Title: Protein interface prediction using solvent accessibility of unbound residues

Authors: Jennifer C. Mortensen, Thom Vreven, and Zhiping Weng*

Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605

*Zhiping Weng

Program in Bioinformatics and Integrative Biology

University of Massachusetts Medical School

ASC-5th floor room 1069

368 Plantation St.

Worcester, MA 01605

Phone: 508-856-8866; Fax: 508-856-2392

E-mail: Zhiping.Weng@umassmed.edu

Keywords:

Protein-protein interactions; docking; solvent accessible surface area; ZDOCK; structure

Short title:

Interface prediction using solvent accessibility

Acknowledgments:

The authors thank S. Vangaveti and T. M. Borrman and other members of the lab for their helpful discussion. This work was supported by the National Institutes of Health grant GM116960.

Abstract

The prediction of protein-protein interfaces requires both the identification of interface residues and the proper spatial orientation of the component proteins. Many methods have been developed to identify interface residues, often using relative interface propensity (RIP), the enrichment of a particular amino acid type at the interface compared to the rest of the protein surface. We aimed to improve RIP for interface identification by incorporating the solvent accessibility of each amino acid. We studied the surface residues of 290 unbound structures corresponding to components of protein complexes and compared the relative solvent accessible surface area (rSASA) distributions of residues that end up in the interface and those that do not. Our results show that the side-chains of amino acids that become interface residues are more solvent exposed than non-interface surface residues on the unbound protein structure. Using this knowledge, we created an rSASA-dependent probability of becoming an interface residue for each amino acid type. Our results show that the solvent accessible surface area of residues should be taken into account when identifying interface residues and can be applied to other interface prediction techniques that use RIP to improve their results.

Introduction

When proteins bind to each other to form a complex, interactions occur between their residues at the binding interface, and residues that were previously solvent exposed become solvent excluded. The ability to predict which residues will become buried in a protein-protein interaction allows for enhanced interface prediction, mutagenesis to confirm complex structure, and targeted drug discovery.[1] Despite the development of numerous algorithms, the prediction of the residues that are involved in the binding process remains a difficult problem.[1–3] One

commonly used parameter in these algorithms is the relative interface propensity (RIP),[4] which measures the observed enrichment of amino acid types at the interface relative to the protein surface in general. The use of RIP as a predictor for interface residues is, however, intrinsically limited as it only considers the type of the amino acid and is unable to distinguish between residues that share that type. Therefore, RIP is typically used in combination with other features, and in that way has shown varying degrees of success.[4–11]

In this work, we show that residues which are more solvent-exposed in the unbound component proteins are more likely to be found in the interface, but the exact relationship depends strongly on the amino acid type. We use this knowledge to redefine RIP as linearly dependent on the solvent-accessible surface area (SASA) of residues. While prior papers have explored SASA-dependent RIP, there has been no analysis of the improvement of including SASA in the calculations. For example, Liang *et al.* used an accessible surface area dependent propensity based on deviations from the average accessibility of an amino acid type and combined this with a side-chain energy score and conservation score to predict surface residues.[6] Hwang *et al.* weighted RIP values by a probability value for residues in specific SASA ranges and used the weighted propensity with other variables to predict surface patches.[11] In both of these works, interface residues were identified with improved accuracy from prior methods but the impact of the SASA-dependent RIP cannot be distinguished from the other scoring variables. In this work, we show that the new SASA-dependent formalism is a better predictor than standard RIP, and we envision that the performance of interface prediction algorithms can be improved by incorporating our SASA-dependent RIP equations.

Methods:

Dataset:

To avoid potential bias, we created a subset of proteins from docking benchmark 5.0[12] for training our models. The docking benchmark contains 230 complexes with NMR or X-ray crystal structures of the bound and unbound components. We first excluded the NMR structures and the small set of legacy antibody-antigen cases that had no unbound structures for the antibodies. Next, we identified component structures with more than 90% sequence identity to remove redundancy—for each set of structures with high sequence identity, we kept the highest-resolution structure and excluded the remaining from subsequent analysis. To limit our studies to amino acid side-chains with similar backbone conformations in the bound and unbound state, we removed complexes with an interface RMSD $\geq$ 2.5Å between the bound and unbound structures. This resulted in 290 component proteins from 185 complexes with 6,618 interface residues and 53,907 non-interface surface residues. We used this subset of proteins for training our models and tested the models on the entire set of 230 complexes.

We calculated the solvent-accessible surface area (SASA) for each residue in each unbound structure using FreeSASA.[13] To obtain SASA values with high accuracy and independent of structure orientation, FreeSASA was run with a resolution of 200 slices per atom (default is 20). Before calculating SASA for each complex, small molecule ligands were removed, nonstandard amino acids were modified to the closest standard amino acid, and residues with missing backbone atoms were removed from the structure. To calculate the relative solvent accessibility (rSASA), we used a reference dataset of 792 high-resolution protein structures[14,15] to find the maximum SASA value of each amino acid type (see **Table S1**). Outliers with large SASA values

due to nearby missing atoms, unusual backbone configurations and other reasons were removed. We also calculated $rSASA_{SC}$ and $rSASA_{BB}$, the relative side-chain and backbone solvent accessibility, using the maximum side-chain and backbone SASA values in the same dataset.

Interface residues were defined as all residues with a change in SASA $\geq$ 1Å between the component proteins in the complex and separated. For each amino acid type, we created two subsets: interface surface residues and non-interface surface residues, where surface residues have rSASA $\geq$ 0.1 in the unbound component protein. Duplicate residues from proteins containing multiple identical chains were removed.

Analysis of SASA distributions:

We investigated the difference between the rSASA distributions of interface and non-interface surface residues, for residues of any type and for each residue type specifically, using a two-sample Kolmogorov-Smirnov test to determine the significance. Significance values were assessed after controlling the FDR to be below 0.05 using the Benjamini-Hochberg procedure.[16] We repeated this analysis for $rSASA_{SC}$ and $rSASA_{BB}$.

To ensure that the observed differences between interface and non-interface rSASA distributions were not caused by secondary structure, we further divided the residues by loop, α-helix or β-sheet backbone secondary structure assignment by DSSP.[17,18] The rSASA distributions were compared for each backbone category as described above.

SASA-dependent interface probability

We tested several different methods for predicting interface residues. We began by calculating the probability of being a surface residue for each amino acid type in our dataset by using the RIP equations from Yan 2008 and Dong 2007.[19,20] For amino acid type *i*, let N_i be the total number of interface surface residues and M_i be the number of non-interface surface residues across all proteins in our dataset. Among all surface residues of type *i*, the fraction that are at the interface (n_i) or surface (m_i) is calculated as

$$n_i = \frac{N_i}{\Sigma_k N_k} \quad (1)$$

$$m_i = \frac{M_i}{\Sigma_k M_k} \quad (2)$$

where Σ_k is summed over all 20 types of amino acids. The relative interface propensity of amino acid type *i* is then defined as

$$RIP_i = \frac{n_i}{m_i} \quad (3)$$

Furthermore, the probability of a surface residue of amino acid type *i* to be an interface (I) residue is defined as

$$P_i(I) = \frac{N_i}{N_i + M_i} \quad (4)$$

When $N_i << M_i$, RIP_i and $P_i(I)$ are correlated:

$$P_i(I) \approx \frac{\Sigma_k N_k}{\Sigma_k M_k} RIP_i \quad (5)$$

In our dataset, N_i is always much less than M_i, resulting in a linear correlation between RIP_i and $P_i(I)$ (**Figure S4** and **Table S3**). Nonetheless, $P_i(I)$ is a probability and thus more convenient to work with than RIP_i.

