The blobulator: a webtool for identification and visual exploration of hydrophobic modularity in protein sequences

Connor Pitman[1]**, Ezry Santiago-McRae**[1]**, Ruchi Lohia**[2]**, Kaitlin Bassi**[1]**, Thomas T. Joseph**[3]**, Matthew E.B. Hansen**[4]**, and Grace Brannigan**[1,5,*]

[1]**Center for Computational and Integrative Biology, Rutgers University–Camden, 201 Broadway, 08103, NJ, USA**
[2]**Department of Physiology, University of Toronto, 1 King's College Circle, M5S 1A8, Toronto, Ontario, Canada**
[3]**Department of Anesthesiology and Critical Care, Perelman School of Medicine, University of Pennsylvania, JMB 305, 3620 Hamilton Walk, 19104, PA, USA**
[4]**Department of Genetics, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd, 19104, PA, USA**
[5]**Department of Physics, Rutgers University–Camden, 201 Broadway, 08103, NJ, USA**

ABSTRACT

Motivation: Clusters of hydrophobic residues are known to promote structured protein stability and drive protein aggregation. Recent work has shown that identifying contiguous hydrophobic residue clusters (termed "blobs") has proven useful in both intrinsically disordered protein (IDP) simulation and human genome studies. However, a graphical interface was unavailable.
Results: Here, we present the blobulator: an interactive and intuitive web interface to detect intrinsic modularity in any protein sequence based on hydrophobicity. We demonstrate three use cases of the blobulator and show how identifying blobs with biologically relevant parameters provides useful information about a globular protein, two orthologous membrane proteins, and an IDP. Other potential applications are discussed, including: predicting protein segments with critical roles in tertiary interactions, providing a definition of local order and disorder with clear edges, and aiding in predicting protein features from sequence.
Availability: The blobulator GUI can be found at `www.blobulator.branniganlab.org`, and the source code with pip installable command line tool can be found on GitHub at `www.GitHub.com/BranniganLab/blobulator`.

Keywords: sequence-structure relationship, single-nucleotide polymorphism, computational methods, hydrophobicity, protein sequence organization, intrinsically disordered proteins

1 INTRODUCTION

Hierarchical frameworks play a critical role in the conceptual understanding of protein structure, function, and evolution. The classic structural hierarchy (primary, secondary, tertiary, quaternary structure) remains central to designing and interpreting biophysical studies that span the proteome. Consequently, many broadly-applicable informatics analyses rely explicitly on structural modularity, either experimentally determined [1] or predicted from sequence [2, 3]. Others implicitly neglect protein modularity by arbitrarily defining a "local" sequence to the residue of interest and neglecting distant residues.

Structural hierarchies present conceptual limitations when detecting the organization implicit within protein sequences. The classical structural hierarchy is best suited for proteins with known structures and excludes intrinsically disordered proteins (IDPs) entirely [4]. Additionally, it discounts salient collective features of tertiary structure, such as solvent-accessible surface area, while also implicitly under-emphasizing unresolvable residues found in random coil and linking regions. Approximately one-third of the eukaryotic proteome does not fit easily into the classic structural hierarchy [5], requiring other approaches to detect organization in protein sequences.

Most tools that extract information innate to protein sequences either utilize the structural hierarchy framework or work without a hierarchical framework. Examples include predicting disease-associated mutations [6–12], predicting solubility [13–15], detecting aggregating regions [16–20], predicting intrinsic disorder [18, 21–25], designing protein sequences with specific features [26–28], and comparing sequences [29, 30]. Some of these tools use a fixed-width moving window to define local sequence context, which artificially places the considered residue at the center of its "local sequence" and ignores any natural boundaries present within proteins. Alternatively, under the classical structural hierarchy, a residue's local sequence context is typically considered to be its associated secondary structure element. This can be defined in several ways: using a protein's solved structure, a secondary structure predictor, or a tertiary structure predictor that predicts secondary structure. However, tools for predicting secondary structure in those proteins without solved structures (as well as IDPs) have limited accuracy [31–34]. Tertiary structure predictors such as AlphaFold do predict secondary structure, but this is not their primary function, and any potential biases in detecting secondary structure remain unclear.

To address some of these limitations, we previously developed an algorithm called "blobulation" to segment protein sequences based on hydrophobicity. We have already used blobulation to meaningfully segment an IDP in order to measure tertiary interactions in simulation data [35], and later demonstrated its use in providing a meaningful definition of "local context" in a study on disease-associated mutations in the human exome [36]. Given an amino acid sequence, blobulation outputs residue clusters, termed "blobs" as inspired by polymer physics [37].

In this paper, we introduce the blobulator, a graphical user interface (GUI) for blobulating protein sequences. This GUI allows users to view protein sequence properties with varying levels of granularity in real-time by interactively adjusting the analysis parameters. Additionally, the user can introduce mutations into the sequence and visualize how these changes affect the blob topology. After giving an overview of the underlying algorithm and the blobulator GUI, we provide examples of how one might use the blobulator to gain insight into various types of proteins, specifically a globular protein, two membrane proteins, and an IDP.

2 THE BLOBULATION ALGORITHM

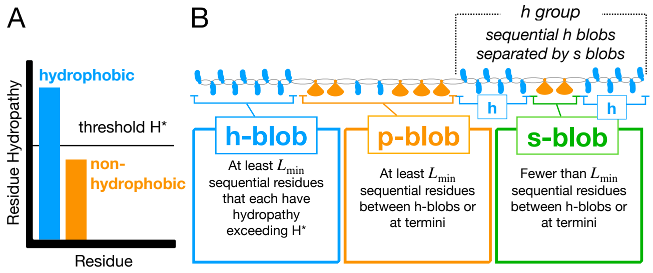

A
B
h group
sequential h blobs
separated by s blobs
hydrophobic
threshold H*
non-
hydrophobic
Residue Hydropathy
Residue
h
h
h-blob
At least L_{min} sequential residues that each have hydropathy exceeding H*
p-blob
At least L_{min} sequential residues between h-blobs or at termini
s-blob
Fewer than L_{min} sequential residues between h-blobs or at termini

Figure 1. Blobulation algorithm. Figure adapted from Lohia et al. 2022 [36]. A) First, the sequence is digitized. Residues are classified as either hydrophobic (blue) or non-hydrophobic (orange) by comparing their hydropathy to the user-selected threshold, H^*. B) The sequence is then segmented into one of three categories: longer than L_{min} and hydrophobic (h-blobs, blue), longer than L_{min} and non-hydrophobic (p-blobs, orange), or shorter than L_{min} and non-hydrophobic (s-blobs, green).

Whole-sequence blobulation consists of two steps: digitization and clustering.

1. In the digitization step, the algorithm defines a given residue as hydrophobic or non-hydrophobic as follows:

   (a) The user selects a normalized (0 to 1) hydrophobicity scale, and sets a threshold hydrophobicity (H^*) on this scale. The default settings are the Kyte-Doolittle hydrophobicity scale [38] and a H^* of 0.4.

   (b) The hydropathy for each individual residue (i) is calculated based on the normalized hydrophobicity scale.

   (c) The hydropathies are smoothed for each residue i and the residues adjacent to it in sequence ($i-1$, i, and $i+1$) , yielding the smoothed hydropathy H_i.

   (d) Residue i is classified as hydrophobic ($H_i > H^*$) or non-hydrophobic ($H_i \leq H^*$), shown in Fig. 1A.

2. In the clustering step, subsequences termed "blobs" are defined as follows:

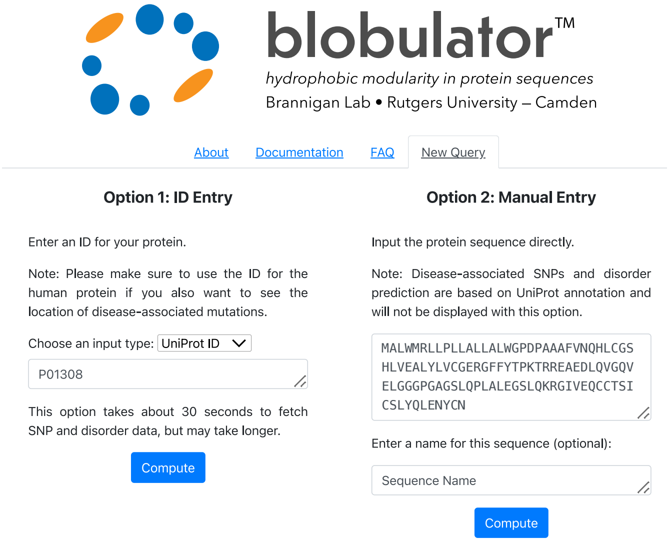
   (a) Subsequences of at least L_{min} sequential hydrophobic residues are classified as hydrophobic blobs (h-blobs)(Fig. 1B).

