International Journal of
Molecular Sciences

MDPI

Article

BIAPSS: A Comprehensive Physicochemical Analyzer of Proteins Undergoing Liquid–Liquid Phase Separation

Aleksandra E. Badaczewska-Dawid [1], Vladimir N. Uversky [2,*] and Davit A. Potoyan [1,3,*]

[1] Department of Chemistry, Iowa State University, Ames, IA 50011, USA; abadacz@iastate.edu
[2] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[3] Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA
[*] Correspondence: vuversky@usf.edu (V.N.U.); potoyan@iastate.edu (D.A.P.)

Abstract: The liquid–liquid phase separation (LLPS) of biomolecules is a phenomenon which is nowadays recognized as the driving force for the biogenesis of numerous functional membraneless organelles and cellular bodies. The interplay between the protein primary sequence and phase separation remains poorly understood, despite intensive research. To uncover the sequence-encoded signals of protein capable of undergoing LLPS, we developed a novel web platform named BIAPSS (Bioinformatics Analysis of LLPS Sequences). This web server provides on-the-fly analysis, visualization, and interpretation of the physicochemical and structural features for the superset of curated LLPS proteins.

Keywords: liquid–liquid phase separation; membraneless organelles; intrinsically disordered proteins; proteins with low complexity

check for updates

Citation: Badaczewska-Dawid, A.E.; Uversky, V.N.; Potoyan, D.A. BIAPSS: A Comprehensive Physicochemical Analyzer of Proteins Undergoing Liquid–Liquid Phase Separation. *Int. J. Mol. Sci.* **2022**, *23*, 6204. https://doi.org/10.3390/ijms23116204

Academic Editor: Vitaly V. Kushnirov

Received: 1 May 2022
Accepted: 27 May 2022
Published: 31 May 2022

1. Introduction

The spatiotemporal organization of biomolecules and biomolecular interactions is essential for the efficient regulation of cellular biochemistry. The underlying biophysical mechanism for membraneless compartmentalization is liquid–liquid phase separation (LLPS). In the past few years, the LLPS of biomolecules has become a unifying physical mechanism for understanding the principles of intracellular compartmentalization, the formation of membraneless organelles (MLOs), and gene regulation [1–14]. In the LLPS process, the relatively well-mixed solution of biomolecules separates into liquid droplets. The ability of proteins to phase separate appears to be encoded primarily in the peculiarities of their primary sequences, which often contain low-complexity regions and intrinsically disordered regions (IDRs) that are enriched in charged and multivalent interaction centers [6–8,10,11,13–19]. While some general sequence trends have emerged, the quantitative aspects of how amino acid sequences encode and decode phase separation still remain largely unknown [20–22]. This is because many different combinations of relevant interactions seem to be contributing to phase separation, without any one being universally necessary [23]. As a consequence (with a few exceptions [24–30]), mostly case-by-case studies of different sequences are performed, with the broader context of many findings, including their statistical significance, remaining unknown.

Following the statistical trends in PubMed, biological LLPS has been gaining widespread attention in the last two decades. The rapidly growing amount of data from both in vitro and in vivo experiments have systematically narrowed the range of the LLPS-promoting conditions [31]. From these studies, we know that the regulatory mechanisms of phase separation appear strongly context-dependent [31]. The key factors include: the physicochemical state of the protein (e.g., posttranslational modifications), the environmental conditions (e.g., temperature, pH), and the concentration of binding partners (e.g., proteins, nucleic acids, carbohydrates, lipids). Many among the recent hypotheses suggest the prevalence of: (i) electrostatics and π-stacking; or (ii) specific sequence decoration in charge or hydrophobicity; and (iii) the role of short sequential (e.g., GARs (glycine–arginine-rich) [32]) or structural (e.g., LARKS (low-complexity amyloid-like reversible kinked segments) [33]) motifs [15]. However, deciphering the interplay between sequence composition and phase separation turns out to be challenging.

In recent years, several databases have emerged that collect LLPS-related protein sequence data and metadata, with prominent examples being PhaSepDB [25], PhaSePro [26], LLPSDB [27,28], and DrLLPS [29]. These databases collect and annotate partially overlapping sets of phase-separating protein sequences, including data on the experimental conditions and significant annotations. In particular, PhaSePro, LLPSDB, and a subset of PhaSepDB contain manually curated proteins, which are recognized for driving the formation of subcellular compartments.

