Revealing protein sequence organization via contiguous hydrophobicity with the blobulator toolkit

Connor Pitman[1], Ezry Santiago-McRae[1], Ruchi Lohia[2], Ryan Lamb[1], Kaitlin Bassi[1], Lindsey Riggs[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 Cancer Biology, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd, 19104, PA, USA
[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
[*]Correspondence: grace.brannigan@rutgers.edu

ABSTRACT 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 within protein sequences (termed “blobs”) has proven useful in both intrinsically disordered protein (IDP) simulation and human genome studies. However, an accessible toolkit was unavailable, and the role that blobs play across the structural context of a variety of protein families remained unclear. Here, we present the *blobulator* toolkit: consisting of a webtool, a command line interface, and a VMD plugin. We demonstrate how identifying blobs using 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. The *blobulator* webtool can be found at www.blobulator.branniganlab.org, and the source code with pip installable command line tool, as well as the VMD plugin with installation instructions, can be found on GitHub at www.GitHub.com/BranniganLab/blobulator.

SIGNIFICANCE Understanding the sequence organization of proteins remains a challenge, especially for those lacking stable structure. Our work introduces the blobulator, a computational toolkit that reveals patterns of contiguous hydrophobicity, offering a new way to analyze protein sequence organization. By identifying hydrophobic "blobs," this approach provides insights into protein stability, interactions, and evolutionary conservation—using sequence alone. We demonstrate its broad applicability across globular, membrane, and intrinsically disordered proteins, highlighting its potential for both fundamental research and practical applications in bioinformatics and molecular biophysics. The blobulator’s interactive web tool, which integrates protein sequences and genomic mutation data, along with the structural visualization plugin make this approach more widely accessible to users from multiple disciplines.

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, and quaternary structure) remains central to designing and interpreting biophysical studies that span the proteome. Under this hierarchy, residues are organized into modules typically represented by secondary structure elements. 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 using a fixed-width moving window.

The classic structural hierarchy has proven extraordinarily powerful for connecting protein sequence, structure, function, and evolution, but is not universally applicable. It is only well-suited for proteins with known structures and excludes intrinsically disordered proteins (IDPs) (4) entirely. Approximately one-third of proteins in the eukaryotic proteome are either entirely disordered or contain large regions of disorder, and consequently do not fit easily into the classic structural hierarchy (5), requiring other approaches to detect organization in protein sequences. Additionally, there is a significant mixing of information across levels in the classic structural hierarchy; while secondary structure can be predicted from primary structure (sequence) with some accuracy, tertiary structure cannot be predicted from secondary structure alone. Informatics methods that rely on

secondary structure frequently neglect predictable features of tertiary interactions, such as solvent-accessible surface area, while also implicitly under-emphasizing unresolvable residues found in random coil and linking regions.

Most tools that extract information innate to protein sequences either utilize the structural hierarchy framework or work without a hierarchical framework. Examples of applications for such tools 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), identifying subsequences in proteins based on amino acid composition (29), and comparing sequences (30, 31). Some of these tools use a fixed-width moving window to define local sequence context, which artificially places every residue at the center of its "local sequence" and ignores any natural boundaries present within proteins, missing an opportunity to detect signal (32).

Alternatively, under the classical structural hierarchy, the local sequence context of a residue is typically considered to be its associated secondary structure element. Secondary structure elements can be identified in several ways: using a solved structure, a secondary structure predictor, or a tertiary structure predictor that predicts secondary structure. However, tools for predicting secondary structure have limited accuracy in the absence of tertiary structure (33–36).

To address some of these limitations, we previously developed an algorithm called "blobulation" to segment protein sequences into hydrophobic ("h-blobs") and non-hydrophobic ("p- and s-blobs") subsequences without requiring a structure or the use of a structure predictor (32, 37). We have already used blobulation to meaningfully segment an IDP in order to measure tertiary interactions in simulation data (37), and later demonstrated its use in providing a meaningful definition of "local context" in a study on disease-associated mutations in the human exome (32). Given an amino acid sequence, blobulation outputs residue clusters, termed "blobs" as inspired by polymer physics (38).

In this paper, we introduce the *blobulator* toolkit: a webtool, a VMD plugin, and a command line interface. While we focus mainly on use of the webtool and VMD plugin in the example applications section, we note that the command line interface can be used to easily scale the amount of proteins analyzed in bioinformatics-type studies.

2 BLOBULATION

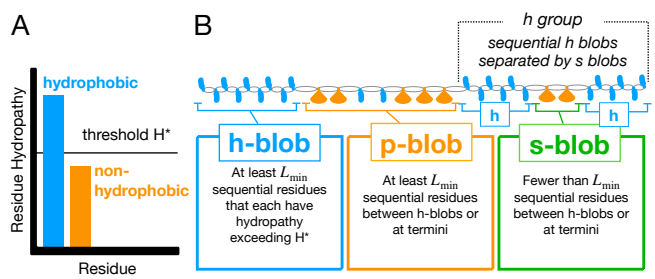

Figure 1: **Blobulation algorithm.** Figure adapted from Lohia et al. 2022 (32). A) First, the sequence is digitized. Residues are classified as either hydrophobic (blue) or non-hydrophobic (orange) by comparing their hydropathy to the user-selected threshold, H^*. B) The sequence is then segmented into h-blobs, s-blobs, and p-blobs based on H^* and $L_{\min}$.

Whole-sequence blobulation consists of two steps: digitization and clustering. We've included a description of the algorithm below, as well as a pseudocode version of the algorithm in section 6.

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.
   (b) The hydropathy for each individual residue (i) is assigned based on the normalized hydrophobicity scale. The hydropathies are then 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.
   (c) 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) The user sets $L_{\min}$ equal to a positive integer representing a threshold number of residues required to form an h- or p-blob.

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

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

Though there are default settings for the initial blobulation, the user can tune various parameters. 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). Alternatively, when using blobulation to adaptively explore blobs within a given protein, 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. Additionally, high $L_{\min}$ will only yield h-blobs with a low to moderate H^*.

We will also consider higher-order organization beyond individual blobs, "blob groups". These are composed of h-blobs separated only by s-blobs, and are labeled as follows: by their type (h, p, s) and group number (1, 2, 3) and a sub-group letter (a, b, c) which form a unique label for every blob (h1a, s1, h1b, etc). Examples and further discussion of blob groups can be found in sections 4.1 and 4.3. We also refer to the connectivity of blobs in a protein sequence as its "blob topology".

3 THE BLOBULATOR

Here, we present three tools by which an amino acid sequence can be blobulated: a webtool, a command line interface (CLI), and a Visual Molecular Dynamics (VMD) plugin. The webtool can be used to interactively explore blobulated sequences, introduce mutations, and investigate various blob properties. The command line interface is the backend for the webtool, but also accepts DNA inputs and can be used to run batch processes. The VMD plugin can be used with molecular structures for analysis and creating images and movies of blobulated proteins in VMD.