We hypothesized that the interface propensity of a particular amino acid is correlated to its solvent accessibility and that SASA in the unbound component protein could be used to improve interface prediction. To test this, we calculated the probability, *P(I)* at a specific SASA, rSASA, or $rSASA_{SC}$ value, across all amino acid types, as well as by amino acid type. To construct the functions, we grouped the residues in ten bins, covering the range from zero to the maximum observed value of SASA, rSASA or $rSASA_{SC}$ for the residue type. Bins containing fewer than 5 interface residues or 20 total surface residues were discarded. We then determined the probability of being an interface residue for each bin and fit a linear function to obtain the following equations:

$$P_{rSASA}(I) = P(I) + 0.10 * rSASA\ - 0.058 \quad (6)$$

$$P_{SASA}(I) = P(I) + \ 0.073 * \frac{SASA}{100}\ + 0.066 \quad (7)$$

where *P(I)* = 0.158, the overall probability of a surface residue being found at an interface. This process was repeated to obtain amino acid dependent equations $P_{i,\,rSASA}(I)$, and $P_{i,\,SC}(I)$, as summarized in Table 1, where $P_{i,SC}(I)$ is the $rSASA_{SC}$ dependent probability. For amino acids without a significant difference in interface and non-interface surface rSASA (or $rSASA_{SC}$) distributions (Figure 1; amino acids Asn, Asp, Gln, Gly, and Ser for rSASA and Ala, Asn, Asp Glu, Gln, Gly, and Ser for $rSASA_{SC}$), the equations for $P_{i,\,rSASA}(I)$ or $P_{i,\,SC}(I)$ were set to a constant value of $P_i(I)$.

Testing prediction accuracy:

We tested the ability of various equations to identify interface residues using the entire dataset of 230 protein complexes. For each equation shown in **Table 1**, probability values were calculated for all surface residues in the unbound component proteins. Surface residues of each protein were ranked by probability and the top T residues selected (T = 1, 2, …, 10). If multiple residues had the same value, they were chosen in a random order. A prediction was considered successful if one or more top residues was a true interface residue in the complex structure. This was repeated for all complexes and the success rate was compared between the probability equations. As a control, T surface residues were also randomly chosen, and the success rate calculated. This process was repeated 50 times and the average success rate and standard deviations were calculated.

Comparison to existing work:

To allow for comparison with prior work, we implemented the methods of Liang *et al.* and Hwang *et al.* and used them to predict interface residues in our dataset. For Liang *et al.*, the residue interface propensity, RIP_L was calculated using

$$RIP_L = C_i * rSASA \quad (8)$$

where C_i is the natural log of the relative interface propensity divided by the average rSASA for residue *i* and is given in Table 1 in ref. Liang 2006.[6] For Hwang *et al.*, RIP_H was calculated using

$$RIP_H = propensity * P_{i,rSASA}(I)_H \quad (9)$$

where the propensity and probability values were taken from Figure 1 in ref. Hwang 2016.[11]

Results and Discussion

The side-chains of interface surface residues are more solvent exposed than non-interface surface residues

Interface forming residues are more solvent exposed in the unbound proteins than other surface residues (**Figure 1A**). All nonpolar residues except Gly showed a significant shift in rSASA distribution of interface residues towards higher rSASA values. Most charged and polar residues had significant shifts in the same direction except for Asn, Asp, Gln, and Ser, for which no significant difference was detected. When the distributions were split by side-chain and backbone rSASA, it became clear that the increase in rSASA was due to an increase in side-chain exposure (**Figure 1B and Figure S1**). Interface residues had a similar shift to higher $rSASA_{SC}$ values as seen for rSASA, with statistically significant shifts for all amino acids seen in the rSASA distributions except for Ala, Gln, and Glu. Gly was omitted in the $rSASA_{SC}$ analysis. In contrast, interface and non-interface surface residues have similar distributions of $rSASA_{BB}$ values for most residues (**Figure S1**).

The increase in solvent exposure of interface residues is not due to an enrichment of residues in a particular secondary structure. Protein-protein interfaces were reported to have different secondary structure composition than the rest of the protein surface, although different studies reported different secondary structure enrichment.[21,22] We split our dataset by backbone type (α-helix, β-sheet, and loop) using DSSP secondary structure assignments (see **Methods**).[17,18] We observed that interface residues were enriched in loops (**Table S2,** $p < 0.00001$). Because different secondary structures can be more solvent exposed than others, we tested whether the higher rSASA values observed in our dataset were due to the enrichment of loops. We observed

a strong difference in rSASA distributions between interface and non-interface surface residues across all three secondary structure types, although some of these shifts were not statistically significant due to small counts for some amino acid types in certain secondary structures, especially β-sheet (**Figure S2**).

The probability of being an interface residue is dependent on both amino acid type and solvent accessibility

The relative interface propensity of amino acids in our dataset shows that interfaces are enriched in large nonpolar and aromatic residues and depleted in some polar and charged amino acids (**Figure 2A**). This agrees well with prior work, and a comparison is provided in **Figure S3**.[5,6,8,9,11,19,20,23–26] Although each prior study found a different set of residues to be enriched at the interface, there are some common themes across all studies. Met, Phe, Trp and Tyr are consistently found to be enriched at the interface while Ala, Asp, Glu, Lys, Pro, Ser and Thr are almost never enriched. Moreover, Arg is found by most studies to be enriched in the interface despite being a charged residue, while the other positively charged residue Lys is not found enriched by any of the studies (**Figure S3**).

To facilitate the use of SASA to improve interface prediction, we have chosen to work with $P_i(I)$, the probability of a residue (i) being found at the interface, instead of directly with RIP_i (**Figure 2B**). For datasets where N_i,the number of interface surface residues of type i, is much smaller than M_i, the number of non-interface surface residues of type i, RIP_i and $P_i(I)$ are linearly related, as follows from EQ5 and shown in **Figure S4**. Because interface residues are shifted to higher rSASA values (**Figure 1**), we began by calculating rSASA- and SASA-dependent interface

probabilities across all amino acids: $P_{rSASA}(I)$ and $P_{SASA}(I)$. As seen in **Figure 3**, there is a clear relationship between rSASA or SASA and the probability of being at an interface. When we consider both amino acid type and solvent accessibility, $P_{i, rSASA}(I)$ and $P_{i, SC}(I)$, the probability of an aromatic or large nonpolar residue being found at the interface is strongly dependent on rSASA while small or charged residues have a very weak relationship (**Figure 4**). (Note: in Figure 4, all lines are plotted, even those without a statistically significant shift in **Figure 1**. However, for predicting interface residues, the equations for these insignificant amino acids will be set to a constant value of $P_i(I)$).

