Figures and figure supplements

Pi-Pi contacts are an overlooked protein feature relevant to phase separation

Robert McCoy Vernon et al
Figure 1. PDB statistics for planar pi-pi interactions. (A) Average number of sp² groups involved in planar pi-pi contacts per 100 protein residues binned by crystal structure resolution. Values are shown for contacts defined by

Vernon et al. eLife 2018;7:e31486. DOI: https://doi.org/10.7554/eLife.31486

Figure 1 continued on next page
the nature of the involved sp\(^2\) groups, with all groups in black, aromatic to non-aromatic sp\(^2\) in blue, non-aromatic to non-aromatic in pink, backbone to backbone in gray, and aromatic to aromatic in orange. Error bars show bootstrap SEM. (B) Planar pi-pi contact interaction frequencies for each residue type, with the average across all residue types shown as a red line, and (C) frequency of each residue type in contributing to planar pi-pi interactions, with bars showing overall frequency colored proportionally by the nature of the contact partners.

Figure 1—source data 1 and 2.
DOI: https://doi.org/10.7554/eLife.31486.002
Figure 1—figure supplement 1. Proportion of sidechain to backbone VDW contacts that satisfy planar contact criterion. To examine relative contact enrichment, sidechain contacts to the backbone are normalized against the total number of contacts satisfying the same VDW criterion (two pairs of atoms within 4.9 Å), with comparison between (left) planar sp² sidechain groups (for W, F, Y, H, R, Q, N, E and D) and (right) selected sp³ planar surfaces (for C, S, M, T, K, L, V, I). The sp³ planar surfaces were chosen as a control by taking sets of atoms describing exposed planar surfaces, as described in the Materials and methods. Comparing relative planar contact frequency, we observe the majority of sp² sidechain types show clear enrichment relative to the sp³ controls.
DOI: https://doi.org/10.7554/eLife.31486.003
Figure 1—figure supplement 2. Selected sidechain-to-sidechain contact frequencies by resolution. Percentage of residues involved in planar contacts are shown in red, and percentage in any other non-planar VDW contact are shown in blue, with panels showing contacts by sidechain group (for panels A-F: R to R, R to K, H to R, H to K, Q to R, and Q to K). We observe that the increase in planar pi-pi contacts to arginine at higher resolution comes at the expense of non-planar VDW contacts (panels A, C and E). In contrast, contacts made to an arbitrary surface plane at the end of lysine sidechains do not show this increase in planar orientation with resolution (panels B, D and F).

DOI: https://doi.org/10.7554/eLife.31486.004
Figure 2. Examples of planar pi-pi contacts in folded protein structures. Pi-pi interactions shown using rods to describe the normal vector of the plane. Rods extend to a carbon VDW radius of 1.7 Å, colored by category with sidechain groups in purple, backbone in blue, small molecule ligands in orange, and RNA in gray. Ligand molecules are green, with relevant water molecules shown as red spheres and hydrogen bonds as yellow lines. (A) Arginine ladder motif in Porin P (PDB:2o4v). (B) Catalytic site from arginine kinase (PDB:1m15). (C) Network of interactions in nitrogenase (PDB: 3u7q). (D) Backbone/sidechain contacts at the ends of secondary structure elements (PDB:4b93). (E) RNA-binding interactions (PDB: 4lgt). (F) Interaction network stacked between disulfide bonds (PDB: 4v2a).

DOI: https://doi.org/10.7554/eLife.31486.007
Figure 3. Correlation of planar pi-pi interactions with solvent and lack of secondary structure. (A) Contact frequency for sidechain groups (red) and backbone (blue) increases with the total number of solved water molecules within 4.9 Å of the residue, based on structures with >1 water oxygen per residue, including all molecules within 8 Å of the chain of interest, including symmetry partners. (B) Representative example of a pi-stacked sidechain in contact with 11 water molecules (PDB:4u98), showing how the interaction does not appear to compete with solvent. (C) Mean contact frequency vs. sequence distance from regular secondary structure and loop/turn regions. (D) Example of the range of interactions found >10 residues from helix/strand secondary structure (PDB:4b4h).

DOI: https://doi.org/10.7554/eLife.31486.008
Figure 3—figure supplement 1. Effect of solvation on pi-pi category frequencies. Effects of solvation, measured by the total number of water molecules within 4.9 Å of a given residue, on the overall frequency of different types of interactions, categorizing contacts by the identities of the solvent contact tested residue and its partner, where the solvated residue is listed first (green for aromatic to aromatic, blue for aromatic to non-aromatic, orange for non-aromatic to aromatic, and pink for non-aromatic to non-aromatic). Note that non-aromatic includes backbone interactions.

