SEQUENCE-BASED PROPERTIES THAT IDENTIFY INTRINSICALLY DISORDERED PHASE-SEPARATING PROTEIN REGIONS

by

Ayyam Y. Ibrahim, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Biochemistry
August 2022

Committee Members:

Karen A. Lewis, Chair

Steven T. Whitten, Co-Chair

L.Kevin Lewis

Xiaoyu Xue

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I am very happy and excited to be completing this chapter in obtaining my MS in Biochemistry at Texas State University as I am writing my thesis document. I could not have done or achieved any of this without the love, support, and encouragement of my friends and family as well as the guidance of the wonderful mentors and advisors I have had throughout my life that enabled this opportunity for me.

I would like to make several shout-outs. The first one is obviously to my mom, Subhia Ibrahim, father, Yousef Mostafa, my siblings Ahlam, Kareem, and Ibrahim for continuous guidance and encouragement to believe I can achieve whatever I set my mind to. They were always there to remind me what I was capable of in my times of doubt. An additional arm of that same support system is my long-term Genentech mentors and friends. In particular, Sarah Robinson, Christina Feder, and David Russell, a huge part of why I achieved and accomplished my current goals, and what I will accomplish in my future aspirations would not be at all possible without you all by my side, especially through the most difficult and continuous challenges I have faced during my time here in Texas and beyond.

I would like to sincerely thank and acknowledge my mentor and PI, Dr. Karen A. Lewis, who truly went above and beyond with her support and guidance that provided me ample room for my scientific inquiry to learn, flourish, and ultimately mold me into the scientist I am today. Your endless support and mentorship would not have made any of this possible. Because of you, I was able to find my niche in protein biochemistry. Most

importantly, you taught me to have self-confidence to believe in myself and what I am capable of especially in all the times I have doubted myself. You are also the very first one of my mentors that not only asked the correct pronunciation of my name but said it correctly! My Co-PI, Dr. Steven Whitten, I am truly grateful and thankful for your guidance, support, and confidence in my ability to excel in my scientific capabilities, despite the challenges that come along the way.

I would like to thank my LewisKA lab team members, especially Ezra Hackler and Olga Petrova, for all your unconditional support and help, I cannot thank everyone enough!

Last, but not least I would like to acknowledge my BRIDGES family. Dr. Nicquet Blake, Dr. Babatunde Oyajobi, and Dr. Carolyn Chang, you all have been tremendous at supporting, guiding, and pushing me to help me reach my goals. You all gave me wisdom and encouragement to push forward and reach higher, thank you! I would also like to thank the BRIDGES scholars for their support and the fun times we have had together as a cohort and family. In specific, I would like to thank Marc Rodriguez, Tionna Scott, and Nakya Diaz, we began this journey together, struggling and supporting one another, but we are here now reaching our goals! I hope we continue to be in touch in the next chapter of our journey. Thank you all!

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRACT x

CHAPTER

I. INTRODUCTION 1

II. METHODS 16

III. REVISED PROTEIN SEQUENCE SETS 27

IV. INTRINSIC PROPERTIES OF PHASE-SEPARATING PROTEIN REGIONS 36

V. PREDICTING FOLDED, ID, AND PS-ID PROTEIN REGIONS FROM SEQUENCE 45

VI. CONCLUSIONS 53

APPENDIX SECTION 56

REFERENCES 65

LIST OF TABLES

Table **Page**

1. Summary of mean ν_{model} in the protein sequence sets.....30

2. Summary of mean β-turn propensity in the protein sequence sets 30

3. Summary of mean values for the protein sequence sets 48

LIST OF FIGURES

Figure **Page**

1. β-turn propensity vs. ν_{model} plot shows the means of each set occupy different regions when using three different β-turn propensity scales 8

2. Predicting protein regions that drive phase separation 9

3. Sliding window calculation applied to whole sequence of six other proteins of varying mechanisms.....12

4. ParSe 1.0 identifies PS-IDRs, but is rare in the human proteome 13

5. Means and standard deviation of ν_{model} and β-turn propensity shows that expanded sets and subsets are not statistically different from each other.....32

6. Normal distribution curves for ν_{model} and β-turn propensity of the sequence sets 34

7. Grouping amino acid property scales by type from AAindexver9.2 37

8. p-values calculated by Welch's t-test indicate most amino acid scales can be used to separate the classes 40

9. p-values calculated by Mann Whitney U-test indicate most amino acid scales can be used to separate the classes.....41

10. PCA indicates two primary modes of variation in the human proteome.....43

11. Three intrinsic sequence-based properties separate the protein classes: PS-IDR, IDR, and Folded.....44

12. Predicting protein regions from the primary sequence using a sliding-window algorithm.....46

13. Sequence-calculated ϕ, α, and ν_{model} determined the window label.....47

14. ParSe 2.0 prediction for the phase-separating yeast protein Sup35.....49

15. ParSe 2.0 predicted phase-separating regions in six proteins verified to exhibit LLPS behavior 50

16. ParSe 2.0 predicted PS-IDRs are rare in the human proteome 51

ABSTRACT

We aimed to investigate the molecular mechanisms that drive liquid-liquid phase separation (LLPS) of intrinsically disordered proteins (IDPs). This phenomenon is critical in many cellular processes (including RNA metabolism, chromatin rearrangement, and signal transduction), and known to be driven primarily, but not exclusively, by IDPs. To fully understand how these processes occur and are regulated, it is important that we understand the interactions and sequence properties underlying phase separation behavior. IDPs are proteins that contain intrinsically disordered regions (IDRs), which are regions that do not adopt stable tertiary or secondary structures. While at least 40% of the human proteome is classified as IDPs, only a subset exhibit phase separation behavior.

Previous work created a computer algorithm called ParSe (partition sequence) that successfully predicts folded, ID, and phase-separating (PS) IDRs from the protein primary sequence. This algorithm uses the polymer scaling exponent, *ν*, and a conformational parameter, the intrinsic beta-turn propensity, to distinguish the three protein classes (folded, F; disordered, D; and phase-separating disordered, P). Here, we confirm that the *ν* and beta-turn propensity values follow a normal distribution in three expanded protein sequence sets (PS-IDR, IDR, and Folded). Next, we determined the ability of 568 intrinsic sequence-based properties to define the F, D, and P populations in the sequence sets. We found that most of these properties yield statistically significant differences in the means of the sequence sets. Principal component analysis identified

two principal modes of variance in the human proteome: one corresponding to physiochemical properties, like hydrophobicity, charge, or *v*, and the other to conformational propensities, like preferences for alpha-helix, beta-turn, or beta-sheet. These results established that a hydrophobicity scale could accurately distinguish between folded and ID populations, and that an alpha-helix scale paired with *v* could optimally identify PS-IDR from IDR. Using those three parameters, a second-generation version of ParSe was developed.

I. INTRODUCTION

Eukaryotic cells contain organelles and compartments that are utilized to facilitate many biochemical reactions, whereby specific molecules and proteins are sequestered within such structures and their associated reaction rates and processes are concomitantly increased.[1] Cells are found to contain two classes of organelles: membrane-bound organelles, such as the endoplasmic reticulum, which is used to synthesize, process, and transport proteins, and membraneless organelles (MLOs), such as the nucleolus, which controls ribosome biogenesis.[1,2,3] The distinction between these two organelle classes is that one has a lipid-based encapsulation, while the other is membrane-free.[3]

Early studies established that MLOs exhibit liquid-like behavior and coalesce through a physicochemical process referred to as liquid-liquid phase separation (LLPS).[2] LLPS in cells is a spontaneous and reversible assembly of proteins and other biomacromolecules where, through a de-mixing transition, two liquid phases are formed, one dense and the other dilute.[4] Also called biological condensates, MLOs form stable droplets that are found to exchange molecules with the surrounding cytoplasm.[3] Phenomenologically, LLPS can be illustrated by the example of oil and vinegar. After oil and vinegar are vigorously mixed, they de-mix over time and ultimately form two distinct phases upon reaching equilibrium. This is easily demonstrated with oil and balsamic vinegar. Mechanistically, MLOs are formed through multivalent interactions between biological macromolecules.[5] MLOs may include protein-protein, protein-nucleic acid, and nucleic acid-nucleic acid interactions; these macromolecular interactions are

mediated at the atomic level by π-π, cation-π, cation-anion, and/or dipole-dipole interactions.[5]

Functionally, these condensates play important roles in a wide variety of cellular processes, including cell cycle regulation, signal transduction, and chromatin rearrangement.[4,5] A classic example is stress granules, which are ribonucleoprotein MLOs formed from mRNAs stalled at translation initiation along with both RNA-binding proteins and non-RNA-binding proteins, including many translation initiation factors.[6] Protein-RNA interactions also form within P granules, an MLO that is associated with RNA metabolism.[6,7]

One well-studied component of stress granules is heterogenous nuclear ribonucleoprotein A1 (hnRNPA1).[7] This protein contains two RNA recognition motifs (RRMs) at its N-terminus that binds to RNA molecules and an intrinsically disordered low complexity domain (LCD) at its C-terminus.[7] LLPS by hnRNPA1 occurs spontaneously in a protein concentration-dependent and temperature-dependent manner without a molecular crowding agent.[7] While phase separation of hnRNPA1 is promoted upon binding to RNA, the LCD is sufficient to promote LLPS even in the absence of RNA ligand.[7] Hexanediol is a compound that disables the selectivity filter of the nuclear pore complex phenylalanine-glycine motifs in natively unfolded domains.[8,9] LLPS by hnRNPA1 was disrupted by hexanediol, suggesting that aromatic residues in the LCD contribute to LLPS.[7] Additionally, lowering the NaCl concentration led to LLPS at lower concentrations, indicating that electrostatic interactions also contributed to LLPS.[7] Furthermore, hnRNPA1 is enriched in the aromatic residues (tyrosine and

phenylalanine) and charged residue (arginine) in the LCD that are critical for cation-π interactions in driving phase separation.[7]

As observed for hnRNPA1, electrostatic properties of macromolecules can be critical for MLO formation.[10] Furthermore, posttranslational modifications (PTMs) altering the proteins charge features is a way to control the multivalent interactions of the proteins, thereby regulating phase separation.[10,11,12] PTMs play a critical role in signaling pathways and contributes to the structural, functional, and dynamic integrity of MLOs.[10] An example of tyrosine phosphorylation is the T-cell receptor (TCR) signaling pathway.[13] The TCR consists of multivalent proteins: protein linker for activation of T-cells (LAT), growth factor receptor-bound protein 2 (GRB2), GRB2-related adaptor protein 2 (Gads), son of sevenless 1 (Sos1), and SH2 domain-containing leukocyte protein of 76kDa (SLP-76).[13] LAT contains multiple tyrosine residues, whereby the tyrosine's get phosphorylated (phosphotyrosines), which then act as binding sites for GRB2 and Gads to activate Sos1 or SLP-76.[13] By mutating the tyrosine residues on LAT to phenylalanine residues, they found that phosphorylation of those sites reduced cluster formation.[13] This suggests that tyrosine phosphorylation induces phase separation of the TCR.[9] Additionally, this indicates that phosphorylation can be used to regulate the mechanisms controlling the assembly and disassembly of MLOs.[13]

Similarly, the addition of acetyl groups to positively charged amino acids such as lysine will not only neutralize the positive charge but will also enhance its hydrophobicity, affecting the inter- and intra-molecular interactions of the polypeptide.[14] Tau protein is an IDP that is enriched in lysine residues , some of which are acetylated.[15] Neutralization of the lysine positive charge by acetylation has been found to have

significant effects on Tau function (microtubule assembly/stabilization) and dysfunction (Tau aggregation) by mediated p300/CREB-binding protein (CBP) HAT.[15-19,20]

Intrinsically disordered proteins are a critical component of many MLOs

The underlying mechanisms that facilitate LLPS of biological macromolecules are not fully understood. However, proteins that contain intrinsically disordered regions (IDRs) have been found to play a major role in LLPS.[22] IDRs are regions within proteins that do not adopt stable tertiary structures or secondary structures. Intrinsically disordered proteins (IDPs) are proteins with IDRs. In general, IDRs are depleted in hydrophobic amino acids relative to folded protein regions, which supports the observation that IDRs cannot spontaneously fold into globular structures stabilized by a hydrophobic core.[24,25]

Rather, IDRs typically contain a higher proportion of charged amino acids than globular proteins, and additionally are often comprised of repetitive and/or low complexity sequences.[24,25] Based on these distinct differences in the compositions of folded and IDRs within proteins, many sequence-based algorithms to predict the presence of IDRs in a protein have been successfully developed.[26-28] When whole genomes are analyzed using these algorithms, IDPs are estimated to be a large percentage ($\geq 40\%$) of eukaryotic proteomes, including the human proteome.[29-31] Such a large proportion of the proteome containing IDRs suggests that IDR-containing proteins may confer useful cellular functions that were able to be positively selected by evolution.[32] Indeed, IDRs can provide many potential functional advantages, particularly within cellular signaling and regulatory pathways. For example, posttranslational modification of residues within the IDR enables regulation of both protein structure and function, while the inherent

conformational flexibility of IDRs could allow a sequence to potentially adopt different conformations when binding to different partners.[32] Another possible benefit that would be subject to natural selection is the ability of IDR sequences to undergo or even drive LLPS.[32]

Supporting this hypothesis, many MLOs involve proteins that contain RNA-binding domains. As described above, P-granules are ribonucleoprotein-rich MLOs that regulate RNA metabolism.[33,34] One characteristic P-granule component is the well-studied LAF-1 protein, which has a disordered N-terminal arginine-glycine-glycine (RGG)-rich domain.[34] This sequence is a mixture of positively charged (arginine) and negatively charged (aspartic acid) residue; electrostatic interactions between these charged residues are sufficient to promote phase separation *in vitro*.[33,34] The RGG domain of LAF-1 is also critical for driving protein-RNA interactions.[33,34] Tau is a microtubule-associated protein also found in MLOs like the nucleolus and stress granules.[35] Tau has a strong propensity to promote phase separation mediated by electrostatic interactions between positively charged residues in the C-terminal/middle regions and negatively charged N-terminal region.[35] Hydrophobic interactions also play a role in promoting LLPS of Tau.[35]

Identifying phase-separating IDRs using only amino acid sequence

While the presence of IDRs in a protein sequence can be accurately predicted using a number of computer algorithms, several research groups have also developed sequence-based predictors to identify LLPS-competent regions.[36-38] Some of the multivalent molecular interactions that are thought to facilitate the formation of protein-

rich droplets, such as hydrophobic, π-π, ion-π, and charge-charge interactions, can be identified from the intrinsic properties of a sequence and, as such, LLPS predictors have been developed for primary sequences.[36-38]

The same interactions that drive LLPS have also been hypothesized to affect the hydrodynamic size (e.g., radius of hydration, R_h, or radius of gyration, R_g) of IDRs in the monomeric state.[39-42] Conceptually, a disordered, flexible polymer that has favorable interactions with the solvent adopts an ensemble of elongated and swollen conformations with an average R_h that is larger than the average R_h of the relatively compacted ensemble that is observed when self-interactions dominate. Similarly, the propensity for a particular protein to drive phase separation is determined by the balance of intramolecular and solvent interactions, with LLPS requiring protein-protein contacts over protein-solvent contacts.[43]

One framework to quantify such a relationship is derived from polymer theories developed for long homopolymers.[44,45] Called the polymer scaling exponent, ν, this metric is obtained experimentally from the dependence of size (e.g., hydrodynamic radius, R_h, or radius of gyration, R_g) on polymer length, N, in the power law relationship, $R_h \propto N^{\nu}$. Small values for ν (~0.3) indicate a net preference for self-interactions, while larger values (~0.6) suggest chain-solvent interactions are preferred instead.[46] Because proteins are heteropolymers, the parameter ν_{model} was introduced as a phenomenological substitute to ν and is used to normalize the protein hydrodynamic size to its chain length:[43]

$$\nu_{\text{model}} = \log (R_h/R_o)/\log (N) \quad (1)$$

where R_o is a constant set to 2.16 Å, and R_h can be calculated from sequence using an equation that has been found to be accurate for monomeric IDPs.[47-50]