3.1 Webtool

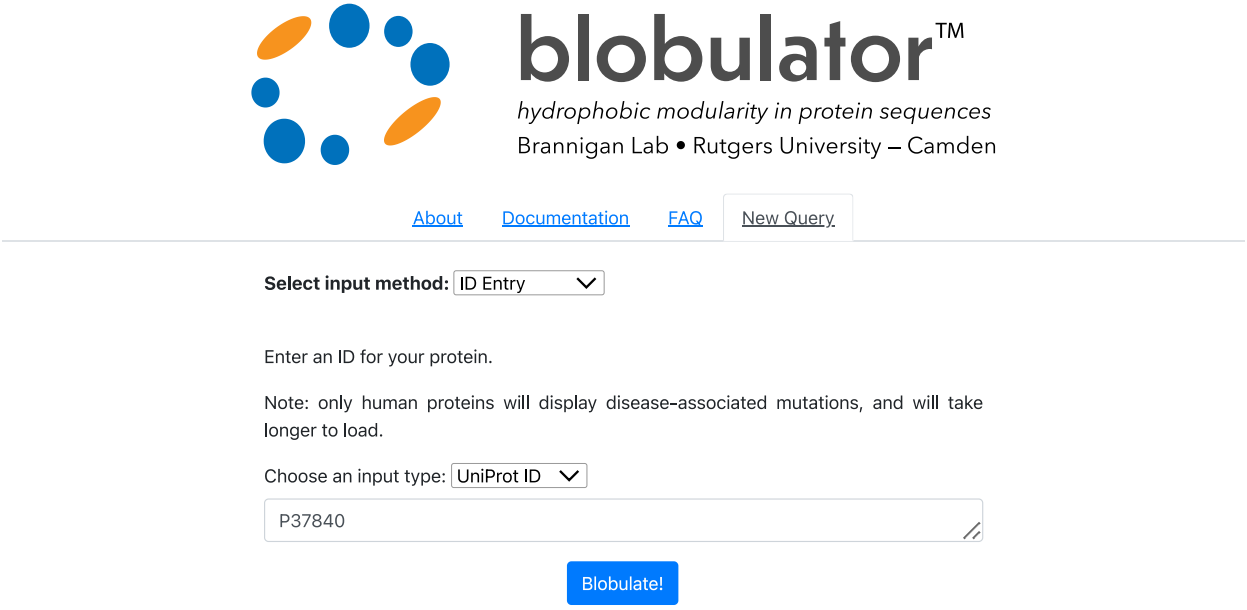

Figure 2: **Screenshot of the *blobulator* webtool home page.** Protein sequences are submitted via UniProt ID or ENSEMBL ID (ID Entry), or the protein sequence and sequence name (Manual Entry). Users can toggle between the two submission options using the dropdown menu. The default values are for alpha-synuclein. Available at `www.blobulator.branniganlab.org`

Blobulation of any amino acid sequence can be achieved using the *blobulator* webtool found at `www.blobulator.branniganlab.org`. On the "New Query" tab of the homepage, 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 the initial blobulation, the user can interactively tune some of the parameters. For example, while previous studies from our lab (32, 37) used the Kyte-Doolittle scale, the user may select the Eisenberg-Weiss scale (39) or the Moon-Fleming scale (40). More hydrophobicity scales will be added in the future. Additionally, 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 and selecting the checkbox (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 webtool output (Fig. 3E).

The smoothed hydropathy of each residue H_i (as defined in section 2) is shown in the first "Results" 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 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 webtool 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, globular tendency, the predicted sensitivity of each blob to deleterious mutations, and two tracks showing the predicted order and disorder of each blob, respectively.

The first two of these tracks are colored by mean net charge per residue and globular tendency, both calculated based solely on charged residues in the blob. (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 (41). 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 (32). This study used a dataset containing about 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 (42) (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 (43) (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 (44). Additionally, if the user does not access the protein sequence using a UniProt ID, this track will not appear.

3.2 Python Scripting and Command Line Interface

For use in high-throughput applications, like those presented in (32), we also provide a stand-alone Python package for blobulation (45). This package also comprises the backend for the blobulator webtool, and is called when a protein is submitted via that interface. The package can be used either in Python scripts or directly via a command line interface (CLI) (45). The input to the CLI can be either amino acid or DNA sequences in plain text or fasta format. The outputs of the CLI are CSV files (one for each input sequence) with each amino acid labeled by blob type and the blob index, identical to the downloadable blobulation output of the webtool (see 3.1). Example scripts are also provided on the blobulator GitHub (45).

3.3 VMD Plugin

While the webtool allows users to view the blobulation of the sequence and various averaged physicochemical properties, the plugin tool for Visual Molecular Dynamics (VMD) (46) allows users to incorporate blob information into visualization and analysis of protein structure and dynamics. Users can assign blobs with the same parameters and algorithm in the webtool, while taking advantage of the existing VMD tools for scriptable analyses and creation of high-quality images and movies. Additionally,

Figure 3: **Partial screenshot of the blobulator webtool results page showing parameter interface.** The user make parameter and sequence adjustments by choosing a hydropathy scale from a dropdown menu (A), setting the hydrophopathy cutoff H^* using numerical entry, the control slider, or by clicking on a 1-letter code (B), selecting the minimum h- or p-blob length $L_{\min}$ using numerical entry or the control slider (C), and by introducing mutations to the sequence by selecting the position and replacement amino acid and clicking the "mutate residue" checkbox (D). Additionally, the user can and perform a variety of operations by clicking buttons: print screen, download data, reset zoom, clear mutation, and lock control panel (E). The hydrophobicity of each residue (gray bars) compared to the threshold H^* (blue, horizontal line) is displayed on the “smoothed hydropathy per residue" track. Residues can be mutated by clicking the black triangles (which represent known disease-associated mutations, only available for blobulation via ID entry of a human sequence) and the sequence can be zoomed (zoom in: click and drag, zoom out: double click or click “Reset Zoom" (F)). Blob type is represented by both bar height (shown on the y-axis) and color on the “blob type" track. Height is preserved in subsequent tracks to indicate blob type (G). The example protein shown in these tracks is insulin (UniProt ID: P01308).

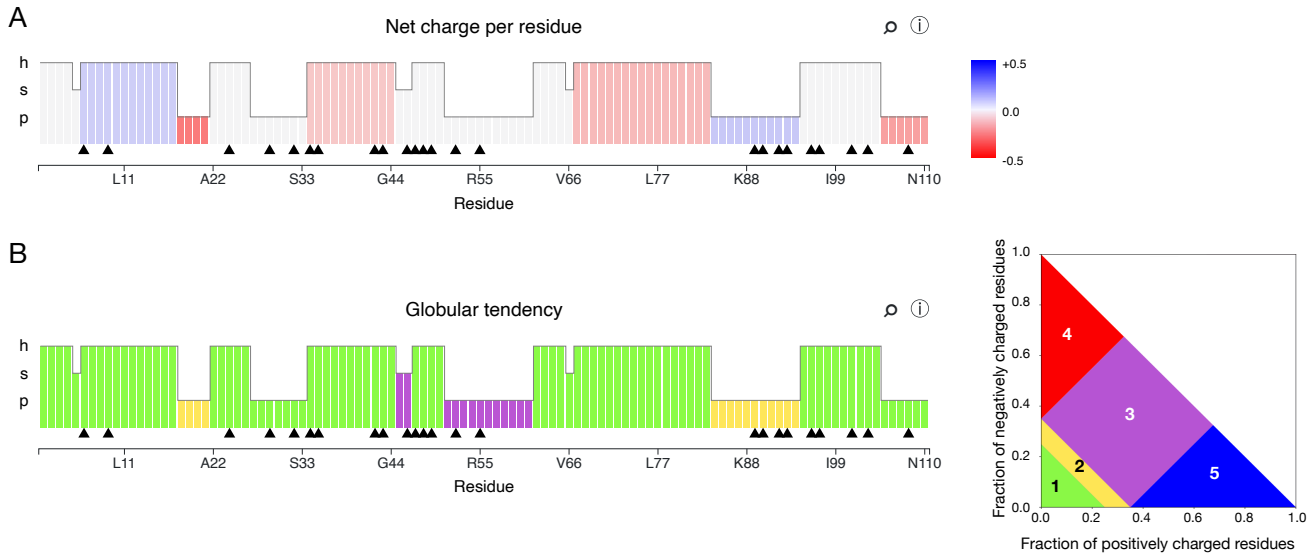

Figure 4: **Charge-based tracks on the blobulator webtool.** Blobulation of insulin as in Fig. 3. A) Blobs colored according to net charge per residue: positive (blue) or negative (red). B) Blobs colored by their Das-Pappu phase (41): phase 1 (globular, green), phase 2 (Janus/boundary, yellow), phase 3 (strong polyelectrolyte, purple), phase 4 (strong polyanion, blue), or phase 5 (strong polycation, red).