SASA-dependent interface probability results in improved prediction of interface residues by creating a spread of probability values

Using the probability of interface equations as shown in **Table 1** and **Table 2**, we predicted interface residues from the unbound component structures in docking benchmark 5.0 (**Figure 5A**).[12] The best results were obtained using the $P_{i, SC}(I)$ equations, for which a true interface residue was selected as the highest-ranking residue 46% of the time. For comparison, randomly selecting a surface residue identifies an interface residue 15% of the time and SASA-independent P(I) calculations 28% of the time. Our new equations result in an increased success rate for all values of T, the number of predicted interface residues. Using $rSASA_{SC}$ resulted in a slight improvement over rSASA. The slope and intercept values for each $rSASA_{SC}$ equation are listed in Table 2 and can be easily implemented to predict interface residues in other proteins.

In contrast to using a SASA-independent RIP or $P_i(I)$, using $P_{i, SC}(I)$ creates a spread of probability values for each amino acid type (**Figure 6**). The spread of $P_{i, rSASA}(I)$ or $P_{i, SC}(I)$ is

typically centered on the rSASA-independent $P_i(I)$ value with a long tail towards high probability values. By creating a rSASA-dependent probability, residues that have a low interface propensity but a high rSASA value can be correctly identified as interface residues which would not occur using rSASA-independent $P_i(I)$. In fact, using rSASA-independent $P_i(I)$ results in almost exclusive selection of Tyr as the most likely interface residue while the rSASA-dependent methods select a wider variety of amino acid types with some variety between methods (**Figure 7**). The variety in amino acid type, along with the spread of SASA-dependent probability values, improves interface prediction and will likely outperform standard RIP methods when used in conjunction with more complex interface prediction software.

When we compare our results to those obtained using two previously published methods (**Methods**), we find similar prediction accuracies (**Figure 5B and 7**).[6,11] For all three rSASA dependent-equations, prediction accuracy is substantially better than SASA-independent RIP. Liang *et al.* is similar to our approach, creating an rSASA dependent equation. The key difference is that our work uses a linear correction to the SASA-independent equation while Liang *et al.* directly scale the propensity. As a result, our equations have a variety of Y-intercepts and slopes, while Liang's equations all start at the origin. In practice, both approaches result in similar accuracy in predicting an interface residue in the top 1 to 10 interface residues (**Figure 5B**) but the different approaches may impact the results of algorithms that use more than the top 10 residues, such as those that create surface patches. In this work, we show that an rSASA-dependent approach outperforms SASA-independent RIP, validating the use of the equations in Liang *et al*. The equation by Hwang *et al*. is nonparametric (i.e., it tabulates by bin of SASA), so it can be more accurate with more training data. However, it needs more parameters and it is

coarse grain (bins), so some accuracy may be lost, and the bins with few residues yield less reliable statistics. Our approach assumes a linear relationship between $P_i(I)$ and rSASA, yielding equal performance with much fewer parameters.

In conclusion, our analysis of rSASA distributions of amino acids shows that interface surface residues are, on average, more solvent exposed in the unbound complex than the rest of the protein surface. This finding is independent of backbone secondary structure and seen most strongly in large nonpolar and aromatic amino acids. By studying side-chains and backbones separately, we show that the difference in rSASA of interface residues is due to the exposure of side-chain atoms. Our side-chain rSASA-dependent equations for predicting interface residues obtain higher accuracy than standard RIP methods, although the performance of our equations are about the same as two other rSASA-dependent methods.[6,11] While these equations are not sufficient to predict protein complexes, they provide a substantial improvement to the typical relative interface propensity equations used in existing methods.

References:

1. Maheshwari S, Brylinski M. Predicting protein interface residues using easily accessible on-line resources. *Brief Bioinform*. 2015;16(6):1025-1034. doi:10.1093/bib/bbv009
2. Zhou H-X, Qin S. Interaction-site prediction for protein complexes: A critical assessment. *Bioinformatics*. 2007;23(17):2203-2209. doi:10.1093/bioinformatics/btm323
3. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel J-C, Deane CM. Progress and challenges in predicting protein interfaces. *Brief Bioinform*. 2016;17(1):117-131. doi:10.1093/bib/bbv027

4. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*. 1997;272(1):133-143. doi:10.1006/jmbi.1997.1233

5. Neuvirth H, Raz R, Schreiber G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*. 2004;338(1):181-199. doi:10.1016/j.jmb.2004.02.040

6. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*. 2006;34(13):3698-3707. doi:10.1093/nar/gkl454

7. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*. 2005;21(8):1487-1494. doi:10.1093/bioinformatics/bti242

8. Negi SS, Braun W. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J Mol Model*. 2007;13(11):1157-1167. doi:10.1038/jid.2014.371

9. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins Struct Funct Genet*. 2005;61(1):21-35. doi:10.1002/prot.20514

10. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins Struct Funct Genet*. 2001;44(3):336-343. doi:10.1002/prot.1099

11. Hwang H, Petrey D, Honig B. A hybrid method for protein-protein interface prediction. *Protein Sci*. 2016;25(1):159-165. doi:10.1002/pro.2744

12. Vreven T, Moal IH, Vangone A, et al. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version

2. *J Mol Biol*. 2015;427:3031-3041. doi:10.1016/j.jmb.2015.07.016

13. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*. 2016;5:189. doi:10.12688/f1000research.7931.1

14. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*. 1997;6:1661-1681. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2143774/pdf/9260279.pdf.

15. Dunbrack RL. Rotamer libraries in the 21 st century Roland L Dunbrack Jr. *Curr Opin Struct Biol*. 2002;12:431-440. http://linkinghub.elsevier.com/retrieve/pii/S0959440X02003445.

16. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. 1995;57(1):289-300.

17. Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*. 2015;43(D1):D364-D368. doi:10.1093/nar/gku1028

18. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637. doi:10.1002/bip.360221211

19. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. *Protein J*. 2008;27:59-70. doi:10.1007/s10930-007-9108-x

20. Dong Q, Wang X, Lin L, Guan Y. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinformatics*. 2007;8:1-13. doi:10.1186/1471-2105-8-147

21. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein*

Sci. 2006;15(5):1017-1029. doi:10.1110/ps.051589106

22. Guharoy M, Chakrabarti P. Secondary structure based analysis and classification of biological interfaces: Identification of binding motifs in protein-protein interactions. *Bioinformatics*. 2007;23(15):1909-1918. doi:10.1093/bioinformatics/btm274

23. Dai W, Wu A, Ma L, Li YX, Jiang T, Li YY. A novel index of protein-protein interface propensity improves interface residue recognition. *BMC Syst Biol*. 2016;10(112):381-392. doi:10.1186/s12918-016-0351-7

24. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci*. 1996;93(1):13-20. doi:10.1073/PNAS.93.1.13

25. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol*. 2003;325:377-387. doi:10.1016/S0022-2836(02)01223-8

26. Talavera D, Robertson DL, Lovell SC. Characterization of protein-protein interaction interfaces from a single species. *PLoS One*. 2011;6(6):e21053. doi:10.1371/journal.pone.0021053

Tables:

Table 1: Probability equations created to identify interface residues.