   (b) All other linking sequences are then classified based on their L_{min} as either non-hydrophobic blobs (p-blobs, $L \geq L_{\text{min}}$) or short blobs (s-blobs, $L < L_{\text{min}}$) (Fig. 1B).

Though there are default settings for the initial blobulation, the user can tune various parameters. For example, while previous studies from our lab [35, 36] used the Kyte-Doolittle scale, the user may select the Eisenberg-Weiss scale [39]. More hydrophobicity scales will be added in the future. Different combinations of H^* and L_{min} will also detect blobs with varying properties, which may be useful depending on one's research question (some examples are shown in the Applications section). While there are no "correct" settings for H^* and L_{min}, choosing biologically relevant parameters is critical. Varying parameters may also reveal different layers of organization. Very high thresholds (H^* approaching 1) will return one large p-blob, and very low thresholds (H^* approaching 0) will return one large h-blob, while moderate thresholds will reveal intrinsic segmentation within the protein sequence. High L_{min} will only give meaningful results with a low to moderate H^*.

The higher-order organization of blobs can also reveal useful information. For example, "blob groups" are h-blobs separated only by s-blobs (some examples of these are shown in the globular protein and IDP application sections). While blob groups are not explicitly labeled in the GUI, they are labeled as follows in the blobulator's CSV output: by their type (h, p, s) and group number (1, 2, 3) and, if necessary, a sub-group letter (a, b, c) which form a unique label for every blob. This is shown in Fig. 3G (h1a, s1, h1b, p1, etc). We also refer to the order, number, and length of blobs in a protein sequence as its "blob topology".

3 THE BLOBULATOR GUI


blobulator™
hydrophobic modularity in protein sequences
Brannigan Lab • Rutgers University – Camden
About
Documentation
FAQ
New Query
Compute
Compute

Option 1: ID Entry

Enter an ID for your protein.

Note: Please make sure to use the ID for the human protein if you also want to see the location of disease-associated mutations.

Choose an input type: UniProt ID

P01308

This option takes about 30 seconds to fetch SNP and disorder data, but may take longer.

Option 2: Manual Entry

Input the protein sequence directly.

Note: Disease-associated SNPs and disorder prediction are based on UniProt annotation and will not be displayed with this option.