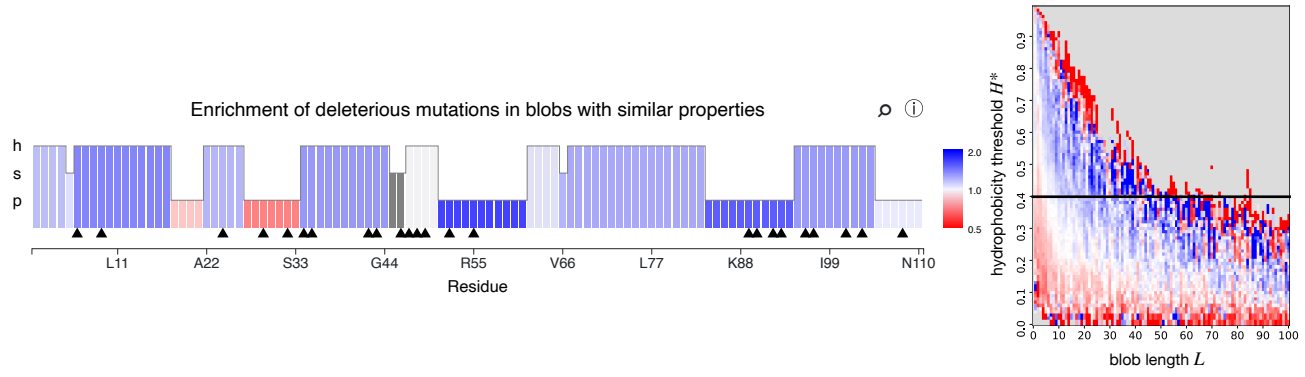

Figure 5: **Predicted enrichment of disease-associated SNPs (dSNPs) track on the blobulator webtool.** Left: Blobulation of insulin as in Fig. 3 colored by dSNP enrichment (blue: enriched, red: depleted) based on panel on right. Right: enrichment of dSNPs in hydrophobic blobs based on analysis of a large dataset of human SNPs (32). The hydrophobicity cutoff ($H^* = 0.4$) used is indicated (black, horizontal line).

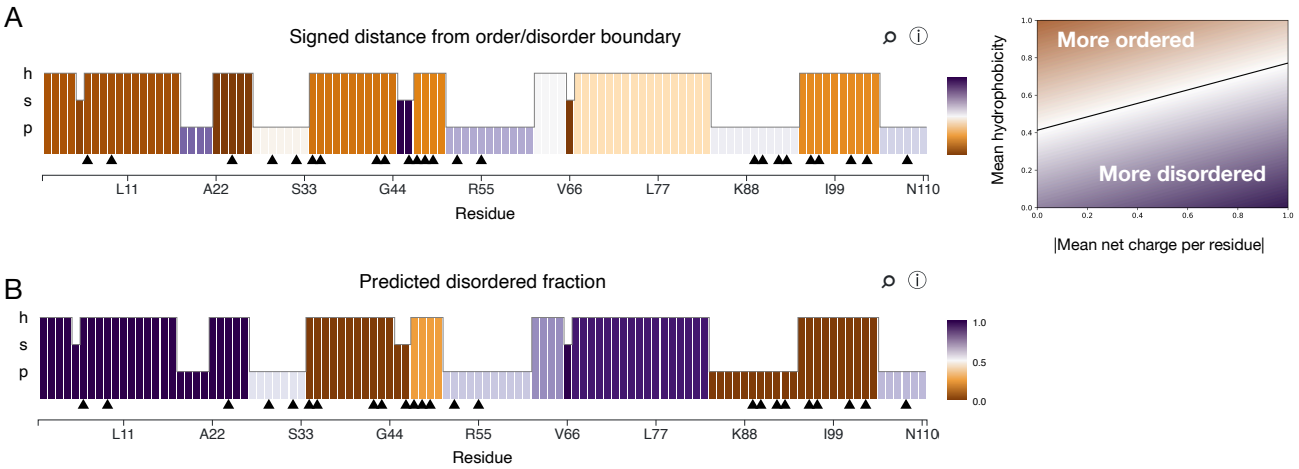

Figure 6: **Disorder-based tracks on the blobulator webtool.** Blobulation of insulin as in Fig. 3. A) Blobs colored according to the Uversky-Gillepse-Fink boundary plot (right)(42). B) Blobs colored according to their predicted fraction of disordered residues, according to PV2 as provided by the Database of Disordered Protein Prediction (D2P2).

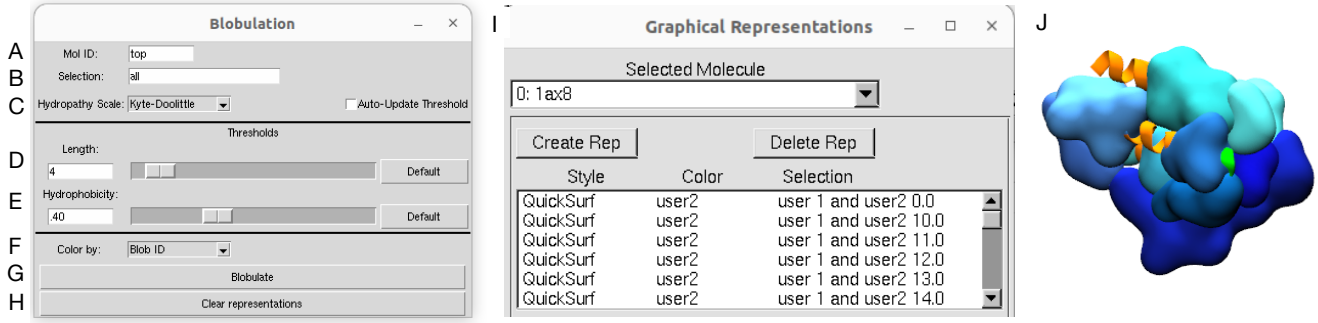

Figure 7: **Screenshot of the VMD blobulator plugin.** The user can select the molecule (A) and provide an atomselection for the display (B), select a hydrophobicity scale and toggle the option to snap the hydrophobicity threshold to a default for that scale (C), set the minimum blob length $L_{\min}$ (D) and hydrophobicity cutoff H^* (E) via eider. They also can toggle visualization options (F). The “Blobulate!" button initiates blobulation and creates graphical representations within VMD (G), all representations can be deleted by pressing the “Clear representations" button (H). Graphical representations control panel (I) with representations showing in an example viewer window (J) showing Leptin (PDB: 1ax8).

the blobulation algorithm was reimplemented in TCL, the native VMD scripting language, for maximum compatibility. The TCL script extracts the sequence information from the provided structure and then detects blobs based on user-provided parameters. The user, user2, and user3 fields associated with each atom store the type of blob (h, p, s), the blob ID, or the group ID, respectively. This information can then be used in atomselections or user-field coloring schemes provided by the VMD software; the molecular images in this manuscript were created by creating an atomselection for each h-blob, representing it with "quick-surf", and coloring each h-blob a slightly different shade of blue. Examples of images created using this plugin tool can be found throughout applications in this paper (Figs. 8, 9, 11, 13). For proteins with many blobs, this approach can require many separate graphical representations; we include an automated approach to create the representations in a batch process.

Since users of VMD typically create representations through GUIs, we also provide a GUI through which the user can provide and adjust parameters, as well as call the scripts that blobulate the sequence and create the representations. This GUI is similar to the control panel found on the blobulator webtool's output page on the webtool, with some additional features. There is an option to blobulate only part of a protein using the Selection field – which may be useful for large proteins, or to "zoom in" on a given set of residues (Fig. 7B). Adjacent to the threshold sliders, there are buttons to restore each to field to its default value: 4 and the equivalent of 0.4 on the selected scale for the length and hydrophobicity thresholds, respectively (Fig. 7D and E). Additionally, in the viewer window, users are able to adaptively tune parameters and view how this changes the blobs found on a protein's structure. Finally, using a dropdown menu (Fig. 7F) users can color blobs by their type (all h-blobs are colored the same shade of blue), or by their ID (all h-blobs are colored distinct shades of blue). Scripts and installation instructions for the VMD plugin can be found in the blobulator github repository: `https://github.com/BranniganLab/blobulator`.

