A model for hydrophobic protrusions on peripheral membrane proteins - Supporting information

Edvin Fuglebakk and Nathalie Reuter

S1 Data sets of proteins with experimentally verified membrane-binding sites

We used a small set of protein structures to establish the definition of protrusions and adjust the parameters c and n (Cf. *Materials and Methods*). The dataset consists of structures of peripheral proteins with striking protrusions at their experimentally-verified membrane-binding site. Table S1 contains the list of PDB codes, the protein family to which they belong and the amino acids forming the membrane-binding site.

We also collected a larger dataset of peripheral proteins with experimentally-identified binding sites. The structures are listed in Table S2. This dataset does not overlap with the one listed in Table S1 and could thus be used for analysis purposes, and in particular for those results reported in Figure 7 of the main manuscript. This set has some overlap with the list provided by Lomize *et al.* [11].

PDB ID	interfacial binding site	family
1RLW	F35, M38, L39, N95, Y96, V97, M98 [34]	C2-domain
1BYN	M173, G174, R233, F234, K235 [40, 41]	C2-domain
1UOV	V304, G305, I367, K369 [42]	C2-domain
1GMI	I89, Y91 [43]	C2-domain
1H6H	F35, Y94, V95 [35]	PX domain
1CZS	W26, W27 [44]	Discodin domain
1D7P	M2199, F2200, L2251, L2251 [12]	Discodin domain
1T6M	W51, Y92, Y208, W246, Y250, Y252 [46, 45, 33]	Bacterial PLC
1H0A	L6, M10 [38]	ENTH domain
1LOX	L195 [47]	Lipoxygenases

Table S1: Peripheral protein structures used for defining and parameterizing the model of hydrophobic protrusions. Family classifications are taken from OPM[23].

PDB ID	interfacial binding site	classification
1DSY	M186, N189, R216, R249, R252 [48]	C2-domain
1O7K	R43 I65 W80 [35]	PX domain
1HYJ	V21, T22 [52]	FYVE PIP_3 domain
1VFY	L185 L186 R193 [36]	FYVE PIP_3 domain
1PTR	L250 W252 L254 [37]	C1 domain
1A8A	T72 S144 W185 S228 S303 [50]	Annexins
1DM5	E142 S144 G145 [51]	Annexins
1IAZ	W112 W116 [53]	Pore-forming Equinatoxin
1NB1	C1 G2 E4 T5 V6 G7 S18 W19 P20 V21 C22 G26 L27 P28 V29 [54]	Cyclotide
1POC	2I 14K 78I [55]	Insect sec. PLA_2
1N28	V3 K10 L19 F23 F63 K115 [56]	Vertebrate sec. PLA_2
1POA	W61 F64 Y110 [39]	Vertebrate sec. PLA_2
1VAP	W20 W109 [57]	Vertebrate sec. PLA_2
4P2P	W3 [62]	Vertebrate sec. PLA_2
1COY	M81 [58]	GMC oxidoreductases
1PFO	W464 W466 [59]	Chol.-dep. Cytolysin
1D1H	W30 [60]	Spider toxins
1PXQ	W30 [61]	Subtilosin A
2FNQ	W413 W449[63]	Lipoxygenases
1G13	T90 L126 N136 [64]	ML domain
1EIN	P42 D96 T123 I252 [32]	Fungal lipases
3PAK	Y164 R216 Y221 R222 [49]	Lectin domain
1F6S	K98 V99 [65]	C-type lysozyme
2DA0	K18 K19 I23 K25 N30 N48 N77 [66]	Pleckstrin-homology domain

Table S2: Protein structures and corresponding membrane-binding sites used for systematic comparison with the *Likely Inserted Hydrophobes*. Family classifications are from OPM[23], except for 3PAK and 1F6S which were taken from SCOPe [31] as the structures are not present in OPM. Quaternary structures are also taken from OPM, except for 3PAK and 1F6S; those were obtained from the literature. Residue numbering corresponds to that used in the listed PDB ID. All structures are either monomers or homo-oligomers where all chains are equally likely to interact with the membrane. Chain identifiers are therefore not provided.