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGS
HLVEALYLVCGERGFFYTPKTRREAEDLQVGQV
ELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
CSLYQLENYCN
```

Enter a name for this sequence (optional):

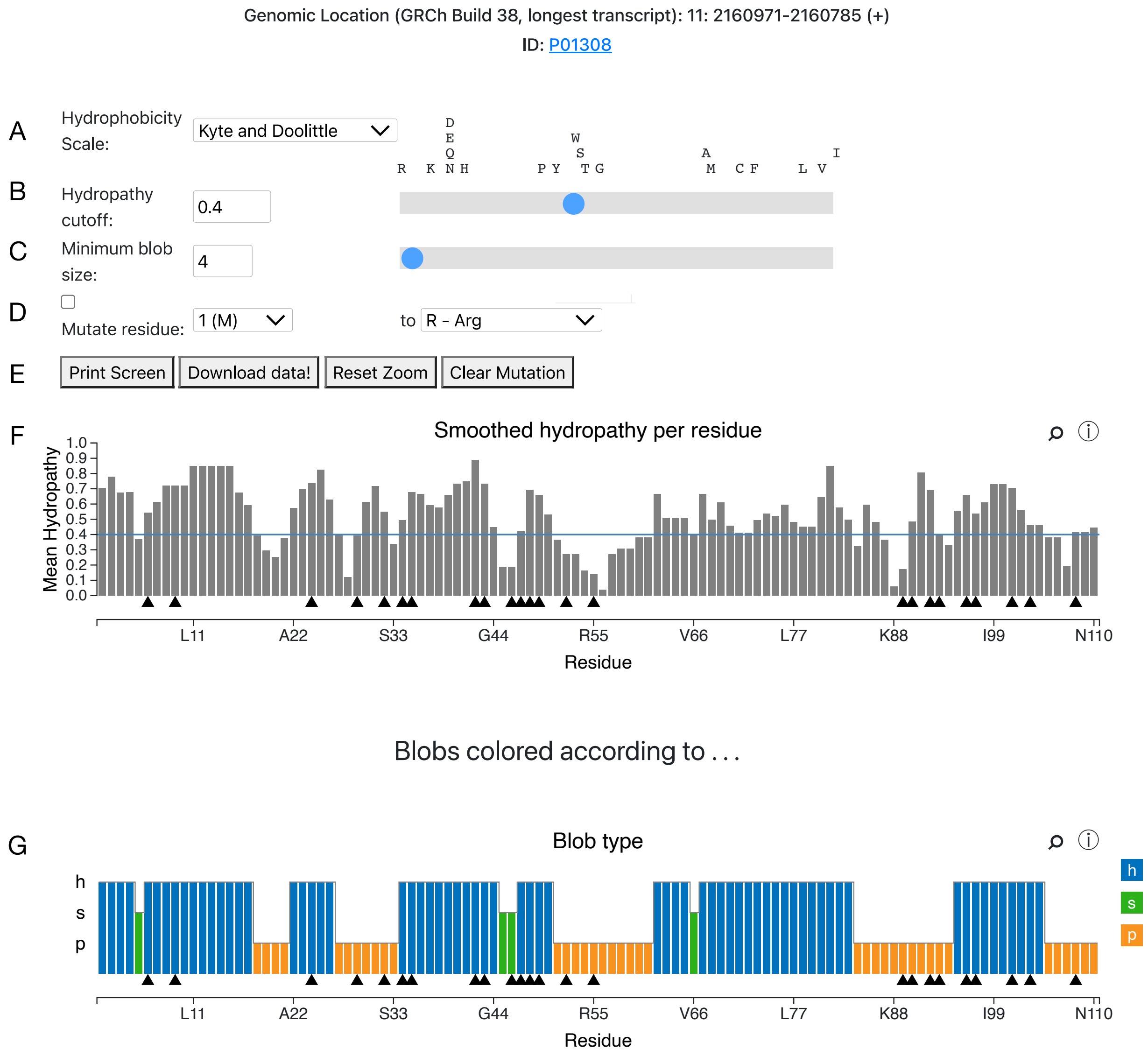
Sequence Name

Figure 2. Screenshot of the blobulator home page. Protein sequences are submitted via UniProt ID, ENSEMBL ID (Option 1: ID Entry), or the protein sequence and sequence name (Option 2: Manual Entry). Default values are for alpha-synuclein. Available at `www.blobulator.branniganlab.org`

Blobulation of any amino acid sequence can be done using the blobulator found at `www.blobulator.branniganlab.org`. On the "New Query" tab of the homepage (Fig. 2), the user can submit either a database ID or manually enter a protein sequence for blobulation (Fig. 2). From here, a user can blobulate any protein or amino acid sequence using an accepted ID or manually entering the sequence directly, as well as toggle to other information tabs. Figure 3 illustrates blobulation of insulin, a peptide hormone that promotes glucose absorption by cells.

After blobulation, the user can interactively tune some of the parameters. For example, they can change the hydrophobicity scale used for digitization (Fig. 3A) or interactively modify the H^* threshold and L_{min} cutoff (here we use H^*= 0.4 and L_{min}= 4, respectively) to segment the protein into components with varying hydrophobicity and length properties. This can be done by adjusting the respective sliders or manually entering a value (Fig. 3B and C). A user may also introduce a mutation by selecting one of the black triangles or manually entering the position and alternate amino acid in the "mutate residue" field (Fig. 3D). All changes update dynamically. Users can download both an image of the output as a PDF and the raw data (in CSV format) used to generate the GUI output (Fig. 3E).

Each residue's smoothed hydropathy H_i (as discussed in section 2) is shown in the first outputted track (Fig. 3F) along with the H^* threshold (blue line). The next track shows the output of the clustering step (Fig. 3G), where blobs are colored by their blob type. Blobulation of insulin using the settings H^*= 0.4 and L_{min}= 4 identifies 8 h-blobs, 5 p-blobs, and 3 s-blobs. If a UniProt ID is provided, known disease-associated mutations are obtained from UniProt and displayed on every track as black



Genomic Location (GRCh Build 38, longest transcript): 11: 2160971-2160785 (+)

ID: P01308

Hydrophobicity Scale: Kyte and Doolittle

Hydropathy cutoff: 0.4

Minimum blob size: 4

□ Mutate residue: 1 (M) to R - Arg

Figure 3. Partial screenshot of the interactive blobulator results page. Several settings are available: A) hydrophobicity scale (drop-down menu); B) hydropathy cutoff (H^*) (text entry, slider, or selecting a residue name); C) minimum blob size (L_{min}) (text entry, or slider); and D) the protein sequence can be mutated (toggle: check box, mutate from: first drop-down, mutate to: second drop-down). E) Download the results page as a pdf, download the underlying data as a csv, or reset the zoom for all tracks, or clear any mutations. F) Each residue's hydrophobicity (grey bars) compared to the threshold H^* (blue, horizontal line). Residues can be mutated with the black triangles (only available with UniProt ID) and the sequence can be zoomed (zoom in: click and drag, zoom out: double click or click "Reset Zoom" (E)). G) Blob type is represented by both height (shown on the y-axis) and color. Height is preserved in subsequent tracks to indicate blob type. Controls are as in F. The example protein is insulin (UniProt ID: P01308).

triangles (Fig. 3). Hovering over these with the cursor will display the effect of the mutation on the amino acid, its reference SNP cluster ID (rsid), and link to the Single Nucleotide Polymorphism Database (dbSNP).

The GUI includes a series of tracks in which blobs are colored by additional information relating to charge composition, hydrophobicity, length, or some combination of the three. As shown in figures 4, 5, and 6, these include tracks displaying blob charge and globular tendency, a track showing the predicted sensitivity of each blob to deleterious mutations, and two tracks showing the predicted order and disorder of each blob, respectively.

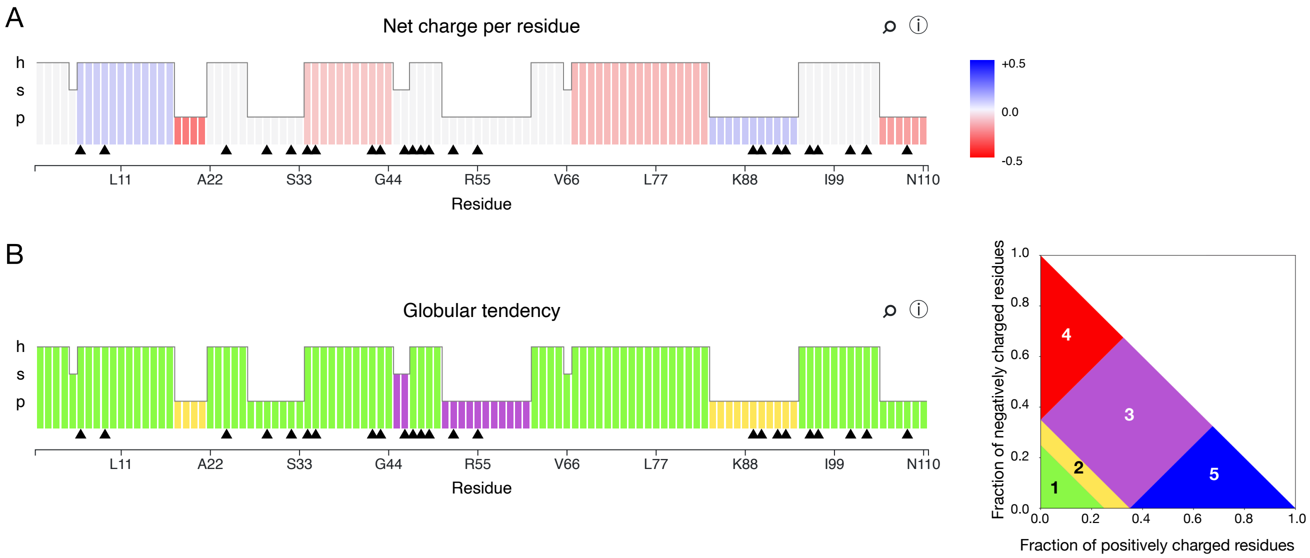


Figure 4. Charge-based tracks. Blobulation of insulin as in Fig. 3. A) Blobs colored according to net charge per residue: positive (blue) or negative (red). B) Blobs colored by their Das-Pappu phase [40]: phase 1 (globular, green), phase 2 (Janus/boundary, yellow), phase 3 (strong polyelectrolyte, purple), phase 4 (strong polyanion, blue), or phase 5 (strong polycation, red).

The first two of these tracks are colored by calculations based solely on blob charge content: mean net charge per residue and globular tendency. (Fig. 4). In the mean net charge per residue track (Fig. 4A), blobs are colored by the average charge of all the residues within each blob. The use of this track is demonstrated in the example in section 4.2. In the globular tendency track (Fig. 4B) blobs are colored based on the region they occupy in the Das-Pappu diagram, which predicts globular tendency based on charge content [40]. The use of this track is demonstrated in section 4.1.

The predicted mutation sensitivity track is colored based on an enrichment calculation that includes both blob hydrophobicity and length (Fig. 5). This track colors blobs based on the enrichment of disease variants in blobs with the same properties found in our previous study [36]. This study used a dataset containing 70,000 human disease-associated and non-disease-associated missense mutations.