Previous studies by Whitten and coworkers demonstrated that sequence-calculated ν_{model} could indeed predict the potential for IDR sequences to de-mix, when ν_{model} was combined with a second parameter, the intrinsic propensity for a sequence to form β-turns.[43] Using a simple, surface area-based molecular model, they proposed a physical mechanism for a β-turn role in promoting LLPS: transient β-turn structures reduce the desolvation penalty of forming a protein-rich phase and increase exposure of atoms involved in π/sp^2 valence electron interactions. By this mechanism, β-turns could act as energetically favored nucleation points, which may explain the increased propensity for turns in IDRs utilized biologically for phase separation. Moreover, the combination of ν_{model} and β-turn propensity composition were distinctive across three sequence sets: folded, ID, and phase-separating (PS) IDR sequences, demonstrated a difference in ν_{model} and intrinsic β-turn propensity (Figure 1). Based on this finding, Whitten and coworkers developed the computer algorithm ParSe (partition sequence) that can accurately identify folded, ID, and PS-IDRs from the protein primary sequence.[43]

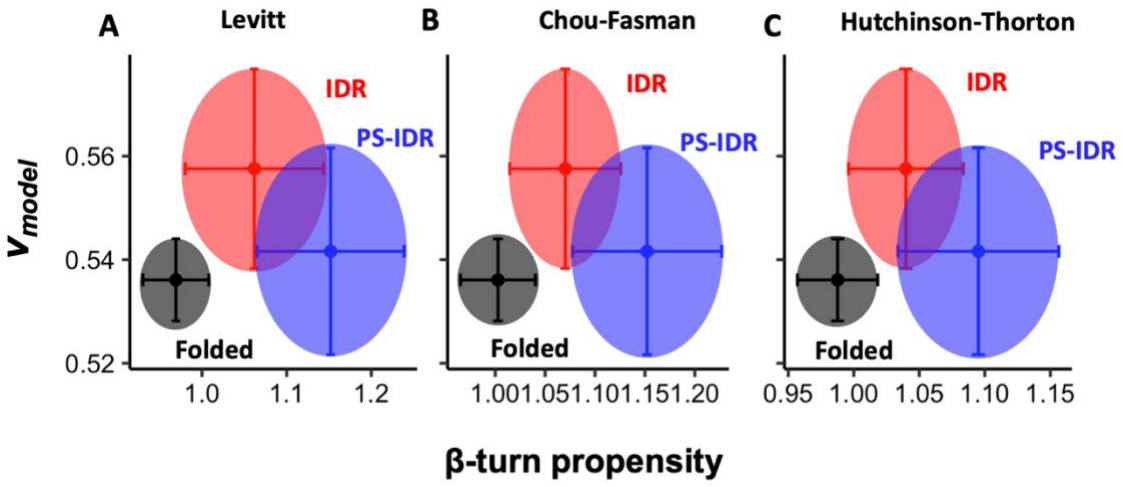


Figure 1. β-turn propensity vs. v_{model} plot shows the means of each set occupy different regions when using three different β-turn propensity scales.[43] A). Levitt scale, **B).** Chou-Fasman Scale, and **C).** Hutchinson-Thorton Scale. Levitt, Chou and Fasman Scales are single-position potential scales. Hutchinson and Thorton are four-position potential scales.[43,51-53]

Figure 2 demonstrates the ability of the ParSe algorithm to identify PS-IDRs in proteins with diverse reported mechanisms driving LLPS.[43] Briefly, to analyze proteins without using predefined boundaries for different regions, the algorithm uses a 25-residue window and then slides this window across the whole sequence in 1-residue steps, as shown schematically in Figure 2A. For each 25-residue window, v_{model} and β-turn propensity are calculated from the amino acid sequence of the window. These values are mapped onto a β-turn propensity *versus* v_{model} plot, which was divided into sectors labeled PS, ID, and Folded (Figure 2B). Sector boundaries were defined by the mean and standard deviations in v_{model} and β-turn propensity in the IDR sequence set (Figure 1). Figure 2, B-C shows the results from using this algorithm on the Sup35 sequence (UniProt accession ID P05453) where each dot in Figure 2B represents a different 25-residue window. The Sup35 primary sequence was then assigned a new three-letter code:

P, D, or F based on window localization into the PS, ID, or Folded sectors. Next, identified regions in the Sup35 sequence of length ≥20 residues that were at least 90% of only one of these labels were color-coded (Figure 2C). Consistent with the ParSe prediction, it has been shown that Sup35 has an N-terminal prion domain (residues 1-125) that mediates phase separation, and ID middle and folded C-terminal domains.[54,55]

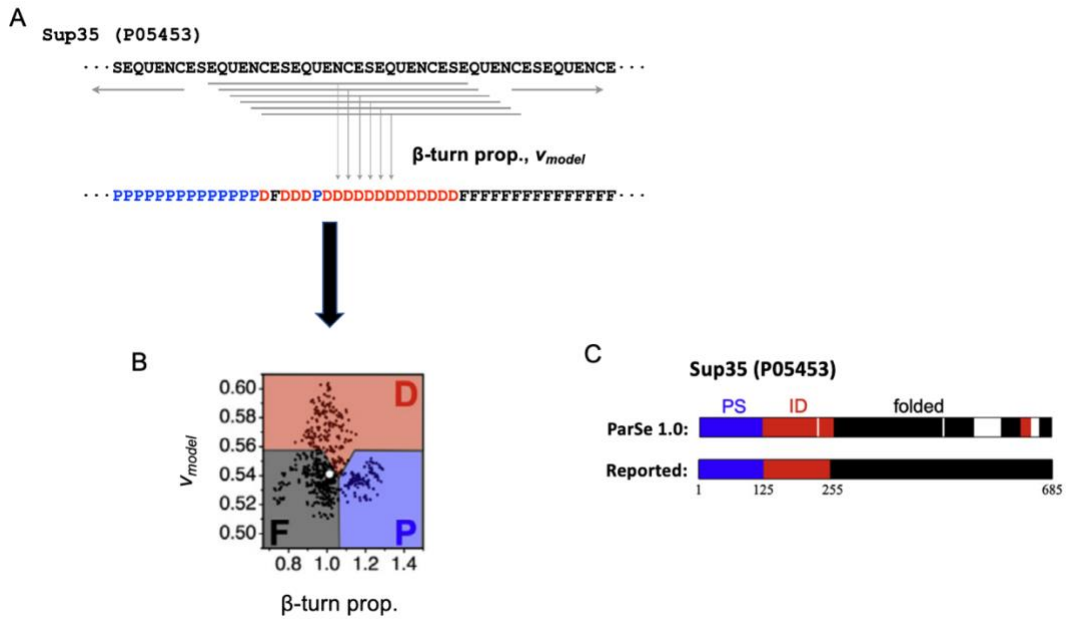


Figure 2. Predicting protein regions that drive phase separation.[43] A). Sliding window calculation that was used to identify protein regions that drive phase separation by using v_{model} and β-turn propensity. 25-residue window were calculated for each in the primary sequence. **B).** Each window is assigned a label represents PS-IDR(blue), D represents IDR (red), and F represents folded regions (black).The label was given to the central residue of the window. N- and C-terminal residues that do not belong to a central window position were assigned to a label of the first and last windows. The larger white dot is the calculated v_{model} and β-turn propensity for the whole protein sequence. **C).** Contiguous regions containing residues greater than or equal to 20 that were 90% of only one label P, D, or F were colored blue, red, or black to represent PS-IDR, IDR, or Folded regions.

Figure 3 A-F shows the results from applying this algorithm to the whole sequences of six multi-domain proteins that are well-characterized and known to have regions that drive LLPS: FUS, LAF-1, spidroin-1, SSB, DDX4, and eIF4G2.[56-66] FUS has multiple domains: an N-terminal PS-IDR (also a low-complexity domain or LC domain),

a C-terminal PS-IDR, three arginine-glycine-glycine (RGG1, RGG2, RGG3) repeat domains, and a folded RNA recognition motif (RRM).[56,57] The folded RRM is a short domain that consists of residues 285-371.[57] The LC domain has been identified as the major driver in promoting LLPS.[57] Recent studies found that phase separation of FUS is driven by the cation-pi interactions between multiple arginine residues in the RGGs and tyrosine residues in the LC domain.[57] As described above, the N terminus of LAF-1 is intrinsically disordered and contains an arginine/glycine rich domain (residues 1-168) that uses electrostatic interactions to both promote phase separation and bind single-stranded RNA.[58] The core of the protein (residues 231-628) contains a RecA-like DEAD box helicase containing ATP and RNA-binding sites.[58] Spidroin-1 is a silk-wrapping protein consisting of highly repetitive sequence that alternates between folded regions and short IDR regions; the IDR sequences drive phase separation via hydrophobic interactions.[59,60]

Single-stranded DNA-binding protein (SSB) contains a highly conserved C-terminal peptide (CTP) that has protein-protein interactions and a less conserved intrinsically disordered linker (IDL) that is thought to drive LLPS.[61,62] LLPS of SSB proteins is thought to be driven by a low sequence complexity ID linker region that connects a highly conserved N-terminus OB fold (residues 1-113) to a C-terminal peptide motif (residues 168-178).[61,62] *In vitro* studies found that salt interactions with the backbone (especially at glycine positions) and at side chain amide groups in the IDL are necessary in regulating the propensity to undergo LLPS.[61,62] Moreover, SSB is an example of a highly multivalent phase-separating system.[61,62] Protein:DNA and protein:protein interactions are necessary for phase separation of SSB, and so not only is

DNA required but both the folded and PS-IDR domains of SSB must be present to facilitate phase-separation.[61,62]

Like LAF-1, DDX4 is DEAD-box helicase.[63,64] However, its N terminus (residue 1-236) uses a far more complex network of charge, hydrophobic, cation-π, and aromatic interactions to drive phase separation, relying heavily on interactions that involve phenylalanine and arginine residues.[63,64] Finally, the eIF4G2 translational regulator protein contains a short, N-terminal PS-IDR region (enriched in glutamine and asparagine) that has been experimentally demonstrated to be sufficient to drive LLPS *in vitro*.[65,66] Also, modeling based on sequence similarity has been used to predict two structured domains in eIF4G2, one of which was identified by ParSe.[65,66] Overall, the ParSe algorithm predicted regions driving LLPS in proteins with a variety of reported mechanisms, indicating that ν_{model} and β-turn propensity may represent a unifying property driving LLPS.

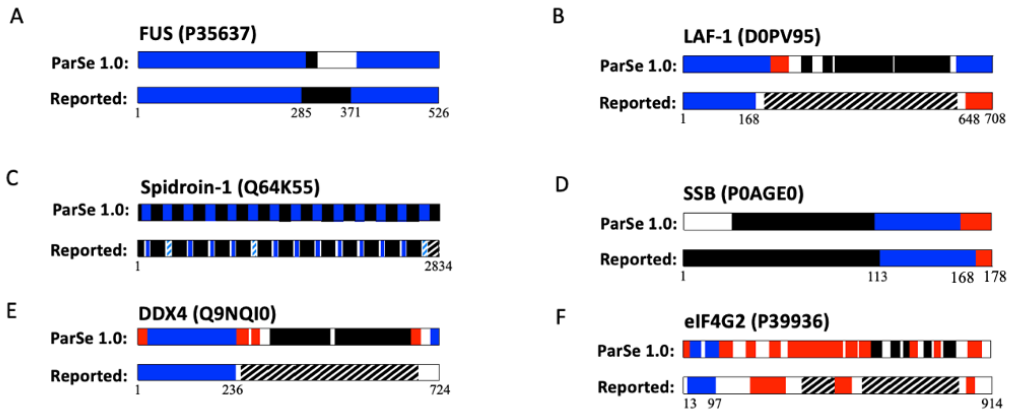


Figure 3. Sliding window calculation applied to whole sequence of six other proteins of varying mechanisms. A-F). Proteins are identified by name and UniProt accession number.[43,56-66] Blue regions are PS-IDR, red regions are IDR, and black regions are Folded. Striped regions represent 80% identify to a known sequences that phase separate (blue) or fold (black).[67]

Moreover, it was noticed that known LLPS proteins (e.g., those in Figure 3) had not just IDRs with high average β-turn propensity and low average ν_{model}, but that they tended to contain long (≥50 residue) stretches labeled by ParSe to be "P". To determine if this feature is unique to proteins driving LLPS, the prevalence of regions predicted from sequence to have high LLPS potential was calculated in the human proteome. These were identified as regions with at least 90% of residue positions labeled as "P" by ParSe. Figure 4 shows that ~70% of the human proteome had a region at least one residue in length with predicted high LLPS potential (i.e., a single P-labeled position), while only ~4% have such a region that is at least 50 residues in length. This result shows that few human proteins possess a region of substantial length (≥50 residues) that combines high β-turn propensity with low ν_{model}.[43] This calculation was repeated for a set of 43 proteins assembled by Vernon *et al* that have been verified *in vitro* to exhibit phase separation behavior, finding that ~90% of these *in vitro* sufficient LLPS protein have a region

predicted by ParSe to have high LLPS potential that is 50 residues in length or longer.[37] The DisProt database, which is a collection of experimentally verified IDPs and IDRs, mirrored the human proteome result, demonstrating that ID alone is not sufficient to trigger LLPS prediction by ParSe.[69,70] The set of proteins listed by SCOPe (Structural Classification of Proteins extended, version 2.07) that represent the globular fold classes across families and superfamilies were mostly devoid of regions predicted to have high LLPS potential by ParSe.[71,72] Thus, while proteins containing long, contiguous P-labeled regions are highly represented in proteins known and verified to undergo LLPS, these regions appear relatively unique to that class of proteins.[43]

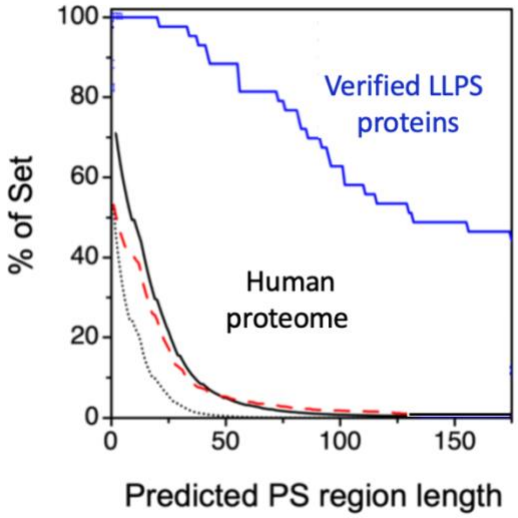


Figure 4. ParSe 1.0 identifies PS-IDRs, but is rare in the human proteome. Solid blue is ParSe 1.0; blue is confirmed LLPS proteins. Solid black line is the human proteome (~75,000 protein sequences), red is the DisProt database (~1,500 IDR sequences), and gray is the SCOPe database (~14,000 folded sequences).[67,69-72]

Thesis goals

Previously, Whitten and coworkers investigated the wide-spread idea that the polymer scaling exponent (ν) could predict LLPS potential in IDRs and found that it could when combined with the β-turn propensity.[43] This result was used to develop a computer algorithm, ParSe, for identifying phase-separating IDRs within proteins, and further suggested a mechanistic role for β-turns in promoting the formation of protein-rich droplets.[43]

The goals for this thesis dissertation are three-fold. First, the set of sequences used to represent non-phase-separating IDRs will be increased. Previously, this set consisted of only 23 IDRs that were selected because experimental R_h were available and thus these IDRs could be used to test the sequence-based equation that calculates ν_{model}.[43] To expand the IDR sequence set, all sequences found in the Biological Magnetic Resonance Bank (BMRB) and DisProt databases classified as ID were added to the original set.[69,70,73] However, any sequence that matched a sequence found in the Protein Data Bank (PDB), a repository of verified folded proteins, were omitted.[67] The set of folded protein sequences used previously was also expanded.[74] The original set was obtained from known LLPS proteins but expanded here to include folded regions from a wider set of proteins, obtained from another study.[74] The set of PS-IDR sequences used previously, which represents 224 unique protein sequences, was not expanded.[43]

Second, the range of amino acid properties that identify PS-IDRs from sequence will be exhaustively investigated to gain insight into the mechanisms and protein features that possibly have a role in phase separation. This will be done using 566 amino acid properties obtained from the Amino Acid Index database which is a curated set of

numerical indices representing various physicochemical and biochemical properties of the amino acids.[75-79] Also included is a newly developed hydrophobicity scale that was designed to predict sequences that drive LLPS.[80] The amino acid properties that identify PS-IDRs will be determined by finding a statistical difference in the means when the three sequence sets, folded, ID, and PS-IDR, are compared. Principal component analysis (PCA) will be used to determine those properties that exhibit different modes of variance in the human proteome, and thus can be combined for predicting protein class.