Probability of being at the interface	Amino acid dependent?	SASA dependent?	Slope values (F: figure; T: table)
$P_i(I)$	Yes	No	$P_i(I)$ values: 0.11 to 0.25, see F2 and T2
$P_{rSASA}(I)$	No	Yes	0.10, see F3
$P_{SASA}(I)$	No	Yes	0.073 / 100 Å^2, see F3
$P_{i,\,rSASA}(I)$	Yes	Yes	0 to 0.52, see F4
$P_{i,\,SC}(I)$	Yes	Yes	0 to 0.49, see F4 and T2
Random	No	No	none

Table 2: Equations for $P_i(I)$ and $P_{i,\,sc}(I)$

		$P_{i,\,sc}(I)$ Equations	
	$P_i(I)$	slope	y-intercept
ALA	0.107	0	0
ARG	0.155	0.213	-0.085
ASN	0.150	0	0
ASP	0.125	0	0
CYS	0.118	0.274	-0.045
GLN	0.132	0	0
GLU	0.123	0	0
GLY	0.136	0	0
HIS	0.160	0.055	-0.013
ILE	0.156	0.257	-0.066
LEU	0.152	0.364	-0.097
LYS	0.112	0.107	-0.051
MET	0.189	0.326	-0.076
PHE	0.188	0.287	-0.075
PRO	0.114	0.152	-0.058
SER	0.112	0	0
THR	0.127	0.086	-0.022
TRP	0.222	0.494	-0.107
TYR	0.250	0.421	-0.113
VAL	0.122	0.186	-0.051

Figure Legends:

Figure 1: rSASA distributions of amino acids on the protein non-interface surface (blue) and interface surface (red) (intersection between distributions in violet) for (A) full residues and (B) side-chains on unbound protein structures. Asterisks indicate statistically significant differences between non-interface and interface residues. Although we restricted our study to surface residues with rSASA ≥ 0.1, $rSASA_{SC}$ values can range from 0 to 1.

Figure 2: Interface propensity of surface amino acids. (A) Relative interface propensity for each amino acid type as defined by EQ3. (B) Probability of being an interface residue, $P_i(I)$, as defined by EQ4.

Figure 3: Relationship between (A) SASA and (B) rSASA and the probability of being an interface residue across all amino acid types.

Figure 4: Relationship between probability of being an interface residue and (A) rSASA and (B) $rSASA_{SC}$ for each amino acid type. Amino acids are grouped by slope.

Figure 5: Success rate of identifying the true interface by selecting the top T residues after ranking residues by their probability using various interface prediction methods. When two or more residues have the same probability values, they are randomly ranked. The error bars show the standard deviation over 50 runs with different random number seeds.

Figure 6: Distribution of probability values for surface residues (blue) using (A) rSASA and (B) $rSASA_{SC}$ dependent equations compared to standard $P_i(I)$ values (black). Amino acids without blue values have a predicted $P_{i,\,rSASA}(I)$ or $P_{i,\,SC}(I)$ set to $P_i(I)$.

Figure 7: Distribution of amino acid types in the top 1 or top 5 predicted interface residues for $P_{i,\,SC}(I)$, $P_i(I)$ and using data from Ref. 11 and Ref. 6.

Figures

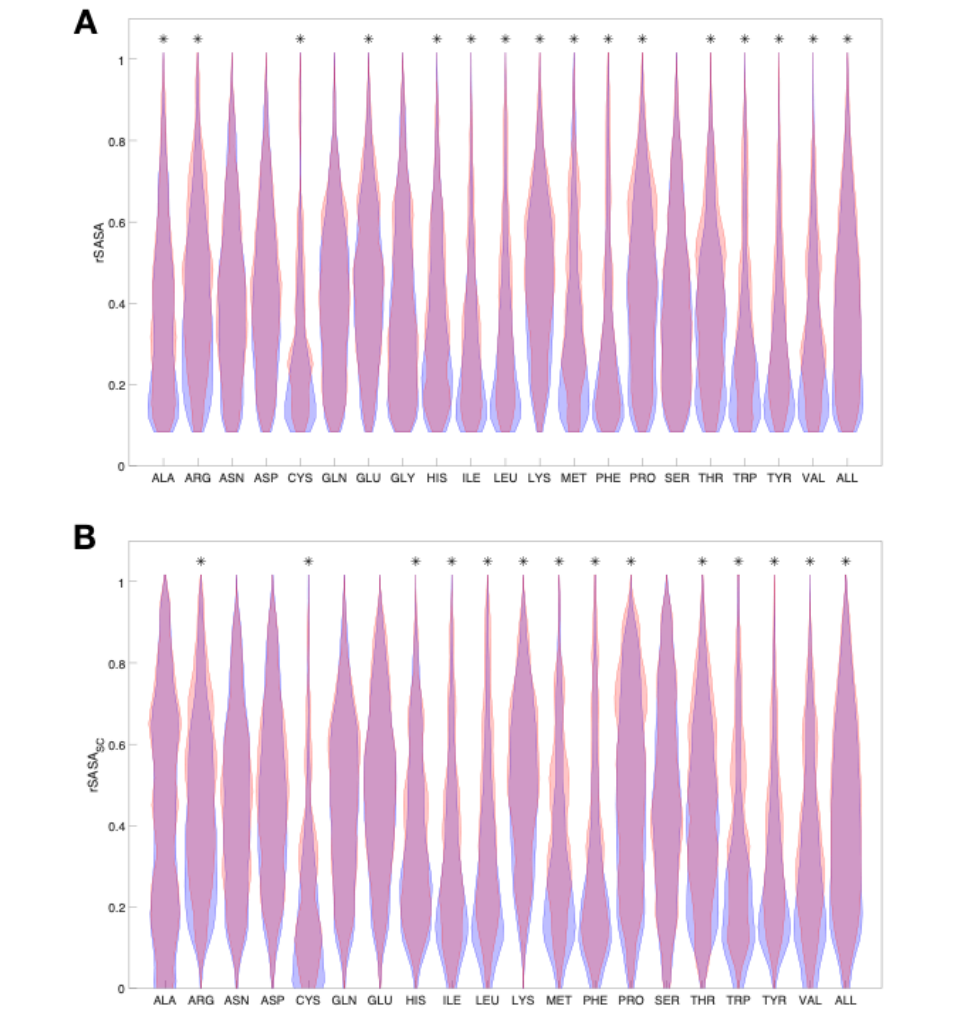

Figure 1: rSASA distributions of amino acids on the protein non-interface surface (blue) and interface surface (red) (intersection between distributions in violet) for (A) full residues and (B) side-chains on unbound protein structures. Asterisks indicate statistically significant differences between non-interface and interface residues. Although we restricted our study to surface residues with rSASA ≥ 0.1, $rSASA_{SC}$ values can range from 0 to 1.