The final pair of tracks discussed here are colored based on calculations that include both blob hydrophobicity and net charge to estimate the disorder of each blob (Fig. 6). The first of these tracks colors blobs by their signed distance from the order/disorder boundary defined by the Uversky-Gillepse-Fink boundary plot [41] (shown on the right) using its mean hydrophobicity and mean net charge per residue (Fig. 6A). The use of this track is examined in section 4.3. The next track colors blobs by the PV2 predicted disordered fraction [42] (Fig. 6B), stored in the database of disordered prediction (D2P2) and retrieved from UniProt. We note that there is generally a lack of consensus among disorder predictors for whether or not a given group of residues are disordered [43]. Additionally, if the user does not access the protein sequence using a UniProt ID, this track will not appear.

4 EXAMPLE APPLICATIONS

In this section, we present the blobulation of three different types of proteins: a globular protein, a membrane protein, and an IDP. For each protein, we demonstrate how blobulation adds additional context to the known properties of these proteins.

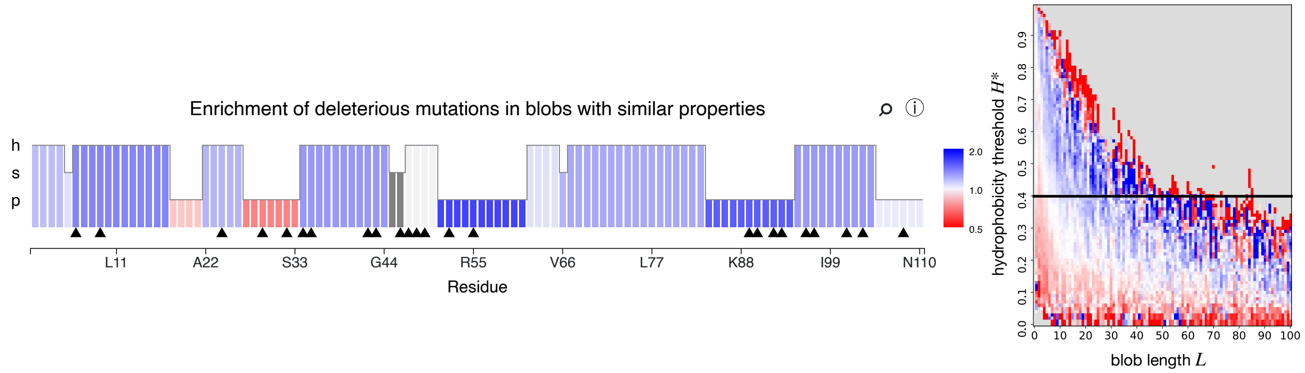


Figure 5. Predicted enrichment of disease-associated SNPs (dSNPs). Left: enrichment of dSNPs in hydrophobic blobs based on analysis of a large dataset of human SNPs [36]. The hydrophobicity cutoff ($H^* = 0.4$) used is indicated (black, horizontal line). Right: Blobulation of insulin as in Fig. 3 colored by dSNP enrichment (blue: enriched, red: depleted) based on analyses from [36].

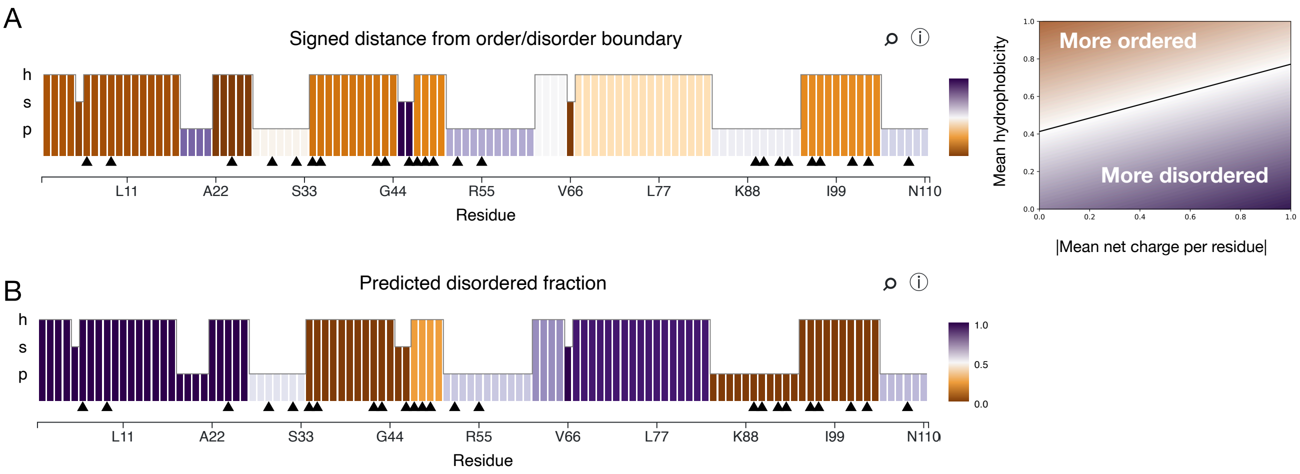


Figure 6. Disorder-based tracks. Blobulation of insulin as in Fig. 3. A) Blobs colored according to the Uversky-Gillepse-Fink boundary plot (right)[41]. B) Blobs colored according to their predicted fraction of disordered residues, according to PV2 as provided by the Database of Disordered Protein Prediction (D2P2).

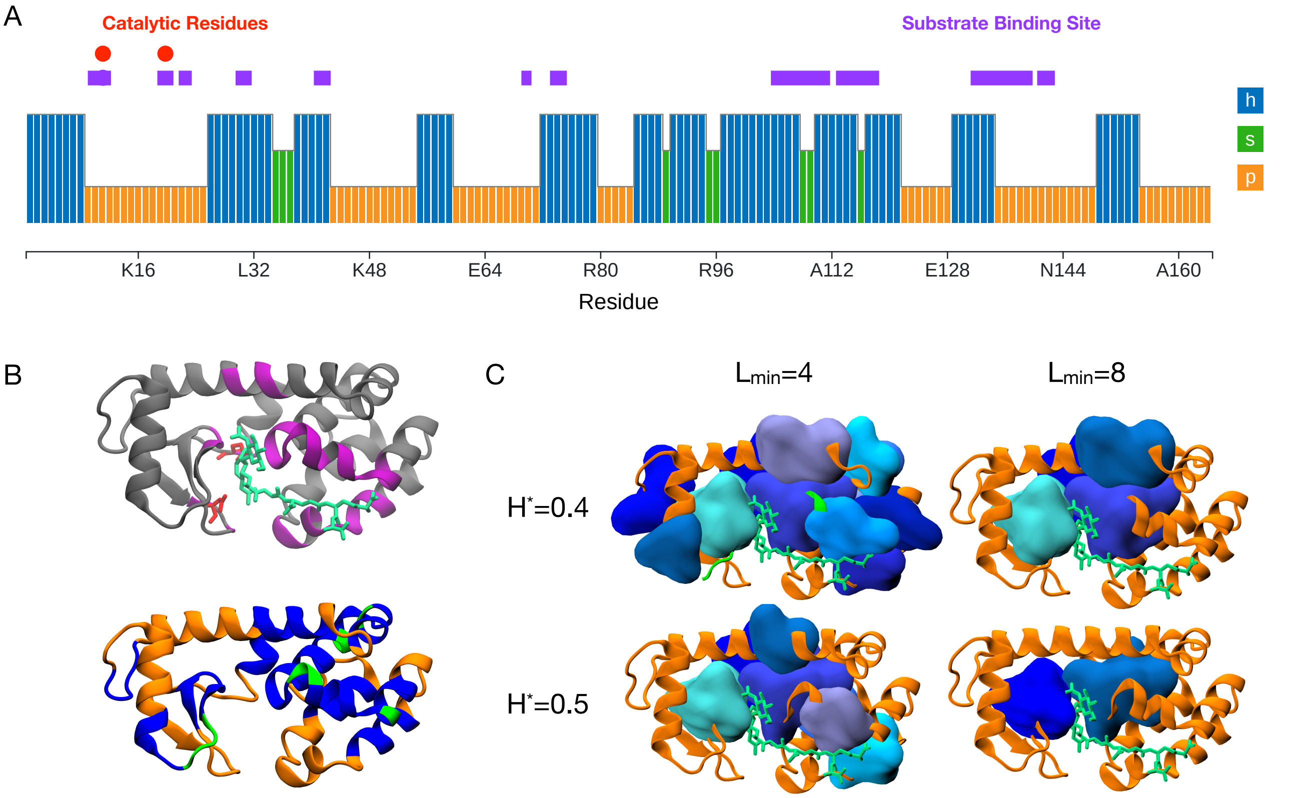


Figure 7. Blobulation of lysozyme (UniProt ID P00720). A) Blobulation track using default settings (H^*= 0.4, L_{min}= 4). Annotations indicate catalytic residues (red) and the substrate binding site (purple). B) Molecular image of lysozyme (PDB:2LZM) with peptidoglycan (green) colored by the substrate binding site (left, residues found within 7Å of peptidoglycan) or by blob type (right, h-blob: blue, p-blob: orange, s-blob: green). C) Blobulation under increasingly stringent settings (H^*= 0.4 and 0.5, L_{min}= 4 and 8). H-blobs are shown as surfaces. Molecular images were generated in VMD [44, 45].

4.1 Globular Protein: Lysozyme

In an aqueous environment, most globular proteins have a highly hydrophobic core surrounded by a solvent-accessible surface. One such protein is lysozyme, an antibacterial enzyme that cleaves peptidoglycan's sugar component from its peptide component. Lysozyme has two catalytic residues near the N-terminus and several substrate contact sites along the sequence (represented by pink and purple dots respectively in Fig. 7A). To provide an example of how one might detect hydrophobic blobs that correspond to the hydrophobic core of a globular protein, we blobulated bacteriophage T4 lysozyme and varied both the hydrophobicity threshold and minimum length (Fig. 7B). Higher H^* and $L_{\min}$ eliminate the h-blobs that are detected at the surface of the protein when using more relaxed settings. Blobulation using the most stringent settings shown here (H^*= 0.5, $L_{\min}$= 8) reveals two h-blobs at the protein's center, away from the solvent-accessible surface. By gradually increasing parameter thresholds, we are able to detect the longest and most hydrophobic parts of a globular protein that correspond to the core, as well as the shorter and less hydrophobic blobs that interact at the protein's surface.

We also identified blobs containing contact residues within 7Å of peptidoglycan (shown in Fig. 7). When using a shorter $L_{\min}$ ($L_{\min}$= 4), a set of the detected h-blobs are found in proximity to peptidoglycan along its entire length. However, when using a longer $L_{\min}$ ($L_{\min}$= 8), the only substrate-binding h-blobs detected are those interacting with peptidoglycan's sugar component. Stabilization of this sugar ring is vital for the enzyme's ability to cleave peptidoglycan [46] and we find this ring wedged between two long hydrophobic blobs (Fig. 7B), providing an example of long hydrophobic blobs with interactions that are critical for function.

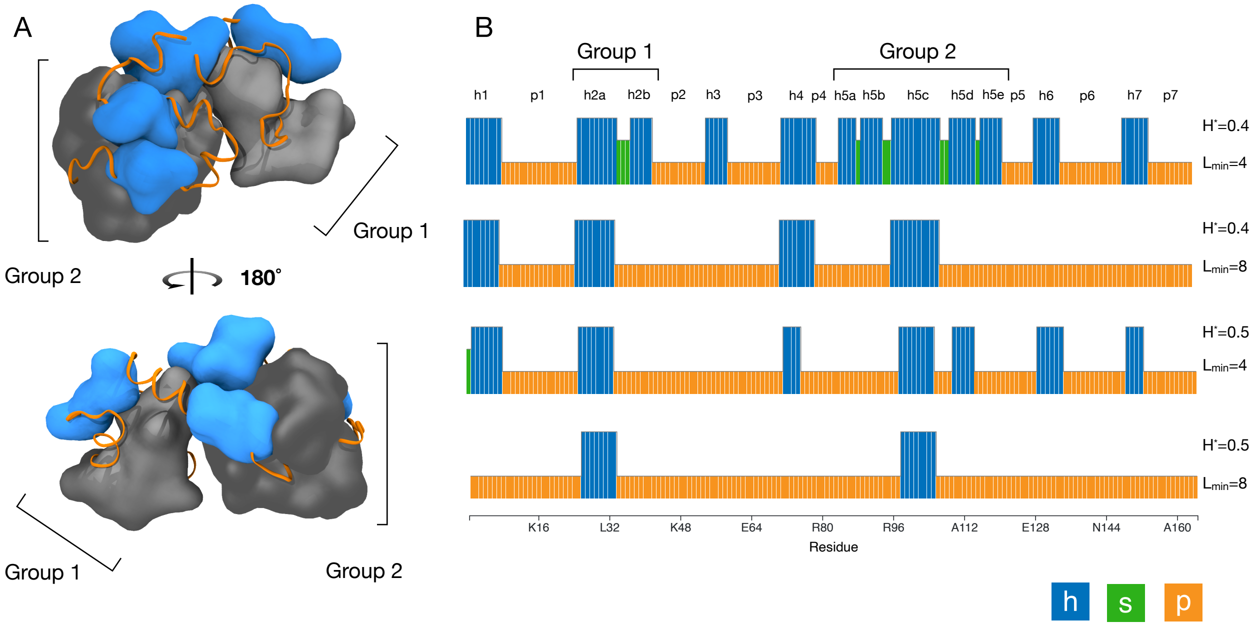


Figure 8. Blob groups in T4 lysozyme. A) Structural view (PDB:2LZM) blobulated using default settings (H^*= 0.4, $L_{\min}$= 4). Groups are grey, ungrouped h-blobs are blue, and p-blobs are orange ribbons. B) Blobulation of T4 lysozyme under increasingly stringent settings. Annotations indicate blob identifiers and blob groups. Molecular images were generated in VMD [44, 45].

Blobulation of lysozyme using "relaxed" settings (H^*= 0.4 and $L_{\min}$= 4) reveals two examples of "blob groups": sets of h-blobs separated only by s-blobs (as defined in section 2, shown in grey in Fig. 8A). Both groups are found near the center of lysozyme, surrounded by ungrouped h-blobs and p-blobs. Additionally, each group contains an h-blob that remains detected under increasingly stringent settings (Fig. 8B). The groups detected here are akin to synergistic tertiary elements formed from secondary structure elements, like alpha-helical bundles. This is an example of hierarchical blob clustering, where highly hydrophobic blobs are found in the protein core surrounded by less hydrophobic blobs within the same group, which are in turn surrounded by individual h-blobs.