The accumulation of high-quality datasets is certainly a necessary condition for making progress towards uncovering the driving forces of protein phase separation. However, one needs a biophysically motivated computational infrastructure to be able to harness the data from carefully and manually curated sets of phase-separating proteins for revealing the molecular features that determine protein phase separation. We argue that providing the concise but informative patterns of various features, all together horizontally stacked along the protein sequence, could improve the identification of the significant yet nontrivial correlations that contribute to the multivalent interactions. On the basis of these premises, we have developed a novel web platform named BIAPSS: Bioinformatics Analysis of Liquid–Liquid Phase-Separating Protein Sequences (available at https://biapss.chem.iastate.edu/ and last accessed on 31 May 2022). BIAPSS combines a high-throughput interactive deep sequence analysis with a comprehensive pre-parsed bioinformatics database containing a wide array of physicochemical and evolutionary features that are relevant for low-complexity, disordered, and ordered proteins. This platform provides scientists working in the field of biomolecular condensates with a versatile tool for the rapid and on-the-fly deep statistical analysis of LLPS-driver protein sequences.

2. Results and Discussion

2.1. Introduction of BIAPSS

Figure 1 represents the features included in the comprehensive BIAPSS analyzer. These features combine sequence composition and biophysical properties. The composition component is represented by: (i) the amino acid content, including frequencies and patterning (i.e., distribution and enriched regions); and (ii) the sequence complexity, which comprises the detection of low-complexity regions, repeats, short motifs, and the on-the-fly calculation of Shannon entropy. The biophysical component covers the physicochemical and structural properties. Specifically, we provide a set of residue-resolution patterns: polarity, hydrophobicity, aromaticity, charge induced interactions and hydrogen bonding. These properties, correlated with the experimentally confirmed LLPS regions, facilitate the identification of the nature and driving forces of interactions. The structural properties aid in filtering out the interactions involved primarily in stabilizing the structure or in identifying regions prone to disorder-to-order transitions.

Previously, such structural switchers were recognized in low-complexity and internally disordered sequences that function via phase separation [34]. Thus, the collected molecular properties incorporate robust sequence-based predictions for the secondary structure, solvent accessibility, intrinsic disorder, and intramolecular contacts. Finally, the evolutionary context derived from the joint outcome of the HMMER-based analysis [35] and Pfam database search [36] highlights the location of functional domains, including those specialized in nucleic acid recognition. The other highly conserved short motifs or individual positions detected through the analysis of multiple sequence alignments may confirm that evolution deliberately preserves phase separation.