S2 Analysis performed on an alternative reference set and alternative quaternary structure determination

The analysis of protruding hydrophobes presented in the main text relies on a dataset of peripheral protein structures and a reference dataset consisting of solvent-exposed regions of transmembrane proteins. The choice of the reference dataset relies on the assumption that solvent-exposed regions of TM proteins are representative of non membrane-binding protein surfaces, i.e. that they have not been subject to evolutionary selection enhancing their affinity for biological membranes. Moreover by using this reference dataset we were comparing protein fragments with models of real proteins. Further, and for both the reference and peripheral proteins dataset, the analysis relies on the quaternary information stored in the OPM-database [23]. In order to assess the sensitivity of our analysis to these two aspects we performed the same analysis with (i) a different reference dataset and (ii) quaternary structure information from a different source.

S2.1 Definition of alternative data sets

We built a reference dataset consisting of a broad selection of proteins excluding proteins classified as membrane binders by either OPM or SCOP. The following criteria were applied:

- We selected all PDB IDs classified in SCOPe [31] in the classes *All alpha proteins* (sunid: 46456), *All beta proteins* (sunid: 48724), *Alpha and beta proteins (a+b)* (sunid: 51349), *Alpha and beta proteins (a/b)* (sunid: 53931) or *Multi domain proteins* (sunid: 56572).
- We removed from this set any PDB ID that has one or more domains classified in the same SCOPe-family as any domain in the OPM-database [23]. This excludes not only the peripheral membrane binders, but also any transmembrane protein found in the reference set used for our primary analysis.
- In order to reduce redundancy, we iteratively removed proteins with domains that share SCOPe-family classification with any other domain in the set, until there were no such shared classifications left. This process ensures that there is at most one representative for each family in the set.
- We removed any structure not determined by X-ray crystallography to be able to treat quaternary structure predictions consistently.
- We generated quaternary structure models using PISA [30].

For the sake of consistency we similarly created a modified version of the set of peripheral proteins.

- We obtained the list of PDB IDs used for the set of peripheral binders in our primary analysis.
- We iteratively removed proteins with domains that share SCOPe-family classification with any other domain in the set. This process ensures that there is at most one representative for each family in the set.

- We removed any structure not determined by X-ray crystallography.
- We generated quaternary structure models using PISA.

After filtering out the few structure files that did not exactly comply to the expected PDB format, the final set of peripheral membrane binders contained 170 proteins, and the final reference set contained 2250 proteins.

SCOPe version 2.06 was used. PISA predictions were obtained through the "Protein interfaces, surfaces and assemblies" service PISA at the European Bioinformatics Institute. (`http://www.ebi.ac.uk/pdbe/prot_int/pistart.html`). Where PISA predicted that the asymmetric unit represents the most stable quaternary structure in solution, we obtained structures from the protein data bank (`http://www.rcsb.org/`)[28].

S2.2 Results and discussion

We here present the results of the same analysis as presented in the main manuscript, using the datasets described above. Note that in contrast to the exposed regions of transmembrane proteins used as reference dataset in our original analysis, we have no firm indication that the proteins in this reference set are not membrane-binding. They are however chosen from the SCOPe classification and all pdb codes classified in OPM are excluded, so we find it reasonable to assume that it is less enriched in peripheral membrane binders than the set of peripheral proteins obtained from OPM.

We analyzed the alternative data sets as described for the original datasets (Cf. main text) but with two differences. Firstly, the alternative sets defined in section S2.1 were used. Secondly, wherever the mean of a property over OPM-families was used in the primary analysis, we instead used the values obtained for individual proteins in the dataset. The dataset was filtered for redundancy using the SCOPe classification of protein domains, as explained in section S2.1. Results obtained with the alternative datasets are presented on Figures S1, S2, S3, S4 and S5 which correspond to Figures 2, 3, 5, 4, 10 and 9 in the main text, respectively. The data representing peripheral membrane binders are colored in blue in all plots, and the data representing the alternative reference set is colored in red.