To provide an example of how blobulation adds context to the effect of mutations, we blobulated the S117V and T157I mutants, which both affect the temperature sensitivity of lysozyme's stability through altering intraprotein hydrophobic interactions. S117 makes the protein more stable at higher temperatures by altering hydrophobic residue interactions in the substrate binding cleft [47], while T157I makes the protein less stable at higher temperatures and disrupts hydrogen bonding at

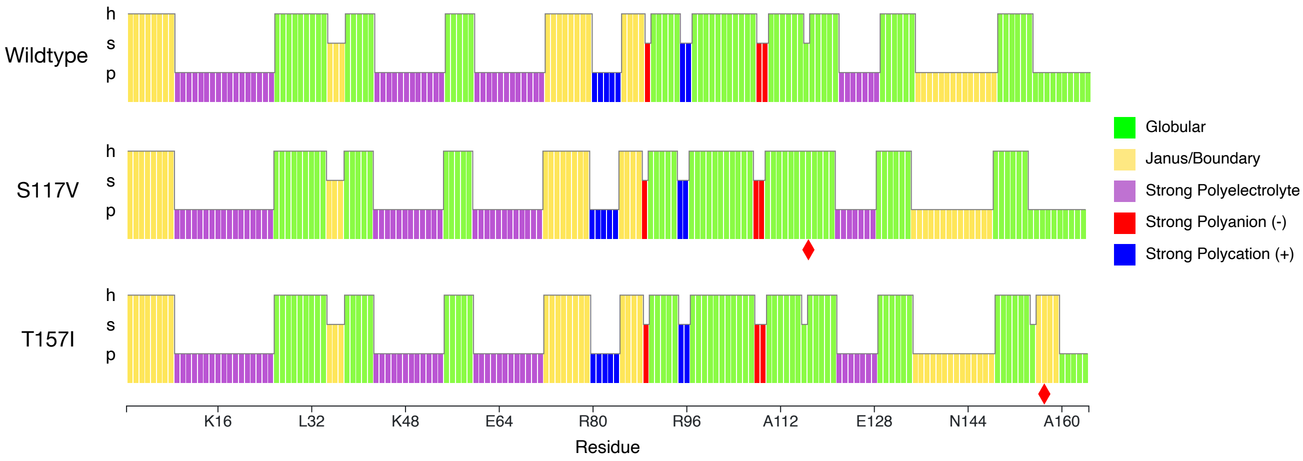


Figure 9. Globular tendency of T4 lysozyme blobs. Blobulation using default parameters (H^*= 0.4, L_{min}= 4). S117V increases thermostability and joins two h-blobs into one. T157I decreases thermostability and creates a s-blob and an additional h-blob. Colored as in Fig. 4B [40]. Red diamonds indicate mutated residues.

the periphery of the protein [48]. We find that both mutations change the blob topology (Fig. 9): S117V merges two h-blobs and T157I creates a new h-blob four residues in length. We have previously found that mutations that split, dissolve, or merge h-blobs are enriched for deleterious mutations [36], and this result is consistent with that finding. Additionally, the h-blob introduced by the T157I mutation is classified as a Janus region on the Das-Pappu diagram. Janus proteins often have degenerate conformations and switch between ordered and disordered depending on their environment [40], which may cause a change to this blob at higher temperatures and lead to less stability overall for the protein. Finally, it is noteworthy that blobulation is able to detect a change from both mutations, whereas tools such as AlphaFold (which predict atomic coordinates and not necessarily interactions) do not detect structural differences between these mutants and the wildtype (average backbone root mean squared deviation of predicted wildtype and mutant structures is 0.09Å).

4.2 Membrane Proteins: Pentameric Ligand-Gated Ion Channels (pLGICs)

Unlike soluble globular proteins, which enclose their hydrophobic residues inside a hydrophilic shell, integral membrane proteins use exposed hydrophobic regions as anchors in the cell membrane. The pentameric ligand-gated ion channel (pLGIC) protein family is responsible for neurotransmitter reception in the nervous system of the bilateria, but representatives are also found in many other eukaryotes and some prokaryotes [49]. Despite relatively low sequence identity and diverse environmental and ligand requirements, pLGICs are highly structurally conserved [50–54]. The N-terminal extracellular domain consists of a beta-sandwich, which is responsible for ligand sensing and much of the substrate selectivity (Fig. 10A-B). When the protein's ligand is bound, the signal is transmitted to the transmembrane helical bundle (Fig. 10D), which changes conformation to permit ions to pass through the channel formed by the five pore-lining helices. In this example, we will compare the blobulation of two structural models of the pLGIC family: first, the *Caenorhabditis elegans* glutamate-gated chloride channel (GluCl); and second, the proton-gated *Gloeobacter* ligand-gated cation channel (GLIC).

Starting from the N-terminus, both GluCl and GLIC contain disordered N-terminal extensions which include signal sequences (Fig. 11). GluCl has an additional hydrophobic blob in the N-terminal extension. P-blobs dominate the extracellular beta-sandwich in both cases. The two additional hydrophobic blobs in GluCl correspond to $\beta1$ and $\beta5$, which contain two of the four residues that contact glutamate during binding. GLIC, in contrast, is proton-gated via titratable residues mostly in the ECD [51] and lacks these extracellular h-blobs. The transmembrane domains of both GluCl and GLIC consist of four transmembrane helices per subunit [49], each about 19 residues long. This length corresponds to the typical thickness of a plasma membrane[55]. These helices are clearly identified using $H^* = 0.33$ and $L_{\text{min}} \approx 19$. The extended p-blob seen in the blobulation of GluCl (Fig. 10) corresponds to a disordered intracellular domain common in eukaryotic pLGICs [51].

As ion channels, electrostatic interactions are important for the function of all pLGICs. Although GluCl has a net positive charge in its ECD, GLIC has a net negative charge 11. This is consistent with, and may contribute to, the anion conductivity of GluCl and the cation conductivity of GLIC. Similarly, the pore-lining M2 helix is positively charged in GluCl, but negative in GLIC. More granular settings (e.g. $H^* = 0.7$, $L_{\text{min}} = 1$) reveal that the charges in the M2 blob are mostly on the intracellular side, close to the M1-M2 loop (data not shown). GLIC has an additional positive charge near the M2-M3 loop due to lysine 290, which can be seen by using the blobulator's zoom feature. Finally, the M4 helix of each protein has a net positive charge,

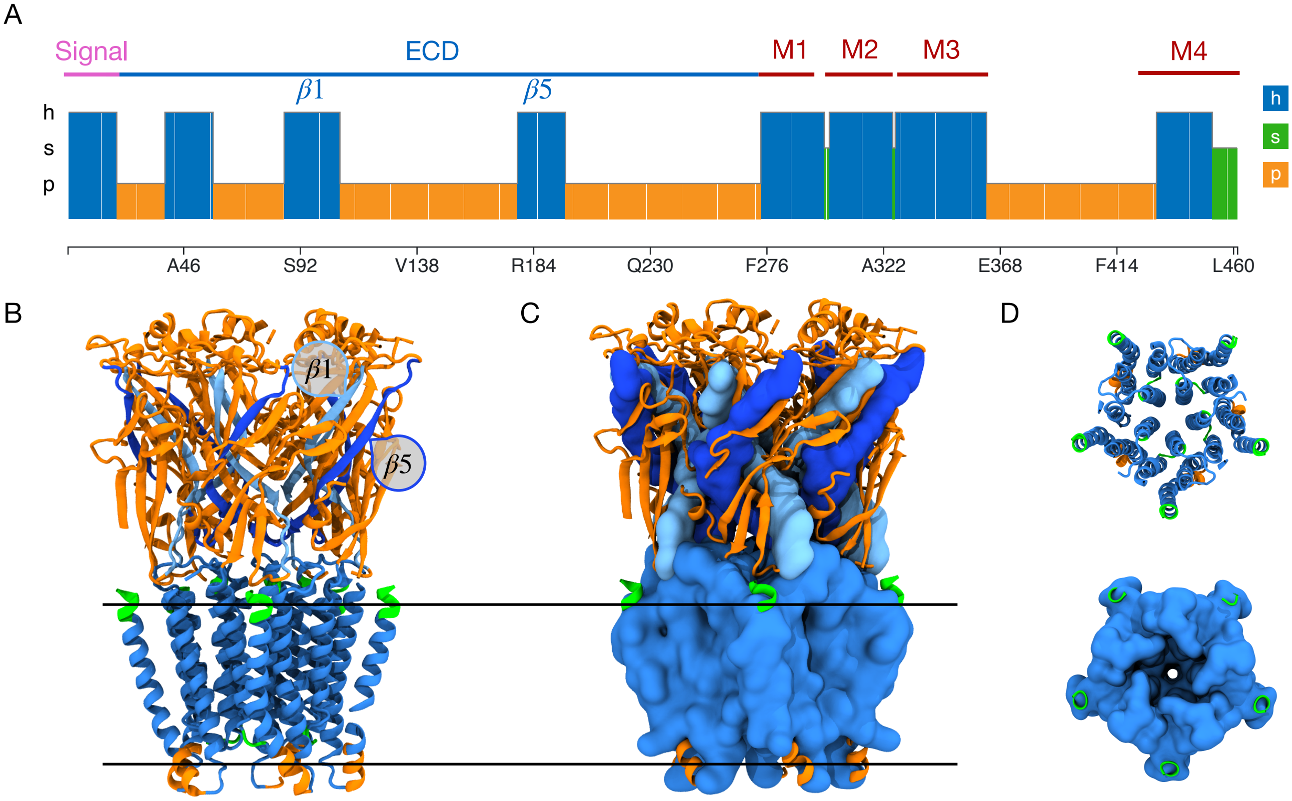


Figure 10. Blobulation of GluCl, a pentameric ligand-gated ion channel. A) Blobulation of UniProt ID G5EBR3 using settings to detect transmembrane regions ($H^* = 0.33$, $L_{\text{min}} = 19$). B) A cartoon representation of GluCl (PDB:3RHW) colored according to blob type and ID: beta 1 blob (light blue), beta 5 blob (dark blue), transmembrane blobs (intermediate blue), terminal s-blobs (green), p-blobs (orange). The first two blobs in the sequence are absent from the structure. The black lines represent the lipid membrane. C) GluCl showing h-blobs in surface view to better show blob-blob contacts. Colored as in B. D) Extracellular views of the TMDs of the proteins shown in B (upper) and C (lower). Molecular images were generated in VMD [44, 45].