DOI: https://doi.org/10.7554/eLife.31486.009
Figure 3—figure supplement 2. Enrichment of pi-pi contacts, relative to overall VDW contacts, as a function of the number of interactions with water. Water contacts are measured to residue A, and the percentage of pi-pi contacts per VDW contact is measured for all contacts from residue A to residue B.
residue B. Panel A shows the change in percentage of pi-pi contacts per VDW contact by number of waters for each sidechain-sidechain interaction, with pi-contact enrichment with solvation being a consistent property of the majority of interactions involving at least one non-aromatic sidechain. Panels B-F show slope measurements for a selection of examples, Phe to Phe, Arg to Arg, Phe to Arg, Arg to Glu and Phe to Glu, respectively.

DOI: https://doi.org/10.7554/eLife.31486.010
Figure 4. Sidechain contacts at interface positions. Contact frequencies are shown for the nine sp²-containing sidechain types, split into three bars based on interface proximity. From left to right, these bars are i) no other chain within 4.9 Å of any sidechain atom, ii) within 4.9 Å VDW contact distance of any atoms in a different chain within the unit cell of the crystal, iii) within 4.9 Å of any atoms in a chain from a neighboring unit cell, as determined by crystal symmetry data. Bars are colored by the proportion of total contacts contributed by three categories, bottom/black corresponding to local (sequence separation ≤ 4 residues) intrachain contacts, middle/blue to non-local intrachain contacts, and top/pink to interchain contacts, showing that overall contact frequencies and local contact frequencies remain similar and that the non-local contacts do not discriminate between intra and interchain.

DOI: https://doi.org/10.7554/eLife.31486.011
Figure 5. Prediction of phase separation based on planar pi-pi interactions. (A) Reliability plot showing average predicted and observed contact frequencies for percentile bins by pi-pi contact prediction for proteins in the PDB, with PDB sequences used for training in blue and the leave out set in red. Bars show SEM. (B) Highest number of contacts predicted, by window, for two phase separation predictor training sets and three test sets, for the unoptimized predictor. (C) Modified ROC curve showing the final predictor’s performance on three test sets vs. the human proteome, with the full set in pink (N = 62), the full set minus the insufficient for phase separation set shown in green (N = 44), and the sufficient for phase separation set in blue (N = 32). (D) Results for the final predictor (as for panel b) plotted with the predictor’s phase separation propensity scores (PScore). Data underlying B-D included in Figure 5—source data 1 and Figure 5—source data 2.

DOI: https://doi.org/10.7554/eLife.31486.012
Figure 5—figure supplement 1. Contrasting behavior of disorder prediction algorithms and the phase separation prediction. Disopred3 (Jones and Cozzetto, 2015) derived disorder predictions are shown on the y axis and PScores are shown on the x axis for four different test sets, (A) our PDB test set, representing a negative set for both phase separation and disorder, (B) a random sample of 4385 sequences from the human proteome, (C) the subset of the human proteome annotated as containing disorder in the Disprot database (Piovesan et al., 2017), representing a positive set for disorder, and (D) our full phase separation test set. Results are split into four categories separated by PScore = 4 and Disorder = 0.8, with the percentage of sequences in each category inset in blue. The majority of known phase-separating proteins are associated with disorder, and are predicted to be disordered, but sequences predicted to phase separate represent a small subset of both the known and the predicted disordered proteins.

DOI: https://doi.org/10.7554/eLife.31486.013
Figure 5—figure supplement 2. Comparison of scores used in generating phase separation predictions. (A) Highest number of short-range backbone contacts predicted, by window, for the PDB test set, the human proteome, the set of disordered human proteins from Disprot, and the full phase separation test set (N = 121), where percentile ranges are shown in colored boxes. (B) Highest number of long-range backbone contacts predicted, as for panel a. (C) Results for the final predictor plotted with the predictor’s phase separation propensity scores (PScore). Prediction of long-range backbone contacts provides the majority of the discrimination seen in the final predictor.

DOI: https://doi.org/10.7554/eLife.31486.014
Figure 6. Association of phase separation propensity scores with protein interactions, splice isoforms, PTMs, and GO localization, process, and function terms. (A) Protein-protein interaction enrichment by the PScore of partner 1 vs. the PScore of partner 2. The color gradient shows the natural logarithm of the observed over expected ratio. (B) Percentage of human proteins at each PScore range that are detected in more than 10% of AP-MS negative control experiments. (C), Score ranges for alternative splicing variants shown as vertical lines sorted by reference sequence values. (D), Number of PTMs vs. average relative PScore, with methylation shown in red, phosphorylation in green, and ubiquitination in blue.