Third, and lastly, the findings from the second goal will be leveraged to develop a second-generation version of the ParSe algorithm and determine if the changes improve the predictive accuracy of ParSe.

II. METHODS

PS-IDR sequence set

Sequences of intrinsically disordered proteins that are known to exhibit LLPS behavior were obtained from Vernon *et al* and two curated databases of experimentally characterized proteins, PhasePro and DisProt database.[37,69-72] These were chosen because each contains lists of proteins that have been manually curated for experimentally verified cases of LLPS.[43] DisProt is a database of intrinsically disordered proteins that are manually curated from literature.[69,70] PhaSePro is a comprehensive database of proteins that are known to drive phase separation in living cells.[68] We began with the IDRs from 43 proteins reported by Vernon *et al* to undergo phase-separation as purified, isolated proteins *in vitro* from those that do not.[37] To identify the IDRs in the Vernon *et al* protein set, we used the GeneSilico MetaDisorder, which is a service online that predicts PS-IDRs in a sequence.[26] We next added IDRs from 59 human proteins to this set listed in the PhaSePro database as showing LLPS behavior and 18 IDRs annotated "liquid-liquid phase separation were found and identified by search using the disorder function ontology identifier for LLPS, IDPO: 00041 in the DisProt database.[68-70] Since DisProt is manually curated for verified cases of IDRs, we assumed that IDRs that drive LLPS lacked folded regions.[69,70] After merging these three subsets of sequences, duplicate entries were removed along with IDRs with N <20.[43] A set of IDPs that are not known to phase separate but with monomeric experimental mean R_h rather than sequence-predicted mean R_h were created from literature reports.[43] The human proteome reference set UP000005640, the Structural Classification of Proteins-extended (SCOPe), and the consensus disordered regions from the DisProt database (06/2021) excluding those

regions with the ontology identifier for phase separation, were used as negative controls representing lists of protein sequences that do not drive phase separation.[67,69-72] Together, this enlarged set contains 224 sequences.

IDR sequence set

The previous IDR set consisted of 23 sequences.[43] To increase the size of the IDR set, we searched the Biological Magnetic Resonance Bank (BMRB), the database of proteins that have been investigated by NMR and found all those regions that are at least twenty residues long that had NMR data consistent with the proteins being disordered and were remaining monomeric in solution.[73] We also used the DisProt database.[69,70] We searched the BMRB database, culled it for sequences that are at least 20 residues in length or longer.[73] For the sequences added to the IDR set, we took all sequences in the BMRB labeled as intrinsically disordered, then removed those that matched sequences found as folded in the protein data bank.[67] Similarly, we took all "consensus" sequences in DisProt, removed those annotated as "liquid-liquid phase separation" and then also removed those that matched sequences found as folded in the protein data bank.[67,69,70] Then we combined these sets, removing duplicates. By expanding the IDR sequence set that do not drive phase separation, we added 98 sequences from the BMRB and the DisProt database.[69,70,73] In total, this set is comprised of 121 sequences (Table 4).

Folded sequence set

The Protein Data Bank (PDB) was used to identify these folded regions (N ≥20). Sequences from folded protein regions were used as a control set (e.g., sequences not

enriched for IDRs that drive LLPS).[67] The previous Folded sequence set consisted of 82 folded sequences.[43] This set was expended from another study.[74] The expanded set contains sequences of 122 human proteins with nonhomologous folded structures, 54 folded extremophile proteins, 53 metamorphic folded proteins, 90 membrane folded proteins, and 32 proteins with small (N = 36) to large (N = 415) folded structures. This sets represents 421 sequences.[74]

Calculation of mean R_h from sequence

The hydrodynamic radius, R_h, is calculated from sequence and is highly predictable from sequence in IDPs and strongly depends on sequence composition.[43] R_h can be accurately predicted from the intrinsic chain bias for the polyproline II (PPII) conformation (Table 6) and protein net charge. The equation for calculating R_h is shown below:

$$R_h = 2.16\text{Å}\cdot N^{(0.503-0.11)\cdot} ln(1-fPPII)) + 0.26\cdot|Qnet| - 0.29\cdot N^{0.5}, \quad (2)$$

where the net charge, Q_{net}, is calculated from the number of lysine and arginine residues minus the number of glutamic and aspartic acid in a protein sequence. N is the number of residues and f_{PPII} is the fractional number of residues in the PPII conformation. f_{PPII} is estimated from $\Sigma\ PPPII,i/N$, where $PPPII$, is the experimental PPII propensity determined for amino acid type i in unfolded peptides and summed over the protein sequence.

Calculation of ν_{model} for each protein sequence

Proteins are heteropolymers and the property, ν_{model} was introduced as a substitution that normalizes protein hydrodynamic size to the chain length from sequence. The equation to calculate ν_{model} is shown in the introduction (equation 1).

Calculation of β-turn propensity

Calculating the propensity for sequences to form β-turns was achieved by the following equation:

$$\sum \frac{scale_i}{N}, \quad (3)$$

where N is the number of amino acids and $scale_i$ is the value for amino acid type i in the normalized frequency for β-turns from Levitt.[51] The normalized frequency for β-turns from Levitt was applied to calculate the specificity in the different turn positions whereby a 4-residue window was applied.[77] With each residue position in the window, a turn position was slid across the protein sequence in 1-residue increments. Next, the summation of the turn potentials in a window was divided by 4, and the overall window sum was divided by the number of windows.

Normal distributions and histograms to evaluate normality of protein sequence sets

A Fortran program, dataset_generation_fromsequence_files.f calculated for every sequence in each of the sets ν_{model} and β-turn propensity. The input files were the previous IDR, expanded IDR, combined Folded, and PS-IDR sequence sets. The output generated from this Fortran script was the sequence number (count), type (F, P, or D), ν_{model}, and β-turn propensity. The output file, generated_dataset.txt was then converted into an excel

spreadsheet. R programming software was applied for different statistical tests in evaluating the normality of each sequence set and comparing sequence sets.

Determining means of two independent groups

We utilized a parametric test, Welch's t-test, and a nonparametric test, Mann-Whitney U-test to compare the means of two-independent groups.[81,82] In other words, we used these statistical tests to also evaluate the normality of our protein sequence sets (results in Tables 1 and 2) for ν_{model} and β-turn propensity.[81,82]

Three additional Fortran programs, Mann_Whitney_U_test_beta_turn.f, Mann_Whitney_U_test_nu.f, and Welch's_T_test.f gave us calculated one-tailed p-values to compare the means of two protein sequence sets at a time. First, we used the Welch's_T_test.f script for the calculation of Welch's t-test that involved two components: t-statistic and degrees of freedom (dof).[81] The normalized frequency of β-turn's by Levitt and Intrinsic PPII bias measured in peptides by Hilser's group were incorporated in ALL the programs including the input files of protein sequence sets (shown above) to obtain ν_{model} and β-turn propensity values (values for each amino acid in Tables 5 and 6).[77-80] The t-statistic is calculated as follows:

$$t - statistics = \frac{(mean1 - mean2)}{\sqrt{(\frac{var1}{N1} + \frac{var2}{N2})}}, \quad (4)$$

where mean1 and mean2 are the means of each group, var1 and var2 are the variance of the two groups, N1 and N2 are the first and second sample sizes.[81] The dof is calculated as follows:

$$dof = \frac{(\frac{var1}{N1} + \frac{var2}{N2})^2}{\frac{(\frac{var1}{N1})^2}{(N1-1)} + \frac{(\frac{var2}{N2})^2}{(N2-1)}}, \quad (5).$$

Furthermore, the one-tail p-values of ν_{model} and β-turn propensity were calculated from the t-statistic and dof for Welch's t-test when comparing two proteins sequence sets at a time.[81]

Next, we used both Mann_Whitney_U_test_beta_turn.f and Mann_Whitney_U_test_nu.f programs to determine the one-tail p-values of ν_{model} and β-turn propensity of using the Mann-Whitney U-test.[82] The set order is important for the U-test, therefore the set with fewer entries go first.[82] In this test, the first step was to assign ranks and to do so we ordered our data from smallest to largest.[82] For example, the scripts read the protein sequence set with the fewest sequence in that set first and the largest second. Again, we can only compare two sequence sets at a time. Second, we calculated the sum of the ranks for the two sequence set groups. Third, we determined the sample size for both groups when we calculated the sum of the ranks in each group. Fourth, we computed U for each group using the following equation:

$$U_a = (n_a \cdot n_b) + \frac{n_a \cdot (n_a+1)}{2} - T_a, \quad (6)$$

where n_a and n_b are the sample sizes in each group, T_a is the sum of the rank for one of the groups.[82] Fifth, we determined the value of U to compare it to the U-critical value. An example is the following instance,

U_a = 32 from equation 6

and $U_b = (n_a \cdot n_b) - U_a = 4$,

we would use $U_b = 4$ because it's the smaller value. Sixth, we computed the standard deviation (Std dev) of U and finally computed the Z-score using the following equation:

$$Z = \frac{U - \left(\frac{n_a \cdot n_b}{2}\right)}{Std\ dev}, \quad (7)$$

Furthermore, the output gave us the one-tail p-values of ν_{model} and β-turn propensity for

Mann-Whitney U-test when comparing two proteins sequence sets at a time.

Searching other amino acid properties

We explored and accessed an Amino Acid Index Database that contains 566 different properties that was used to separate PS-IDRs, IDRs, and Folded sequence sets.[75-79] These amino acid property scales consist of various physiochemical and biochemical properties of all amino acid type. The link for the database can be accessed using the following link: https://www.genome.jp/aaindex/.[75-79] Also, we included a new hydrophobicity scale that was designed to predict protein sequences that drive protein phase separation (Table 7).[80] The amino acid index was then downloaded as a text file to be used in the Fortran programs for the identification of amino acid property scales (aaindex1_new.text).

Classifying amino acid property scales from database

Each amino acid property scale was individually evaluated and classified and grouped the scales by type: conformational, physiochemical, and other. Scales that had identifications such as graph theory, hydrostatic pressure, stability, slopes tripeptide, etc were classified as other (Figure 7). Furthermore, we have grouped the hydrophobicity scales as two types: structural and solution scales (Figure 7). Structural scales were produced based on the physical location of the amino acid in three-dimensional structures; for example, propensity to be buried within a globular protein core versus surface-exposed. Almost a third of the 566 scales represent structure-based hydrophobicity scales (Figure 7). The solution-based hydrophobicity scales were defined

based on the partitioning of each amino acid between aqueous and organic solvents. BreakDown, readxl, ggplot2, and tidyverse packages in the R environment were used to extrapolate each amino acid scale into groups (Figure 10).

Identifying new amino acid property scales from database to separate sequence sets

Each scale from the database was calculated individually for each sequence in the PS-IDR, IDR, and Folded sets by the sequence sum divided by *N*. We also performed Welch's t-test and Mann Whitney U-test to obtain one-tail p-values to identify the scales that showed statistical differences in the means of the sequence sets.[81,82] Welch's t-test and Mann-Whitney U-test compares two sets and only one property at a time.[81,82] Two Fortran programs were used for this analysis:
Find_best_sequence_property_Welch_T_test.f and
Find_best_sequence_property_Mann_Whitney_U_test.f. The new hydrophobicity scale from Table 7 was incorporated in this analysis.[80] These programs were used to obtain p-values of ALL amino acid property scales when comparing two sets at a time: PS-IDR from IDR, IDR from Folded, and PS-IDR from Folded. One of the inputs is the amino index in txt format (aaindex1_new.text). The other input file contains the protein sequences in fasta format (shown above) that read two sets at a time (previously mentioned) for the comparison and then computes the one-tail p-values for all amino acid property scales including ν_{model} and the new hydrophobicity scale for a total of 568 properties.

The output from each of these statistical tests is the one-tail p-value. The property with the smallest p-value in that property will generate the greatest statistical difference

between the two sets. To interpret the amino acid property scales when comparing the protein sequence sets, the one-tail p-values obtained from both statistical tests were saved in a text file, which then was converted into an excel spreadsheet. P-values are so small that they are sometimes logged because they can span several orders of magnitude and the log makes the values easier to plot, interpret, and transform the p-values. By using -log, this enabled us to make the p-values positive, and the bigger the value the smaller the underlying p-value (equals more significance). Hence, our p-values are very small in all these compared protein sequence set distributions. Each scale was calculated individually for each sequence by taking the sum for each sequence and dividing it by N in the protein sequence sets: PS-IDR, IDR, and folded.

The following packages in the R environment were used to interpret the p-values when comparing two sets at a time: ggplot2, magrittr, tidyverse, readxl, and ggforce. From these R packages, we then developed box plots to interpret the one-tail p-values of all the amino acid properties for both Welch's t-test and Mann-Whitney U-test.[81,82]

Principal component analysis (PCA) across the human proteome

To identify which amino acid property scale to be used at separating the three protein sequence sets: PS-IDR, IDR, and Folded, we computed the principal components across the human proteome. Principal Component Analysis (PCA) is a statistical procedure that allows us to summarize our information content in the human proteome of the different amino acid properties.[83] This will enable us to capture the variance in the human proteome of the different amino acid properties.[83] We took this approach by taking three amino acid property scales from each category that had the three smallest p-

values. The Fortran program, PCA_dataset_generation_top3_scalees_human_proteome.f read the input file of the Human Proteome in fasta format (sequences.fasta) and the aaindex1_new.txt for this step of our amino acid property analysis.

In the Fortran program, we used a simple sliding window calculation of each of the three scales in order to calculate the sequence sum via a sliding 25-residue window that will be applied to proteins in the human proteome of length of at least 100 amino acid residues to 500 residues (approximately spanning 55,000 human proteins in the sequences.fasta input file). Proteins have modular structure, meaning proteins have disordered and folded regions. By applying the sliding window calculation, we captured differences in sequences of proteins to identify possible protein that have modular structure. Each sequence sum calculated for each of the three property scales was divided by the window length and then ν_{model} was calculated for each window (25-residues each). The output file extrapolated from the Fortran script was of the scale-calculated properties of each 25-residue window as PCA_dataset.txt. PCA_dataset.txt was then read in the R environment to perform PCA across the human proteome. PCA was obtained from the following R packages: ggfortify, ggplot2, factoextra, MetBrewer, and tidyverse.

ParSe 2.0 algorithmic steps

In this step, we took the data obtained from PCA, where we ultimately found two amino acid properties: hydrophobicity structure scale from Vendruscolo and the alpha-helix propensity scale from Scheraga.[84,85] Both of these scales gave means in PS-IDR, IDR, and Folded sets with the smallest p-values when comparing IDR from Folded and PS-IDR from IDR. By modifying our algorithm to make a second version of ParSe. This

version, like ParSe 1.0, an input primary sequence is read to determine its length, N, and the number of each amino acid type (only to the 20 types). R_h is then calculated by Equation 2, incorporating Q_{net} and f_{PPII} (mentioned previously), which then is used to obtain ν_{model} by Equation 1. The hydrophobicity structure and alpha-helix propensity scales are added now for this calculation (Tables 8 and 9).[84,85] ParSe 2.0 uses a sliding window calculation to compute ν_{model}, hydrophobicity, and alpha-helix propensity for every 25-residue window of the primary sequence. Every 25-residue is then slid across a whole sequence in 1-residue increments. The values of ν_{model}, hydrophobicity, and alpha-helix propensity calculated for a window determines the window's localization to a PS-IDR, IDR, and Folded region in a 3D plot. The sector or region boundaries (cutoff) are determined based on the mean ± standard deviation in ν_{model}, hydrophobicity, and alpha-helix propensity at separating PS-IDR from IDR and IDR from Folded.