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 can be used to provide a useful framework for identifying protein modularity using only sequence alone.

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 the sugar component of peptidoglycan from the peptide component. 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. 8). Lysozyme has two catalytic residues near the N-terminus and several substrate contact sites along the sequence (Fig. 8A). Higher hydrophobicity and length thresholds (H^* and $L_{\min}$) eliminate the h-blobs that are detected at the surface of the protein when using more relaxed settings (Fig. 8B). 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 isolate the components 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. This is consistent with our previous findings that h-blobs tend to be buried in structured proteins (32).

To investigate whether h-blobs are also involved in substrate binding in lysozyme, we identified blobs containing contact residues within 7Å of peptidoglycan (shown in Fig. 8). 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 the peptidoglycan sugar component. Stabilization of this sugar ring is vital for the enzyme's ability to cleave peptidoglycan (49), and we find this ring wedged between two long hydrophobic blobs (Fig. 8B), providing an example of long hydrophobic blobs with interactions that are critical for function.

Blobulating 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 gray in Fig. 9A). Both blob 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 and is also found at the core of lysozyme (Fig. 9B, shown also in the molecular image in 8). The groups detected here are akin to synergistic tertiary elements formed from secondary structure elements, such as alpha-helical bundles. This is an example of hierarchical blob clustering, where hydrophobic blobs detected at restrictive parameters 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 of lysozyme, which both affect the temperature on the stability of lysozyme through altering intraprotein hydrophobic interactions. S117 makes the protein more stable at higher temperatures by altering hydrophobic residue interactions in the

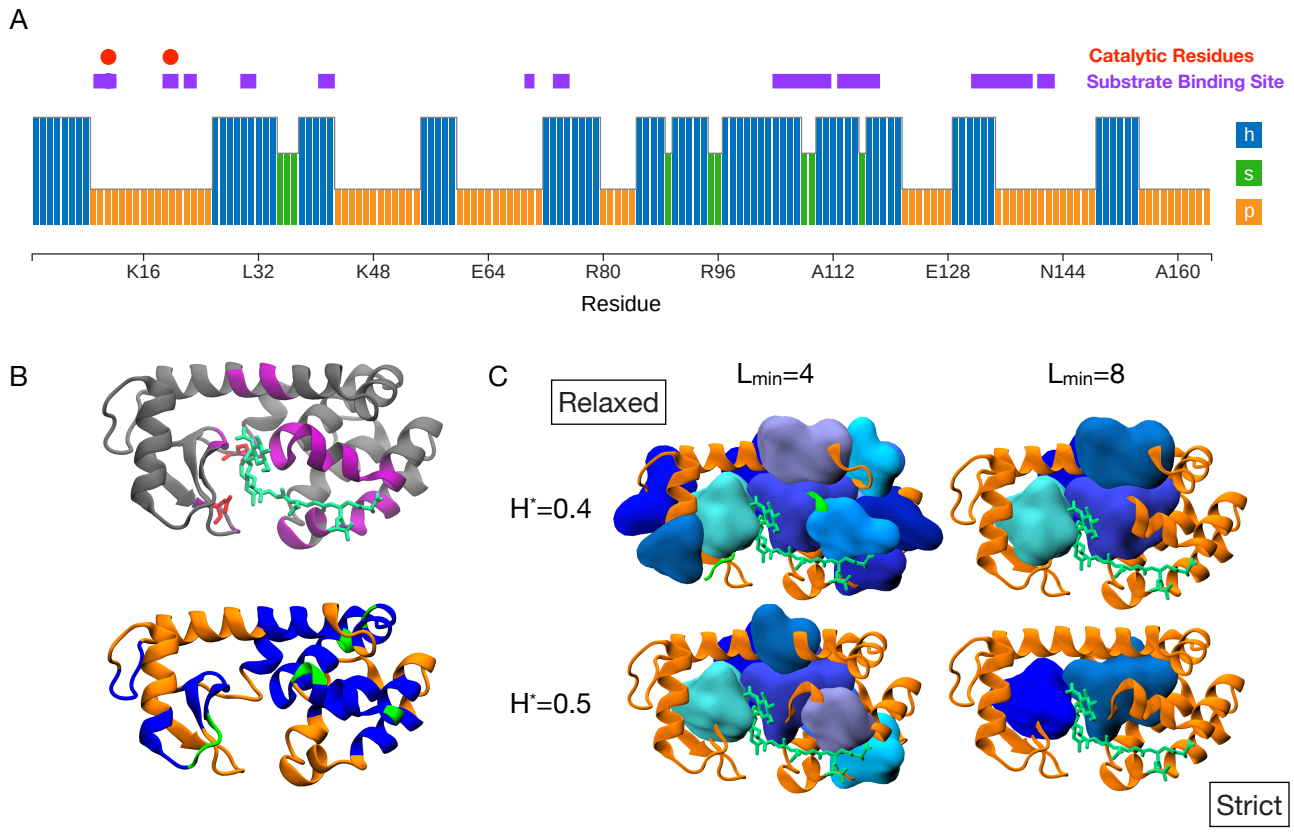

Figure 8: **Blobulation of lysozyme.** A) Blobs colored according to blob-type, as outputted from the blobulator webtool, and produced using default settings (H^*= 0.4, $L_{\min}$= 4). Annotations indicate catalytic residues (red) and the substrate binding site (purple). B) Molecular image of lysozyme (PDB:148L) 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 (47, 48) using the VMD plugin introduced in 3.3.

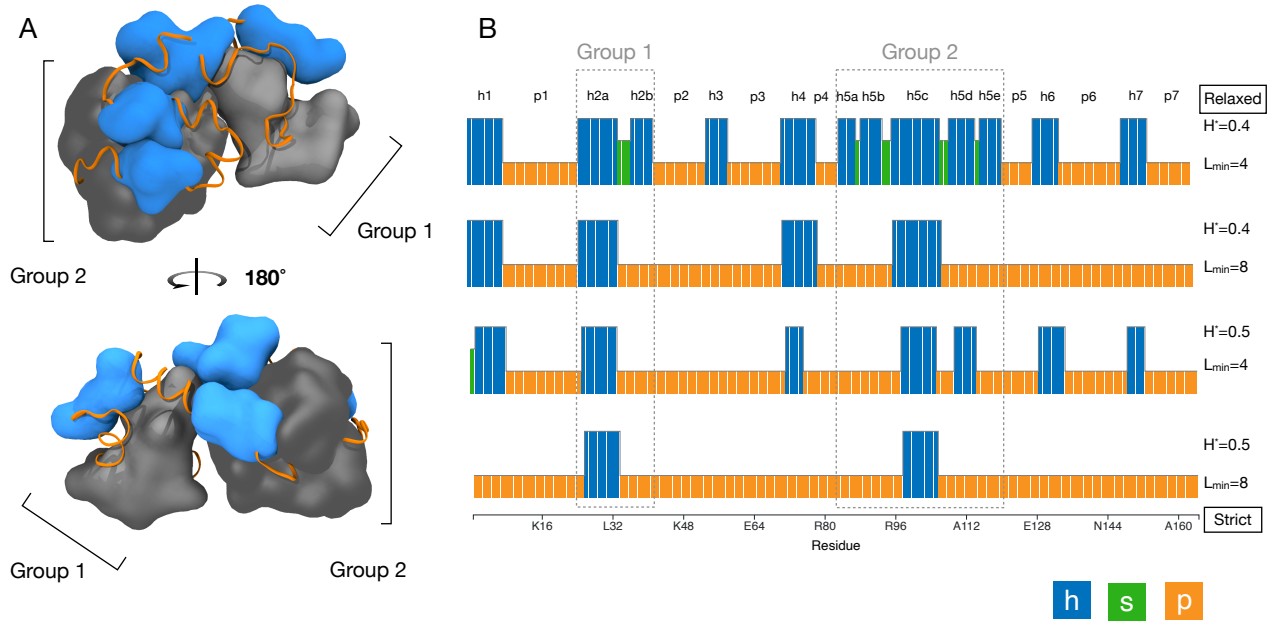

Figure 9: **Blob groups in T4 lysozyme.** A) Structural view (PDB:2LZM) blobulated using default settings (H^*= 0.4, $L_{\min}$= 4). Groups are gray, ungrouped h-blobs are blue, and p-blobs are orange ribbons. B) Blobulation of T4 lysozyme under increasingly stringent settings. Annotations indicate blob identifiers and blob groups. Molecular images were generated in VMD (47, 48).