DOI: https://doi.org/10.7554/eLife.31486.017
Figure 7. PScore enrichment by gene ontology annotation for subcellular localization (A), biological process (B), and molecular function (C). The color gradient shows the natural logarithm of the observed over expected ratio. Heatmaps show enrichment in vertebrate sequences across six defined score ranges, with the highest score range (PScore ≥4) labeled with human enrichment values calculated using PANTHER (see Materials and methods). DOI: https://doi.org/10.7554/eLife.31486.018
Figure 8. Visual confirmation of phase separation. (A) Test tubes containing transparent or turbid solutions of 1 mM FMR1 C-terminus (residues 445–632) along with their corresponding DIC microscopy images taken at room temperature or 4°C, respectively. (B) 1 mM FMR1 C-terminus forms droplets exhibiting liquid fusion properties at 4°C. (C) 40 μM solutions of Human Cytomegalovirus pAP along with corresponding microscopy images taken at room temperature or 80°C, respectively.

DOI: https://doi.org/10.7554/eLife.31486.019
Figure 8—figure supplement 1. Visual confirmation of phase separation, using 20 mg/ml ficoll as a crowding agent. (A) 200 µM FMR1 C-terminus shows reversible droplet formation between 2°C and RT, (B) 220 µM engrailed-2 shows reversible droplet formation between 2°C and 35°C. DIC Images taken at 63x magnification, where shading reflects the differences in position relative to the focal plane of the free floating droplets. Scale shown as black bars sized to 10 µm.

DOI: https://doi.org/10.7554/eLife.31486.020
Appendix 1—figure 1. Contact definitions. (A) Contacts are identified first as sp² planes in which at least two pairs of atoms come within 4.9 Å of one another, and then by restricting to the subset with (B) planar surfaces (at the carbon VDW radius of 1.7 Å) with points along the planar normal vectors coming within 1.5 Å of one another and (C) a planar orientations for which the absolute value of the dot product of normal vectors is ≥0.8. (D) Shows the rationale for these restrictions, where binning sidechain-sidechain interactions by the relative orientation between planes shows that planar (same-orientation) interactions, primarily in the 0.8 to 1.0 range (angles between the planes from 0 deg to 36 deg), show enrichment relative to the uniform distribution expected for random orientations. Of these, interactions with only one atom-atom pair within VDW contact (shown in blue) have no bias. Enrichment comes entirely from contacts with either two pairs of planar surfaces within 1.5 Å of each other (shown in purple) or two distinct pairs of atoms within 4.9 Å but without the planar surface contact (shown in green). (E) Minimum distance measurements between pairs of atoms found in separate sp² groups, measured from the closest pairing for each atom. Gray shows all sidechain-sidechain measurements, and green/purple show distances corresponding to the groups in D. (F) Representative examples of sidechain-sidechain and sidechain-backbone pi-contacts are shown as sticks (PDB: 1gde), with carbon atoms in gray, oxygen in red, and nitrogen in blue. Planar normal vectors extended to the carbon VDW radius, representing pi-orbital locations, are shown as purple rods for sidechain groups and blue rods for backbone groups, and the yellow line denotes a hydrogen bond where both donor and acceptor atoms are in pi-contact distance to a third sidechain. (G) A space-filling representation of the sp² atoms in F, with gray lines between normal vector rods used to show the planar surface measurements taken for defining pi-contacts.

DOI: https://doi.org/10.7554/eLife.31486.025
Appendix 1—figure 2. Cross validation against NMR restraints and X-ray structure resolution. (A), The relationship between contact frequency and experimental data quality is not unique to crystallography, as shown by the effect of increasing the number of restraints on sidechain specific contact frequencies over 2589 structures solved by NMR. For each sidechain/protein combination we calculated the average number of distance restraints involving sidechain atoms (from the first sp² atom onward), and then binned residues into five categories, with red for structures without any sidechain distance restraints for that residue type, and ranking quartiles from light gray to black by order of increasing restraints, where the consistent increase in contact frequency from left to right confirms that more restraints result in higher planar pi-contact frequencies. For Glu and Asp, less than 1% of the structures were derived using distance restraints to the carboxyl’s lone carbon atom so we did not split them into quartiles. (B), To control for potential sample bias we also tested the relationship between resolution and contact frequency for crystallographic structures that have been solved at least three different times at different resolutions, with bars showing contact frequencies over identical populations of residues for the highest (blue), median (black), and lowest resolution (red) structures. Error bars show standard error of the mean (SEM).