If a sequence window, based on its ν_{model} and alpha-helix propensity values is high, the central residue in that window is labeled "D" for IDR. If a window, based on ν_{model} and alpha-helix propensity values is low, the central residue in that window is labeled "P" for PS-IDR. If a window, based on its hydrophobicity values, the central residue in that window is labeled "F" for Folded. Protein regions predicted by ParSe 2.0 to be PS-IDR, IDR, or Folded are determined by finding contiguous residues in regions of a sequence with a length greater than or equal to 20 that are at least 90% of only one label P, D, or F.

ParSe 2.0 was then applied to predict protein regions that drive protein phase separation of all mechanisms that was used previously used in ParSe 1.0. We used 7 proteins: Sup35, FUS, LAF-1, Spidroin-1, SSB, DDX4, eIF4G2.

III. REVISED PROTEIN SEQUENCE SETS

Many proteins show modular characteristics, with some regions folded into stable, globular structures, and other regions intrinsically disordered, or ID.[43] Clear, compositional differences are found between folded and ID regions (IDRs) when they are surveyed.[43] For example, differences in hydrophobicity, charge, and sequence complexity.[43] Based upon these differences, predictive algorithms have been developed that identify folded and IDRs within proteins from the primary sequence.[36-38] Among IDRs, some have been found to drive liquid-liquid phase separation (LLPS), while most do not. To better understand the features and properties of proteins that promote LLPS, we have constructed a novel database consisting of three sets of sequences, all derived from experimentally validated biological proteins. The first sequence set is comprised of intrinsically disordered sequences that are known to not undergo phase separation (hereafter called "Non-PS-IDR". The second set is also comprised of intrinsically disordered sequences but is composed of only IDRs that are confirmed to spontaneously phase separate (hereafter called "PS-IDR").[43] The third set was constructed using sequences from folded protein regions (hereafter called "Folded"); this set is used as a control (i.e., to represent sequences that are not enriched for intrinsic disorder). Our hypothesis is that by analyzing the sequence-based differences found between these three subsets, we can identify molecular mechanisms that underlie LLPS behavior.

Defining the sequence sets

This work builds upon a previous study where 224 IDRs were extracted from lists of proteins verified to exhibit LLPS behavior.[43] These protein sequences were obtained

from Vernon *et al*, the PhaSePro database, and the DisProt database.[37,68-72]These sources were chosen because each contains protein lists that have been manually curated for experimentally verified cases of LLPS. The resulting "phase-separating IDR" ("PS-IDR") set of sequences is therefore a set of protein sequences enriched for LLPS behavior. The protein sequences in this set have been published elsewhere.[43]

A set of 23 IDRs not known to phase separate were used as a comparison set.[43] These sequences were chosen because they have been confirmed to remain monomeric in solution (i.e., to not undergo phase-separation) and the hydrodynamic size of each has been measured, allowing for direct tests of the ν_{model} calculation.[43] This original Non-PS IDR set is not extensive; in fact, it is only ~10% the size of the PS-IDR set. Specifically, the small sample size raises the concern that a sample of 23 is not an accurate representation of IDRs in general.

Therefore, we sought to expand the number of sequences in the Non-PS-IDR set. To do this, we started with all sequences in the Biological Magnetic Resonance Bank (BMRB) from proteins or protein fragments exhibiting spectroscopic hallmarks of ID, and thus that have been classified as IDPs (104 sequences).[68] Next, we added the consensus disordered regions (~1,500 sequences) from the DisProt database (2021_06), excluding those regions with the ontology identifier for LLPS, IDPO:00041.[64,65] We found substantial overlap between these two sets. Duplicate sequences (111 sequences) were removed. Additionally, there were some sequences that matched sequences found in the Protein Data Bank (PDB), representing sequences that fold in some contexts.[61] Those sequences were also removed, leaving 98 new IDR sequences that can be added to the original set of 23 IDRs. This expands the original Non-PS IDR set to 121 unique protein

sequences, hereafter referred to as the "Expanded Non-PS-IDR Set" (Table 1).

Simultaneously, similar work was carried out to expand the original set of 82 sequences representing the folded regions within confirmed LLPS proteins ("Folded Set").[43] As detailed elsewhere, an additional 339 nonhomologous sequences were identified in human proteins, membrane proteins, extremophile proteins, small-to-large proteins, and metamorphic proteins.[74] The protein sequences in both the original (82 sequences) and expanded (421 sequences) Folded Sets have been published elsewhere.[43,74]

Comparing ν_{model} and β-turn propensity between the original and expanded sequence sets

These revisions to the composition of these sequence sets might also alter our previous conclusions that the parameters ν_{model} and β-turn propensity could be used to identify folded, non-phase-separating intrinsically disordered, and phase-separating intrinsically disordered regions from primary sequence alone. To determine if the sequence set changes significantly alters the ν_{model} and β-turn propensity of each training set, we evaluated the mean values in ν_{model} (Table 1) and β-turn propensity (Table 2) and compared them in several combinations. First, the mean values were compared between the three expanded sets. These values were also compared between the previous and expanded versions of the Non-PS-IDR and Folded sets (as the PS-IDR set was not revised).

Table 1. Summary of mean ν_{model} in the protein sequence sets.

			vs PS-IDR Test Set		*vs* previous set	
Set	**Number**	ν_{model} [a]	*t*-test [b]	*U*-test [b]	*t*-test [b]	*U*-test [b]
PS-IDR Test Set	**224**	0.542 ± 0.020	-	-	-	-
IDR Null	**121**	0.558 ± 0.022	$2.5e^{-10}$	$1.6e^{-11}$	-	-
Previous IDR Set	23	0.558 ± 0.019	$3.8e^{-4}$	$2.4e^{-4}$	-	-
BMRB & DisProt	98	0.558 ± 0.023	-	-	0.44	0.48
Folded	**421**	0.537 ± 0.008	$1.2e^{-3}$	$1.5e^{-3}$	-	-
Previous Folded Set	82	0.536 ± 0.008	$2.5e^{-4}$	$7.8e^{-3}$	-	-
Human	122	0.536 ± 0.007	-	-	0.40	0.32
Small-to-large	32	0.537 ± 0.009	-	-	0.36	0.41
Extremophile	54	0.542 ± 0.011	-	-	$1.2e^{-4}$	$2.4e^{-4}$
Membrane	90	0.537 ± 0.006	-	-	0.17	0.21
Metamorphic	53	0.537 ± 0.006	-	-	0.15	0.18

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where values <0.05 indicate the compared sets are statistically different in their means.

Table 2. Summary of mean β-turn propensity in the protein sequence sets.

			vs PS-IDR Test Set		*vs* previous set	
Set	**Number**	**β–turn propensity** [a]	*t*-test [b]	*U*-test [b]	*t*-test [b]	*U*-test [b]
PS-IDR Test Set	**224**	1.152 ± 0.087	-	-	-	-
IDR Null	**121**	1.101 ± 0.075	$4.6e^{-8}$	$4.9e^{-9}$	-	-
Previous IDR Set	23	1.062 ± 0.082	$1.4e^{-5}$	$9.7e^{-7}$	-	-
BMRB & DisProt	98	1.110 ± 0.071	-	-	$6.5e^{-3}$	$9.3e^{-4}$
Folded	**421**	0.971 ± 0.040	$2.0e^{-33}$	$1.1e^{-89}$	-	-
Previous Folded Set	82	0.969 ± 0.039	$8.0e^{-31}$	$1.7e^{-38}$	-	-
Human	122	0.980 ± 0.039	-	-	0.03	0.07
Small-to-large	32	0.968 ± 0.027	-	-	0.42	0.34
Extremophile	54	0.983 ± 0.030	-	-	0.01	0.03
Membrane	90	0.956 ± 0.046	-	-	0.02	0.02
Metamorphic	53	0.972 ± 0.040	-	-	0.30	0.48

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where values <0.05 indicate the compared sets are statistically different in their means.

The means of the previous and expanded Folded sets were overall similar. This was determined by one-tail *p*-values calculated using Welch's unequal variances *t*-test, which assumes a normal distribution, and the nonparametric Mann-Whitney *U*-test, which does not.[81,82] For ν_{model}, sequences from folded regions within extremophiles gave *p*-values <0.05 when compared to the previous folded set, indicating that there is a statistical difference between the means of the folded regions of LLPS proteins (0.538 ± 0.008) and folded regions of extremophile proteins (0.542 ± 0.011). Similarly, statistical differences were found in mean β-turn propensity (i.e., *p*-values <0.05) between the previous PS-IDP set (0.969 ± 0.039) and three subsets of folded sequences: extremophile (0.983 ± 0.030), membrane (0.956 ± 0.046), and human protein (0.980 ± 0.039) folded regions. Generally, these results support that the subsets that were added to the previous Folded set.[43,74]

For the Non-PS-IDR sets, mean ν_{model} is statistically similar between the previous set (0.558 ± 0.019) to the newly-added subset of sequences (i.e., the sequences from the BMRB and DisProt subsets; 0.558 ± 0.023). However, a statistical difference in β-turn propensity was observed between the newly-added sequences (1.110 ± 0.071) and the previous Non-PS-IDR set (1.062 ± 0.082), yielding *p*-values from the *t*- and *U*-tests that were <0.05.[81,82] Overall, however, when comparing means between the sets (Tables 1 and 2) or visually assessing the means in a β-turn propensity versus ν_{model} plot (Figure 5), the differences between the three classes (Folded, Non-PS-IDR, and PS-IDR), were more pronounced than to comparisons of subsets within a class, i.e., between previous and new subsets of the same class type. Moreover, when plotted, the subsets were clearly grouped by class in terms of Folded, Non-PS-IDR, and PS-IDR (Figure 5).

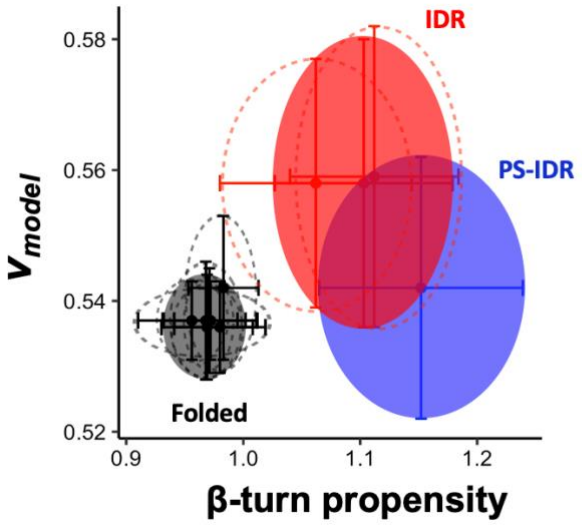


Figure 5. Means and standard deviations of v_{model} and β-turn propensity show that expanded sets and subsets are not statistically different from each other. The dashed lines represent the subsets. The subsets were clearly separated by class: Folded, Non-PS-IDR, and PS-IDR.

Distributions of v_{model} and β-turn propensity in the sequence sets are normal

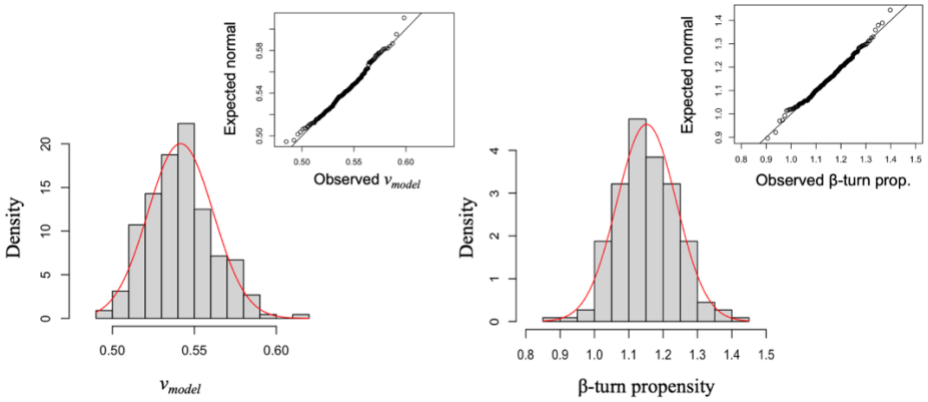
Welch's unequal variances *t*-test assumes that the values in two compared sets are normally distributed.[81] To determine if this is the case in the expanded sequence sets, Figure 6 shows the histogram distribution in v_{model} and β-turn propensity in each set, and then compares the observed histogram distribution to the probability density function of the normal distribution,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad (8)$$

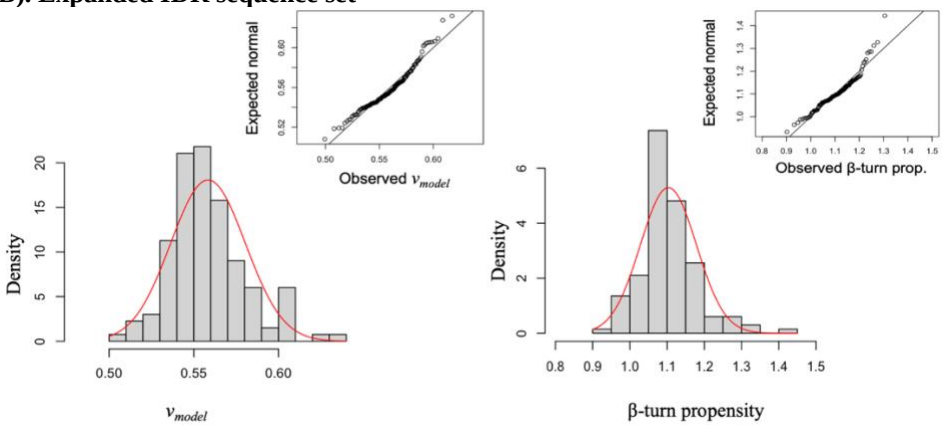
where x is a value in v_{model} or β-turn propensity, and μ and σ are the distribution mean and standard deviation, respectively.[86] Also shown in this figure is a quantile-quantile plot, which directly compares two distributions by plotting the quantiles against each other, in this case the observed and the probability density function. When the trend in a quantile-quantile plot follows the identity line, then this is evidence that the compared

distributions are similar. Because this is observed for each sequence set in both ν_{model} and β-turn propensity, this suggest that the sets are similar to normal. Accordingly, *p*-values calculated by Welch's unequal variances *t*-test are appropriate for statistically comparing the means in ν_{model} and β-turn propensity between the three sequence sets, folded, ID, and PS-IDR.

A). PS-IDR sequence set



B). Expanded IDR sequence set



C). Combined Folded sequence set

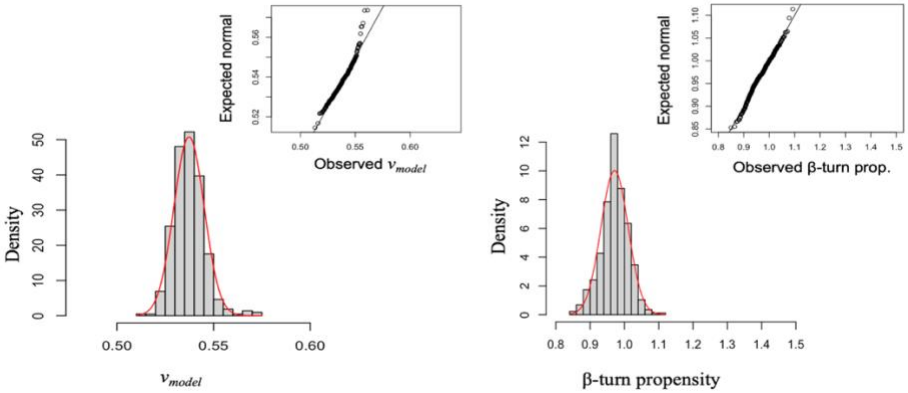


Figure 6. Normal distribution curves for ν_{model} and β-turn propensity of the sequence sets. A).PS-IDR B). Expanded IDR, and C). Combined Folded sequence sets. Red solid lines indicate the normal distribution of the same mean and median values of each of the sets. The quantile-quantile plots (Insets) show the variance from expected, further demonstrating the normality of each sequence set.

Conclusions

In summary, the sequence sets that cumulatively represent folded, ID, and PS-IDRs were expanded to better represent proteins in general. The expanded sets showed differences in mean ν_{model} and β-turn propensity when the folded, ID, and PS-IDR sets were compared, similar to prior results with smaller sequence sets.[38] Moreover, the observed distributions in ν_{model} and β-turn propensity in the expanded sets were similar to normal.