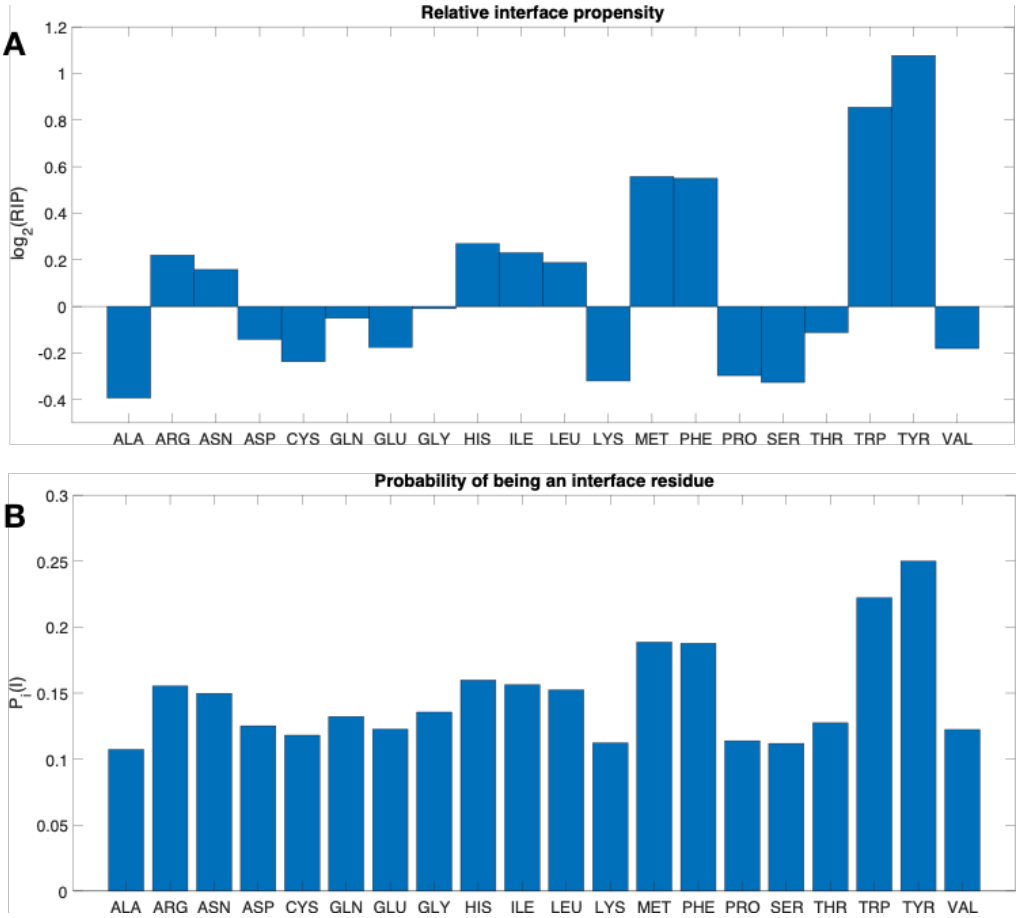

Figure 2: Interface propensity of surface amino acids. (A) Relative interface propensity for each amino acid type as defined by EQ3. (B) Probability of being an interface residue, $P_i(I)$, as defined by EQ4.

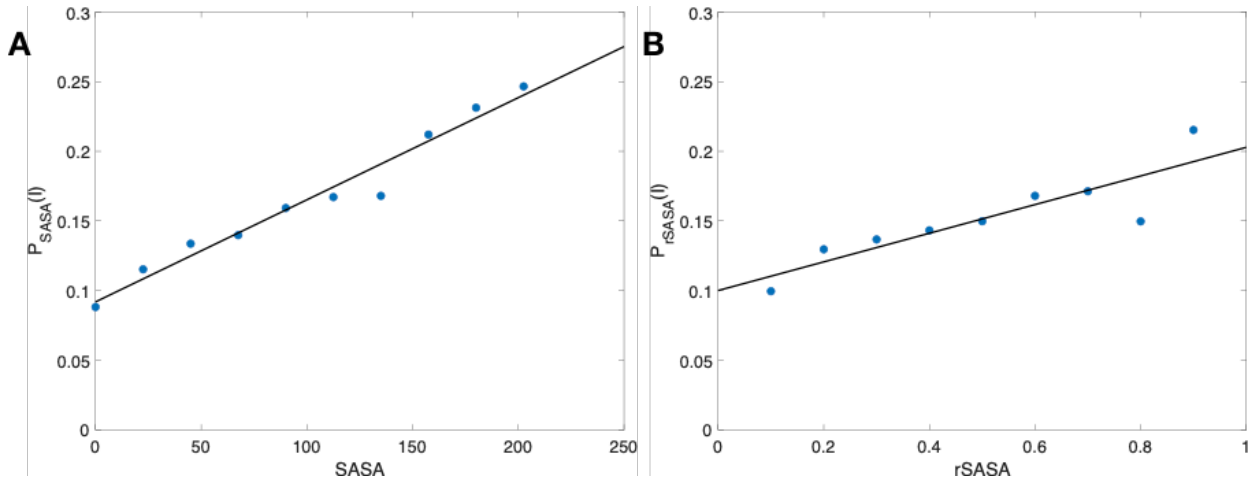

Figure 3: Relationship between (A) SASA and (B) rSASA and the probability of being an interface residue across all amino acid types.