DOI: https://doi.org/10.7554/eLife.31486.026
Appendix 1—figure 3. Pi-pi interactions underestimated by some energy functions. (A), Contact frequency during molecular dynamics simulations of 100 proteins, made available through Dynameomics (Kehl et al., 2008), shows a rapid initial loss of >80% of sidechain pi-contacts which continues to decline throughout the simulation (blue points). By comparison, sidechain hydrogen bonding shows a stable loss of only 20% of interactions (red points). (B), Minimization of 762 crystal structures against the Talaris2014 energy function by Rosetta3.4 (Leaver-Fay et al., 2011; O’Meara et al., 2015), with starting contact frequencies (left bars) decreasing after minimization (right bars). (C–F), Analysis of the relationship between the energetic effects of point mutations (ΔΔG) and pi-contacts for experimental ΔΔGs (blue bars) and ΔΔGs predicted by simulation against the FOLDX force field (Schymkowitz et al., 2005) (C,E) and Rosetta (D,F). Panels C,D show predicted ΔΔG values vs. observation for residues that are not involved in pi-contacts in black, and residues that are involved in pi-contacts in blue, with lines of best fit colored the same. Panels E,F show how correlation values change as outliers are removed, with correlation consistently worse for mutations involving pi-contacts (blue lines) relative to those that don’t (black lines).

DOI: https://doi.org/10.7554/eLife.31486.027
Appendix 1—figure 4. Hydrogen bonding correlates with planar-pi contacts. Percentage of sidechains involved in at least one hydrogen bond is shown for sidechains that are not in a planar-pi contact in black, and for sidechains that are in a planar-pi contact in green, with panel (A) showing the hydrogen bond frequency across all groups, including ligands and water, (B) showing the hydrogen bond frequency to backbone atoms, and (C) showing the frequency of hydrogen bonding to a sidechain. Hydrogen bond frequency consistently increases with planar pi-pi contacts for all sidechains but Trp and Tyr.

DOI: https://doi.org/10.7554/eLife.31486.028
Appendix 1—figure 5. Backbone pi-pi contacts in secondary structure motifs. Examples of secondary structure motifs showing enrichment for local backbone pi-contacts (contacts made to sidechains within 5 residues of the peptide bond) are displayed. Bar graphs show contact frequency at each position in a motif, as defined by DSSP (Kabsch and Sander, 1983) abbreviated residue class ('E', 'S', 'T', 'H', 'G', and ' '), with bars colored by the associated residues, with green for peptide bonds between two residues classified as turns, blue for bonds in strands, red in helices, and black for bonds that are either unclassified or present at the transition point between classifications. Gray horizontal lines represent the decile values across all backbone contact frequencies, showing that the bonds most likely to end up in the top decile come primarily from transition points between secondary structures (ranging from 2x to 20x enrichment, relative to the median of 1.7%). Protein structures show representative examples of each motif with contacts found at the most enriched position, taken from (A), PDB:1aap, (B), PDB:1gte, (C), PDB:1k5c, (D), PDB:1nhc, (E), PDB:1k3i, (F), PDB:1i8k, (G), PDB:2c4w, and (H), PDB:1kwf.
DOI: https://doi.org/10.7554/eLife.31486.029
Appendix 1—figure 6. Peptide sequence effects on contact frequency. Heatmaps show enrichment in the total proportion of planar pi-pi contact involvements observed for peptide bonds between two residues (the first, N-terminal residue on the x-axis and the second, C-terminal residue on the y-axis) relative to the proportion of peptide bonds. Enrichment for (A) short-range contacts (sequence separation <5) and (B) long-range contacts (separation ≥ 5 or a different chain), respectively, to the peptide bond itself. (C), Enrichment for finding residues within 5 residues of a sidechain that makes a pi-contact to any group in the structure, demonstrating general sequence effects on the contact propensity of neighboring residues. The color gradient shows the natural logarithm of the observed over expected ratio.

