Supporting Information

We list the Protein Databank codes for each protein used in our studies. Table S1 lists the 182 proteins selected from the Dunbrack 1.0 Å database that comprise the Dun1.0 dataset used in our studies. (Proteins with ligands and modified residues were removed from the dataset.) Table S2 lists the 149 protein-protein complexes in the PPI dataset. Table S3 contains the 19 transmembrane proteins in the TM dataset.

In Fig. S1, we show the relationship between the prediction accuracy (using the hard-sphere model) and relative solvent accessible surface area, rSASA, for the Ile, Leu, Phe, Ser, Thr, Trp, Tyr and Val residues in the Dun1.0 database. We find that for all residues (except Ser) the prediction accuracy decreases as rSASA increases.

Table S1: Protein Databank codes for each protein in Dun1.0

1A6M	1N4W	1ZK4	2JHF	3DHA	3SOJ
1AHO	1NKI	1ZLB	2O9S	3DK9	3TEU
1BYI	1NLS	1ZUU	2OV0	3E4G	3U7Q
1C75	1NQJ	1ZZK	2P5K	3EA6	3UI4
1C7K	1NWZ	2A6Z	2PND	3F1L	3V1A
1EB6	1O7J	2B97	2PNE	3FSA	3VII
1EXR	1OAI	2BF6	2PWA	3FYM	3VOR
1F94	1OD3	2BT9	2QCP	3G21	3VRC
1G4I	1OK0	2BW4	2QSK	3G46	3ZR8
1G66	1P1X	2CE2	2QXI	3GOE	3ZSJ
1G6X	1PQ7	2CHH	2R31	3H31	3ZUC
1GCI	1Q6Z	2CWS	2RBK	3HGP	3ZZP
1GQV	1R6J	2DDX	2RH2	3IP0	4A02
1GWE	1RTQ	2DSX	2V8T	3JU4	4A7U
1IQZ	1TG0	2E4T	2VB1	3JUD	4ACJ
1IX9	1TQG	2ERL	2VHA	3JYO	4AR5
1IXH	1TT8	2F01	2VHK	3KFF	4AXO
1J0P	1U2H	2FDN	2VXN	3KLR	4AYO
1JFB	1UCS	2FMA	2XFR	3KS3	4DPB
1K4I	1UFY	2FVY	2XJP	3M5Q	4EA9
1K5C	1UG6	2FWH	2XOD	3NE0	4EGU
1KTH	1US0	2G6F	2XOM	3NIR	4F1V
1KWF	1V0L	2GGC	2XU3	3NOQ	4G9S
1L9L	1V6P	2GKG	2Y78	3O4P	4GA2
1LNI	1VBW	2H3L	3A02	3O5Q	4HNO
1M1Q	1VYR	2H5C	3A38	3PSM	4I8H

1M40	1W0N	2HS1	3A4R	3PUC	7A3H
1MC2	1X6Z	2I4A	3AGN	3Q46	
1MJ5	1X8Q	2IIM	3AJ4	3QR7	
1MN8	1XMK	2IXT	3BWH	3RQ9	
1MUW	1Y55	2JFR	3CCD	3RWN	

Table S2: Protein Databank codes for each protein in PPI

1AAP	1U07	2H2R	3KGK	4J78	4YNH
1CKA	1UTI	2HQX	3KTP	4JVU	4Z27
1D4T	1UZ3	2IPR	3L32	4K12	4Z8J
1DJT	1V8H	2OEI	3M8J	4K5A	4ZGW
1DQZ	1VH5	2PKF	3MAB	4K8Y	5B08
1EZG	1W5R	2PV1	3NDD	4KN8	5C04
1F46	1WMH	2Q20	3NSO	4LEB	5D38
1G2Q	1X2I	2R1U	3OBQ	4LLD	5DDZ
1IJY	1X6I	2W2A	3PSM	4LN2	5DWP
1IRQ	1ZRS	2W6A	3PTL	4LNP	5EPW
1KTN	2A35	2XHF	3RQ9	4M91	5GT5
1KYF	2A8F	3AZD	3SO6	4NPU	5GTU
1MFG	2AB0	3BZZ	3SR3	4OHJ	5HEY
1MKK	2BPD	3C8P	3VZ9	4ONL	5HHE
1MTP	2C61	3CT6	3ZIT	4P61	5IMM
1MY7	2CAR	3CZZ	3ZRX	4PRS	5J4F
1NXM	2DPL	3DRF	4AVR	4Q9B	5K2I
1OAI	2E10	3DS2	4C18	4QLP	5K3D
1QKD	2ETX	3F1L	4DO2	4RDJ	5KWN
1SH8	2FHZ	3F1P	4ERY	4UU3	5LND
1SQE	2FLU	3G1S	4FZO	4UUL	5N8A
1SSH	2GEC	3GMG	4G6C	4WJO	5TZ5
1T6F	2GOM	3HJ2	4G7X	4WW1	5XAV
1T7H	2GRR	3I2Z	4IHE	4X9Z	5XN3
1TVN	2GU9	3IVV	4IHN	4XO9	