The conclusions drawn from the primary datasets are supported by the analysis of the alternative datasets. The relative importance of large aliphatic residues on protruding locations in peripheral proteins is reproduced (Figure S5). The higher frequency of hydrophobes on protrusions on the surface of peripheral membrane binders compared to the reference dataset is shown on Figure S1. The tendency of protruding hydrophobes from the peripheral dataset to be co-insertable is different from that of the reference dataset (Figure S4). There is still a stronger contrast between the data sets when the analysis is restricted to vertex residues of low protein density (Figure S2) and when restricted to co-insertable pairs of protruding hydrophobes (Figure S3). The analysis of secondary structure elements also yields a result similar to what was obtained for the primary datasets (Figure S6).

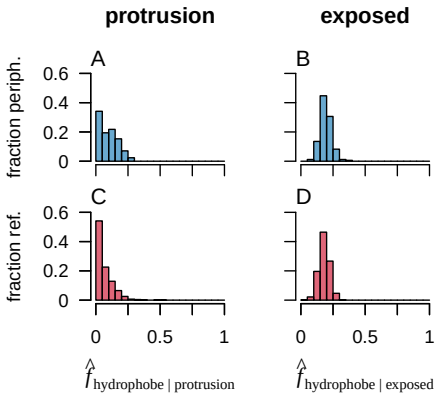

Figure S1: Frequencies of hydrophobes on surface amino acids, both on protrusions (A and C) and among all solvent exposed amino acids (B and D). Compare peripheral proteins (blue) and the reference set (red). The horizontal axes show the mean fraction of protrusions or solvent exposed amino-acids that are hydrophobic. The vertical axis shows the fraction of protein families. See caption of corresponding Figure 2 in main text.

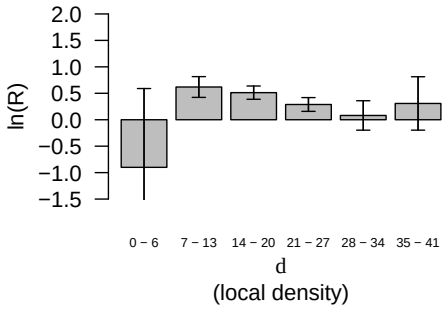

Figure S2: The plot shows the logarithm of the odds-ratio comparing the frequency of hydrophobes on *vertex* residues in peripheral proteins and the reference set. Positive values reflect higher frequencies in the peripheral proteins. See caption of corresponding Figure 3 in main text.

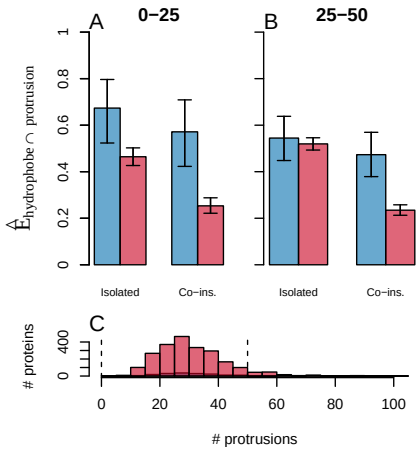

Figure S3: Occurrence of *co-insertable protruding hydrophobes* on protein surfaces. Panels A and B show the weighted fraction of proteins that have protruding hydrophobes, in the peripheral proteins (blue) and the reference set (red). Panel C shows the frequency distribution of the total number of protruding residues ("# protrusions") for all proteins. The selections analysed in panel A and B are found between the dashed lines in panel C. See caption of corresponding Figure 5 in main text.

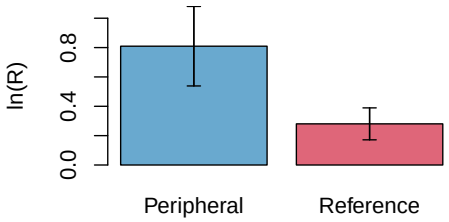

Figure S4: The tendency for protrusions to be co-insertable is quantified by the weighted frequency of co-insertion, and is compared between each data set and a null model using the odds ratio. Positive values reflect higher frequencies of co-insertion than in the null model.See caption of corresponding Figure 4 in main text.