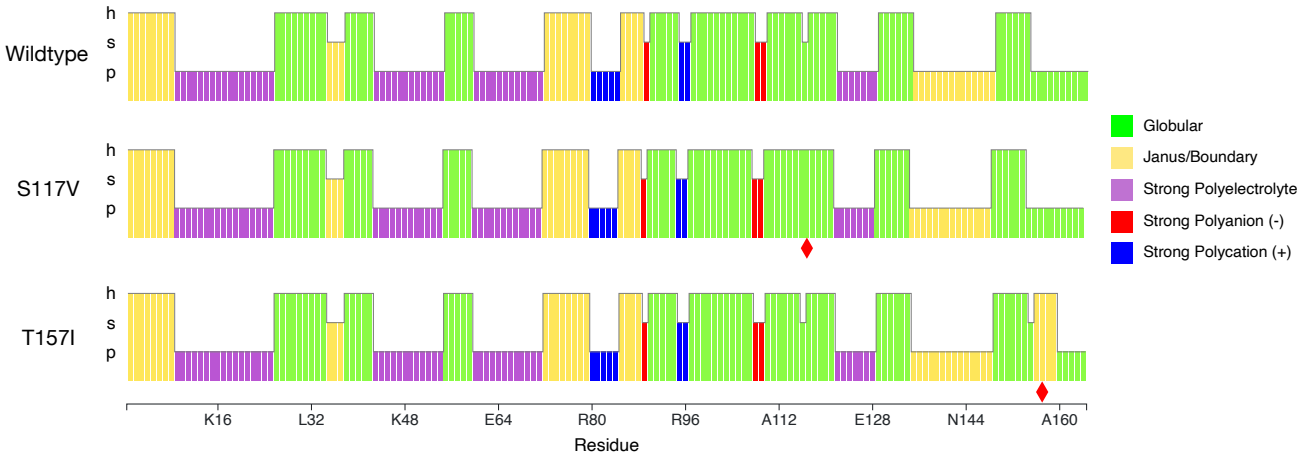

Figure 10: **Globular tendency of T4 lysozyme blobs.** Blobulation using default parameters (H^*= 0.4, $L_{\min}$= 4). Blobs for each sequence are colored by Das-Pappu phase (41), as in Fig. 4B. Red diamonds indicate mutated residues. S117V, which increases thermostability, joins two h-blobs into one. T157I, which decreases thermostability, creates a new h-blob (as well as an s-blob).

substrate binding cleft (50), while T157I makes the protein less stable at higher temperatures and disrupts hydrogen bonding at the periphery of the protein (51). We find that both mutations change the blob topology (Fig. 10): 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 (32), 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 phase diagram (41). Janus proteins often have degenerate conformations and switch between ordered and disordered depending on their environment (41), 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 (52) (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)

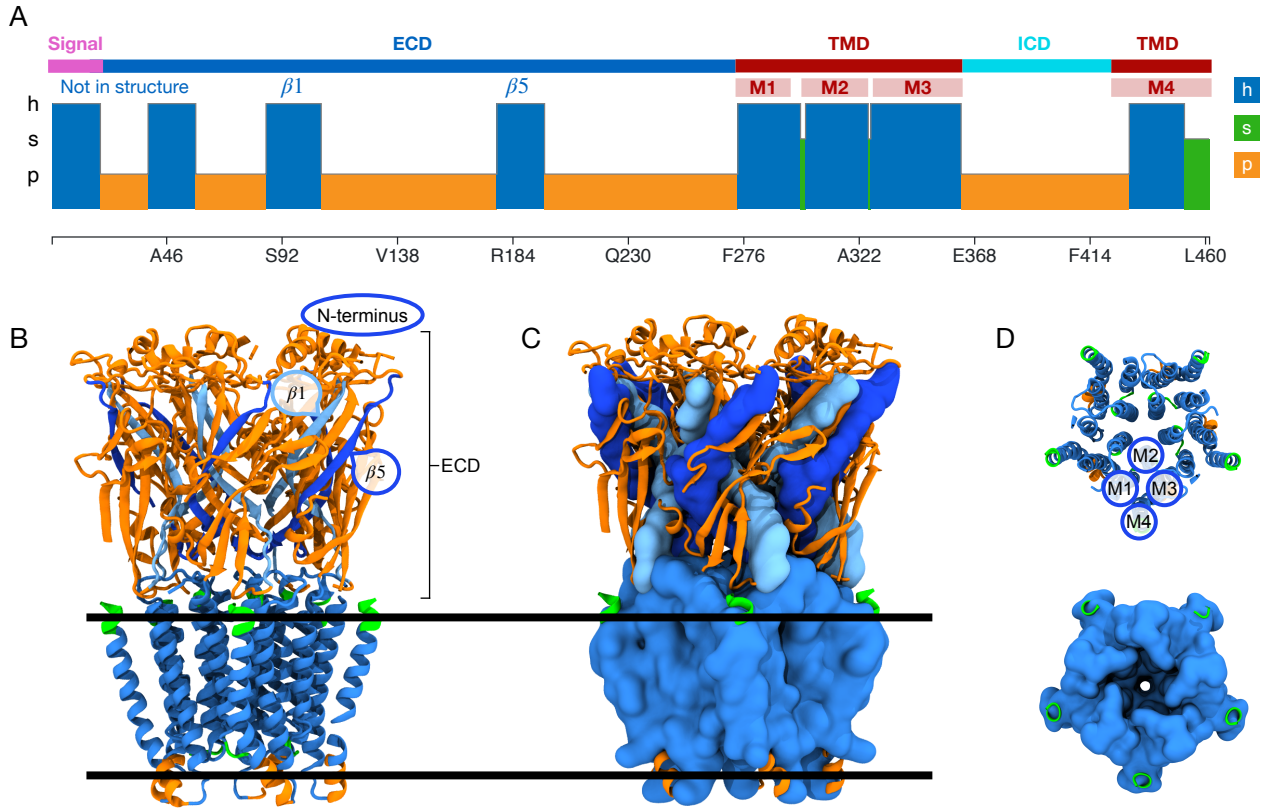

Figure 11: **Blobulation of GluCl, a pentameric ligand-gated ion channel.** A) Blobulation of GluCl (UniProt ID: G5EBR3) using settings to detect transmembrane regions ($H^* = 0.33$, $L_{\min} = 19$). B) A cartoon representation of GluCl (PDB: 3RHW) colored according to blob type and ID: $\beta 1$ blob (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 (47, 48).

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 bilateria, but representatives are also found in many other eukaryotes and some prokaryotes (53). Despite relatively low sequence identity and diverse environmental and ligand requirements, pLGICs are highly structurally conserved (54–58). The N-terminal extracellular domain consists of a

beta-sandwich, which is responsible for ligand sensing and much of the substrate selectivity (Fig. 11A-B). The transmembrane domain consists of four transmembrane helices per subunit (Fig 11D)(53). When the ligand is bound in the ECD, a signal is transmitted to the transmembrane helical bundle (Fig. 11D), 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).