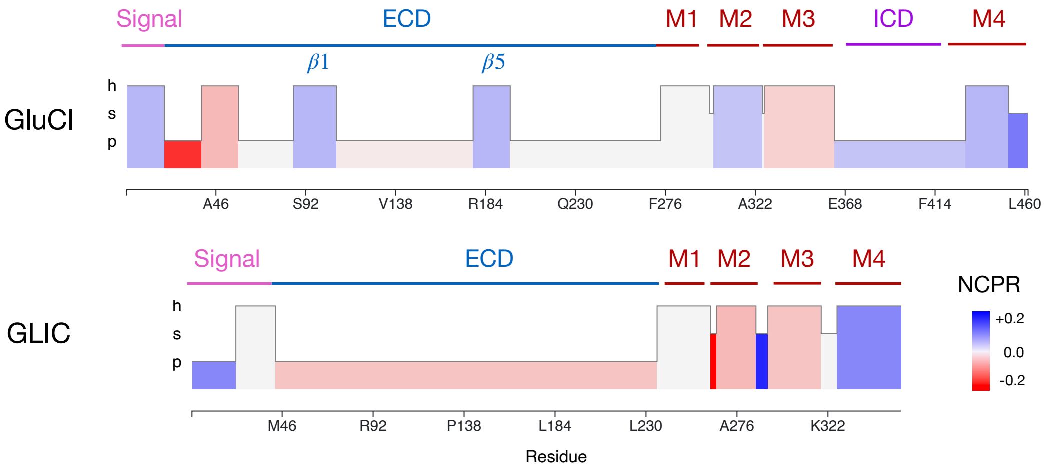


Figure 11. Comparative blobulation of GluCl, an anion-conducting pLGIC, and GLIC, a cation-conducting pLGIC. Net Charge Per Residue (NCPR) tracks for GluCl (UniProt: G5EBR3) and GLIC (UniProt: Q7NDN8) blobulated using settings to detect transmembrane regions ($H^* = 0.33$, $L_{\text{min}} = 19$). Tracks are aligned by the beginning of the h-blob containing the M2 (pore-lining) helix. Annotations indicate important sequence features available on UniProt. Signal sequences are indicated in pink, ECD regions are indicated in blue, and transmembrane helices (M1 to M4) are indicated in red.

consistent with experimental observations that M4 binds anionic lipids [56, 57]. These results suggest the blobulator's utility in interpreting membrane proteins either individually or within a larger evolutionary context. At present, only one protein can be blobulated at a time via the GUI, but future plans include the ability to compare two or more blobulated sequences simultaneously.

4.3 IDP: α-synuclein

Intrinsically disordered proteins (IDPs) fulfill critical roles despite lacking stable tertiary structure (as reviewed in Ref[58]). One such IDP is α-synuclein. Although its function is not fully understood, it forms aggregates thought to play a role in Parkinson's disease [59, 60]. It contains a helix-turn-helix motif at its N-terminal and a random coil near its C-terminal. Although residues throughout the helical region of this protein interact with other α-synucleins, residues 71 to 82 are the necessary aggregating motif [61] (Fig. 12 B). We blobulated α-synuclein using the settings H^*= 0.4 and L_{min}= 4 and find that α-synuclein's h-blob group coincides with this aggregating motif (Fig. 12 A). This result demonstrates that the blobulation of IDPs with transient secondary structure elements can reveal hydrophobic clusters that may coincide with impactful tertiary interactions.

Blobulation can aid in predicting the effects of sequence mutations that change the interactions of IDPs with their environments. The helix-turn-helix motif of α-synuclein (Fig. 12B) also interacts with cell membranes [62], but the mechanism driving this interaction is unknown [62]. We consider the known disease-causing mutations A30P and A53T. When expressed in yeast, A53T and wildtype α-synuclein initially bind to the membrane before forming cytoplasmic aggregates, whereas A30P α-synuclein is dispersed throughout the cytoplasm [63]. The A30P mutant has a shorter and less ordered third h-blob than A53T and wildtype α-synuclein (Fig. 13 C). Because the order classification considers residue hydrophobicity, and proline is less hydrophobic than alanine, a decrease in order is expected because predicted order increases with hydrophobicity on the Uversky-Gillepse-Fink boundary plot [41]. This is also consistent with the finding that while this mutant's helical domain is partially disrupted, its N-terminus is more dynamic than the wildtype [64]. In contrast to the wildtype, no blobs are shortened in the A53T mutant protein, and the order of the blob containing the mutated residue is only minimally altered.

5 CONCLUSION

The classical structural hierarchy has both conceptual and practical limitations, some of which we have sought to address with our web tool - the blobulator. The blobulation algorithm can be used either with the webtool or the command-line tool found on our GitHub `https://GitHub.com/BranniganLab/blobulator`, though we note that the interactive webtool is helpful for building an intuition for blobulation. The blobulator works by detecting intrinsic organization in protein sequences using contiguous hydrophobicity and can be used in various research contexts. Here, we introduced the blobulator and showed example applications in various proteins where it can provide salient information. Examples included tuning parameters to

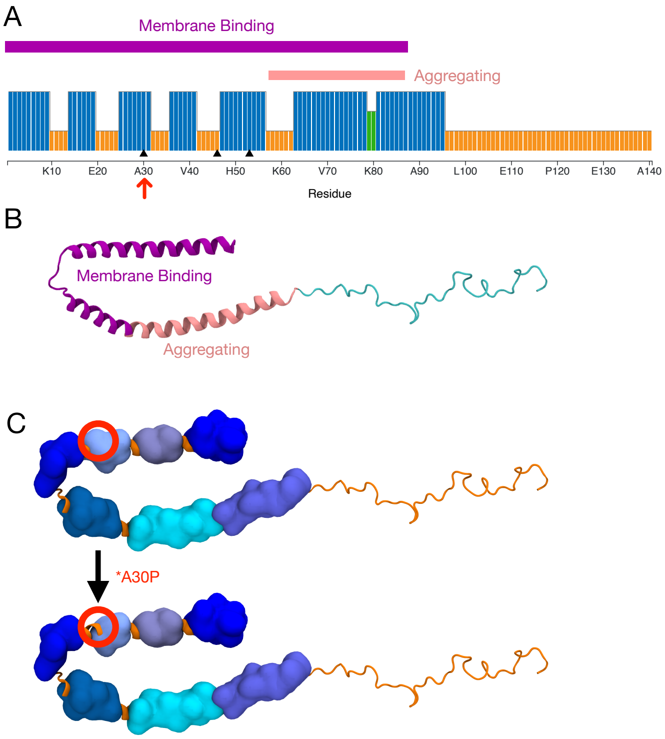


Figure 12. Blobulation of α-synuclein, an intrinsically disordered protein. A) α-synuclein (UniProt ID: P37840) blobulated using default settings (H^{*}= 0.4, L_{min}= 4). Annotations indicate the membrane-interacting region (purple) and the protein-interacting region (pink). B) α-synuclein structure (PDB: 1XQ8), labeled by the membrane-interacting region (purple and pink) and the aggregating region (pink). C) Changes to blobs (blue and purple surfaces) caused by the A30P mutant compared to the wildtype. Molecular images were generated in VMD [44, 45].

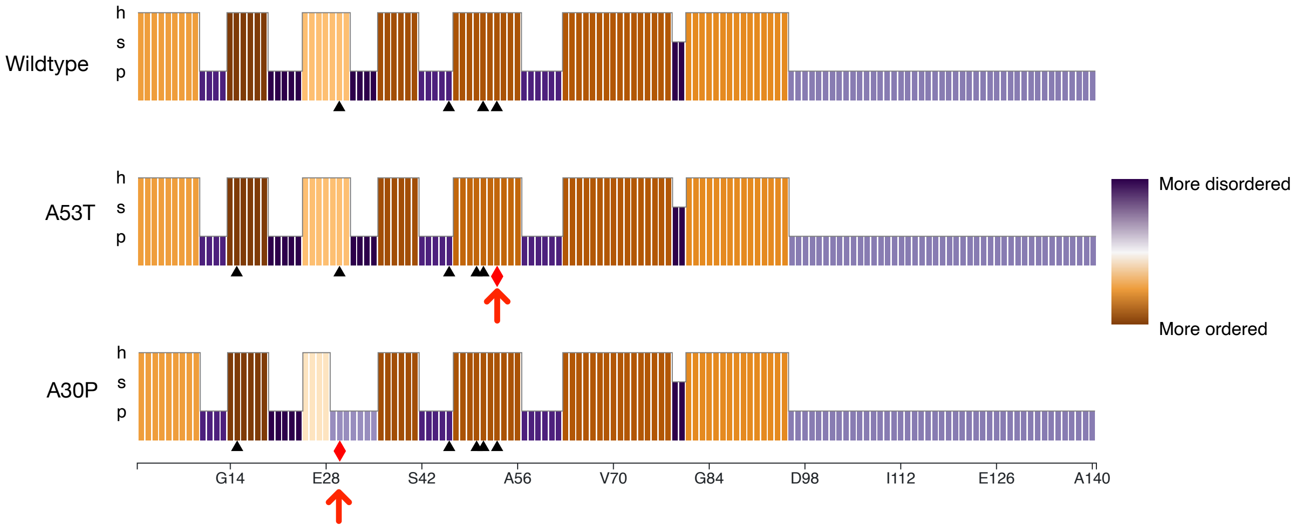


Figure 13. Order predictions for blobs of α-synuclein. Wildtype, A53T, and A30P mutants colored by blob disorder as calculated using each blob's signed distance from the order/disorder boundary of the Uversky-Gillepse-Fink boundary plot [41]. Mutations are indicated: A53T and A30P (red diamonds and arrows), and other known disease-associated mutations (black triangles). Blobulation done using default settings (H^{*}= 0.4, L_{min}= 4).

detect blobs with certain properties, comparing blobs between orthologs with low sequence but high structural homology, and investigating whether known function-altering mutations also alter blob topology. We believe that future studies may benefit from taking these same principles and using the blobulator on much larger datasets than the handful of examples shown here. These studies could include predicting protein segments with critical roles in tertiary interactions, defining local order and disorder with clear edges, and predicting protein features from sequence. Future development plans include options for viewing nucleotide sequences and molecular structures, comparing two blobulated sequences within the same output window, and outputting results in BED format.

