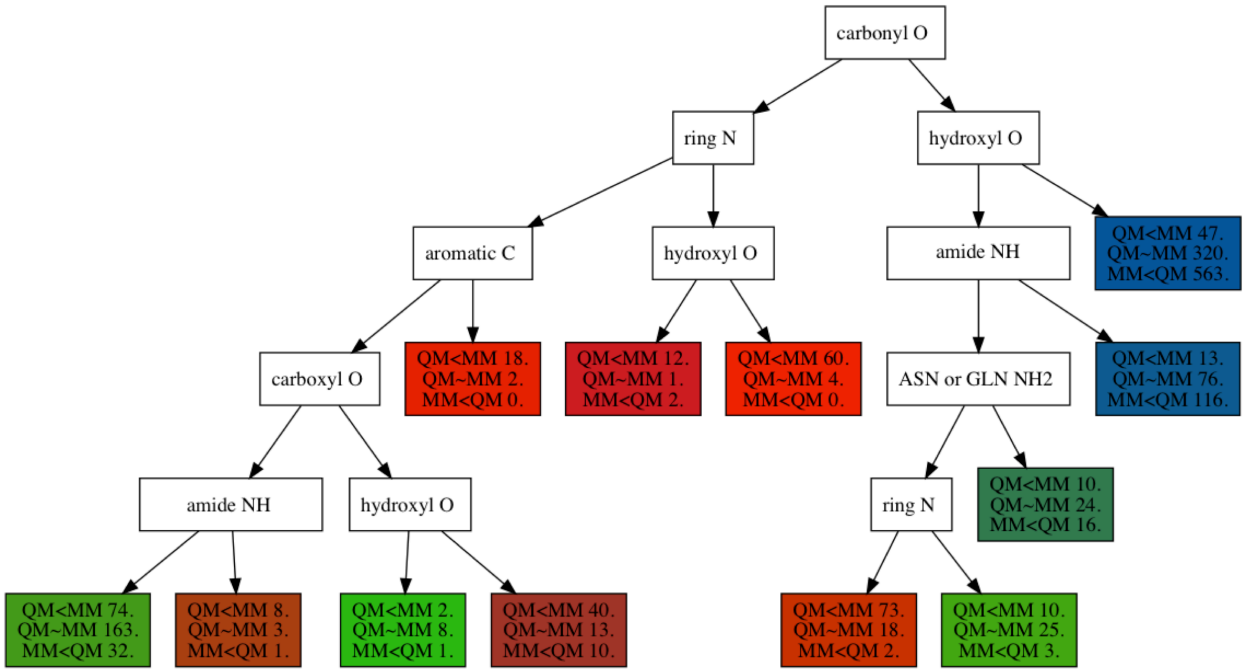


Figure S37. Decision tree based on atom type of 1,770 neutral residue CCs for which both SAPT and MM crystal structure, H-optimized interaction energies were available and the SAPT0 interaction energy was no more than +3 kcal/mol. Each case is classified by MM<QM, QM>MM, or QM~MM, where QM~MM corresponds to cases within 1.5 kcal/mol of each other. The decision tree was created with a maximum depth of 5 and a minimum number of 10 samples in each leaf. The possible divisions were by atom types for S, O (carboxyl, hydroxyl, carbonyl), N (amide NH, Asn or Gln NH2, or ring N) and C (aromatic, sp2, or sp3). These types correspond roughly to those in the earlier atom type figure, Figure S16. The decision tree should be read from top to bottom with the right arrow corresponding to true and left corresponding to false Each leaf lists the number of samples in each category and is colored according to MM-lean (blue), mixed (green), or QM lean (red).

Table S5. Val, Ile, or Leu close contacts with Asp or Glu statistics in the **FD** data set and **HR** subset.

Category	FD set (#, % of total)	HR subset (#, % of total)
VIL-DE containing CCs	758 (1.0%)	26 (1.1%)
VIL-E C to O^- CCs	60 (0.1%)	6 (0.3%)
VIL-D C to O^- CCs	32 (0.04%)	4 (0.17%)
VIL-DE C to O^- CCs	92 (0.12%)	10 (0.42%)
Subset computed with QM and MM	14	10

Table S6. Val, Ile, or Leu CCs with Asp or Glu characterized with QM and MM. Bolded lines are favorable interactions with SAPT0.