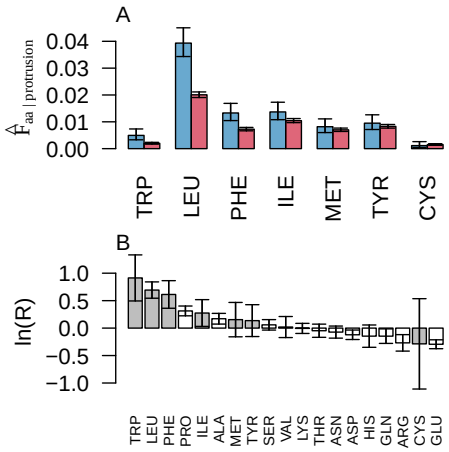

Figure S5: Panel A shows the weighted fractions of hydrophobic amino acids on protrusions from peripheral proteins (blue) and from proteins in the reference set (red). In panel B, the contrast between the two sets is quantified by the odds ratio, so that positive values reflect higher frequencies in the set of peripheral proteins than in the reference set. See caption of corresponding Figure 10 in main text.

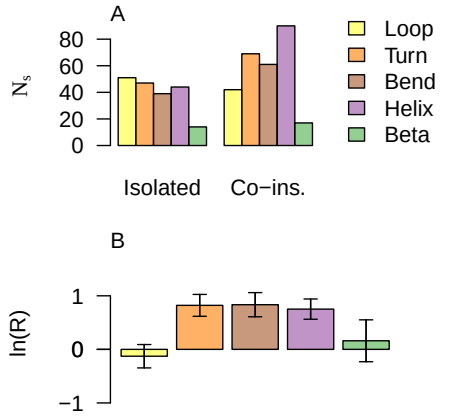

Figure S6: Panel A shows the weighted number of *protruding hydrophobes* associated with the different types of secondary structure elements. We have differentiated between protrusions that have at least one co-insertable protruding hydrophobe (right, labeled "Co-ins."), and those that have not (left, labeled "Isolated"). Panel B compares the weighted frequencies of hydrophobes on protruding secondary structures between the peripheral membrane proteins and the reference set, using the odds ratio. Positive values reflect higher frequencies in the peripheral proteins.See caption of corresponding Figure 9 in main text.

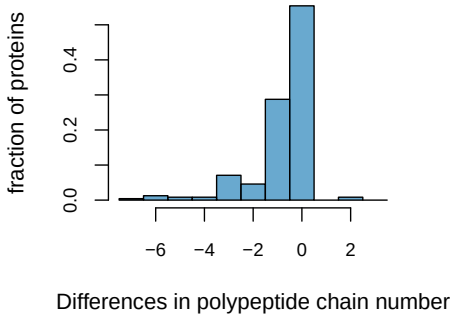

Figure S7: Differences in number of polypeptide chains between the protein models present in the original dataset (retrieved from OPM) and the models used to build the alternative dataset (predicted by PISA). The histogram is calculated for each of the PDB IDs occurring in both datasets. When more chains are present in the PISA models, The difference (x-axis) is negative when the oligomeric state from PISA counts more chains than the one retrieved from OPM.

Interestingly the contrast between peripheral proteins and reference data sets is lower with the alternative data sets than in our primary analysis; it is particularly striking when comparing Figure S1 with Figure 2 and Figure S3 with Figure 5. This can partly be explained by the lack of negative assertions for the reference set which might contain some peripheral membrane binders. It is also evident that the prevalence of protruding hydrophobes is lower in the set of peripheral proteins considered in this analysis than in the primary analysis. We believe that this is due to the difference in the determination of quaternary structures or oligomeric states. Consistent with this, we find that PISA does tend to predict higher order complexation than the quaternary structure models chosen by the OPM curators (see Figure S7). We do expect the curated models used in OPM to be more accurate in this respect. We find it important to note that the analysis is sensitive to how quaternary structure is modelled.

References

See main text