IV. INTRINSIC PROPERTIES OF PHASE-SEPARATING PROTEIN REGIONS

To better understand how IDRs are utilized to drive protein LLPS, we have analyzed protein databases containing subsets of IDRs from proteins not known to phase separate, IDRs from proteins confirmed to spontaneously phase separate, and folded regions from a diverse, but general, set of proteins. The idea is to identify those sequence features and properties that are unique to the phase-separating IDRs, as these possibly represent molecular characteristics selected by evolution to drive the formation of protein-rich droplets.

To do this, we identified from a large, curated set of amino acid property scales, those sequence-based properties yielding statistically significant differences in the means of the three sequence sets.[75-79] To characterize these differences, we applied principal component analysis (PCA) to the human proteome, uncovering the principal modes of variance arising from the identified sequence properties.[85] Results from the PCA indicate that proteomic variance associated with properties that distinguish IDRs and phase-separating (PS) IDRs has two principal modes: one corresponding to physiochemical properties, like hydrophobicity, charge, or ν_{model}, and the other to backbone conformational propensities, like preferences for α-helix, β-turn, or β-sheet. We found that most amino acid scales yield statistically different means in the IDR and PS-IDR sequence sets, suggesting robust instead of discrete property differences.

A curated set of amino acid property scales

We sought to understand how the sequence sets defined in chapter III, i.e., sets representing the folded, ID, and PS-ID protein classes, are different in their physical properties. Amino acid property scales were obtained from the Amino Acid Index database, which manually curates the scientific literature for numerical indices representing the various physicochemical and biochemical properties of the amino acids.[75-79] This database contains 566 amino acid property scales.[75-79] To this list, we added a newly developed hydrophobicity scale designed to predict sequences that drive protein LLPS.[80]

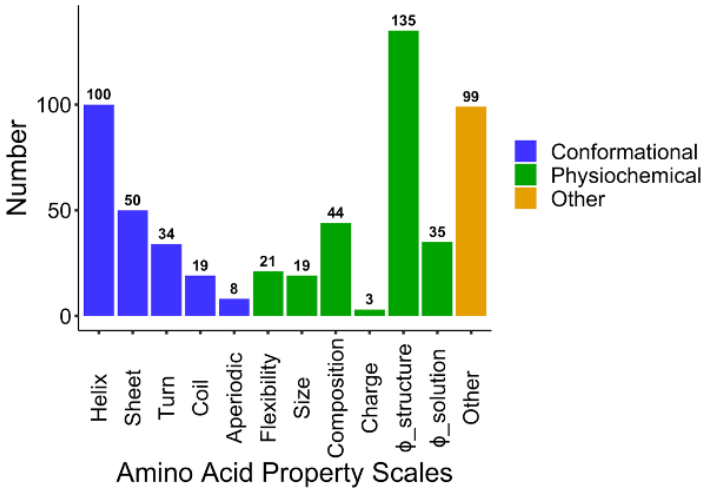


Figure 7. Grouping amino acid property scales by type from AAindex ver9.2.[75-80] 567 amino acid property scales grouped by type depicted in the x-axis and the number of the amino acid scales depicted on the y-axis.

In Figure 7, the amino acid scales are grouped by type and color-coded according to conformation-based scales and physicochemical-based scales. Property scales that do not easily map into a conformation-based or physicochemical-based group (e.g., refractivity, crystal melting point) were combined separately into "other".[75-79] Overall,

hydrophobicity (φ) scales represented almost a third of the amino acid property scales used to analyze the sequence sets.[75-79] Of these, there were two types: those hydrophobicity scales that are structure-based (φ_struct), where the scale is derived from a structural metric like burial or contact frequency in surveys of high-resolution protein structures, and those that are solution-based (φ_sol), where the scale is obtained from solution studies like measuring the transfer free energy of the amino acids from water to an organic solvent.[75-79] Together, conformational propensity scales represented ~37% of the amino acid property scales, with α-helix propensities represented the most.[75-79]

Property differences of the protein sequence sets

For each amino acid property scale, the sequence sum divided by N was calculated individually for each sequence in the three sets: PS-IDR test, IDR null, and folded. The sequence sets are defined in Chapter III, where the null and folded sets refer to the combined IDR null and combined folded sets. ν_{model} also was calculated for each sequence using Equation I, yielding a total of 568 sequence-calculated properties that were investigated.[75-80] To compare the means in any one property between any two of the sequence sets, one-tail p-values were calculated by Welch's unequal variances t-test.[81] By this test, a p-value <0.05 indicates two sets have statistically different means in the compared property (i.e., the probability is less than 5% that the two sets are the same).[81] In contrast, large p-values mean the opposite - the probability is high that the two sets are the same in that specific property.[81]

Figure 8, panel A shows that a scale of α-helix propensities from Scheraga and coworkers gave the smallest one-tail p-value when comparing sequence-calculated

means in the PS-IDR test and IDR null sets, and 82% of scales give p-values <0.05 (indicating means that are statistically different), among the 567 scales and ν_{model}.[84] Moreover, 10% and 25% of scales yield p-values smaller than the p-values obtained from ν_{model} and β-turn propensity, respectively, that are used in ParSe. Each scale type (e.g., α-helix propensity, β-turn propensity, hydrophobicity, etc.) had some scales with very low p-values and some with p-values ≥ 0.05, suggesting that, overall, most, but not all, conformational- and physicochemical-based scales could be used to distinguish IDRs and PS-IDRs from sequence.[75-79]

When comparing means in the PS-IDR test and folded sets (Figure 8, panel B) and in the IDR null and folded sets (Figure 8, panel C), we find that a structure-based hydrophobicity scale from Vendruscolo and coworkers had the smallest one-tail p-value in both cases.[85] Hydrophobicity scales with the lowest p-values when comparing means in the folded and ID sets had among the highest p-values when comparing means in the test and null sets (and vice versa). Also, 95% and 94% of scales produced p-values <0.05 when means were compared between the test and folded sets and the null and folded sets, respectively, showing that almost all amino acid scales yield a statistical difference in mean properties when comparing ID and folded sequences. As before, most scale types (e.g., α-helix propensity, β-turn propensity, hydrophobicity, etc.) had some scales with very low p-values and some with p-values ≥ 0.05, when comparing ID and folded set means.

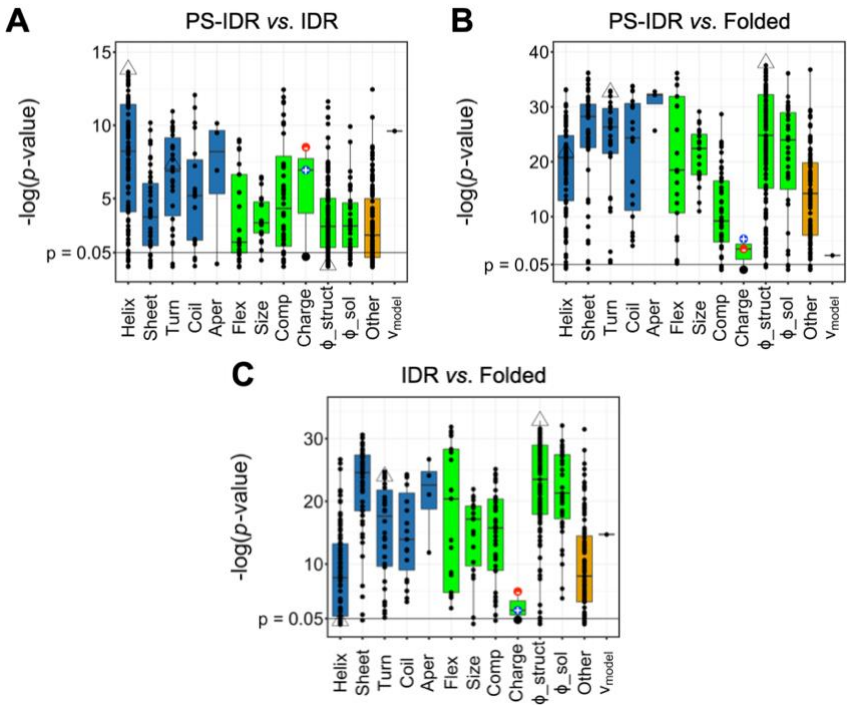


Figure 8. p-values calculated by Welch's t-test indicate most amino acid scales can be used to separate the classes. A). PS-IDR *vs.* IDR, B). PS-IDR *vs.* Folded, and C). IDR *vs.* Folded. Significance level (p-value = 0.05) limit is represented with a solid line across all three box plots. Triangles represent the top three amino acid scales: Turn (by Levitt), Helix (by Scheraga), and Hydrophobicity structure (by Vendrusculo) by their p-values.[51,84,85] The three charged scales are labeled by red circle (negative charge inside circle), blue circle (positive charge inside circle circle), and solid black circle (neutral charge).

While ν_{model} and β-turn propensity are normally distributed in the folded, ID, and PS-IDR sets (see Chapter III), it is not known if the other sequence-calculated properties also show a normal distribution. Because the p-values calculated by Welch's unequal variance t-test assume a normal distribution, p-values also were calculated using the nonparametric Mann-Whitney U-test.[81,82] Figure 9, panels A-C show that overall similar results were obtained with the U-test generated p-values, whereby most of the sequence-based properties show a statistically significant difference in mean values when the three sequence sets are compared. The α-helix propensity and structure-based hydrophobicity scales were the best at separating ID

from PS-ID and ID from folded, respectively.

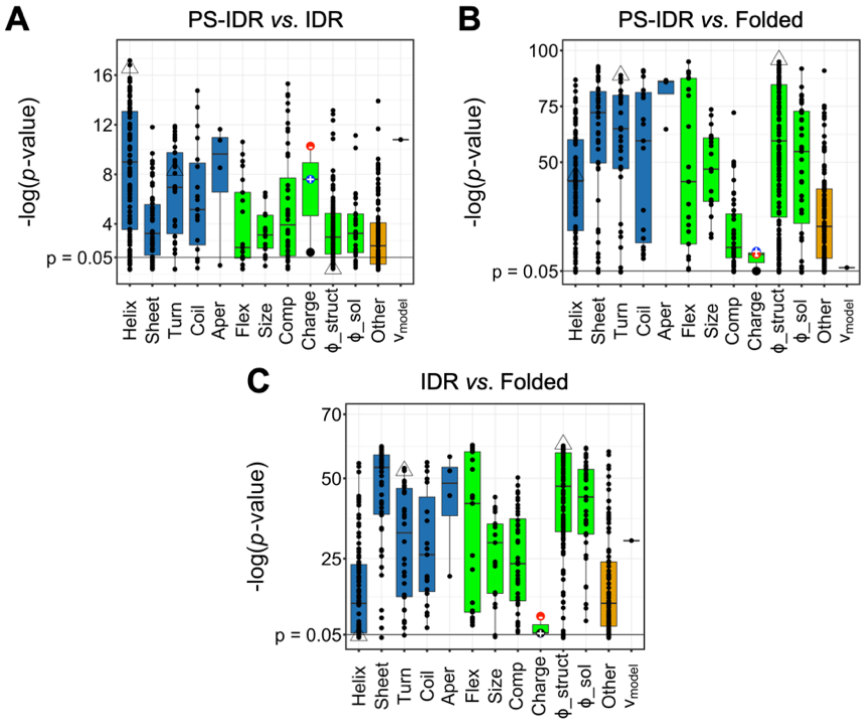


Figure 9. p-values calculated by Mann Whitney U-test indicate most amino acid scales can be used to separate the classes. A). PS-IDR *vs.* IDR, B). PS-IDR *vs.* Folded, and C). IDR *vs.* Folded. Significance level (p-value = 0.05) limit is represented with a solid line across all three box plots. Triangles represent the top three amino acid scales: Turn (by Levitt), Helix (by Scheraga), and Hydrophobicity structure (by Vendrusculo) by their p-values.[51,84,85] The three charged scales are labeled by red circle (negative charge inside circle), blue circle (positive charge inside circle circle), and solid black circle (neutral charge).

PCA indicates two principal modes of variation in the human proteome

To aid in the interpretation of the property differences among the three sequence sets, we applied PCA to the human proteome.[83] PCA applies a coordinate rotation on a data set such that the transformed axes become aligned with the directions of maximum variance.[83] To assess proteomic variance, we first selected three scales from each scale type, wherein the scales represent those with the three smallest p-values for a scale type. One-tail p-values calculated by comparing means

in the PS-IDR test and IDR null sets were used here (i.e., taken from Figure 8, panel A), because a primary goal is to understand the property differences between ID and PS-IDRs. Each scale, three per type, was then used to calculate sequence sums via a sliding 25-residue window applied to proteins in the human proteome with lengths from 50 to 500 residues (representing ~55,000 proteins). A sliding window scheme was used to capture differences within a sequence owing to the possibility of modular protein characteristics. Each sequence sum was divided by the window length (i.e., 25), and ν_{model} also was calculated for each 25-residue window.

The results of the PCA on this data set show that most of the variance (~70%) is captured by two principal modes of variation, and at somewhat equivalent amounts (38% and 30%; Figure 10, panels A and B). One mode of variation trends mostly, but not exclusively, with the variance arising from conformational propensity scales (panel B, blue arrows), and a second trends mostly, but not exclusively, with physicochemical metrics like charge, hydrophobicity, and other compositional details (panel B, green arrows). Thus, while many sequence-calculated properties can be used to discern phase-separating from nonphase-separating IDRs (Figure 8, panel A), or ID from folded (Figure 8, panels B and C), many of these properties exhibit similar variance patterns in the human proteome (Figure 10, panel B), meaning that they partition sequences similarly. As such, the predictive capabilities of amino acid scales are limited. Specifically, turn and coil scales applied to human sequences yield strongly correlated modes of variation that also are mostly anti-correlated with the variance produced from α-helix propensity scales (Figure 10, panel B). In contrast, the proteomic variance arising from hydrophobicity, charge, or ν_{model} have patterns

that, in general, are disparate to the variance arising from turn, coil, and α-helix conformational propensities.

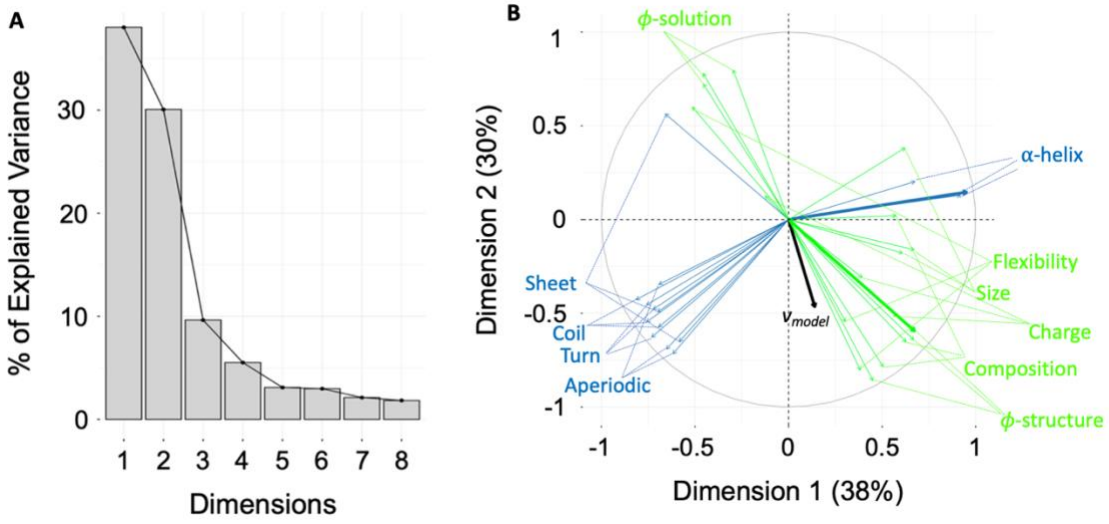


Figure 10. PCA indicates two primary modes of variation in the human proteome. A). Scree plot indicates most of the variation in the first two dimensions. There are two modes of variation when all the amino acid property scales were applied to the human proteome in that 70% of the variance in the human proteome is captured by two modes. **B).** PCA plot shows two primary modes: green as physiochemical and blue as conformational. Hydrophobicity separates IDR from folded and the two properties from the two modes are alpha-helix propensity (thickened blue arrow) and orthogonal (roughly 90 degrees) to that is v_{model}. The "human proteome" was defined as human proteome sequences between 100 and 500 amino acid residues, inclusive.