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance		
PDBID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	total	ele	vdW	total	C-O	H···O	C-H···O
1yqs	**HR**	**ILE**	**ASP**	**36**	**31**	**-9.8**	**20.1**	**-7.3**	**-5.2**	**-2.3**	**-3.7**	**22.3**	**18.6**	**2.71**	**1.90**	**128.4**
1e9g	**HR**	**ILE**	**GLU**	**69**	**63**	**-9.4**	**20.2**	**-6.7**	**-5.4**	**-1.3**	**-1.5**	**10.1**	**8.6**	**2.63**	**2.20**	**125.7**
4eir	**HR**	**VAL**	**ASP**	**121**	**117**	**-16.9**	**39.8**	**-12.8**	**-10.3**	**-0.3**	**-1.1**	**26.8**	**25.6**	**2.65**	**1.90**	**123.7**
1p9i	**HR**	**LEU**	**GLU**	**8**	**12**	**-10.3**	**28.2**	**-8.5**	**-9.6**	**-0.2**	**0.8**	**11.4**	**12.2**	**2.60**	**2.05**	**107.7**
5btw	HR	VAL	GLU	21	25	-30.6	54.8	-11.2	-10.6	2.3	-12.4	50.5	38.1	2.40	1.93	103.6
1n3l	HR	LEU	GLU	245	314	-17.1	37.6	-8.4	-9.2	2.9	-5.2	26.0	20.8	2.46	1.94	98.7
4kef	HR	VAL	ASP	136	82	-6.3	22.3	-6.3	-4.9	4.7	1.3	12.3	13.6	2.54	2.10	101.0
3ayj	HR	VAL	ASP	292	290	-49.4	90.0	-17.3	-16.0	7.3	-17.1	100.1	83.0	2.25	1.78	138.8
1n3l	HR	VAL	GLU	241	314	-25.6	64.9	-13.1	-10.6	15.5	-1.2	78.7	77.5	2.21	1.95	153.6
1w0n	HR	ILE	GLU	12	128	-33.1	74.7	-12.7	-12.7	16.1	-3.5	83.2	79.7	2.26	1.84	98.7
3g7g	**FD**	**ILE**	**ASP**	**32**	**30**	**-28.3**	**42.8**	**-13.6**	**-10.8**	**-9.8**	**-15.5**	**35.1**	**19.6**	**2.62**	**1.89**	**149.8**
2nx4	**FD**	**VAL**	**GLU**	**152**	**139**	**-13.2**	**23.9**	**-7.8**	**-6.2**	**-3.3**	**-4.9**	**14.5**	**9.6**	**2.49**	**2.16**	**94.4**
2wz9	**FD**	**LEU**	**GLU**	**28**	**24**	**-13.7**	**27.3**	**-8.2**	**-8.7**	**-3.3**	**-5.7**	**6.2**	**0.5**	**2.71**	**1.94**	**98.3**
1xte	FD	LEU	GLU	31	38	-23.1	53.0	-10.0	-9.7	10.2	-1.9	45.7	43.8	2.34	1.95	97.0

Table S7. Tyr close contacts with Asn or Gln statistics for the **FD** data set and **HR** subset.

Category	FD set (#, % of total)	HR subset (#, % of total)
Y-N/Q CCs overall	1,179 (1.5%)	32 (1.4%)
Y OH to N/Q O= CCs	609 (0.8%)	16 (0.7%)
Y OH to N/Q NH_2 CCs	245 (0.3%)	8 (0.3%)
Y-N/Q SC-SC CCs total	854 (1.1%)	24 (1.0%)
QM characterized SC-SC CCs	82 (24 N / 58 O)	24 (8 N / 16 O)
O-N CCs favorable/unfavorable	13 / 11	5 / 3
O-O CCs favorable/unfavorable	26 / 32	7 / 9

Table S8. Tyr CCs with Asn or Gln characterized with QM and MM. Bolded lines are favorable interactions with SAPT0. The nature of the closest contact atom (Asn/Gln O vs. N) is indicated as "X" in the CC distance column along with the distance to the Tyr hydroxyl (O-X, in Å) and the distance to the closest hydrogen atom. The X-H-O angle involved is indicated in ° as well. In cases where a secondary species (i.e., N from Asn/Gln in addition to O from Asn/Gln) forms a possible HB, the non-covalent (Y-O) distance is indicated in Å along with the Y-H-O angle in °.

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance						
PDB ID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	total	ele	vdW	total	X	O-X	X/O…H	X-H-O	Y	Y-O	Y-H-O
2w5q	**HR**	**TYR**	**GLN**	**455**	**399**	**-24.3**	**27.2**	**-10.4**	**-5.6**	**-13.1**	**-13.0**	**2.2**	**-10.8**	**O**	**2.58**	**1.60**	**166.2**	**N**	**2.4**	**127.1**
3uaw	**HR**	**ASN**	**TYR**	**222**	**27**	**-23.9**	**28.1**	**-10.6**	**-5.8**	**-12.2**	**-13.6**	**2.4**	**-11.2**	**O**	**2.57**	**1.58**	**171.5**	**N**	**2.6**	**123.6**
4bpz	**HR**	**ASN**	**TYR**	**346**	**222**	**-19.2**	**24.8**	**-8.6**	**-6.6**	**-9.7**	**-9.9**	**0.8**	**-9.2**	**O**	**2.58**	**1.62**	**163.3**			
3vor	**HR**	**GLN**	**TYR**	**56**	**25**	**-22.4**	**29.4**	**-10.5**	**-6.2**	**-9.7**	**-12.8**	**4.5**	**-8.3**	**O**	**2.48**	**1.54**	**159.7**			
4ue0	**HR**	**TYR**	**GLN**	**499**	**489**	**-16.7**	**21.9**	**-7.5**	**-6.8**	**-9.1**	**-9.1**	**0.7**	**-8.4**	**O**	**2.58**	**1.65**	**155.0**			
4usa	**HR**	**TYR**	**GLN**	**766**	**578**	**-15.6**	**20.3**	**-7.5**	**-4.8**	**-7.7**	**-9.2**	**1.7**	**-7.5**	**O**	**2.58**	**1.64**	**157.6**			
4b5o	**HR**	**TYR**	**GLN**	**171**	**141**	**-27.9**	**43.9**	**-14.8**	**-8.3**	**-7.1**	**-13.5**	**7.8**	**-5.6**	**O**	**2.43**	**1.45**	**168.0**			
3a72	**HR**	**GLN**	**TYR**	**204**	**188**	**-15.9**	**24.8**	**-6.3**	**-5.8**	**-3.3**	**-7.6**	**5.8**	**-1.7**	**N**	**2.59**	**1.66**	**151.5**			