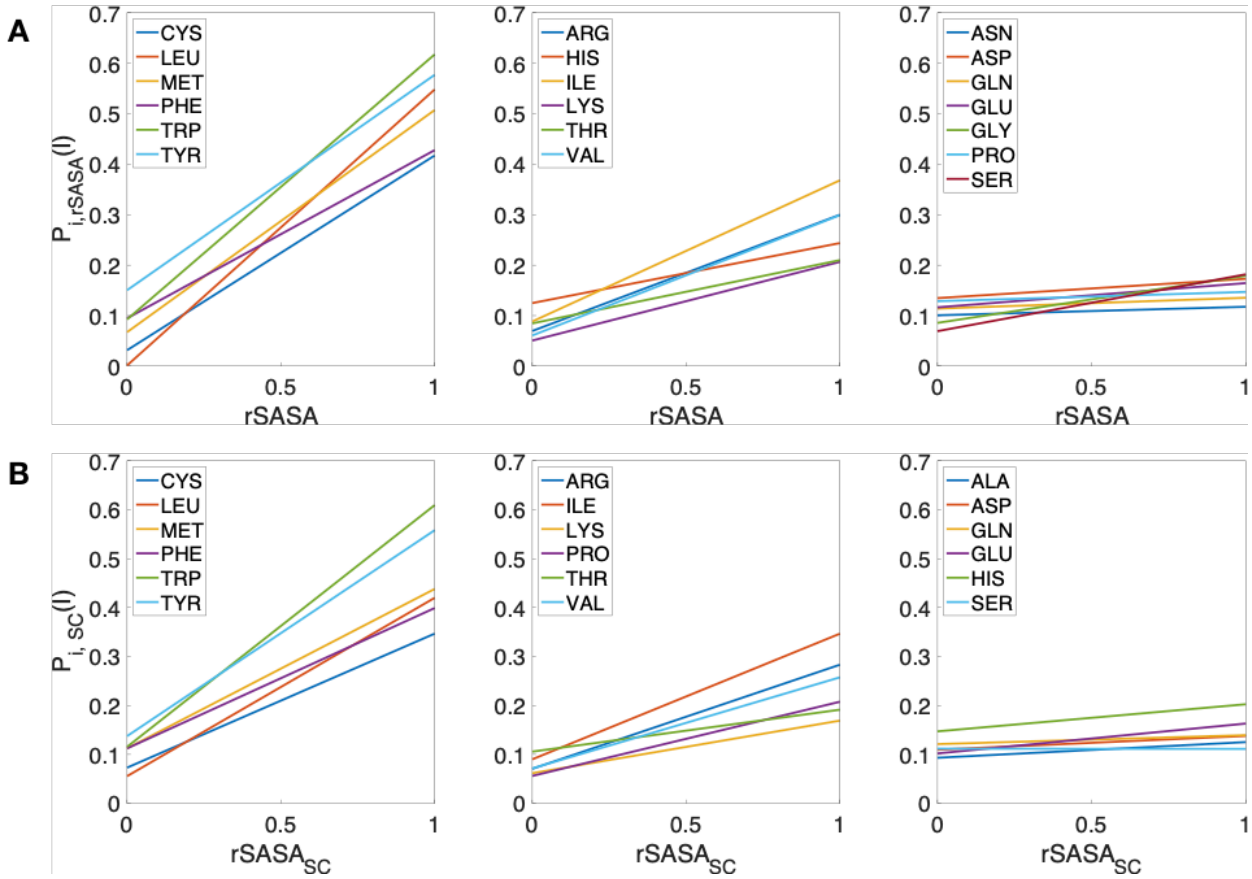

Figure 4: Relationship between probability of being an interface residue and (A) rSASA and (B) $rSASA_{SC}$ for each amino acid type. Amino acids are grouped by slope.

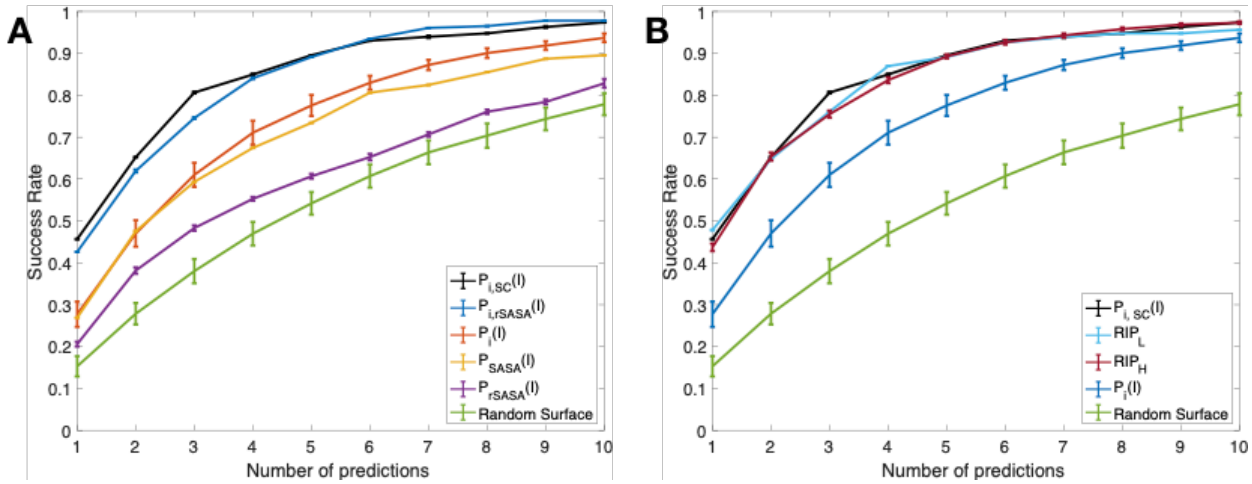

Figure 5: Success rate of identifying the true interface by selecting the top T residues after ranking residues by their probability using various interface prediction methods. When two or more residues have the same probability values, they are randomly ranked. The error bars show the standard deviation over 50 runs with different random number seeds.

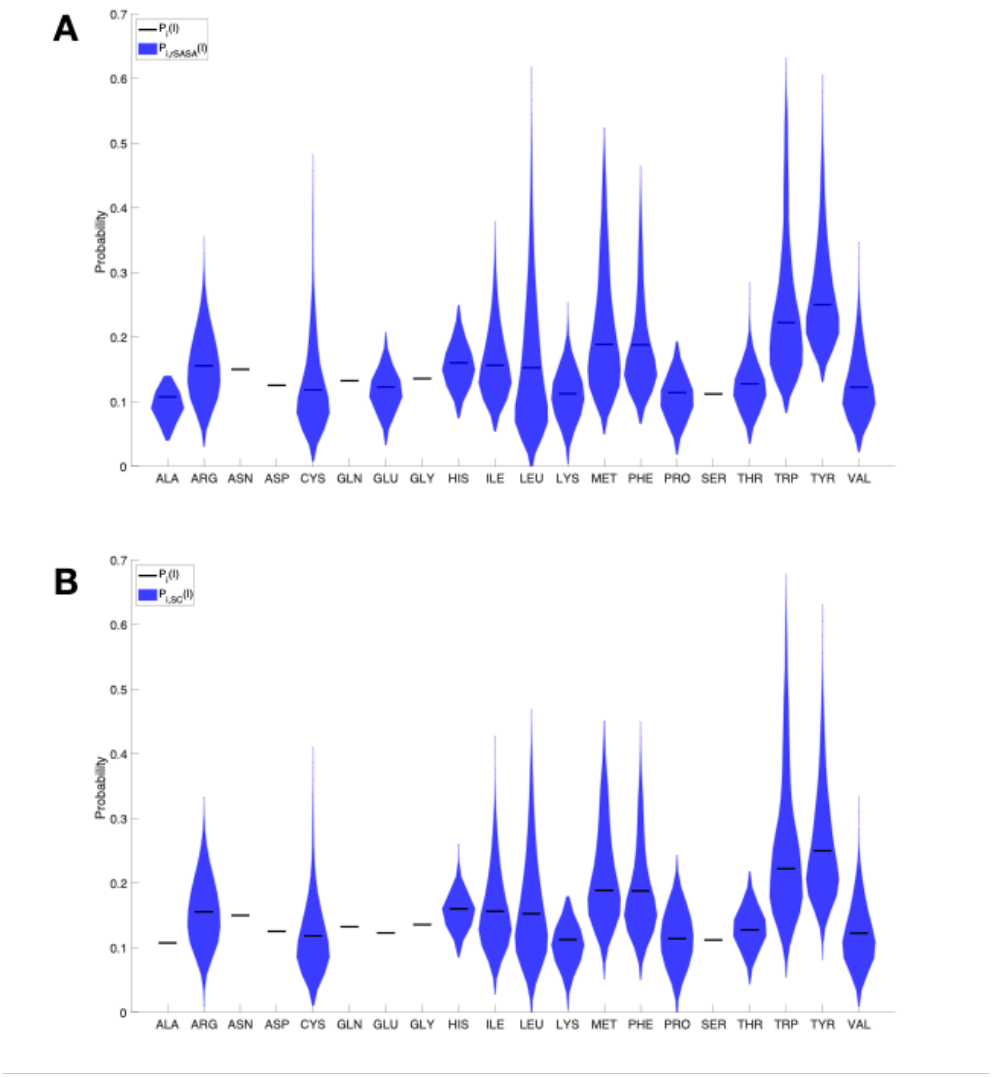

Figure 6: Distribution of probability values for surface residues (blue) using (A) rSASA and (B) $rSASA_{SC}$ dependent equations compared to standard $P_i(I)$ values (black). Amino acids without blue values have a predicted $P_{i,\,rSASA}(I)$ or $P_{i,\,SC}(I)$ set to $P_i(I)$.