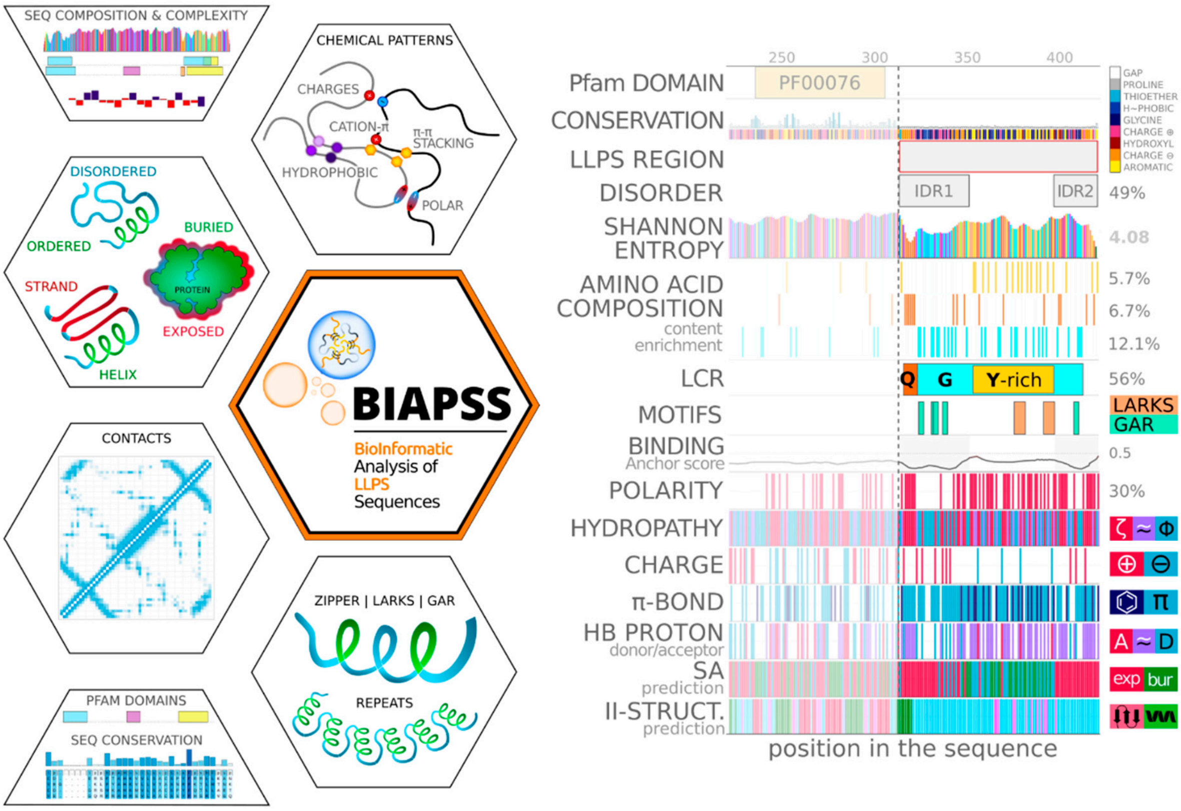


Figure 1. The comprehensive BIAPSS analyzer incorporates the compositional, evolutionary, physicochemical, and structural properties of LLPS proteins. All characteristics can be easily compared and correlated on the horizontally stacked multirow graphs. The interactive exploration helps to filter out sequence signals relevant for phase separation.

The comprehensive approach adopted in BIAPSS (Bioinformatics Analysis of LLPS Sequences) consists of integrating multiple third-party tools and high-performance computing, followed by in-house biostatistical analysis and the extraction of meaningful results (see the Materials and Methods section). The protocol was successfully applied for 501 proteins with experimental evidence of phase separation. Moreover, to the best of our knowledge, the resulting platform represents the broadest database with physicochemically characterized LLPS proteins (see Figure 2). In particular, the pool of entries completely covers the contents of several primary databases of curated LLPS deposits (PhaSePro, PhaSepDB.v1, LLPSDB), which collect annotations and experimental conditions. High interest in the phase-separation phenomenon has already spurred the growth of experimental data repositories. However, the deficiency of the computational infrastructure that targets the integrated biophysical and statistical analysis of phase-separating systems still hampers progress in the field. Therefore, in addition to the open access to the raw yet standardized and well-documented results of our extensive work, we have developed a web-based BIAPSS platform for interactive customized exploration and easy interpretation.

As a user-friendly web server, BIAPSS (https://biapss.chem.iastate.edu/, last accessed on 31 May 2022) is billing itself as a central resource for the systematic and standardized statistical analysis of the biophysical characteristics of the known LLPS sequences.

The web service provides users with:

(i) A database of the superset of experimentally evidenced LLPS-driver protein sequences.
(ii) A repository of precomputed bioinformatics and statistics data.
(iii) Two sets of web applications supporting the interactive analysis and visualization of the physicochemical and biomolecular characteristics of LLPS proteins.