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance						
PDB ID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	total	ele	vdW	total	X	O-X	X/O…H	X-H-O	Y	Y-O	Y-H-O
2ii2	**HR**	**TYR**	**GLN**	**223**	**202**	**-16.8**	**28.1**	**-6.9**	**-7.2**	**-2.7**	**-8.4**	**6.9**	**-1.5**	**N**	**2.59**	**1.62**	**159.3**			
2ij2	**HR**	**TYR**	**GLN**	**305**	**109**	**-20.2**	**33.7**	**-8.5**	**-7.7**	**-2.6**	**-9.2**	**9.2**	**-0.1**	**N**	**2.59**	**1.60**	**162.0**			
2c0r	**HR**	**TYR**	**GLN**	**227**	**80**	**-10.6**	**17.7**	**-3.0**	**-4.7**	**-0.6**	**-5.9**	**4.7**	**-1.2**	**N**	**2.59**	**1.85**	**127.5**			
4yl4	**HR**	**ASN**	**TYR**	**107**	**82**	**-17.9**	**32.1**	**-7.4**	**-7.2**	**-0.4**	**-9.4**	**9.1**	**-0.3**	**N**	**2.54**	**1.61**	**151.7**			
4rj2	HR	GLN	TYR	218	27	-12.2	22.9	-4.6	-5.0	1.1	-4.2	5.5	1.3	N	2.58	1.73	139.1			
4rj2	HR	ASN	TYR	222	27	-15.5	29.7	-7.4	-5.7	1.1	-7.2	10.2	3.0	N	2.56	1.58	162.6			
5kar	HR	GLN	TYR	418	287	-3.0	13.4	-2.1	-6.0	2.3	2.7	2.0	4.7	O	2.54	2.39	32.5	N	121.2	
3w4p	HR	TYR	ASN	251	208	-15.3	28.4	-5.2	-5.0	2.8	-6.5	11.6	5.0	N	2.48	1.67	134.9			
2jju	HR	GLN	TYR	52	50	-0.9	11.5	-1.9	-5.9	2.8	3.1	3.0	6.1	O	2.49	2.40	28.3			
3tu8	HR	TYR	ASN	155	114	0.9	8.0	-1.4	-3.4	4.1	3.7	1.6	5.3	O	2.56	2.56	32.7			
2osx	HR	TYR	ASN	444	406	2.0	7.2	-1.4	-2.9	5.0	4.3	1.8	6.2	O	2.58	2.58	29.7			
5l2v	HR	GLN	TYR	129	92	3.5	6.4	-1.4	-2.8	5.7	5.0	1.9	6.9	O	2.55	2.55	45.0			
4mzc	HR	ASN	TYR	95	42	1.9	9.3	-1.7	-3.1	6.4	4.9	4.9	9.8	O	2.48	2.48	35.5			
4ck4	HR	TYR	GLN	42	35	-3.6	18.8	-2.6	-4.5	8.1	4.5	11.7	16.2	O	2.35	2.35	34.2			
4ypo	HR	GLN	TYR	57	22	-6.2	24.2	-3.0	-4.6	10.4	2.0	26.0	28.0	O	2.22	2.22	67.6			
4qyn	HR	ASN	TYR	79	73	-31.5	67.3	-13.5	-10.3	12.0	-9.8	56.7	46.9	O	2.10	1.44	119.5			
4bqe	**FD**	**TYR**	**GLN**	**788**	**671**	**-25.9**	**31.5**	**-12.3**	**-6.0**	**-12.8**	**-13.5**	**3.1**	**-10.4**	**O**	**2.53**	**1.54**	**171.6**	**N**	**2.7**	**121.4**
3kru	**FD**	**TYR**	**GLN**	**157**	**56**	**-24.6**	**28.8**	**-10.4**	**-6.5**	**-12.7**	**-13.3**	**2.3**	**-11.0**	**O**	**2.56**	**1.60**	**160.7**			
2cgq	**FD**	**TYR**	**ASN**	**66**	**6**	**-23.8**	**27.5**	**-10.8**	**-5.7**	**-12.7**	**-13.2**	**2.0**	**-11.2**	**O**	**2.57**	**1.58**	**174.0**			
3lyn	**FD**	**TYR**	**ASN**	**102**	**54**	**-24.3**	**29.4**	**-11.0**	**-5.9**	**-11.8**	**-13.1**	**2.7**	**-10.4**	**O**	**2.55**	**1.57**	**171.7**	**N**	**2.8**	**119.2**
1orr	**FD**	**ASN**	**TYR**	**265**	**129**	**-22.5**	**27.5**	**-10.3**	**-6.3**	**-11.5**	**-11.5**	**1.5**	**-9.9**	**O**	**2.58**	**1.60**	**167.7**			
4qdc	**FD**	**TYR**	**GLN**	**244**	**131**	**-26.2**	**34.0**	**-12.1**	**-6.8**	**-11.1**	**-14.4**	**4.4**	**-9.9**	**O**	**2.52**	**1.55**	**163.3**			
5tdx	**FD**	**GLN**	**TYR**	**215**	**201**	**-21.2**	**28.0**	**-10.1**	**-7.7**	**-10.9**	**-10.0**	**0.8**	**-9.2**	**O**	**2.58**	**1.60**	**167.3**			
2je8	**FD**	**GLN**	**TYR**	**418**	**406**	**-20.6**	**26.7**	**-10.0**	**-5.5**	**-9.4**	**-12.7**	**3.4**	**-9.3**	**O**	**2.52**	**1.56**	**164.9**			
1uqr	**FD**	**TYR**	**ASN**	**132**	**8**	**-17.2**	**20.5**	**-8.0**	**-4.6**	**-9.2**	**-9.8**	**1.9**	**-7.9**	**O**	**2.57**	**1.63**	**159.9**			
3u80	**FD**	**TYR**	**ASN**	**135**	**8**	**-15.8**	**18.7**	**-7.3**	**-4.3**	**-8.8**	**-10.8**	**1.9**	**-8.9**	**O**	**2.58**	**1.62**	**165.1**			