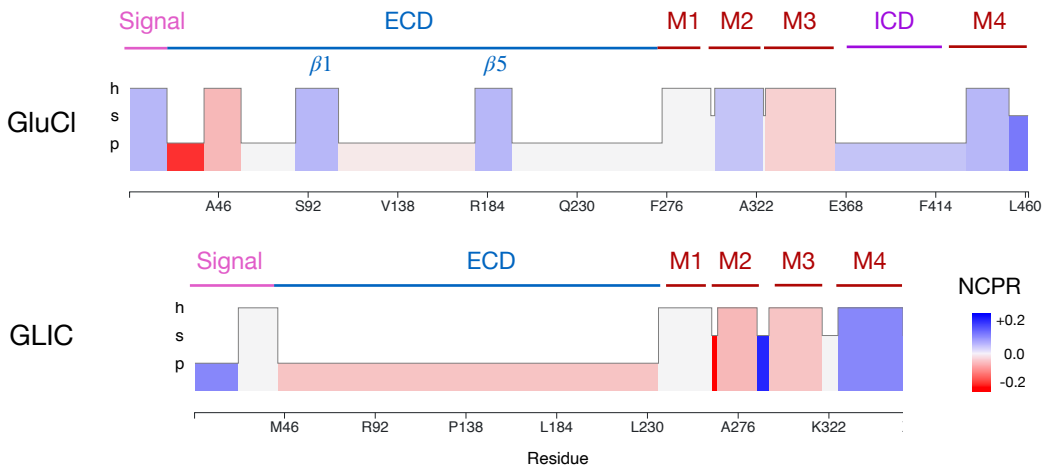

Figure 12: **Comparative blobulation of GluCl (as in Fig. 11), an anion-conducting pLGIC, and GLIC, a cation-conducting pLGIC.** Net Charge Per Residue (NCPR) tracks for GluCl (UniProt: G5EBR3) and GLIC (UniProt: Q7NDN8) blobulated using settings to detect transmembrane regions ($H^* = 0.33$, $L_{\min} = 19$). Tracks are aligned by the beginning of the h-blob containing the M2 (pore-lining) helix. Annotations indicate important sequence features available on UniProt and indicated in Fig. 11. Signal sequences are indicated in pink, ECD regions are indicated in blue, and transmembrane helices (M1 to M4) are indicated in red.

Starting from the N-terminus, both GluCl and GLIC contain disordered N-terminal extensions which include signal sequences (Fig. 12). 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 $\beta 1$ and $\beta 5$, 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 (55) and lacks these extracellular h-blobs. The transmembrane helices are clearly visible using $H^* = 0.33$ and $L_{\min} \approx 19$. This length corresponds to the typical thickness of a plasma membrane(59). The extended p-blob seen in the blobulation of GluCl (Fig. 11) corresponds to a disordered intracellular domain common in eukaryotic pLGICs (55).

As ion channels, the electrostatic potential of a pLGIC's pore plays a central role in both conductivity and selectivity. GluCl has a net positive or net neutral charge along the ion conduction pathway in its ECD and along the pore-lining M2 helix. GLIC, in contrast, has a net negative charge in these regions (Fig. 12). This is consistent with the anion conductivity of GluCl and the cation conductivity of GLIC. The charges in the M2 blob are on the intracellular side, close to the M1-M2 loop. GluCl has a single positive charge (R306), while GLIC has several negative charges in this region. GLIC has an additional positive charge near the M2-M3 loop due to K290, which can be seen by using the blobulator's zoom feature. Finally, the M4 helix of each protein has a net positive charge, consistent with experimental observations that M4 binds anionic lipids (60, 61). These results suggest the blobulator's utility in interpreting membrane proteins either individually or within a larger evolutionary context.

4.3 IDP: α-synuclein

Intrinsically disordered proteins (IDPs) fulfill critical roles despite lacking stable tertiary structure (as reviewed in Ref (62)). 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 (63, 64). 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 (65) (Fig. 13 B). We blobulated α-synuclein using the settings H^*= 0.4 and $L_{\min}$= 4 and find that the h-blob group in α-synuclein coincides with this aggregating motif (Fig. 13 A). This result demonstrates that the blobulation of IDPs with

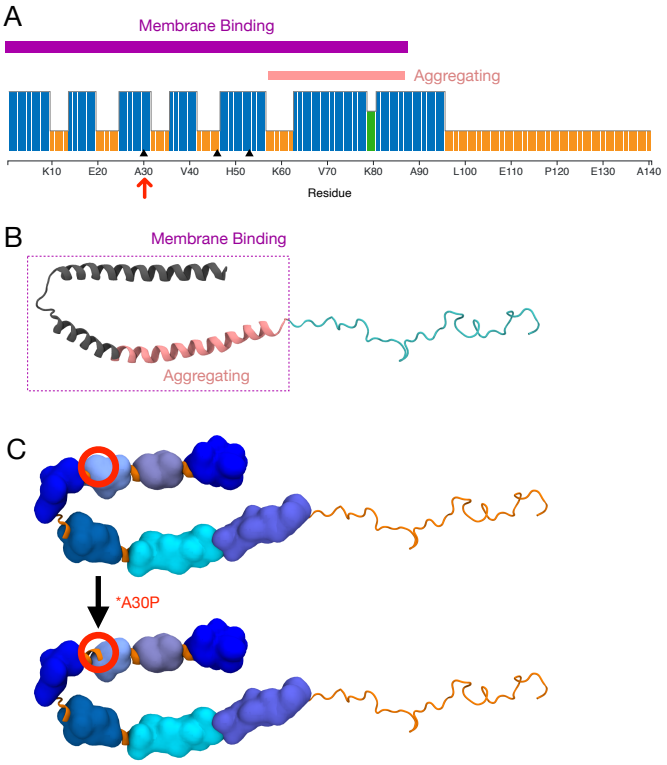

Figure 13: **Blobulation of α-synuclein, an intrinsically disordered protein.** A) Webtool "blob type" track of α-synuclein (UniProt ID: P37840) blobulated using default settings (H^*= 0.4, $L_{\min}$= 4). Annotations indicate the membrane-interacting region (purple) and the protein-interacting region (pink). B) α-synuclein structure (PDB: 1XQ8), labeled by the membrane-interacting region in cartoon. Aggregating in pink, non-aggregating in purple. C) Changes to blobs (blue surfaces, change indicated by red circles) caused by the A30P mutant compared to the wildtype. Molecular images were generated in VMD (47, 48).

transient secondary structure elements can reveal hydrophobic clusters that may coincide with tertiary interactions.

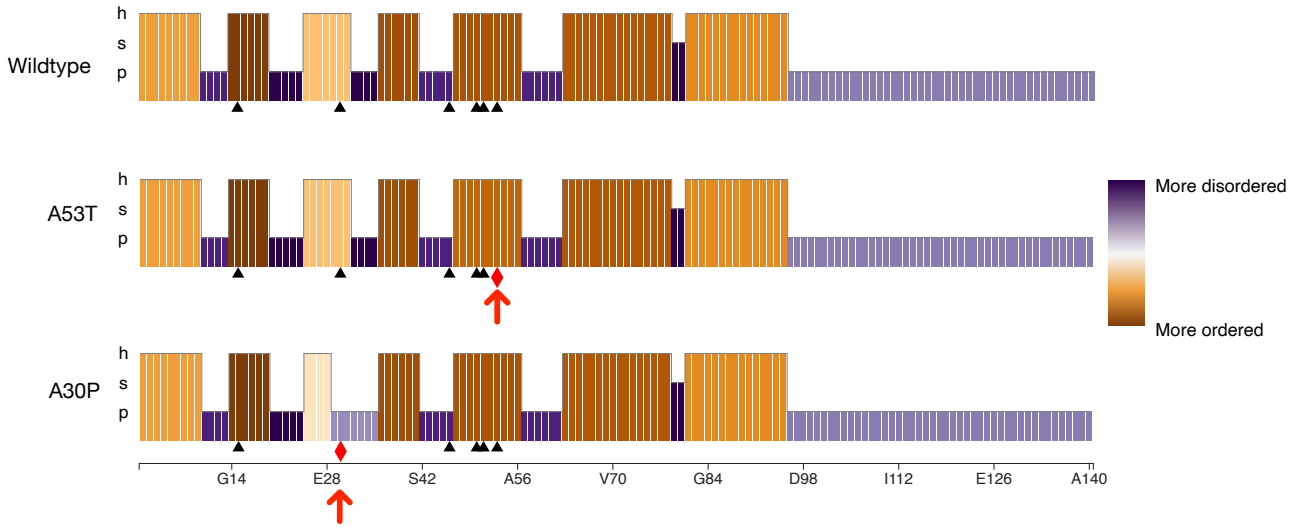

Figure 14: **Order predictions for blobs of α-synuclein.** Wildtype, A53T, and A30P mutants colored by blob disorder as calculated using each blob's signed distance from the order/disorder boundary of the Uversky-Gillepse-Fink boundary plot (42). Mutations are indicated: A53T and A30P (red diamonds and arrows), and other known disease-associated mutations (black triangles). Blobulation uses default settings (H^*= 0.4, $L_{\min}$= 4).