To illustrate these results, hydrophobicity was calculated for each sequence in the three sequence sets, test, null, and folded, using the scale from Vendruscolo mentioned above.[85] This property separates sequences according to folded versus ID (Figure 11). However, there are many sequence-based ID predictors already available that could have been used for this purpose instead.[43] Next, to separate ID sequences according to phase separation potential, the α-helix propensity of each sequence was calculated using the scale from Scheraga mentioned above, as well as a second property that exhibits a mode of variation orthogonal to that of the helix propensity

scale in the human proteome, which is satisfied by ν_{model} (Figure 10, panel B).[84] The test, null, and folded set means each are well separated by these three intrinsic sequence-based properties (Figure 11).

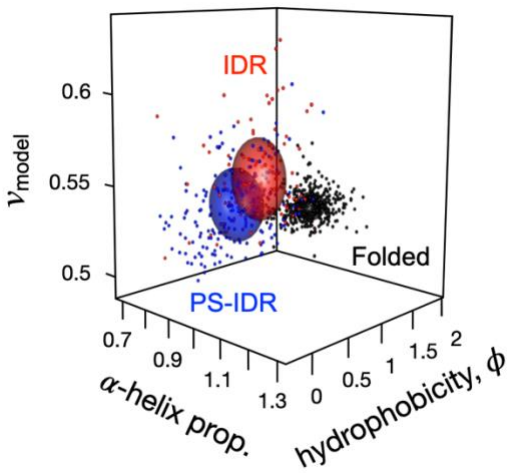


Figure 11. Three intrinsic sequence-based properties separate the protein classes: PS-IDR, IDR, and Folded. Vendruscolo hydrophobicity scale separates IDR (P or D) from Folded (F).[85] Alpha-helix propensity scale by Scheraga and ν_{model} separate PS-IDR (P) from IDR (D).[84]

Conclusions

We have analyzed protein databases containing subsets of proteins that are folded, ID, or ID and enriched for phase separation behavior. We found robust differences in the sequence-calculated properties of these three subsets, when compared. Out of 568 sequence-calculated properties, most yield statistically significant differences in the means of the sequence sets. We applied PCA to the human proteome and found that the myriad of intrinsic, sequence-calculated properties results in two principal modes of proteomic variance; one corresponding to physiochemical properties and the other conformation propensities.

V. PREDICTING FOLDED, ID, AND PS-ID PROTEIN REGIONS FROM SEQUENCE

In Chapter IV, we demonstrated that a structure-based hydrophobicity scale from Vendruscolo, an α-helix propensity scale from Scheraga, and ν_{model} (eq 1 in Chapter I) could be used in combination to distinguish folded, IDR null, and PS-IDR test sequence sets (see Figure 11).[84,85] The Vendruscolo scale gave the smallest *p*-value when comparing means in the folded and IDR sets (relative to the other sequence-calculated properties tested), and thus this scale efficiently distinguishes folded and ID regions from sequence.[85] Likewise, the Scheraga scale gave the smallest *p*-value when comparing means in the ID and PS-IDR sets.[84] The PCA showed that intrinsic sequence properties that best separate the means in the IDR null and PS-IDR test sets exhibit two principal modes of variance in the human proteome.[83] Thus, a second scale can be used with the Scheraga α-helix scale to distinguish ID and PS-IDRs from sequence.[84] We selected ν_{model} for this purpose because ν_{model} has a mode of variance in the human proteome that is unlike the variance arising from α-helix propensities (see Figure 10).

A second-generation version of the ParSe algorithm

Here, we test if these three sequence-calculated properties can be used to improve the ability of the ParSe algorithm to predict protein regions that drive LLPS. To do this, we modified the algorithm making a second-generation version, ParSe 2.0. In this version, as with the original, we apply a 25-residue window and then slide this window across a protein sequence in 1-residue steps (Figure 12).[43] For each 25-

residue window, the hydrophobicity (ϕ), α-helix propensity (α), and ν_{model} are calculated from the amino acid sequence of the window. Based upon the window-specific sequence-calculated values, the central residue of each window is labeled with a one-letter code, F, D, or P.

Figure 12. Predicting protein regions from the primary sequence using a sliding-window algorithm. For a given protein sequence (shown is a segment of the Sup35 primary sequence), the ParSe algorithm slides a 25-residue window across the whole sequence and calculates the hydrophobicity (ϕ), α-helix propensity (α), and ν_{model} of each window. The central residue of a window is assigned the label F for high ϕ, while low ϕ coupled with high α and high ν_{model} is assigned the label D, and low ϕ coupled with low α and low ν_{model} is assigned the label D. Protein regions that are predominantly labeled F are predicted to be folded. Those regions mostly labeled D or P are predicted to be ID or PS-ID, respectively.

The label F is assigned to windows with high ϕ values. This is shown schematically in Figure 13, panel A, plotting ϕ calculated for every 25-residue window in the phase-separating yeast protein Sup35 (UniProt accession ID P05453). In ParSe 2.0, windows are labeled F when ϕ ≥ folded set mean - 2σ; the mean ± σ for the folded set is shown by the black-filled ellipse in the figure for reference. Windows with ϕ < folded set mean - 2σ are assigned either P or D, depending on the values of α and ν_{model} calculated for the window. For the low-ϕ windows, those with high α and high ν_{model} are labeled D, while those with low α and low ν_{model} are labeled P (Figure 13, panel B).

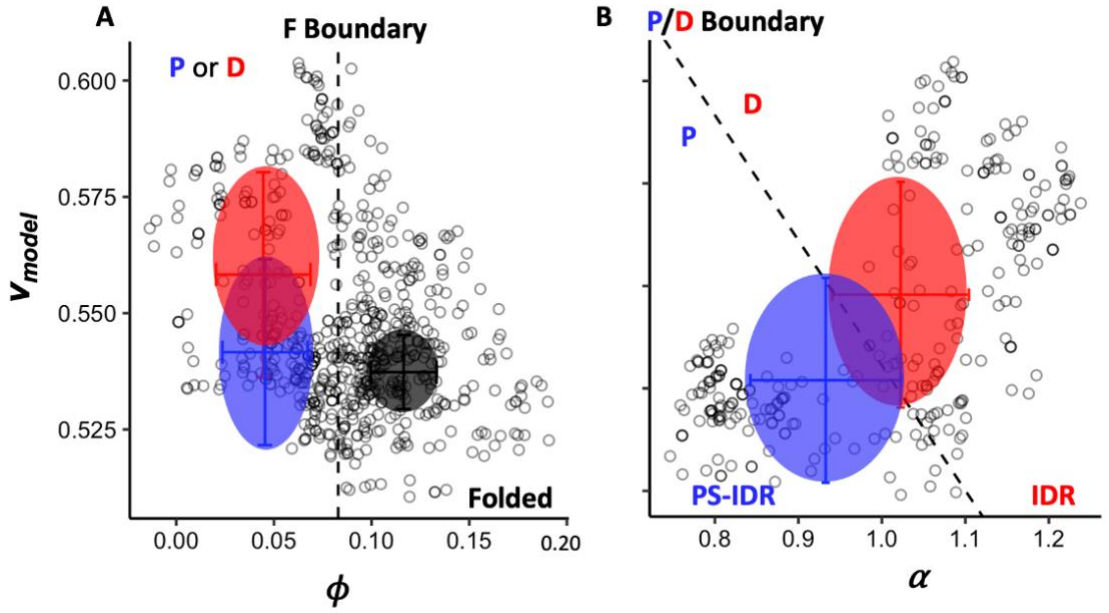


Figure 13. Sequence-calculated ϕ, α, and ν_{model} determined the window label. A) Small, open circles show ϕ and ν_{model} for each 25-residue window in the Sup35 sequence. Black-, red-, and blue-filled ellipses show the mean ± σ in the folded, IDR null, and PS-IDR test sequence sets, respectively. Windows with ϕ above the cutoff boundary (shown by the dashed line) are assigned the label F. **B)** Small, open circles show α and ν_{model} for each 25-residue window in the Sup35 sequence that was not labeled F (i.e., those with low ϕ). Windows with α and ν_{model} below the cutoff boundary (shown by the dashed line) are assigned the label P, while those with α and ν_{model} above the cutoff boundary are assigned the label D.

The P/D boundary is determined by the means and standard deviations in the PS-IDR test and IDR null sets for α-helix propensity and ν_{model} (shown in the figure by the blue- and red-filled ellipses). Note that the P/D boundary bisects the distribution overlap in the two sequence sets in α-helix propensity and ν_{model}. One point on this boundary line was determined by averaging the two points defined by (x_1, y_1) = (mean - σ in α for the IDR null set, mean ν_{model} for the IDR null set) and (x_2, y_2) = (mean α for the PS-IDR test set, mean + σ in ν_{model} for the PS-IDR test set). A second point on this boundary line was determined by averaging the two points defined by (x_1, y_1) = (mean α for the IDR null set, mean - σ in ν_{model} for the IDR null set) and (x_2, y_2) = (mean + σ in α for the PS-IDR test set, mean ν_{model} for the PS-IDR test set). The P/D boundary defined by these two

points is,

$$\nu_{model} = (-0.244)*\alpha + 0.789. \quad (9)$$

The means ± σ in ϕ, α, and ν_{model} for the three sequence sets, folded, IDR null, and PS-IDR test, are given in Table 3.

Table 3. Summary of mean values for the protein sequence sets.

Sequence Set	Hydrophobicity (ϕ)[a]	α-helix[a]	ν_{model} [a]
Folded	0.116 ± 0.016	1.031 ± 0.043	0.537 ± 0.007
IDR Null	0.044 ±0.024	1.027 ± 0.067	0.558 ± 0.022
PS-IDR Null	0.045 ± 0.021	0.960 ± 0.065	0.542 ± 0.020

[a] mean ± σ (standard deviation).

Identifying phase-separating regions in proteins using ParSe 2.0

Residues in regions within a protein sequence of length ≥ 20 that are at least 90% of only one of these labels F, D, or P are predicted by ParSe to be a folded, ID, and PS-IDR, respectively. This definition was not changed in ParSe 2.0, and Figure 14 shows its application to Sup35. This 685-residue protein is known to have three domains; the ID N-terminal prion domain (residues 1–124), the ID middle domain (residues 125–254), and the folded C-terminal catalytic domain (residues 255–685).[53,54] Of these domains, only the N-terminal prion domain mediates phase separation, which matches the ParSe 2.0 prediction.[53,54]

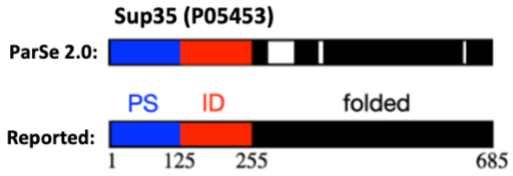


Figure 14. ParSe 2.0 prediction for the phase-separating yeast protein Sup35. In the top color bar, blue, red, and black regions are those regions within the protein sequence that are predicted to by PS-ID, ID, and folded respectively. White-colored regions are segments of the sequence that have mixed F, D, and/or P labels. The bottom color bar shows those regions that have been reported by experiment.[43]

To evaluate whether ParSe 2.0 can predict and identify regions of proteins that drive phase separation as well as or better than ParSe 1.0, we applied the algorithm to the same six model proteins that were analyzed by ParSe 1.0.[43] Figure 15 shows the results from applying ParSe 2.0 to the whole sequences of additional proteins with diverse reported mechanisms driving LLPS. As described in the introduction, LLPS of these proteins is driven via different mechanisms, such as protein-protein interactions driven by hydrophobic-hydrophobic, cation-π, charge-charge, and aromatic-aromatic contacts.[56-66] The proteins are identified by name and UniProt accession number in this figure. The domain-level structure of each is outlined below. Overall, ParSe 2.0 accurately predicted regions that drive LLPS in proteins with a variety of reported mechanisms.

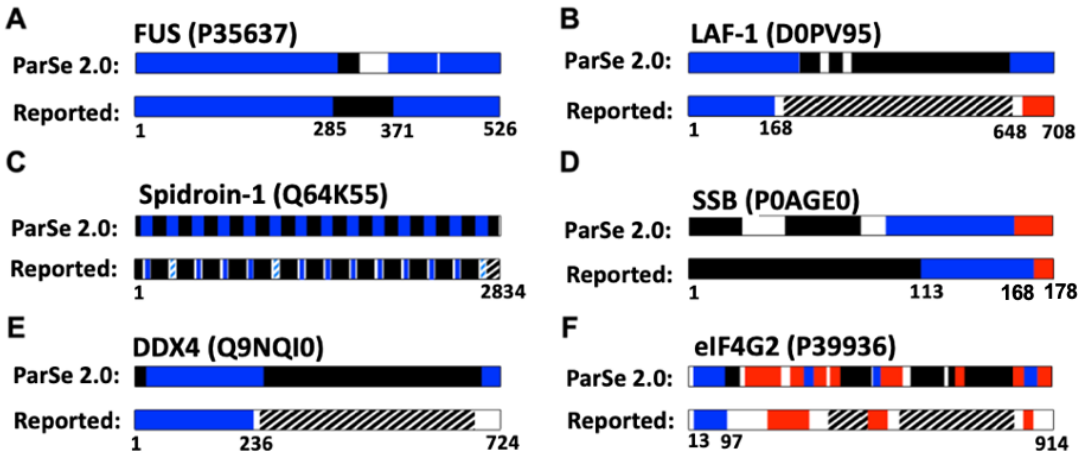


Figure 15. ParSe 2.0 predicted phase-separating regions in six proteins verified to exhibit LLPS behavior. A-F). Proteins are identified by name and UniProt accession number.[56-66] Blue regions are PS-IDR, red regions are IDR, and black regions are Folded. Striped regions represent 80% identify to a known sequences that phase separate (blue) or fold (black).[67]

Predicted PS regions by ParSe 2.0 are rare in the human proteome

We noticed that the proteins in Figures 14 and 15 had predicted phase-separating (PS) regions that tended to be long (≥50 residues). To determine if this feature is unique to proteins driving LLPS, we measured the prevalence of regions predicted from sequence to have high LLPS potential in the human proteome using ParSe 2.0 (Figure 5.5). These were identified as regions with at least 90% of residue positions labeled as P by the algorithm. We found that, like the result from using the original ParSe (~70% of the human proteome had a region at least one residue in length with predicted high LLPS potential (i.e., a single P-labeled position), while only ~5% have such a region that is at least 50 residues in length. This result shows that few human proteins possess a predicted PS region of substantial length (≥50 residues).

Next, we repeated this calculation for the set of 43 proteins assembled by *Vernon et al* that have been verified *in vitro* to exhibit phase separation behavior.[64] We find that

almost 90% of these "*in vitro* sufficient" LLPS proteins have a region predicted by ParSe 2.0 to have high LLPS potential that is 50 residues in length or longer. The DisProt database, minus the LLPS annotated IDPs, mirrored the human proteome result, demonstrating that ID alone is not sufficient to trigger LLPS prediction.[69,70] The set of proteins in SCOPe (Structural Classification of Proteins extended, version 2.07) that represent the globular fold classes across families and superfamilies, were mostly devoid of regions predicted to have high LLPS potential by ParSe 2.0[71,72] Thus, while proteins containing long, contiguous P-labeled regions are highly represented in proteins known to undergo LLPS, these regions appear relatively unique to this class of proteins.