3wq6	**FD**	**TYR**	**ASN**	**460**	**417**	**-19.6**	**27.5**	**-9.8**	**-6.7**	**-8.7**	**-9.2**	**1.9**	**-7.3**	**O**	**2.54**	**1.58**	**163.3**			
2iu5	**FD**	**GLN**	**TYR**	**122**	**73**	**-17.4**	**23.2**	**-8.5**	**-6.0**	**-8.7**	**-10.5**	**1.4**	**-9.1**	**O**	**2.58**	**1.62**	**162.6**			
4uop	**FD**	**GLN**	**TYR**	**388**	**386**	**-16.6**	**24.1**	**-8.7**	**-7.3**	**-8.5**	**-10.6**	**1.3**	**-9.4**	**O**	**2.54**	**1.60**	**158.8**			
2j1g	**FD**	**ASN**	**TYR**	**249**	**242**	**-15.9**	**19.7**	**-7.6**	**-4.6**	**-8.5**	**-9.8**	**1.9**	**-7.9**	**O**	**2.58**	**1.63**	**160.1**			
2j6l	**FD**	**ASN**	**TYR**	**178**	**123**	**-17.1**	**21.7**	**-7.6**	**-5.3**	**-8.3**	**-9.3**	**2.2**	**-7.1**	**O**	**2.55**	**1.64**	**151.9**			
4m7h	**FD**	**ASN**	**TYR**	**418**	**411**	**-16.3**	**20.6**	**-7.9**	**-4.6**	**-8.2**	**-10.2**	**2.5**	**-7.8**	**O**	**2.55**	**1.61**	**158.7**			
1t3i	**FD**	**TYR**	**ASN**	**393**	**34**	**-17.6**	**23.8**	**-8.9**	**-5.5**	**-8.1**	**-11.2**	**3.1**	**-8.0**	**O**	**2.52**	**1.57**	**163.6**			
4qp5	**FD**	**TYR**	**GLN**	**327**	**325**	**-17.7**	**26.4**	**-8.8**	**-7.2**	**-7.3**	**-12.8**	**2.9**	**-9.9**	**O**	**2.48**	**1.55**	**156.8**			
3ahx	**FD**	**TYR**	**ASN**	**396**	**353**	**-16.6**	**22.7**	**-7.2**	**-6.1**	**-7.2**	**-9.2**	**2.8**	**-6.4**	**O**	**2.53**	**1.67**	**143.3**			
3jsy	**FD**	**TYR**	**ASN**	**199**	**77**	**-17.8**	**21.3**	**-4.9**	**-5.0**	**-6.4**	**-9.6**	**5.1**	**-4.5**	**N**	**2.60**	**1.81**	**132.6**			
3sma	**FD**	**TYR**	**ASN**	**261**	**189**	**-29.8**	**46.2**	**-12.6**	**-7.4**	**-3.6**	**-13.7**	**22.0**	**8.3**	**N**	**2.44**	**1.50**	**151.8**			
2xvm	**FD**	**TYR**	**ASN**	**160**	**130**	**-19.1**	**33.2**	**-8.5**	**-7.1**	**-1.4**	**-8.8**	**10.8**	**2.0**	**N**	**2.55**	**1.56**	**166.1**			
3k86	**FD**	**TYR**	**GLN**	**1171**	**1096**	**-16.5**	**28.3**	**-5.9**	**-7.1**	**-1.2**	**-7.9**	**7.4**	**-0.5**	**N**	**2.58**	**1.66**	**148.0**			
2rci	**FD**	**TYR**	**GLN**	**169**	**149**	**-13.3**	**21.7**	**-4.6**	**-4.9**	**-1.1**	**-6.6**	**5.6**	**-0.9**	**N**	**2.60**	**1.71**	**145.4**			
3phs	**FD**	**TYR**	**GLN**	**68**	**66**	**-14.3**	**29.2**	**-7.6**	**-8.4**	**-1.1**	**-6.0**	**2.4**	**-3.6**	**N**	**2.60**	**1.70**	**151.7**			
2c61	**FD**	**TYR**	**GLN**	**320**	**299**	**-15.1**	**25.2**	**-5.7**	**-5.5**	**-1.1**	**-7.3**	**7.2**	**-0.2**	**N**	**2.59**	**1.66**	**150.9**			
5nn4	**FD**	**TYR**	**ASN**	**438**	**403**	**-16.8**	**30.2**	**-7.1**	**-7.3**	**-1.0**	**-8.0**	**8.2**	**0.2**	**N**	**2.60**	**1.61**	**164.1**			
1ex0	FD	GLN	TYR	313	283	-7.4	13.0	-1.8	-3.6	0.2	-2.4	4.2	1.7	N	2.60	2.06	111.4			
3rv1	FD	TYR	ASN	112	28	-18.9	34.3	-8.9	-5.8	0.6	-8.9	14.3	5.4	N	2.52	1.52	169.5			
4hvc	FD	TYR	GLN	1193	1170	-16.0	28.9	-6.4	-5.6	1.0	-7.0	9.8	2.8	N	2.55	1.60	153.4			
2osa	FD	GLN	TYR	444	360	-22.6	45.2	-12.5	-8.7	1.3	-8.8	10.0	1.2	N	2.46	1.54	154.1			
1yzf	FD	TYR	GLN	150	125	-19.4	39.4	-9.4	-7.4	3.1	-9.0	18.1	9.1	N	2.50	1.52	162.4			
3a68	FD	GLN	TYR	102	41	1.0	7.4	-1.7	-3.3	3.5	4.2	2.1	6.3	O	2.58	2.44	29.2			
3dal	FD	TYR	ASN	148	141	1.0	7.3	-1.6	-2.7	4.0	4.1	2.4	6.5	O	2.57	2.57	30.3			
5jbo	FD	TYR	ASN	431	385	0.2	10.1	-1.5	-4.7	4.1	2.9	1.4	4.3	O	2.56	2.56	36.3			
4ts4	FD	ASN	TYR	81	73	-1.5	11.2	-1.9	-3.5	4.3	3.4	2.3	5.7	O	2.57	2.33	129.9			
4ge6	FD	TYR	ASN	357	331	0.9	7.8	-1.4	-2.9	4.5	3.7	2.0	5.8	O	2.57	2.57	34.4			