One application for blobulation is in quantifying the effects of mutated residues to their local sequence context. The helix-turn-helix motif of α-synuclein (Fig. 13B) also interacts with cell membranes (66), but the mechanism driving this interaction is unknown (66). Here, 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 (67). The A30P mutant has a shorter and less ordered third h-blob than A53T and wildtype α-synuclein (Fig. 14 C). Because the order prediction 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 (42). This is also consistent with the finding that while the helical domain of the mutant is partially disrupted, its N-terminus is more dynamic than the wildtype (68). 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

Here we presented the *blobulator* toolkit consisting of a webtool, a command line interface, and a VMD plugin. We used these tools to provide a new perspective on the underlying organization of classic proteins from multiple families. Using only protein sequences, we were successfully able to identify blobs corresponding to the hydrophobic core and transmembrane domains, and further understand known disease-causing mutations by viewing how they alter blob topology. We expect that future studies may benefit from taking these same principles and using the *blobulator* CLI on much larger datasets than the handful of examples shown here, and in particular, cases where a clear definition of local contexts within a protein sequence is necessary. 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. Ultimately, we hope to fill in some of the gaps inherent to the structural hierarchy by providing an interactive tool that detects the intrinsic modularity in any protein sequence based on hydrophobicity.

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 LR, ESM, CP, and GB are supported by NSF DGE 2152059. We are also grateful to Dr. Jérôme Hénin, Mr. Jesse Sandberg, and Mr. Jahmal Ennis for testing and feedback.

REFERENCES

1. Porollo, A. A., R. Adamczak, and J. Meller, 2004. POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics* 20:2460–2462.

2. Porollo, A., and J. Meller, 2007. Prediction-based fingerprints of protein–protein interactions. *PROTEINS: Structure, Function, and Bioinformatics* .

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

4. Dyson, H. J., and P. E. Wright, 2002. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology* .

5. Ward, J., J. Sodhi, L. McGuffin, B. Buxton, and D. Jones, 2004. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* .

6. Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 46:310–315.

7. Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, 2010. A method and server for predicting damaging missense mutations. *Nature Methods* .

8. Zhang, H., M. S. Xu, X. Fan, W. K. Chung, and Y. Shen, 2022. Predicting functional effect of missense variants using graph attention neural networks. *Nature Machine Intelligence* .

9. Ng, P. C., and S. Henikoff, 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* .

10. Hopf, T. A., J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, and D. S. Marks, 2017. Mutation effects predicted from sequence co-variation. *Nature Biotechnology* 35:128–135.

11. Laine, E., Y. Karami, and A. Carbone, 2019. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Molecular Biology and Evolution* .

12. Frazer, J., P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks, 2021. Disease variant prediction with deep generative models of evolutionary data. *Nature* .

13. Khurana, S., R. Rawi, K. Kunji, G.-Y. Chuang, H. Bensmail, and R. Mall, 2018. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* .

14. Hebditch, M., M. A. Carballo-Amador, S. Charonis, and J. Warwicker, 2017. Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* .

15. Wu, J., J. Wang, Z. Wu, S. Zhang, Y. Deng, Y. Kang, D. Cao, C.-Y. Hsieh, and T. Hou, 2022. ALipSol: An Attention-Driven Mixture-of-Experts Model for Lipophilicity and Solubility Prediction. *Journal of Chemical Information and Modeling* 62:5975–5987.

16. Garbuzynskiy, S. O., M. Y. Lobanov, and O. V. Galzitskaya, 2009. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 26:326–332.

17. Conchillo-Solé, O., N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, 2007. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 8.

18. Walsh, I., F. Seno, S. C. Tosatto, and A. Trovato, 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research* 42:W301–W307.

19. Louros, N., G. Orlando, M. De Vleeschouwer, F. Rousseau, and J. Schymkowitz, 2020. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nature Communications* 11.

20. Tartaglia, G. G., A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo, 2008. Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology* .

21. Shimizu, K., 2014. POODLE: Tools Predicting Intrinsically Disordered Regions of Amino Acid Sequence. *Methods in Molecular Biology* .

22. Jones, D. T., and D. Cozzetto, 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* .

23. Erdos, G., M. Pajkos, and Z. Dosztanyi, 2021. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research* .

24. Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, 2003. Protein Disorder Prediction: Implications for Structural Proteomics. *Structure* .

25. Holehouse, A. S., R. K. Das, J. N. Ahad, M. O. Richardson, and R. V. Pappu, 2017. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal* 112:16–21.

26. Huang, B., T. Fan, K. Wang, H. Zhang, C. Yu, S. Nie, S. Nie, Y. Qi, W.-M. Zheng, J. Han, Z. Fan, S. Sun, S. Ye, H. Yang, and D. Bu, 2023. Accurate and efficient protein sequence design through learning concise local environment of residues. *Bioinformatics* .

27. Castorina, L. V., R. Petrenas, K. Subr, and C. W. Wood, 2023. PDBench: evaluating computational methods for protein-sequence design. *Bioinformatics* .

28. Madani, A., B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* .

29. McConnell, B. S., and M. W. Parker, 2023. Protein intrinsically disordered regions have a non-random, modular architecture. *Bioinformatics* 39:btad732. https://doi.org/10.1093/bioinformatics/btad732.

30. Deléage, G., 2017. ALIGNSEC: viewing protein secondary structure predictions within large multiple sequence alignments. *Bioinformatics (Oxford, England)* 33:3991–3992.

31. Akon, M., M. Akon, M. Kabir, M. S. Rahman, and M. S. Rahman, 2021. ADACT: a tool for analysing (dis)similarity amongnucleotide and protein sequences using minimal andrelative absent words. *Bioinformatics* .

32. Lohia, R., M. E. B. Hansen, and G. Brannigan, 2022. Contiguously hydrophobic sequences are functionally significant throughout the human exome. *PNAS* .

33. Yang, Y., J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, 2018. Sixty-five years of the long march in protein secondarystructure prediction: the final stretch? *Briefings in Bioinformatics* .

34. Jiang, W., C. Chipot, and B. Roux, 2019. 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* 59:3794–3802.

35. Ho, C.-T., Y.-W. Huang, T.-R. Chen, C.-H. Lo, and W.-C. Lo, 2021. Discovering the Ultimate Limits of Protein Secondary Structure Prediction. *Biomolecules* .

36. Ismi, D. P., R. Pulungan, and Afiahayati, 2022. Deep learning for protein secondary structure prediction: Pre and post-AlphaFold. *Computational and Structural Biotechnology Journal* 20:6271–6286.

37. Lohia, R., R. Salari, and G. Brannigan, 2019. Sequence specificity despite intrinsic disorder: How a disease-associated Val/Met polymorphism rearranges tertiary interactions in a long disordered protein. *PLOS Computational Biology* 15:e1007390.

38. Pincus, P., 1976. Excluded Volume Effects and Stretched Polymer Chains. *Macromolecules* .

39. Eisenberg, D., R. M. Weiss, T. C. Terwilliger, and W. Wilcox, 1982. Hydrophobic Moments and Protein Structure. *Faraday Symp. Chem. Soc* .