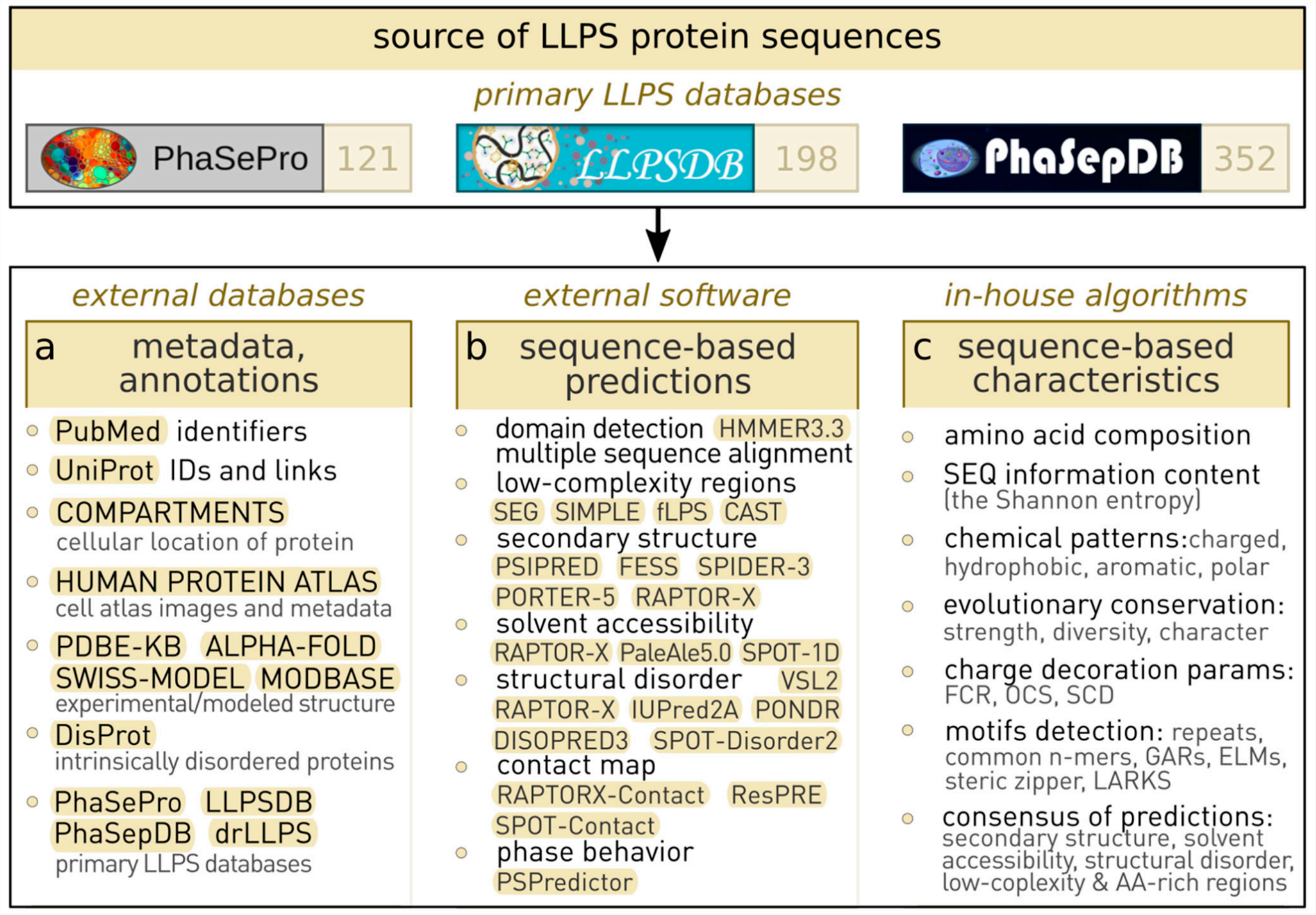

source of LLPS protein sequences
primary LLPS databases
PhaSePro 121
LLPSDB 198
PhaSepDB 352
external databases
a metadata, annotations
PubMed identifiers
UniProt IDs and links
COMPARTMENTS
cellular location of protein
HUMAN PROTEIN ATLAS
cell atlas images and metadata
PDBE-KB ALPHA-FOLD
SWISS-MODEL MODBASE
experimental/modeled structure
DisProt
intrinsically disordered proteins
PhaSePro LLPSDB
PhaSepDB drLLPS
primary LLPS databases
external software
b sequence-based predictions
domain detection HMMER3.3
multiple sequence alignment
low-complexity regions
SEG SIMPLE fLPS CAST
secondary structure
PSIPRED FESS SPIDER-3
PORTER-5 RAPTOR-X
solvent accessibility
RAPTOR-X PaleAle5.0 SPOT-1D
structural disorder VSL2
RAPTOR-X IUPred2A PONDR
DISOPRED3 SPOT-Disorder2
contact map
RAPTORX-Contact ResPRE
SPOT-Contact
phase behavior
PSPredictor
in-house algorithms
c sequence-based characteristics
amino acid composition
SEQ information content
(the Shannon entropy)
chemical patterns:charged, hydrophobic, aromatic, polar
evolutionary conservation: strength, diversity, character
charge decoration params: FCR, OCS, SCD
motifs detection: repeats, common n-mers, GARs, ELMs, steric zipper, LARKS
consensus of predictions: secondary structure, solvent accessibility, structural disorder, low-coplexity & AA-rich regions

Figure 2. The BIAPSS repository collects the largest dataset of known LLPS proteins that have been identified from carefully curated primary LLPS databases. The computational framework starts from protein sequences downloaded from the UniProt database. The approach builds on three complementary components: (**a**) integration of metadata, annotation, and cross-links from the external databases; (**b**) comprehensive sequence-based bioinformatic analysis of evolutionary and biomolecular properties using state-of-the-art third-party software; and (**c**) meticulous physicochemical and compositional analysis and robust data integration using the in-house algorithms. The BIAPSS interactive web applications enable exploration through the distilled essence of the crafted characteristics.