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance						
PDB ID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	total	ele	vdW	total	X	O-X	X/O…H	X-H-O	Y	Y-O	Y-H-O
1b2p	FD	TYR	GLN	45	37	0.9	8.5	-1.9	-3.0	4.5	4.4	3.2	7.6	O	2.57	2.39	29.9			
4ptx	FD	TYR	ASN	398	355	0.4	9.8	-1.5	-4.1	4.6	2.9	2.4	5.4	O	2.55	2.16	48.8			
1on3	FD	TYR	GLN	396	367	-1.1	12.3	-1.8	-4.8	4.7	3.7	2.8	6.6	O	2.55	2.54	31.0			
5ix8	FD	TYR	GLN	127	102	2.0	7.9	-1.4	-3.8	4.7	4.0	1.3	5.3	O	2.57	2.34	43.6			
3t6v	FD	TYR	GLN	96	85	-0.3	10.3	-1.9	-3.3	4.7	4.0	3.4	7.4	O	2.54	2.54	28.5			
2jgn	FD	TYR	GLN	524	416	0.4	9.9	-1.8	-3.6	5.0	4.3	3.3	7.5	O	2.52	2.49	28.1			
3eja	FD	GLN	TYR	138	2	1.4	9.0	-1.6	-3.8	5.1	5.0	2.1	7.2	O	2.56	2.31	29.7			
4duh	FD	TYR	GLN	145	143	1.5	11.6	-1.9	-5.9	5.3	4.1	1.7	5.8	O	2.50	2.50	44.8			
4e2u	FD	TYR	ASN	28	12	-1.9	13.6	-1.9	-4.1	5.7	3.2	6.7	9.9	O	2.46	2.46	31.9			
4dr0	FD	TYR	GLN	142	69	2.7	8.3	-1.4	-3.0	6.6	5.2	1.7	6.9	O	2.56	2.56	40.9			
3ezw	FD	GLN	TYR	422	265	2.3	8.6	-1.5	-2.8	6.6	4.5	2.9	7.4	O	2.54	2.54	43.3			
3tg7	FD	TYR	ASN	916	650	1.8	12.1	-2.2	-4.8	6.9	5.8	4.3	10.1	O	2.51	2.04	30.3			
3dhz	FD	TYR	GLN	152	80	1.9	10.0	-1.6	-3.3	7.0	5.4	3.2	8.7	O	2.51	2.51	35.4			
3qwb	FD	TYR	ASN	93	42	1.4	10.4	-1.7	-3.0	7.1	4.8	3.5	8.3	O	2.52	2.52	34.8	N	2.7	123.4
4gco	FD	TYR	ASN	166	151	-22.2	47.0	-9.0	-8.8	7.1	-7.7	27.6	19.9	N	2.36	1.51	141.0			
3g39	FD	TYR	ASN	137	110	-6.2	20.1	-2.3	-4.4	7.2	1.0	16.4	17.3	O	2.31	2.31	70.8			
2e7v	FD	ASN	TYR	66	61	0.1	13.9	-2.2	-4.6	7.2	5.0	8.3	13.3	O	2.40	2.40	30.3			
5crb	FD	TYR	ASN	83	52	2.5	10.2	-1.6	-2.6	8.5	5.4	6.6	12.0	O	2.44	2.44	49.1			
5a9b	FD	ASN	TYR	196	189	-0.7	17.9	-2.3	-3.9	11.0	5.7	13.7	19.4	O	2.32	2.32	50.7			
2nx4	FD	ASN	TYR	123	61	-18.5	45.1	-5.6	-8.3	12.7	-5.3	38.4	33.1	N	2.25	1.65	114.7			
2d69	FD	GLN	TYR	309	248	-23.9	57.2	-5.3	-8.6	19.4	-2.1	71.8	69.6	N	2.14	1.72	100.8			