40. Moon, C. P., and K. G. Fleming, 2011. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *PNAS* .

41. Das, R. K., and R. V. Pappu, 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* 110:13392–13397.

42. Uversky, V. N., J. R. Gillespie, and A. L. Fink, 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics* .

43. Oates, M., P. Romero, T. Ishida, M. Ghalwash, M. Mizianty, B. Xue, S. Dosztányi, V. Uversky, Z. Obradovic, L. Kurgan, A. Dunker, and J. Gough, 2013. D2P2: Database of Disordered Protein Predictions. *Nucleic Acids Research* .

44. Liu, Y., X. Wang, and B. Liu, 2017. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics* .

45. Lohia, R., C. Pitman, E. Santiago-McRae, R. Lamb, L. Riggs, K. Bassi, M. E. Hansen, T. T. Joseph, and G. Brannigan, 2024. https://github.com/BranniganLab/blobulator.

46. Humphrey, W., A. Dalke, and K. Schulten, 1996. VMD: visual molecular dynamics. *Journal of molecular graphics* 14:33–38.

47. Humphrey, W., A. Dalke, and K. Schulten, 1996. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14:33–38.

48. Stone, J., 1998. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla.

49. Ohsumi, K., C. Katagiri, and T. Kishimoto, 1993. A Covalent Enzyme-Substrate Intermediate with Saccharide Distortion in a Mutant T4 Lysozyme. *Science* .

50. Shoichet, B. K., W. A. Baase, R. Kuroki, and B. W. Matthews, 1995. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences* .

51. Alber, T., S. Dao-pin, K. Wilson, J. A. Wozniak, S. P. Cook, and B. W. Matthews, 1987. Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature* .

52. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

53. Jaiteh, M., A. Taly, and J. Hénin, 2016. Evolution of Pentameric Ligand-Gated Ion Channels: Pro-Loop Receptors. *PLOS ONE* 11:e0151934.

54. Cecchini, M., and J.-P. Changeux, 2015. The nicotinic acetylcholine receptor and its prokaryotic homologues: Structure, conformational transitions & allosteric modulation. *Neuropharmacology* 96:137–149.

55. Howard, R. J., 2021. Elephants in the Dark: Insights and Incongruities in Pentameric Ligand-gated Ion Channel Models. *Journal of Molecular Biology* 433:167128.

56. Nemecz, , M. Prevost, A. Menny, and P.-J. Corringer, 2016. Emerging Molecular Mechanisms of Signal Transduction in Pentameric Ligand-Gated Ion Channels. *Neuron* 90:452–470.

57. Sauguet, L., A. Shahsavar, and M. Delarue, 2015. Crystallographic studies of pharmacological sites in pentameric ligand-gated ion channels. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850:511–523.

58. daCosta, C., and J. Baenziger, 2013. Gating of Pentameric Ligand-Gated Ion Channels: Structural Insights and Ambiguities. *Structure* 21:1271–1283.

59. Baeza-Delgado, C., M. A. Marti-Renom, and I. Mingarro, 2012. Structure-based statistical analysis of transmembrane helices. *European Biophysics Journal* 42:199–207.

60. Carswell, C. L., C. M. Hénault, S. Murlidaran, J. D. Therien, P. F. Juranka, J. A. Surujballi, G. Brannigan, and J. E. Baenziger, 2015. Role of the Fourth Transmembrane α Helix in the Allosteric Modulation of Pentameric Ligand-Gated Ion Channels. *Structure* 23:1655–1664. https://doi.org/10.1016/j.str.2015.06.020.

61. Althoff, T., R. E. Hibbs, S. Banerjee, and E. Gouaux, 2014. X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors. *Nature* 512:333–337.

62. Wright, P. E., and H. J. Dyson, 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews* .

63. Spillatini, M. G., R. A. Crowther, R. Jakes, M. Hasegawa, and M. Goedert, 1998. 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* .

64. Galvin, J. E., K. Uryu, V. M.-Y. Lee, and J. Q. Trojanowski, 1999. Axon pathology in Parkinson's disease and Lewy body dementia hippocampus contains a-, b-, and g-synuclein. *PNAS* .

65. Guerrero-Ferreira, R., N. M. Taylor, A.-A. Arteni, P. Kumari, D. Mona, P. Ringler, M. Britschgi, M. E. Lauer, A. Makky, J. Verasdonck, R. Riek, R. Melki, B. H. Meier, A. Bockmann, L. Bousset, and H. Stahlberg, 2019. Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. *eLIFE* .

66. Das, T., and D. Eliezer, 2019. Membrane interactions of intrinsically disordered proteins: The example of alpha-synuclein. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1867:879–889.

67. Vamvaca, K., M. J. Volles, and P. T. Lansbury, 2009. 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* 389:413–424.

68. Lemkau, L. R., G. Comellas, K. D. Kloepper, W. S. Woods, J. M. George, and C. M. Rienstra, 2012. Mutant Protein A30P α-Synuclein Adopts Wild-type Fibril Structure, Despite Slower Fibrillation Kinetics. *Journal of Biological Chemistry* 287:11526–11532.

SUPPLEMENTARY MATERIAL

6.1 Appendix: Pseudocode Blobulation Algorithm

While the blobulation algorithm was described in section 2, we have included a pseudocode representation here.

6.1.1 Digitization

In step 1 of the algorithm, the hydrophobicity of each residue is digitized and smoothed. The required inputs for this step are a hydrophobicity scale (hydroscale), a hydrophobicity threshold (H_{star}), a length minimum (L_{min}), and a protein sequence (sequence).

```
sequence_length = length(sequence)
residue_hydrophobicities = empty list of length sequence_length
digitized_residues = empty list of length sequence_length
for i = 1 to sequence_length do
    residue_hydrophobicities[i] = hydroscale[sequence[i]]
end for
for i = 1 to sequence_length do
    if i == 1 then
        smoothed_hydrophobicity = (residue_hydrophobicities[i] + residue_hydrophobicities[i + 1]) / 2
    else if i == sequence_length then
        smoothed_hydrophobicity = (residue_hydrophobicities[i] + residue_hydrophobicities[i - 1]) / 2
    else:
        smoothed_hydrophobicity = (residue_hydrophobicities[i – 1] + residue_hydrophobicities[i] +
            residue_hydrophobicities[i + 1]) / 3
    end if
    if smoothed_hydrophobicity > H_star then
        digitized_residues[i] = “hydrophobic”
    else
        digitized_residues[i] = “non-hydrophobic”
    end if
end for
return digitized_residues
```

6.1.2 Clustering

After all residues are digitized, they are clustered into blobs. The digitized_residues, sequence_length, and length minimum (L_{min}) variables from step 1 are used as the input for step 2, in which residues are clustered into blobs. The blobulated sequence is returned.

```
hydrophobic_subseq = empty string
non-hydrophobic_subseq = empty string
blobulated_seq = empty string
for i = 1 to sequence_length do
    if digitized_residues[i] == "non-hydrophobic" then
        if length(hydrophobic_subseq) ≥ L_min then
            blobulated_seq += hydrophobic_subseq
        else if length(hydrophobic_subseq) != 0 then
            non-hydrophobic_subseq += “p” * length(hydrophobic_subseq)
            non-hydrophobic_subseq += "p"
        end if
        hydrophobic_subseq = empty string
    end if
    if digitized_residues[i] == “hydrophobic” then
        if length(non-hydrophobic_subseq) ≥ L_min then
            blobulated_seq += non-hydrophobic_subseq
        else if length(non-hydrophobic_subseq) != 0 then
            blobulated_seq += “s” * length(non-hydrophobic_subseq)
```

```
        end if
        hydrophobic_subseq += "h"
        non-hydrophobic_subseq = empty string
    end if
    if i == sequence_length then
        if length(hydrophobic_subseq) ≥ L_min then
            blobulated_seq += hydrophobic_subseq
        else if length(non-hydrophobic_subseq) ≥ L_min then
            blobulated_seq += non-hydrophobic_subseq
        else if length(non-hydrophobic_subseq) != 0 then
            blobulated_seq += "s" * length(non-hydrophobic_subseq)
        end if
    end if
end for
return blobulated_seq
```

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.