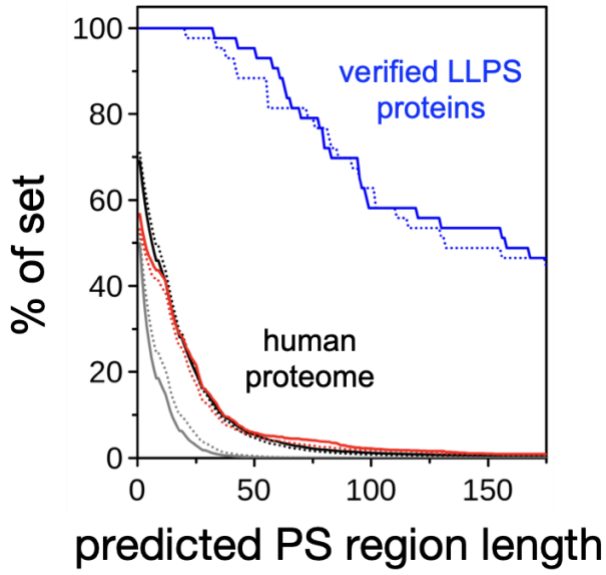


Figure 16. ParSe 2.0 predicted PS-IDRs are rare in the human proteome. The dashed lines are calculations from ParSe 2.0; the solid lines are from the original ParSe for comparison.[37] Blue is a set of confirmed LLPS proteins. Black line is the human proteome.[67] Red is the set of consensus IDP sequences in the DisProt database minus those annotated for LLPS.[69,70] Gray is the set of folded sequences obtained from the SCOPe database.[71,72]

Conclusions

We have evaluated three intrinsic sequence-based properties to predict protein regions that drive phase separation. We achieved this by modifying the ParSe algorithm, ParSe 2.0, where we applied the same 25-residue window calculation (same as ParSe 1.0) and calculated the amino acid sequence for each window hydrophobicity (ϕ), α-helix, and ν_{model}. This was applied to protein regions that drive LLPS, across different LLPS mechanisms. ParSe 2.0 generated similar predictions (compared to ParSe 1.0) of protein regions that drive LLPS.[43] Interestingly, the addition of the hydrophobicity (ϕ) as a property yielded a slightly more accurate prediction of the sequence boundaries between Folded, PS-ID, and ID regions. Additionally, ParSe 2.0 continued to identify long PS-IDRs that are rare in the human proteome. Furthermore, these findings support our hypothesis that other intrinsic properties of amino acids are associated with phase separation (beyond β-turns propensity and ν_{model}) that were applied in ParSe 1.0.[43] The many properties that can be used to predict phase-separating IDRs may reflect the variety of molecular mechanisms that drive LLPS.

VI. CONCLUSIONS

The primary goals of this research study were to 1) expand the previous IDR sequence set, 2) determine whether amino acid characteristics other than ν_{model} and β-turn propensity can define a phase-separating polypeptide, and 3) test whether these amino acid characteristics that are associated with phase separating proteins provide insight into the mechanisms underlying phase separation.

ParSe 1.0 distinguished PS-IDR from non-PS-IDR with the application of just two properties: β-turn propensity and ν_{model}. The ultimate goal of this study was to identify other characteristics that can enable us to separate these protein classes. Performing a pair-wise comparison between the protein sequence sets and computing a *p*-value of these comparisons helped us identify the probability of two sets to be separated or indistinguishable in that property. Furthermore, we have thoroughly characterized and identified that, surprisingly, most amino acid properties from the database could separate PS-IDRs, IDRs, and Folded regions beyond ν_{model} and beta-turn propensity.

To identify trends and/or shared characteristics within the multitude of properties that distinguish folded, ID, and PS-IDR sequences, we used principal component analysis (PCA) to analyze the variances in these properties in the human proteome. If a particular group of amino acid characteristic scales captured similar variances, then we could interpret those scales as separating the protein sequence sets in a similar way. PCA enabled us to identify two primary modes of variation: conformational and physiochemical. A structure-based (physiochemical) hydrophobicity structure scale best separated IDR from Folded sequence sets (Vendruscolo).[85] Additionally, an α -helix

(conformational) propensity scale and ν_{model} (also physiochemical) efficiently separated PS-IDR from IDR.[84]

Finally, we evaluated whether ParSe 2.0 was an improvement over ParSe 1.0 at predicting domain-level protein structure for phase-separating IDRs. When compared to ParSe 1.0, ParSe 2.0 produced similar predictions of protein regions that drive LLPS, across different LLPS mechanisms. This may be expected, as there was not a fundamental change in the methodology of prediction; ParSe 2.0 employed both physiochemical (ν_{model} and hydrophobicity) and conformational (alpha-helix propensity) characteristics of polypeptides. Consistent with these results, ParSe 2.0 continues to identify long regions of predicted PS-IDR as rare within the human proteome (~5% of the human proteome contains PS-IDR regions of at least 50 residues). In contrast, 90% of the experimentally verified proteins that drive phase separation *in vitro* have PS-IDR regions of at least 50 residues.

These results support the idea that additional intrinsic properties of amino acids and polypeptide sequences can be associated with LLPS, beyond the β-turn propensity and ν_{model} that were used in ParSe 1.0. Additionally, those properties found to be associated with phase-separating proteins may give insight into the physical mechanisms underlying LLPS. For example, the polymer scaling exponent describes the balance of self-interaction and solvent interaction of a polypeptide. Additionally, ParSe 2.0 now incorporates two orthogonal classes of amino acid properties that were identified by PCA: a physiochemical dimension (i.e., hydrophobicity, charge, flexibility) and a conformational dimension (i.e., measures of secondary structure propensity, such as beta-turn propensity or alpha-helix propensity).

IDPs are frequently involved in promoting protein phase separation. However, not all IDPs, nor all IDRs within an IDP, necessarily contribute to this behavior. Many IDRs that phase separate consist of low-complexity sequences. Electrostatic interactions, pi-pi contacts, hydrophobic-driven burial are examples of various multivalent interactions of IDRs that promote phase separation. These interactions can be related to two of the sequence-based properties of ParSe 2.0: ν_{model} and hydrophobicity. As previously mentioned IDPs are highly predictable from sequence, in part, because many of them contain low-complexity sequences. The contribution of hydrophobic, electrostatic, and pi-pi contacts have significant impact on folded protein structures, conformational distributions of IDPs, and phase separation properties.

Mutations in IDRs have been implicated in a wide variety of disease states, including neurodegeneration and carcinogenesis. Because IDRs lack a fixed three-dimensional structure, it is difficult to study how mutations in IDRs lead to disease states. While not every mutation in an IDR will affect its propensity for driving phase separation, ParSe 2.0 is a tool that can now be tested for its ability to predict how a particular mutation may alter the phase-separation propensity of an IDR.

This predictive power can be used not only to characterize disease-associated mutations, but also to systematically probe how IDR sequence can give rise to phase-separation behavior, as well as to investigate what sequence changes in PS-IDRs are sufficient to disrupt LLPS. Furthermore, the presence or absence of these interactions can enable us to understand the biological consequences of changing the phase-separation propensity of IDRs This will ultimately enable us to better understand both the underlying biophysics and disease-associated properties of IDRs.

APPENDIX SECTION

Table 4. Expanded IDR sequence set (IDPs that are not known to exhibit phase separation).

Name	Database [a]	Entry number	UniProt accession number	ID region (*N*)
pknG	BMRB	26027	P9WI73	1-75 (75)
HCK	BMRB	27554	P08631	2-79
SIC1	BMRB	16657	P38634	1-90 (90)
SLC9A1	BMRB	26557	P19634	680-815 (136)
ERD14	BMRB	16876	P42763	1-185 (185)
Spp1	DisProt	DP01448	P10923	17-294 (278)
PAGE4	DisProt	DP01435	O60829	1-102 (102)
MAP2K4	DisProt	DP01400	P45985	1-86 (86)
Sufu	DisProt	DP01397	Q9Z0P7	279-359 (81)
HCN1	DisProt	DP01317	O60741	1-93 (93)
SUFU	DisProt	DP01312	Q9UMX1	279-360 (82)
PQBP1	DisProt	DP01308	O60828	82-265 (184)
HIRD11	DisProt	DP01300	Q9SLJ2	1-98 (98)
LEA18	DisProt	DP01299	Q96273	1-97 (97)
PSEN1	DisProt	DP01292	P49768	1-77 (77)
Prothymosin a14	DisProt	DP01228	Q9UMZ1	1-101 (101)
Ppp1r10	DisProt	DP01202	O55000	309-433 (125)
NOLC1	DisProt	DP01178	Q14978	1-699 (699)
Gja4	DisProt	DP01175	P28235	233-333 (101)
DCLRE1C	DisProt	DP01162	Q96SD1	480-575 (96)
ptkA	DisProt	DP01160	P9WPI9	1-81 (81)
H1-0	DisProt	DP01156	P07305	105-194 (90)
Ttn-1	DisProt	DP01090	A0A2I2LG13	2793-6678 (3886)
PM28	DisProt	DP01088	Q9XES8	1-89 (89)

YRB2	DisProt	DP01079	P40517	1-203 (203)
Ahn-1	DisProt	DP01074	Q7YUB9	1-86 (86)
MSA2	DisProt	DP01067	P19599	21-238 (218)
LMP2A	DisProt	DP01060	A8CDV5	1-118 (118)
Omega gliadin storage protein	DisProt	DP01040	Q9FUW7	1-280 (280)
SLE2	DisProt	DP01036	I1JLC8	1-105 (105)
pscP	DisProt	DP00993	Q9I332	1-253 (253)
Small delta antigen	DisProt	DP00965	P0C6L3	60-195 (136)
SBDS-like protein	DisProt	DP00957	C0J347	264-464 (201)
GAP43	DisProt	DP00955	P06836	1-242 (242)
N	DisProt	DP00948	P59595	182-259 (78)
Ppp1r9b	DisProt	DP00943	O35274	1-154 (154)
BASP1	DisProt	DP00930	P80723	1-227 (227)
NABP2	DisProt	DP00864	Q9BQ15	110-211 (102)
trm10	DisProt	DP00798	O14214	1-83 (83)
CNGB1	DisProt	DP00768	Q28181-4	14-99 (86) 272-590 (319)
Smtnl1	DisProt	DP00742	Q99LM3	1-341 (341)
dre4	DisProt	DP00721	Q8IRG6	889-1044 (156)
Ssrp	DisProt	DP00720	Q05344	437-554 (118) 625-723 (99)
N	DisProt	DP00698	O89339	400-532 (133)
RYBP	DisProt	DP00694	Q8N488	1-228 (228)
L1CAM	DisProt	DP00666	P32004	1144-1257 (114)
GMPM1	DisProt	DP00664	Q01417	1-173 (173)

ALB3	DisProt	DP00662	Q8LBP4	339-462 (124)
MAC-41A	DisProt	DP00659	P16458	233-385 (153)
COR47	DisProt	DP00657	P31168	1-265 (265)
N	DisProt	DP00640	Q89933	400-525 (126)
ERD10	DisProt	DP00606	P42759	1-260 (260)
Genome polyprotein	DisProt	DP00588	P27958	1-82 (82)
stm	DisProt	DP00584	A2VD23	1-613 (613)
SEPTIN4	DisProt	DP00537	O43236	1-119 (119)
DHN1	DisProt	DP00530	P12950	1-168 (168)
MYOM1	DisProt	DP00517	P52179	836-931 (96)
NUPR1	DisProt	DP00510	O60356	1-82 (82)
UBA2	DisProt	DP00486	Q9UBT2	551-640 (90)
HY5	DisProt	DP00469	O24646	1-77 (77)
cna	DisProt	DP00461	P08083	1-90 (90)
Chm	DisProt	DP00458	P37727	108-208 (101)
PPP1R1B	DisProt	DP00421	P07516	1-202 (202)
JAG1	DisProt	DP00418	P78504	1094-1218 (125)
URE1	DisProt	DP00353	P23202	1-90 (90)
DNAJC6	DisProt	DP00351	Q27974	547-813 (267)
col	DisProt	DP00342	P09883	1-83 (83)
Trl	DisProt	DP00328	Q08605	368-444 (77)
PPP1R1A	DisProt	DP00325	P01099	1-166 (166)
ADD2	DisProt	DP00241	P35612	409-726 (318)
ADD1	DisProt	DP00240	P35611	430-737 (308)
SSB	DisProt	DP00229	P05455	326-408 (83)

Nucleoplasmin	DisProt	DP00217	P05221	120-200 (81)
CAST	DisProt	DP00196	P20810	137-277 (141)
HMGN2	DisProt	DP00195	P02313	1-89 (89)
Late embryogenesis abundant protein 1	DisProt	DP00186	Q95V77	1-143 (143)
CTDP1	DisProt	DP00177	Q9Y5B0	879-961 (83)
TCF7L2	DisProt	DP00175	Q9NQB0	1-130 (130)
zipA	DisProt	DP00161	P77173	86-185 (100)
RAD23A	DisProt	DP00156	P54725	79-160 (82)
NEFL	DisProt	DP00151	P02547	444-549 (106)
Slbp	DisProt	DP00144	Q9VAN6	97-175 (79)
PTHLH	DisProt	DP00138	P12272	68-144 (77)
H1-4	DisProt	DP00136	P15865	1-217 (217)
PRB4	DisProt	DP00119	P10163	17-310 (294)
H1-0	DisProt	DP00097	P10922	96-193 (98)
TOP2	DisProt	DP00076	P06786	1178-1428 (251)
TOP1	DisProt	DP00075	P11387	1-214 (214)
Structural polyprotein	DisProt	DP03350	P03316	1-113 (113)
RPA1	DisProt	DP00061	P27694	105-180 (76)
HMGA1	DisProt	DP00040	P17096	1-107 (107)
HMGN2	DisProt	DP00039	P05204	1-90 (90)
RAP1	DisProt	DP00020	P11938	1-123 (123)

[a] The Biological Magnetic Resonance Data Bank (BMRB) and DisProt were used to identify IDPs not known to exhibit phase separation behavior.[69,70,73] This list of verified IDPs, where duplicates were removed and was combined with a list of 23 verified IDPs that are identified and published elsewhere.[43]

Table 5. Normalized frequency for β-turn.

Amino Acid	Scale value [a]
Alanine	0.770
Arginine	0.880
Asparagine	1.280
Aspartic Acid	1.410
Cysteine	0.810
Glutamine	0.980
Glutamic Acid	0.990
Glycine	1.640
Histidine	0.680
Isoleucine	0.510
Leucine	0.580
Lysine	0.960
Methionine	0.410
Phenylalanine	0.590
Proline	1.910
Serine	1.320
Threonine	1.040
Tryptophan	0.760
Tyrosine	1.050
Valine	0.470

[a] From Levitt.[51]

Table 6. Intrinsic PPII bias measured in peptides.

Amino Acid	Scale value [a]
Alanine	0.37
Arginine	0.38
Asparagine	0.27
Aspartic Acid	0.30
Cysteine	0.25
Glutamine	0.53
Glutamic Acid	0.42
Glycine	0.13
Histidine	0.20
Isoleucine	0.39
Leucine	0.24
Lysine	0.56
Methionine	0.36
Phenylalanine	0.17
Proline	1.00
Serine	0.24
Threonine	0.32
Tryptophan	0.25
Tyrosine	0.25
Valine	0.39

[a] From Hilser Group.[87]

Table 7. Hydrophobicity scale used for optimized simulation methods for LLPS.

Amino Acid	Scale value [a]
Alanine	0.51507
Arginine	0.24025
Asparagine	0.78447
Aspartic Acid	0.30525
Cysteine	0.46169
Glutamine	0.29516
Glutamic Acid	0.342621
Glycine	01.24153
Histidine	0.55537
Isoleucine	0.83907
Leucine	0.51207
Lysine	0.47106
Methionine	0.64648
Phenylalanine	1.17854
Proline	0.34128
Serine	0.11195
Threonine	0.27538
Tryptophan	0.97588
Tyrosine	1.04266
Valine	0.55645

[a] From Robert Best.[80]

Table 8. Normalized frequency of alpha-helix.

Amino Acid	Scale value [a]
Alanine	1.29
Arginine	0.83
Asparagine	0.77
Aspartic Acid	1.00
Cysteine	0.94
Glutamine	1.10
Glutamic Acid	1.54
Glycine	0.72
Histidine	1.29
Isoleucine	0.94
Leucine	1.23
Lysine	1.23
Methionine	1.23
Phenylalanine	1.23
Proline	0.70
Serine	0.78
Threonine	0.87
Tryptophan	1.06
Tyrosine	0.63
Valine	0.97

[a] From Scheraga.[84]

Table 9. Structure-based Hydrophobicity scale.