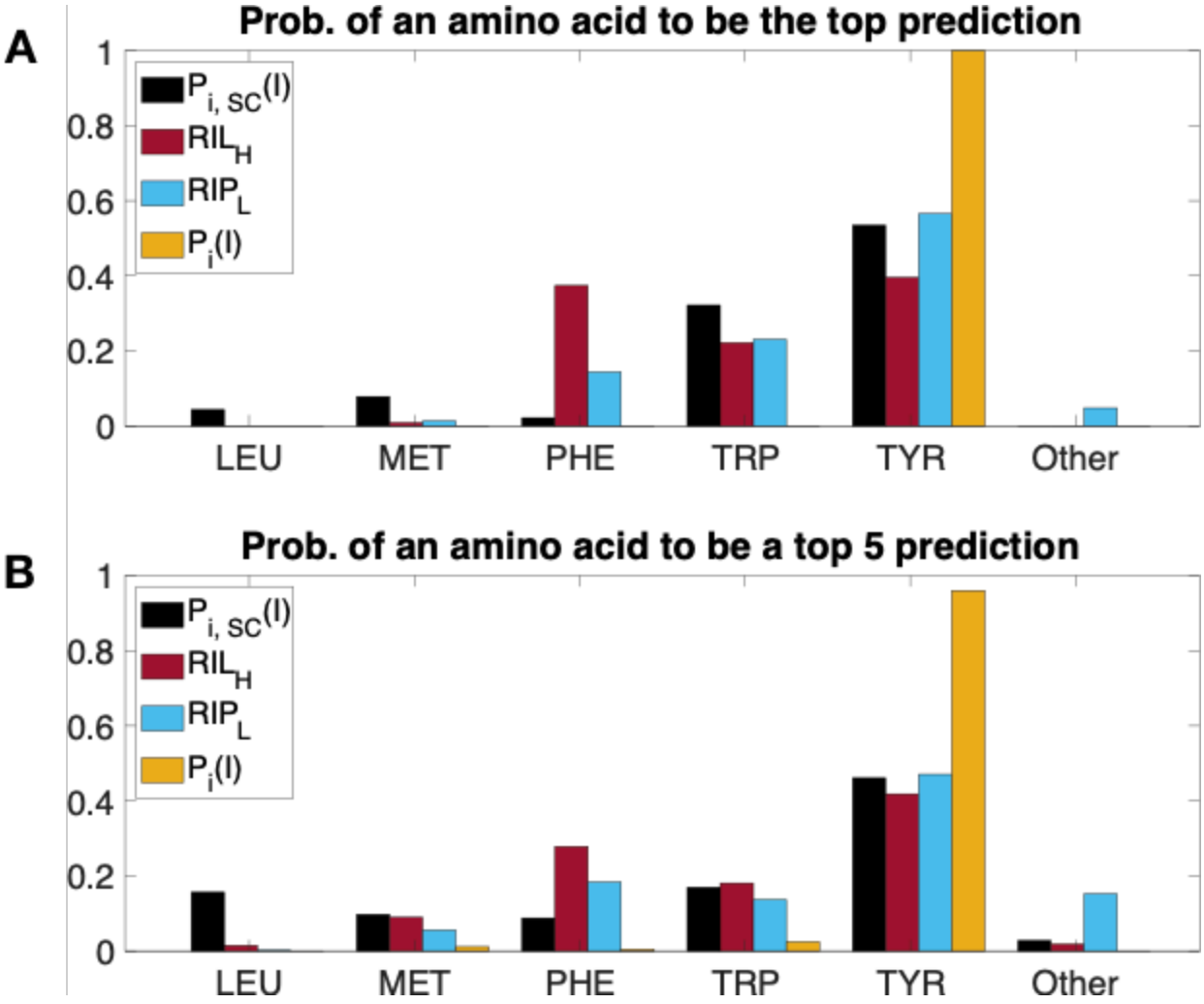

Figure 7: Distribution of amino acid types in the top 1 or top 5 predicted interface residues for $P_{i,\,SC}(I)$, $P_i(I)$ and using data from Ref. 6 and Ref. 11.

Supporting Information

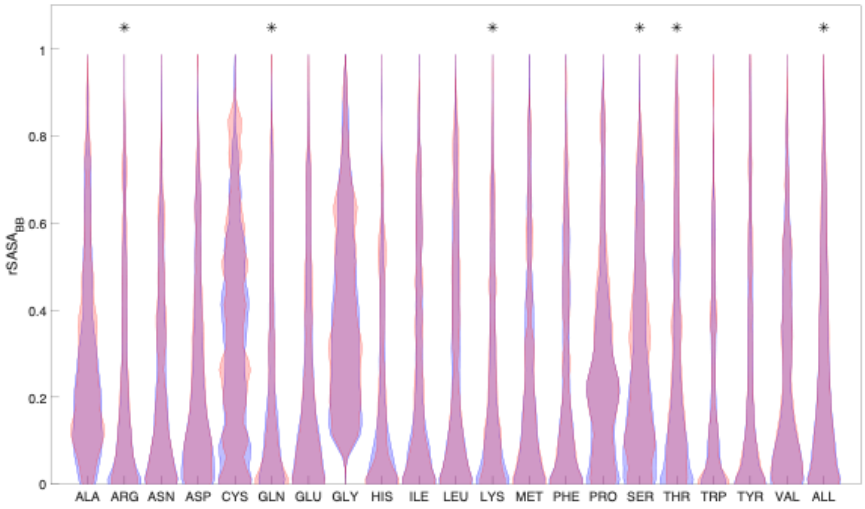

Figure S1: $rSASA_{BB}$ distribution of amino acids in unbound monomers on the protein non-interface surface (blue) and interface surface (red). Asterisks indicate statistically significant differences between non-interface and interface distributions. Although we restricted our study to surface residues with rSASA $\geq$ 0.1, $rSASA_{BB}$ values can range from 0 to 1.

Table S1: Maximum SASA, $SASA_{SC}$ and $SASA_{BB}$ values ($Å^2$) found for each amino acid type in a dataset of 792 high resolution structures after removing outliers.