6 ACKNOWLEDGEMENTS

The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources. TJ is supported by NIH 1K08GM139031, MEBH is supported by NIH 1R35GM134957, and ESM, CP, and GB are supported by NSF DGE 2152059. We are also grateful to Dr. Jérôme Hénin, Mr. Jesse Sandberg, Mr. Jahmal Ennis, and Mr. Ryan Lamb for testing and feedback, and grateful to Ms. Lindsey Riggs for contributions to generation of molecular images.

REFERENCES

[1] Aleksey A. Porollo, Rafal Adamczak, and Jaroslaw Meller. POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics*, 20(15):2460–2462, 2004.

[2] Aleksey Porollo and Jarosław Meller. Prediction-based fingerprints of protein–protein interactions. *PROTEINS: Structure, Function, and Bioinformatics*, 2007.

[3] G. Deleage, C. Combet, C. Blanchet, and C. Geourjon. Antheprot: An integrated protein sequence analysis software with client/server capabilities. *Computers in Biology and Medicine*, 2001.

[4] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 2002.

[5] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, and Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 2004.

[6] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[7] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 2010.

[8] Haicang Zhang, Michelle S. Xu, Xiao Fan, Wendy K. Chung, and Yufeng Shen. Predicting functional effect of missense variants using graph attention neural networks. *Nature Machine Intelligence*, 2022.

[9] Pauline C. Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 2003.

[10] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017.

[11] Elodie Laine, Yasaman Karami, and Alessandra Carbone. Gemme: A simple and fast global epistatic model predicting mutational effects. *Molecular Biology and Evolution*, 2019.

[12] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

[13] Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 2018.

[14] Max Hebditch, M. Alejandro Carballo-Amador, Spyros Charonis, and Jim Warwicker. Protein–sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, 2017.

[15] Jialu Wu, Junmei Wang, Zhenxing Wu, Shengyu Zhang, Yafeng Deng, Yu Kang, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Alipsol: An attention-driven mixture-of-experts model for lipophilicity and solubility prediction. *Journal of Chemical Information and Modeling*, 62(23):5975–5987, 2022.

[16] Sergiy O. Garbuzynskiy, Michail Yu. Lobanov, and Oxana V. Galzitskaya. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332, 2009.

[17] Oscar Conchillo-Solé, Natalia S de Groot, Francesc X Avilés, Josep Vendrell, Xavier Daura, and Salvador Ventura. AGGRESCAN: a server for the prediction and evaluation of ”hot spots” of aggregation in polypeptides. *BMC Bioinformatics*, 8(1), 2007.

[18] Ian Walsh, Flavio Seno, Silvio C.E. Tosatto, and Antonio Trovato. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, 42(W1):W301–W307, 2014.

[19] Nikolaos Louros, Gabriele Orlando, Matthias De Vleeschouwer, Frederic Rousseau, and Joost Schymkowitz. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nature Communications*, 11(1), 2020.

[20] Gian Gaetano Tartaglia, Amol P. Pawar, Silvia Campioni, Christopher M. Dobson, Fabrizio Chiti, and Michele Vendruscolo. Prediction of aggregation-prone regions in structured proteins. *Journal of Molecular Biology*, 2008.

[21] Kana Shimizu. Poodle: Tools predicting intrinsically disordered regions of amino acid sequence. *Methods in Molecular Biology*, 2014.

[22] David T. Jones and Domenico Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 2015.

[23] Gabor Erdos, Matyas Pajkos, and Zsuzsanna Dosztanyi. Iupred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, 2021.

[24] Rune Linding, Lars Juhl Jensen, Francesca Diella, Peer Bork, Toby J. Gibson, and Robert B. Russell. Protein disorder prediction: Implications for structural proteomics. *Structure*, 2003.

[25] Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O.G. Richardson, and Rohit V. Pappu. Cider: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical Journal*, 112(1):16–21, 2017.

[26] Bin Huang, Tingwen Fan, Kaiyue Wang, Haicang Zhang, Chungong Yu, Shuyu Nie, Shuyu Nie, Yangshuo Qi, Wei-Mou Zheng, Jian Han, Zheng Fan, Shiwei Sun, Sheng Ye, Huaiyi Yang, and Dongbo Bu. Accurate and efficient protein sequence design through learning concise local environment of residues. *Bioinformatics*, 2023.

[27] Leonardo V. Castorina, Rokas Petrenas, Kartic Subr, and Christopher W. Wood. Pdbench: evaluating computational methods for protein-sequence design. *Bioinformatics*, 2023.

[28] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023.

[29] Gilbert Deléage. Alignsec: viewing protein secondary structure predictions within large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 33:3991–3992, December 2017.

[30] Mujtahid Akon, Muntashir Akon, Mohimenul Kabir, M. Saifur Rahman, and M. Sohel Rahman. Adact: a tool for analysing (dis)similarity amongnucleotide and protein sequences using minimal andrelative absent words. *Bioinformatics*, 2021.

[31] Yuedong Yang, Jianzhao Gao andJihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondarystructure prediction: the final stretch? *Briefings in Bioinformatics*, 2018.

[32] Wei Jiang, Christophe Chipot, and Benoît Roux. Computing relative binding affinity of ligands to receptor: An effective hybrid single-dual-topology free-energy perturbation approach in NAMD. *Journal of chemical information and modeling*, 59(9):3794–3802, 2019.

[33] Chia-Tzu Ho, Yu-Wei Huang, Teng-Ruei Chen, Chia-Hua Lo, and Wei-Cheng Lo. Discovering the ultimate limits of protein secondary structure prediction. *Biomolecules*, 2021.