Table S9. Average and standard deviation of energetic (SAPT0, MM) all four "double HB" configurations as well as for the four lowest energy N or O HB configurations.

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance				
PDB ID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	total	vdW	ele	total	O-X	X/O…H	X-H-O	Y-H	Y-H-O
Double HB																		
2w5q	HR	TYR	GLN	455	399	-24.3	27.2	-10.4	-5.6	-13.1	2.2	-13.0	-10.8	2.58	1.60	166.2	2.4	127.1
3uaw	HR	ASN	TYR	222	27	-23.9	28.1	-10.6	-5.8	-12.2	2.4	-13.6	-11.2	2.57	1.58	171.5	2.6	123.6
4bqe	FD	TYR	GLN	788	671	-25.9	31.5	-12.3	-6.0	-12.8	3.1	-13.5	-10.4	2.53	1.54	171.6	2.7	121.4
3lyn	FD	TYR	ASN	102	54	-24.3	29.4	-11.0	-5.9	-11.8	2.7	-13.1	-10.4	2.55	1.57	171.7	2.8	119.2
					avg	**-24.6**	**29.0**	**-11.1**	**-5.8**	**-12.5**	**2.6**	**-13.3**	**-10.7**	**2.56**	**1.57**	**170.3**	**2.6**	**122.8**
					stdev	**0.8**	**1.6**	**0.7**	**0.2**	**0.5**	**0.3**	**0.3**	**0.3**	**0.02**	**0.02**	**2.3**	**0.1**	**2.9**
N HB																		
3a72	HR	GLN	TYR	204	188	-15.9	24.8	-6.3	-5.8	-3.3	5.8	-7.6	-1.7	2.59	1.66	151.5		
2ii2	HR	TYR	GLN	223	202	-16.8	28.1	-6.9	-7.2	-2.7	6.9	-8.4	-1.5	2.59	1.62	159.3		
3jsy	FD	TYR	ASN	199	77	-17.8	21.3	-4.9	-5.0	-6.4	5.1	-9.6	-4.5	2.60	1.81	132.6		
3sma	FD	TYR	ASN	261	189	-29.8	46.2	-12.6	-7.4	-3.6	22.0	-13.7	8.3	2.44	1.50	151.8		
					avg	**-20.1**	**30.1**	**-7.7**	**-6.3**	**-4.0**	**10.0**	**-9.8**	**0.1**	**2.56**	**1.65**	**148.8**		
					stdev	**5.7**	**9.6**	**2.9**	**1.0**	**1.4**	**7.0**	**2.4**	**4.9**	**0.07**	**0.11**	**9.9**		
O HB																		
3kru	FD	TYR	GLN	157	56	-24.6	28.8	-10.4	-6.5	-12.7	2.3	-13.3	-11.0	2.56	1.60	160.7		
2cgq	FD	TYR	ASN	66	6	-23.8	27.5	-10.8	-5.7	-12.7	2.0	-13.2	-11.2	2.57	1.58	174.0		
1orr	FD	ASN	TYR	265	129	-22.5	27.5	-10.3	-6.3	-11.5	1.5	-11.5	-9.9	2.58	1.60	167.7		
4qdc	FD	TYR	GLN	244	131	-26.2	34.0	-12.1	-6.8	-11.1	4.4	-14.4	-9.9	2.52	1.55	163.3		
					avg	**-24.3**	**29.5**	**-10.9**	**-6.3**	**-12.0**	**2.6**	**-13.1**	**-10.5**	**2.55**	**1.58**	**166.4**		
					stdev	**1.4**	**2.7**	**0.7**	**0.4**	**0.7**	**1.1**	**1.0**	**0.6**	**0.02**	**0.02**	**5.0**		

Table S10. Met or Cys close contact statistics in the **FD** data set and **HR** subset.

Category	FD set (#, % of total)	HR subset (#, % of total)
Met/Cys containing CCs	2,086 (2.7%)	65 (2.7%)
Met/Cys SC-X CCs	904 (1.1%)	36 (1.5%)
Subset computed with QM and MM	238	17
Met/Cys S in the CC	423 (0.5%)	17 (0.7%)
n+ (Cys/Met-Lys/Arg)	19	2
n- (Cys/Met-Asp/Glu)	17	1
nn (Cys/Met-Other)	202	14
Met/Cys S in closest CC for QM and MM	55	17
n+ (Cys/Met-Lys/Arg)	4	2
n- (Cys/Met-Asp/Glu)	2	1
nn (Cys/Met-Other)	49	14
S-C	20	9
S-O	6	1
S-N	24	7
S-S	5	0

Table S11. Met or Cys CCs characterized with QM and MM. Bolded lines are favorable interactions with SAPT0.