	SASA	**$SASA_{SC}$**	**$SASA_{BB}$**
ALA	122.26	70.28	59.01
ARG	252.47	213.40	49.71
ASN	168.76	130.26	55.78
ASP	163.89	121.36	54.70
CYS	127.27	100.28	43.21
GLN	197.95	157.59	49.83
GLU	191.60	153.29	51.36
GLY	97.38	—	97.38
HIS	197.27	164.55	45.43
ILE	189.75	154.52	41.85
LEU	197.51	157.40	42.54
LYS	217.46	177.37	51.97
MET	185.35	163.53	42.22
PHE	202.84	176.31	45.82
PRO	154.06	114.05	51.01
SER	134.22	83.82	58.02
THR	152.62	113.89	44.41
TRP	206.28	180.28	45.71
TYR	228.69	196.37	41.00
VAL	164.76	130.39	44.91

Table S2: The fraction of each backbone type on the non-interface surface and interface of proteins in our dataset.

	Loop	α-helix	β-sheet
Surface	0.41	0.32	0.27
Interface	0.53	0.28	0.19

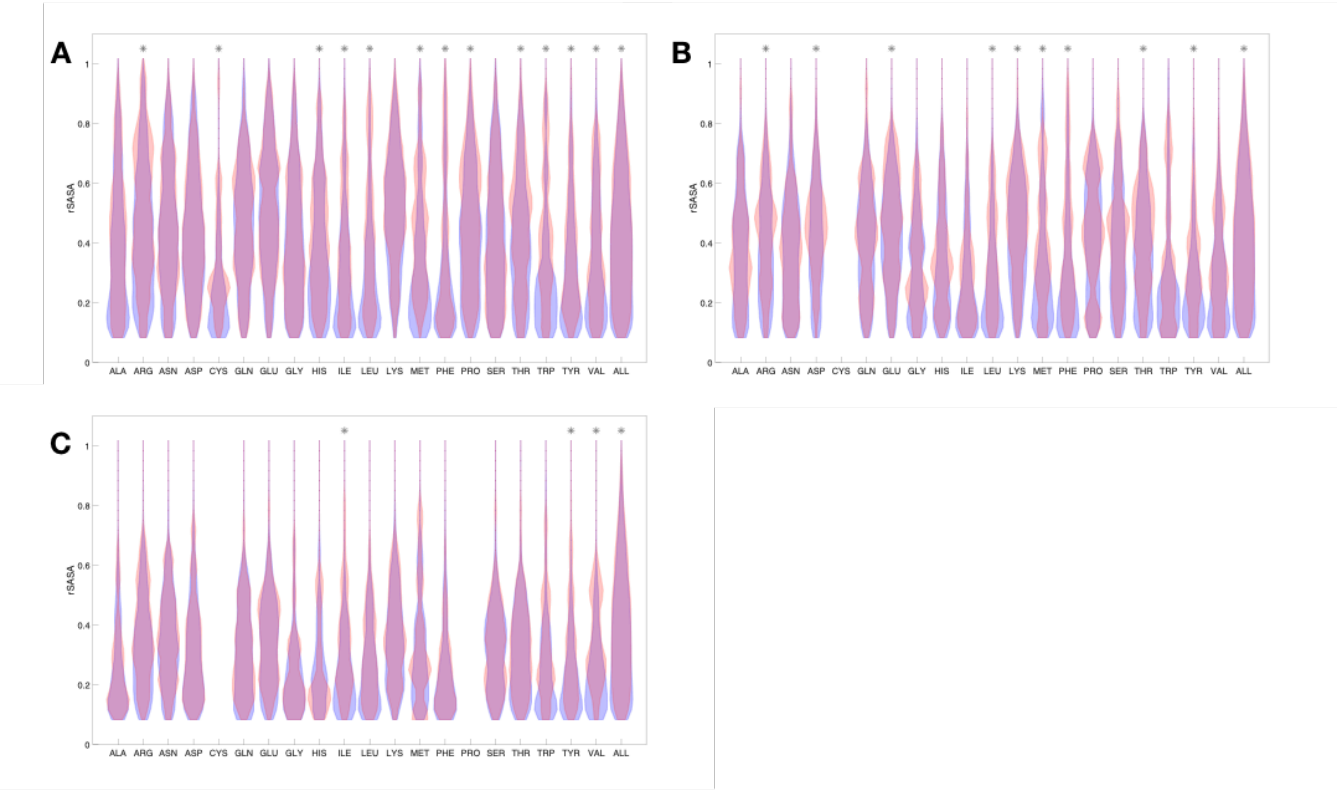

Figure S2: rSASA distribution of amino acids in unbound structures on the protein non-interface surface (blue) and interface surface (red) split by backbone type: (A) loop, (B) α-helix and (C) β-sheet. Asterisks indicate statistically significant differences between non-interface and interface distributions. Residues with fewer than 25 interface residues in a given backbone type (Cys for α-helix and Cys and Pro for β-sheet) were excluded.

	Jones 1996	Ofran 2003	Neuvirth 2004	Chen 2005	Liang 2006	Negi 2007	Dong 2007	Yan 2008	Talavera 2011	Hwang 2016	Dai 2016	**This work**
ALA												
ARG	X	X		X	X	X	X	X	X	X	X	X
ASN												X
ASP												
CYS	X		X	X	X	X		X	X	X	X	
GLN	X	X										
GLU												
GLY	X		X		X							
HIS	X	X	X	X	X	X	X	X	X		X	X
ILE	X		X	X	X	X	X	X	X	X	X	X
LEU				X	X	X	X	X	X	X	X	X
LYS												
MET	X	X	X	X	X	X	X	X	X	X	X	X
PHE	X	X	X	X	X	X	X	X	X	X	X	X
PRO						X					X	
SER			X									
THR												
TRP	X	X	X	X	X	X	X	X	X	X	X	X
TYR	X	X	X	X	X	X	X	X	X	X	X	X
VAL	X			X	X	X	X	X	X	X		

Figure S3: Amino acids identified with enrichment at the protein interface versus the protein surface, calculated using relative interface propensity

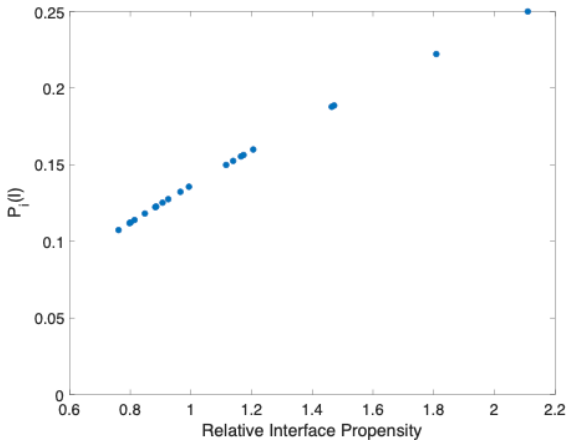

Figure S4: Relationship between $P_i(I)$ and relative interface propensity.

Table S3: Values of N_i and M_i for each amino acid type

Amino Acid	N_i	M_i
ALA	263	2187
ARG	405	2201
ASN	358	2031
ASP	397	2773
CYS	56	418
GLN	299	1961
GLU	446	3189
GLY	410	2613
HIS	170	893
ILE	197	1063
LEU	337	1873
LYS	421	3327
MET	113	486
PHE	197	852
PRO	275	2139
SER	397	3151
THR	384	2628
TRP	116	406
TYR	362	1086
VAL	229	1643