[34] Dewi Pramudi Ismi, Reza Pulungan, and Afiahayati. Deep learning for protein secondary structure prediction: Pre and post-alphafold. *Computational and Structural Biotechnology Journal*, 20:6271–6286, 2022.

[35] Ruchi Lohia, Reza Salari, and Grace Brannigan. Sequence specificity despite intrinsic disorder: How a disease-associated val/met polymorphism rearranges tertiary interactions in a long disordered protein. *PLOS Computational Biology*, 15(10):e1007390, 2019.

[36] Ruchi Lohia, Matthew E. B. Hansen, and Grace Brannigan. Contiguously hydrophobic sequences are functionally significant throughout the human exome. *PNAS*, 2022.

[37] P. Pincus. Exculded volume effects and stretched polymer chains. *Macromolecules*, 1976.

[38] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 1982.

[39] David Eisenberg, Robert M. Weiss, Thomas C. Terwilliger, and William Wilcox. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc*, 1982.

[40] Rahul K. Das and Rohit V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, 2013.

[41] Vladimir N. Uversky, Joel R. Gillespie, and Anthony L. Fink. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, 2000.

[42] M.E. Oates, P. Romero, T. Ishida, M. Ghalwash, M.J. Mizianty, B. Xue, S. Dosztányi, V.N. Uversky, Z. Obradovic, L. Kurgan, A.K. Dunker, and J. Gough. D2p2: Database of disordered protein predictions. *Nucleic Acids Research*, 2013.

[43] Yumeng Liu, Xiaolong Wang, and Bin Liu. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics*, 2017.

[44] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[45] John Stone. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[46] Keita Ohsumi, Chiaki Katagiri, and Takeo Kishimoto. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant t4 lysozyme. *Science*, 1993.

[47] Brian K. Shoichet, Walter A. Baase, Royta Kuroki, and Brian W. Matthews. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences*, 1995.

[48] Tom Alber, Sun Dao-pin, Keith Wilson, Joan A. Wozniak, Sean P. Cook, and Brian W. Matthews. Contributions of hydrogen bonds of thr 157 to the thermodynamic stability of phage t4 lysozyme. *Nature*, 1987.

[49] Mariama Jaiteh, Antoine Taly, and Jérôme Hénin. Evolution of pentameric ligand-gated ion channels: Pro-loop receptors. *PLOS ONE*, 11(3):e0151934, 2016.

[50] Marco Cecchini and Jean-Pierre Changeux. The nicotinic acetylcholine receptor and its prokaryotic homologues: Structure, conformational transitions &amp; allosteric modulation. *Neuropharmacology*, 96:137–149, September 2015.

[51] Rebecca J. Howard. Elephants in the dark: Insights and incongruities in pentameric ligand-gated ion channel models. *Journal of Molecular Biology*, 433(17):167128, August 2021.

[52] Ákos Nemecz, Marie S. Prevost, Anaïs Menny, and Pierre-Jean Corringer. Emerging molecular mechanisms of signal transduction in pentameric ligand-gated ion channels. *Neuron*, 90(3):452–470, May 2016.

[53] Ludovic Sauguet, Azadeh Shahsavar, and Marc Delarue. Crystallographic studies of pharmacological sites in pentameric ligand-gated ion channels. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(3):511–523, March 2015.

[54] Corrie J.B. daCosta and John E. Baenziger. Gating of pentameric ligand-gated ion channels: Structural insights and ambiguities. *Structure*, 21(8):1271–1283, August 2013.

[55] Carlos Baeza-Delgado, Marc A. Marti-Renom, and Ismael Mingarro. Structure-based statistical analysis of transmembrane helices. *European Biophysics Journal*, 42(2-3):199–207, 2012.

[56] Casey L. Carswell, Camille M. Hénault, Sruthi Murlidaran, J.P. Daniel Therien, Peter F. Juranka, Julian A. Surujballi, Grace Brannigan, and John E. Baenziger. Role of the fourth transmembrane α helix in the allosteric modulation of pentameric ligand-gated ion channels. *Structure*, 23(9):1655–1664, September 2015.

[57] Thorsten Althoff, Ryan E. Hibbs, Surajit Banerjee, and Eric Gouaux. X-ray structures of GluCl in apo states reveal a gating mechanism of cys-loop receptors. *Nature*, 512(7514):333–337, 2014.

[58] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews*, 2015.

[59] Maria Grazia Spillatini, R. Anthony Crowther, Ross Jakes, Masato Hasegawa, and Michel Goedert. α-synuclein in filamentous inclusions of lewy and bodies from and parkinson's disease and dementia with lewy and bodies. *Proceedings of the National Academy of Sciences*, 1998.

[60] James E. Galvin, Kunihiro Uryu, Virginia M.-Y. Lee, and John Q. Trojanowski. Axon pathology in parkinson's disease and lewy body dementia hippocampus contains a-, b-, and g-synuclein. *PNAS*, 1999.

[61] Ricardo Guerrero-Ferreira, Nicholas MI Taylor, Ana-Andreea Arteni, Pratibha Kumari, Daniel Mona, Philippe Ringler, Markus Britschgi, Matthias E Lauer, Ali Makky, Joeri Verasdonck, Roland Riek, Ronald Melki, Beat H Meier, Anja Bockmann, Luc Bousset, and Henning Stahlberg. Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. *eLIFE*, 2019.

[62] Tapojyoti Das and David Eliezer. Membrane interactions of intrinsically disordered proteins: The example of alpha-synuclein. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1867(10):879–889, 2019.

[63] Katherina Vamvaca, Michael J. Volles, and Peter T. Lansbury. The first n-terminal amino acids of α-synuclein are essential for α-helical structure formation in vitro and membrane binding in yeast. *Journal of Molecular Biology*, 389(2):413–424, 2009.

[64] Luisel R. Lemkau, Gemma Comellas, Kathryn D. Kloepper, Wendy S. Woods, Julia M. George, and Chad M. Rienstra. Mutant protein a30p α-synuclein adopts wild-type fibril structure, despite slower fibrillation kinetics. *Journal of Biological Chemistry*, 287(14):11526–11532, 2012.