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance		
PDBID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	tot	ele	vdw	total	X	S-X	H…X
5u64	HR	MET	PHE	82	67	-9.9	26.9	-3.2	-10.0	3.8	-1.3	9.3	8.0	C	2.92	2.18
2vy8	HR	MET	LEU	603	599	-14.1	33.3	-4.1	-9.4	5.8	-3.4	12.0	8.6	O	2.77	2.07
3b12	HR	PHE	MET	264	262	-13.3	34.9	-3.7	-11.4	6.4	-1.0	7.2	6.2	C	2.84	2.18
3szh	HR	ASN	MET	24	10	-1.5	18.9	-3.0	-6.8	7.5	3.2	0.4	3.5	O	2.79	2.47
3w42	HR	LEU	MET	38	31	-8.5	25.3	-2.2	-6.2	8.5	0.2	12.4	12.6	C	2.88	2.29
3r87	HR	SER	MET	79	1	-7.8	27.3	-3.1	-7.5	8.9	2.3	5.2	7.6	O	2.73	2.02
1nww	HR	LEU	MET	103	78	-10.7	29.7	-2.5	-7.3	9.1	0.7	12.3	12.9	C	2.88	2.29
3wwl	HR	ASP	MET	38	1	-14.3	37.7	-5.2	-7.8	10.4	-1.5	5.9	4.4	N	2.76	2.08
3ned	HR	LYS	MET	168	141	-37.5	88.2	-13.0	-19.0	18.7	-6.2	111.9	105.8	C	2.71	1.81
4yaa	HR	MET	GLN	328	320	-28.1	75.6	-8.1	-16.6	23.0	-0.8	92.2	91.4	C	2.68	1.94
5lun	**HR**	**CYS**	**THR**	**99**	**74**	**-6.9**	**11.5**	**-1.8**	**-3.1**	**-0.3**	**-0.3**	**6.8**	**6.5**	**O**	**2.70**	**2.70**
4oo4	HR	CYS	ILE	62	5	-6.1	13.9	-2.3	-3.5	2.0	0.2	4.5	4.6	O	2.72	2.54
1q6o	HR	THR	CYS	114	88	-6.5	17.4	-2.5	-5.8	2.6	0.4	2.1	2.5	O	2.82	2.18
4yaa	HR	LEU	CYS	427	418	-5.3	16.3	-1.9	-4.5	4.5	0.1	9.6	9.7	C	2.89	2.25
1y55	HR	CYS	CYS	81	4	-9.1	25.0	-2.9	-8.4	4.5	-1.7	8.9	7.2	C	2.87	2.00
5lun	HR	CYS	SER	99	76	-9.6	27.1	-3.2	-6.0	8.3	1.9	15.7	17.6	O	2.55	2.43
2bog	HR	CYS	GLY	80	77	-8.7	27.4	-4.0	-6.0	8.6	4.4	19.4	23.8	O	2.47	2.47
4iv6	FD	MET	MET	358	79	-13.3	33.2	-3.8	-7.2	8.8	1.3	9.0	10.3	S	2.87	2.77
3txs	FD	MET	MET	104	57	-17.5	45.4	-3.3	-9.6	15.0	-0.6	15.1	14.5	C	2.75	2.25
1ytl	FD	MET	THR	125	92	-9.6	20.2	-3.0	-7.1	0.5	2.4	9.9	12.3	O	2.63	2.33
1ne9	FD	GLN	MET	28	18	-3.4	12.0	-2.0	-4.4	2.2	1.6	2.1	3.7	O	2.77	2.31