Table S3: Protein Databank codes for each protein in TM

1Q16	2SQC	3LDC	3V5U	4Y9H
1U7G	2XOV	3M7L	3WG7	5AEZ
2A65	3B9W	3PCV	4AL0	5G28
2O4V	3GD8	3S8G	4EIY	

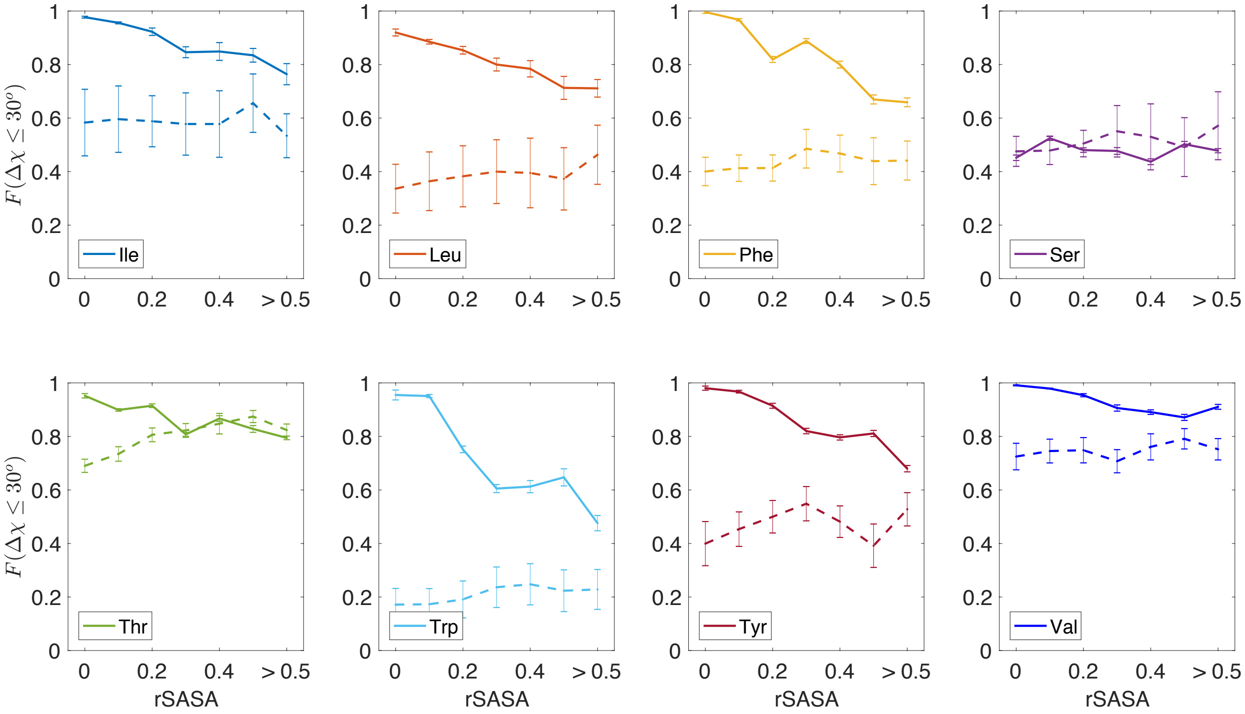

Figure S1: Fraction of residues predicted (using the hard-sphere model) within 30°, F(Δχ ≤ 30°), for Ile, Leu, Phe, Ser, Thr, Trp, Tyr, and Val residues in the Dun1.0 database (solid line) and their corresponding dipeptide mimetics (dotted line) as a function of rSASA. The dotted line provides lower bounds for the prediction accuracy for the residues in each rSASA bin. Due to the low frequency of uncharged residues in the non-core region, we have combined all residues with rSASA > 0.5 into one bin.