Amino Acid	Scale value [a]
Alanine	0.0728
Arginine	0.0394
Asparagine	0.0390
Aspartic Acid	0.0552
Cysteine	0.3557
Glutamine	0.0126
Glutamic Acid	0.0295
Glycine	0.0589
Histidine	0.0874
Isoleucine	0.3805
Leucine	1.23
Lysine	0.0053
Methionine	0.1613
Phenylalanine	0.4201
Proline	0.0492
Serine	0.0282
Threonine	0.0239
Tryptophan	0.4114
Tyrosine	0.3113
Valine	0.2947

[a] From Vendruscolo and coworkers.[85]

REFERENCES

1. Protter, D., & Parker, R. **(2016)**. Principles and Properties of Stress Granules. *Trends in cell biology*, *26*(9), 668–679.

2. Wolf, N., Priess, J., & Hirsh, D. **(1983)**. Segregation of germline granules in early embryos of Caenorhabditis elegans: an electron microscopic analysis. *Journal of embryology and experimental morphology*, *73*, 297–306.

3. Hyman, A. A., Weber, C. A., & Jülicher, F. **(2014)**. Liquid-Liquid Phase Separation in Biology. *Annual Review of Cell and Developmental Biology*, *30*(1), 39-58.

4. Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., & Hyman, A. A. **(2009)**. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science (New York, N.Y.)*, *324*(5935), 1729–1732.

5. Fung, Ho Yee Joyce et al. "IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies." *Current opinion in structural biology* vol. 49 **(2018)**: 36-43.

6. Mitrea, D.M., Kriwacki, R.W. Phase separation in biology; functional organization of a higher order. *Cell Commun Signal* **14,** 1 **(2016)**.

7. Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., & Taylor, J. P. **(2015)**. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, *163*(1), 123–133.

8. Patel, S. S., Belmont, B. J., Sante, J. M., & Rexach, M. F. **(2007).** Natively unfolded nucleoporins gate protein diffusion across the nuclear pore complex. *Cell*, *129*(1), 83–96.

9. Ribbeck, K., & Görlich, D. **(2002).** The permeability barrier of nuclear pore complexes appears to operate via hydrophobic exclusion. *The EMBO journal*, *21*(11), 2664–2671.

10. Ismail, H., Liu, X., Yang, F., Li, J., Zahid, A., Dou, Z., Liu, X., & Yao, X. **(2021)**. Mechanisms and regulation underlying membraneless organelle plasticity control. *Journal of molecular cell biology*, *13*(4), 239–258.

11. Saito, M., Hess, D., Eglinger, J., Fritsch, A. W., Kreysing, M., Weinert, B. T., Choudhary, C., & Matthias, P. **(2019)**. Acetylation of intrinsically disordered regions regulates phase separation. *Nature chemical biology*, *15*(1), 51–61.

12. Li P., Banjade S., Cheng H.-C., et al. **(2012)**. Phase transitions in the assembly of multivalent signalling proteins. *Nature* 483, 336–340.

13. Su, X., Ditlev, J. A., Hui, E., Xing, W., Banjade, S., Okrut, J., King, D. S., Taunton, J., Rosen, M. K., & Vale, R. D. **(2016)**. Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science (New York, N.Y.), 352*(6285), 595–599.

14. Patel J., Pathak R.R., Mujtaba S. **(2011)**. The biology of lysine acetylation integrates transcriptional programming and metabolism. *Nutr. Metab.* 8, 12.

15. Ferreon J.C., Jain A., Choi K.-J., et al. **(2018)**. Acetylation disfavors tau phase separation. *Int. J. Mol. Sci.* 19, 1360.

16. Kamah A., Huvent I., Cantrelle F.X., Qi H., Lippens G., Landrieu I., Smet-Nocca C. Nuclear magnetic resonance analysis of the acetylation pattern of the neuronal tau protein. *Biochemistry.*

17. Cohen T.J., Guo J.L., Hurtado D.E., Kwong L.K., Mills I.P., Trojanowski J.Q., Lee V.M. The acetylation of tau inhibits its function and promotes pathological tau aggregation. *Nat. Commun.* 2011;2:252.

18. Min S.W., Cho S.H., Zhou Y., Schroeder S., Haroutunian V., Seeley W.W., Huang E.J., Shen Y., Masliah E., Mukherjee C., et al. **(2010)** Acetylation of tau inhibits its degradation and contributes to tauopathy. *Neuron*;67:953–966.

19. Cook C., Carlomagno Y., Gendron T.F., Dunmore J., Scheffel K., Stetler C., Davis M., Dickson D., Jarpe M., DeTure M., et al. **(2014).** Acetylation of the kxgs motifs in tau is a critical determinant in modulation of tau aggregation and clearance. *Hum. Mol. Genet* ;23:104–116.

20. Evich M., Stroeva E., Zheng Y.G., et al. **(2016)**. Effect of methylation on the side-chain pKa value of arginine. *Protein Sci.* 25, 479–486.

21. Mitrea, D. M., & Kriwacki, R. W. **(2016)**. Phase separation in biology; functional organization of a higher order. *Cell communication and signaling : CCS*, *14*, 1.

22. Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K., & Zaslavsky, B. **(2015)**. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS letters*, *589*(1), 15–22.

23. Chong, P. A., & Forman-Kay, J. D. **(2016)**. Liquid-liquid phase separation in cellular signaling systems. *Current opinion in structural biology*, *41*, 180–186.

24. Oldfield, C. J., & Dunker, A. K. **(2014).** Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry, 83*(1).

25. Dunker A. K.; Brown C. J.; Lawson J. D.; Iakoucheva L. M.; Obradovic Z. **(2002)**. Intrinsic Disorder and Protein Function.*Biochemistry*, 41, 6573.

26. Kozlowski, L. P., & Bujnicki, J. M. **(2012)**. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC bioinformatics*, *13*(1), 1-11.

27. Emenecker, R. J., Griffith, D., & Holehouse, A. S. (2021). Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical Journal*, *120*(20), 4312-4319.

28. Erdős, G., & Dosztányi, Z. (2020). Analyzing protein disorder with IUPred2A. *Current Protocols in Bioinformatics*, *70*(1), e99.

29. Oates M.E., Romero P., Ishida T., Ghalwash M., Mizianty M.J., Xue B. et al. **(2013)**. D^2P^2: database of disordered protein predictions.

30. Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F. and Jones D.T. **(2004)**. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*

31. Potenza E., Domenico T.D., Walsh I. and Tosatto S.C.E. **(2015)**. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*

32. Macossay-Castillo, M., Marvelli, G., Guharoy, M., Jain, A., Kihara, D., Tompa, P., & Wodak, S. J. **(2019)**. The Balancing Act of Intrinsically Disordered Proteins: Enabling Functional Diversity while Minimizing Promiscuity. *Journal of molecular biology*, *431*(8), 1650–1670.

33. Fung, H., Birol, M., & Rhoades, E. **(2018)**. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Current opinion in structural biology*, *49*, 36–43.

34. Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C. C., Eckmann, C. R., Myong, S., & Brangwynne, C. P. **(2015)**. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7189–7194.

35. Kanaan, N.M., Hamel, C., Grabinski, T. et al. Liquid-liquid phase separation induces pathogenic tau conformations in vitro. *Nat Commun* **11,** 2809 **(2020)**.

36. Vernon, R. M., & Forman-Kay, J. D. **(2019)**. First-generation predictors of biological protein phase separation. *Current opinion in structural biology*, *58*, 88–96.

37. Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., & Forman-Kay, J. D. **(2018)**. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife*, *7*, e31486.

38. Klus, P., Bolognesi, B., Agostini, F., Marchese, D., Zanzoni, A., & Tartaglia, G. G. **(2014)**. The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics (Oxford, England)*, *30*(11), 1601–1608.

39. Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C., & Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, *115*(40), 9929-9934.

40. Lin, Y. H., & Chan, H. S. (2017). Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophysical Journal*, *112*(10), 2043-2046.

41. Lin, Y. H., Brady, J. P., Chan, H. S., & Ghosh, K. (2020). A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes. *The Journal of chemical physics*, *152*(4), 045102.

42. Zeng, X., Holehouse, A. S., Chilkoti, A., Mittag, T., & Pappu, R. V. (2020). Connecting Coil-to-Globule Transitions to Full Phase Diagrams for Intrinsically Disordered Proteins. *Biophysical journal*, *119*(2), 402–418.

43. Paiz EA, Allen JH, Correia JJ, Fitzkee NC, Hough LE, Whitten ST. **(2021)** Nov. Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. J Biol Chem. 297(5):101343.

44. Flory, P. J. (1949). The configuration of real polymer chains. *The Journal of Chemical Physics*, *17*(3), 303-310.

45. Flory, P. J., & Volkenstein, M. (1969). Statistical mechanics of chain molecules.

46. Hofmann H, Soranno A, Borgia A, Gast K, Nettels D, Schuler B. **(2012)**. Oct. 2. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. Proc Natl Acad Sci U S A.09(40):16155-60.

47. Tomasso, M. E., Tarver, M. J., Devarajan, D., & Whitten, S. T. (2016). Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS computational biology*, *12*(1), e1004686.

48. Perez, R. B., Tischer, A., Auton, M., & Whitten, S. T. **(2014)**. Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. *Proteins*, *82*(12), 3373–3384.

49. English, L. R., Tilton, E. C., Ricard, B. J., & Whitten, S. T. **(2017)**. Intrinsic α helix propensities compact hydrodynamic radii in intrinsically disordered proteins. *Proteins*, *85*(2), 296–311.

50. English, L. R., Voss, S. M., Tilton, E. C., Paiz, E. A., So, S., Parra, G. L., & Whitten, S. T. **(2019)**. Impact of Heat on Coil Hydrodynamic Size Yields the Energetics of Denatured State Conformational Bias. *The journal of physical chemistry. B, 123*(47), 10014–10024.

51. Levitt M. **(1978)**. Conformational preferences of amino acids in globular proteins. *Biochemistry*, *17*(20), 4277–4285.

52. Chou, P. Y., and Fasman, G. D. **(1978)** Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45–148 39.

53. Hutchinson, E. G., and Thornton, J. M. **(1994)** A revised set of potentials for beta-turn formation in proteins. Protein Sci. 3, 2207–2216.

54. Franzmann, T. M., Jahnel, M., Pozniakovsky, A., Mahamid, J., Holehouse, A. S., Nüske, E., Richter, D., Baumeister, W., Grill, S. W., Pappu, R. V., Hyman, A. A., & Alberti, S. **(2018)**. Phase separation of a yeast prion protein promotes cellular fitness. *Science (New York, N.Y.)*, *359*(6371), eaao5654.

55. Preis, A., Heuer, A., Barrio-Garcia, C., Hauser, A., Eyler, D. E., Berninghausen, O., Green, R., Becker, T., & Beckmann, R. **(2014)**. Cryoelectron microscopic structures of eukaryotic translation termination complexes containing eRF1-eRF3 or eRF1-ABCE1. *Cell reports*, *8*(1), 59–65.

56. Wang, J., Choi, J. M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S., & Hyman, A. A. **(2018)**. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, *174*(3), 688–699.e16.

57. Loughlin, F. E., Lukavsky, P. J., Kazeeva, T., Reber, S., Hock, E. M., Colombo, M., Von Schroetter, C., Pauli, P., Cléry, A., Mühlemann, O., Polymenidou, M., Ruepp, M. D., & Allain, F. H. **(2019)**. The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Molecular cell*, *73*(3), 490–504.e6.

58. Kim, Y., & Myong, S. **(2016)**. RNA Remodeling Activity of DEAD Box Proteins Tuned by Protein Concentration, RNA Length, and ATP. *Molecular cell*, *63*(5), 865–876.

59. Tremblay, M. L., Xu, L., Lefèvre, T., Sarker, M., Orrell, K. E., Leclerc, J., Meng, Q., Pézolet, M., Auger, M., Liu, X. Q., & Rainey, J. K. **(2015)**. Spider wrapping silk fibre architecture arising from its modular soluble protein precursor. *Scientific reports*, *5*, 11502.

60. Muiznieks, L. D., & Keeley, F. W. **(2016)**. Phase separation and mechanical properties of an elastomeric biomaterial from spider wrapping silk and elastin block copolymers. *Biopolymers*, *105*(10), 693–703.

61. Harami, G. M., Kovács, Z. J., Pancsa, R., Pálinkás, J., Baráth, V., Tárnok, K., Málnási-Csizmadia, A., & Kovács, M. **(2020)**. Phase separation by ssDNA binding protein controlled via protein-protein and protein-DNA interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 26206–26217.

62. Matsumoto, T., Morimoto, Y., Shibata, N., Kinebuchi, T., Shimamoto, N., Tsukihara, T., & Yasuoka, N. **(2000)**. Roles of functional loops and the C-terminal segment of a single-stranded DNA binding protein elucidated by X-Ray structure analysis. *Journal of biochemistry*, *127*(2), 329–335.

63. Brady, J. P., Farber, P. J., Sekhar, A., Lin, Y. H., Huang, R., Bah, A., ... & Kay, L. E. **(2017)**. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proceedings of the National Academy of Sciences*, *114*(39), E8194-E8203.

64. Das, S., Lin, Y. H., Vernon, R. M., Forman-Kay, J. D., & Chan, H. S. **(2020).** Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, *117*(46), 28795-28805.

65. Lin, Y., Protter, D. S., Rosen, M. K., & Parker, R. **(2015).** Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Molecular cell*, *60*(2), 208-219.

66. Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. **(2017).** The SWISS-MODEL Repository-new features and functionality. *Nucleic acids research*, *45*(D1), D313–D319.

67. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:D158–69. 10.1093/nar/gkw1099

68. Mészáros, B., Erdős, G., Szabó, B., Schád, É., Tantos, Á., Abukhairan, R., Horváth, T., Murvai, N., Kovács, O. P., Kovács, M., Tosatto, S., Tompa, P., Dosztányi, Z., & Pancsa, R. (2020). PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic acids research*, *48*(D1), D360–D367.

69. Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., & Dunker, A. K. **(2007)**. DisProt: the Database of Disordered Proteins. *Nucleic acids research*, *35*(Database issue), D786–D793.

70. Hatos, A., et al. **(2020)** DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276.

71. Fox N.K., Brenner S.E., Chandonia J.-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*

72. Chandonia J.-M., Fox N.K., Brenner S.E. SCOPe: Classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.*

73. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., & Markley, J. L. **(2008)**. BioMagResBank. *Nucleic acids research, 36*(Database issue), D402–D408.

74. Khaodeuanepheng, N. Polymer properties that predict protein structure class from the primary sequence (Master's Thesis). Dep. Chem. Biochem. Grad. Coll. Sci. Eng. Texas State Univ.**2022**.

75. Nakai, K., Kidera, A., and Kanehisa, M. **(1988)**. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **2**, 93-100.

76. Tomii, K. and Kanehisa, M. **(1996)**. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9, 27-36.

77. Kawashima, S., Ogata, H., and Kanehisa, M. **(1999)**. AAindex: amino acid index database. *Nucleic Acids Res.* **27**, 368-369.

78. Kawashima, S. and Kanehisa, M. **(2000)**. AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374.

79. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. **(2008)**. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202-D205.

80. Dannenhoffer-Lafage, T., & Best, R. B. **(2021)**. A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins. *The Journal of Physical Chemistry B, 125*(16), 4046-4056.

81. Welch, B. L. **(1947)**. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika, 34*(1/2), 28–35.

82. H. B. Mann. D. R. Whitney. March, **(1947)**."On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." Ann. Math. Statist. 18 (1) 50 – 60.

83. Jolliffe, I. T., & Cadima, J. **(2016)**. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *374*(2065), 20150202.

84. Burgess, A.W., Ponnuswamy, P.K. and Scheraga, H.A. **(1974)**, Analysis of Conformations of Amino Acid Residues and Prediction of Backbone Topography in Proteins. Isr. J. Chem., 12: 239-286.

85. Bastolla, U., Porto, M., Roman, H.E. and Vendruscolo, M. **(2005)**, Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins, 58: 22-30.

86. Priyadarshi, H., Das, R., Kumar, S., Kishore, P., & Kumar, S. **(2017)**. Analysis of variance, normal quantile-quantile correlation and effective expression support of pooled expression ratio of reference genes for defining expression stability. *Heliyon*, *3*(1), e00233.

87. Elam, W.A.; Schrank, T.P.; Campagnolo, A.J.; Hilser, V.J. **(2013)**. Evolutionary Conservation of the Polyproline II Conformation Surrounding Intrinsically Disordered Phosphorylation Sites. *Protein Sci.22*, 405–417.