						SAPT (kcal/mol)					MM (kcal/mol)			CC distance		
PDBID	type	Res1	Res2	#1	#2	elec	ex	ind	disp	tot	ele	vdw	total	X	S-X	H…X
2j6y	FD	GLN	MET	85	82	-8.0	24.6	-4.8	-8.9	2.9	0.0	3.9	3.9	O	2.70	1.93
3kzu	FD	MET	PRO	239	200	-3.8	14.7	-2.1	-5.4	3.4	3.0	1.4	4.4	O	2.78	2.71
3q9d	FD	CYS	MET	115	108	-10.1	23.7	-3.1	-6.5	4.1	1.8	4.7	6.5	S	2.99	2.50
4k02	FD	MET	PHE	62	22	-10.2	25.1	-2.3	-7.4	5.2	0.6	4.1	4.7	C	2.91	2.72
4wum	FD	CYS	MET	341	337	-10.6	28.3	-4.2	-8.4	5.2	-1.6	2.8	1.2	N	2.82	2.07
2r6u	FD	MET	GLU	63	9	-8.3	36.3	-10.2	-10.8	7.0	9.0	15.0	24.1	O	2.64	1.92
5f30	FD	ILE	MET	447	438	-12.7	35.8	-3.3	-8.9	11.0	1.0	18.9	19.9	C	2.90	1.98
1kmt	FD	MET	VAL	126	123	-13.4	36.4	-3.1	-8.2	11.7	-0.8	23.4	22.6	C	2.70	2.29
2gz4	FD	LYS	MET	100	95	-37.6	77.0	-13.8	-11.7	13.8	-9.6	38.4	28.7	N	2.43	2.04
2y4y	FD	GLY	MET	146	104	-9.3	35.3	-4.4	-7.3	14.3	3.7	12.5	16.2	O	2.60	2.24
3t30	FD	MET	CYS	90	50	-25.3	64.0	-5.3	-12.3	21.2	-0.2	24.4	24.2	S	2.84	1.61
3a54	**FD**	**CYS**	**GLN**	**156**	**47**	**-11.5**	**17.9**	**-3.1**	**-8.1**	**-4.8**	**-4.8**	**2.5**	**-2.3**	**C**	**2.96**	**2.19**
2wpg	**FD**	**LYS**	**CYS**	**321**	**174**	**-22.5**	**45.3**	**-17.1**	**-7.5**	**-1.9**	**-4.9**	**7.9**	**2.9**	**N**	**2.78**	**1.89**
2gsj	FD	CYS	GLY	67	63	-4.6	11.1	-2.1	-4.0	0.4	1.2	2.7	3.8	O	2.82	2.33
1shm	FD	CYS	CYS	92	22	-12.4	25.1	-3.2	-7.4	2.2	0.3	4.7	5.1	S	3.02	2.39
1j7g	FD	CYS	ILE	113	110	-13.6	27.1	-3.9	-7.3	2.3	-1.3	6.7	5.4	O	2.67	2.32
3sb4	FD	CYS	PHE	292	266	-3.2	12.2	-2.1	-3.5	3.4	2.8	5.8	8.6	O	2.68	2.68
4b89	FD	ARG	CYS	941	899	-20.9	41.5	-5.3	-11.6	3.7	-5.8	23.0	17.2	C	2.61	2.61
3qpb	FD	SER	CYS	195	98	-18.5	35.5	-4.4	-8.4	4.2	-3.1	9.0	5.9	O	2.67	2.32
2v5m	FD	CYS	CYS	370	329	-15.2	31.7	-4.2	-7.9	4.5	0.6	8.2	8.8	S	2.90	2.41
5ifi	FD	CYS	SER	207	163	-9.7	22.4	-2.6	-5.6	4.5	-2.6	2.9	0.4	O	2.79	2.03
2oit	FD	VAL	CYS	198	168	-9.2	23.7	-2.7	-7.0	4.9	-0.9	12.5	11.6	C	2.86	2.16
4nac	FD	CYS	THR	100	96	-15.2	30.7	-3.2	-7.2	5.0	-1.5	12.2	10.7	O	2.65	2.20
1o0w	FD	CYS	PHE	31	27	-4.3	21.9	-2.9	-8.7	6.0	4.6	1.9	6.5	C	2.94	2.63
3qli	FD	CYS	PRO	235	231	-4.5	18.8	-2.8	-4.1	7.3	3.9	10.1	14.0	O	2.58	2.58
5f0v	FD	ILE	CYS	259	4	-8.0	23.3	-2.1	-5.6	7.7	0.1	12.6	12.7	C	2.76	2.44
3qpb	FD	ALA	CYS	194	98	-9.5	25.3	-1.8	-5.8	8.1	-0.7	5.5	4.8	C	2.97	2.16
2zej	FD	CYS	LEU	1465	1421	-7.7	23.8	-1.8	-6.0	8.4	0.8	6.2	7.0	C	2.96	2.17
5dii	FD	CYS	SER	182	135	-14.5	33.0	-3.8	-4.9	9.8	1.4	27.0	28.4	O	2.46	2.41
2waa	FD	CYS	CYS	183	144	-45.7	98.8	-11.9	-15.8	25.4	0.9	50.6	51.5	S	2.54	2.14
2v5m	FD	CYS	CYS	287	234	-43.8	95.8	-7.1	-12.6	32.3	1.1	34.3	35.4	S	2.62	2.13
2nw0	FD	CYS	VAL	170	5	-38.2	99.4	-11.4	-14.6	35.1	0.3	156.4	156.7	C	2.31	1.98
2v5m	FD	CYS	CYS	182	125	-55.1	120.3	-8.1	-13.9	43.2	2.1	50.5	52.6	S	2.54	2.24
3bzt	FD	CYS	SER	269	256	-79.2	172.6	-31.3	-15.4	46.7	5.1	874.0	879.2	O	1.87	1.87

References

1. Yang, H.; Peisach, E.; Westbrook, J. D.; Young, J.; Berman, H. M.; Burley, S. K., Dcc: A Swiss Army Knife for Structure Factor Analysis and Validation. *Journal of Applied Crystallography* **2016**, *49*, 1081-1084.
2. Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S., Covalent Radii Revisited. *Dalton Transactions* **2008**, 2832-2838.
3. Gilli, P.; Pretto, L.; Bertolasi, V.; Gilli, G., Predicting Hydrogen-Bond Strengths from Acid–Base Molecular Properties. The pKa Slide Rule: Toward the Solution of a Long-Lasting Problem. *Accounts of Chemical Research* **2009**, *42*, 33-44.

Other files

energiesandstructures.zip (14.02 MiB) view on ChemRxiv • download file