DOI: https://doi.org/10.7554/eLife.31486.030
Appendix 1—figure 7. Phase separation propensity predictor testing. (A), ROC curve comparisons of predictor quality for scores made at different points during the training process, measuring ranking against the full test set (N = 62) vs. the human proteome (only sequences with length ≥140) with green showing the results for the highest number of pi-contacts predicted for any 100 residue window, without any weighting for type (AUC:0.82 ± 0.03), pink and orange showing the same measurement split between long-range (AUC:0.85 ± 0.03) and short-range contacts (AUC:0.62 ± 0.04), respectively, and blue showing the final predictor, which uses weighted combinations of both short- and long-range contact predictions (AUC: 0.88 ± 0.02). (B), the final score tested against 59 phase-separating sequences designed by the Chilkoti lab (Quiroz and Chilkoti, 2015; MacEwan et al., 2017; Simon et al., 2017) (detailed in Figure 5—source data 1C), with comparisons against the full set shown (N = 59) in blue (AUC: 0.86 ± 0.03), and then split into green for 18 proteins shown to phase separate from soluble to insoluble as temperature decreases (AUC:0.99 ± 0.01) and pink for the remaining 41 proteins which phase separate from soluble to insoluble as temperature increases (AUC:0.80 ± 0.04). (C), Fraction of sequences at or above a given PScore, with the combined pool of phase separation test set proteins (N = 121), in black, being compared to three reference proteome sets, with human in pink, S. cerevisiae in blue, and E. coli in green. (D), Enrichment plot for data shown in (C), with ≥PScore frequency for the test set shown relative to proteome frequencies. Analysis based on Figure 5-source data 1 and 2.
DOI: https://doi.org/10.7554/eLife.31486.031
Appendix 1—figure 8. Sequence comparisons of high PScore proteins. Panel (A) shows compositional bias, relative to the human average, for the high PScore disordered proteins (x-axis) and low PScore disordered proteins (y-axis) used in panel B. High PScore disordered proteins are enriched primarily in Pro and Gly, while low PScore disordered proteins are not enriched in either, but enriched primarily in Lys and Glu, matching our observation that Arg to Lys mutations abrogate phase separation propensity. Panel (B) shows similarity to the training set measured by minimum dipeptide profile distance to any training set protein, as described in the methods. High PScore (≥4.0) human sequences (in pink) are on average closer to the training set than are all human proteins (in black) or PDB sequences (in green), but the range overlaps with both, and is distinct from the similarity seen in blast level homologs of the training set (in blue). Panel (C) shows Shannon entropy distributions of the human proteome (in black), the PDB (in green), and of a set of human proteins proteins predicted to have long stretches of disorder (Disprot3 ≥0.8) split into those with high PScore (≥4, N = 310) (in pink) and low PScore (<1.0, N = 1044) (in orange), showing that PScore but not disorder results in a bias towards lower sequence entropy, suggesting a compositional bias in phase-separating sequences. Panel (D) shows Shannon entropy values for our natural-protein phase separation test set (N = 62) in pink and the disorder-containing human proteins found in Disprot (N = 205) in orange, confirming the observation in panel C that lower Shannon entropy sequences are associated with phase separation.

DOI: https://doi.org/10.7554/eLife.31486.032
### Appendix 1—figure 9. Prediction examples.

Per-residue PScores used to calculate the final full sequence PScore are shown for a selection of human proteins, with residues colored from purple (PScore ≤ −2) to white (PScore = 0) to green (PScore ≥ 4.0). Black triangles denote residues annotated by PhosphoSitePlus as targets of PTMs, blue triangles denote modification sites with known regulatory significance, and red circles denote modification sites with known disease relevance. Proteins are annotated with the percentage of GO terms (with at least 10 human proteins) and high PScore-enriched GO terms (Panther analysis, PScore ≥ 4, with O/E > 1) of which the protein is a member, as well as the total number of each for which the annotated protein has the highest PScore in the set. Examples are grouped by (A), involvement in synaptic plasticity and neuronal behavior, showing synaptic functional regulator FMR1, and synaptophysin; (B), intracellular biomaterials and related structural proteins, showing focal adhesion kinase 1, vimentin, and keratin type I cytoskeletal 10; (C), proteins involved in signaling pathways, showing CCR4-NOT transcription complex subunit 3, β-catenin, vitamin D3 receptor, and Smoothened homolog; and (D), proteins involved in extracellular biomaterials, showing fibrinogen alpha chain and dentin sialophosphoprotein. (E) The cystic fibrosis transmembrane conductance regulator is shown as an example of a negative prediction, even though containing a large region of intrinsic disorder (residues ~650–840).

**Data Citation:**

DOI: https://doi.org/10.7554/eLife.31486.033