The applications integrate the results from our comprehensive computational approach. The SingleSEQ module includes a residue-resolution biophysical analyzer for interrogating individual protein sequences. The complementary analyses are organized in nine web applications that toggle between a generalized summary view and details specific to a given characteristic. The latter allows users to correlate regions prone to phase separation with an array of physicochemical attributes, structural properties, detected domains, and various sequential or structural motifs. Many characteristics provided by applications in the SingleSEQ pipeline are qualitative and show a profile or pattern of the feature along the amino acid sequence. Examples include distributions of physicochemical characteristics, such as polarity, hydrophobicity, charge, residues forming hydrogen bonds, and pi-stacking. In these cases, the assignment is binary, and the numerical value is the percentage of residues in the sequence that meet the criterion. The second group of characteristics includes structural features predicted based on the amino acid sequences with top-ranked tools. Here, examples include the secondary structure, the solvent accessibility, the tendency to disorder, and low-complexity regions. The visual representation is developed to assign each position along the amino acid sequence a discrete consensus value (e.g., helical or extended, or coil for the secondary structure). The numerical value is the percentage of residues that meet the given criterion (e.g., % of helical). Figure 1 is a concept image, while, in the interactive graphs, there is a label of what the given value refers to. Furthermore, for those interested in in-depth analyses, the individual applications offer an on-the-fly exploration of the results from the original tools, which typically provide the fractional probabilities for each variant of a feature (e.g., p(helical), p(extended), p(coil)) for each position along the protein sequence.

BIAPSS also includes the MultiSEQ module. One of its aims is to obtain insight into the overall characteristics of the sufficient nonredundant set of LLPS-driver protein sequences. The comparison to the benchmarks of various protein groups enables a statistical inference of specific phase-separating affinities. Finally, BIAPSS incorporates an extensive cross-reference section that links all entries to primary LLPS databases and other external resources, thereby serving as a central navigation hub for the phase-separation community. All the data used by BIAPSS are freely available for download as well-formatted files with detailed descriptions, facilitating rapid implementation in user-defined computational protocols. The long-term plan for BIAPSS is for it to serve as a unifying hub for the experimental and computational community. Thus, it provides a comprehensive set of analytic tools, biophysically featured data, and standardized protocols that facilitate the identification of the sequence signals that drive the LLPS, which altogether can support applications for designing new sequences of biomedical interest.

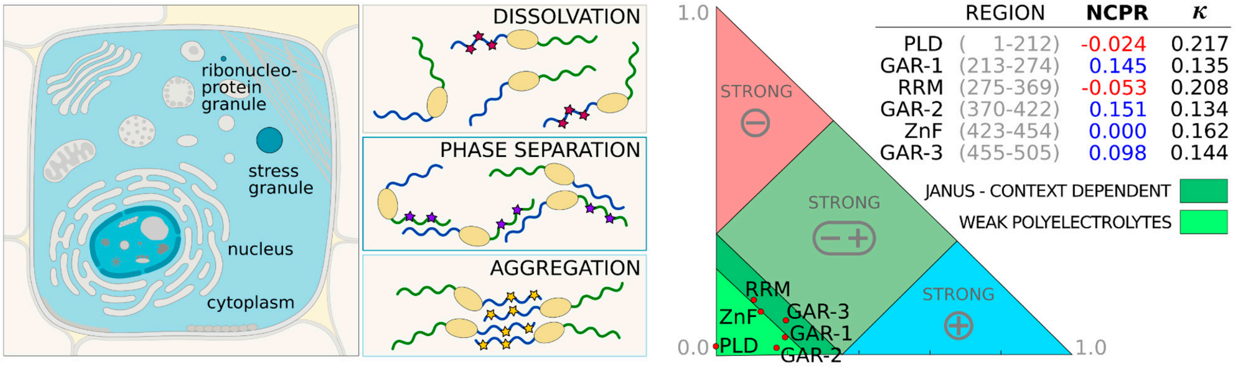
2.2. Case Study and Tutorial: Fused in Sarcoma (FUS)

To illustrate the practical utility of BIAPSS, we carefully interpreted the results for fused in sarcoma (FUS) (UniProt ID: P35637), which is a widely used model system to study biological phase separation [37]. We provide below the details of the BIAPSS-based analysis, combined with a handy tutorial on the BIAPSS functionalities.

Fused in sarcoma (FUS) is one of the early discovered biological systems that undergoes self-organization by liquid–liquid phase separation (LLPS) [37]. Since then, the protein has been the subject of extensive experimental and computational research to understand the molecular mechanisms and interactions that drive this phenomenon. FUS can be found in the SingleSEQ module of the BIAPSS service by the UniProt identifier (P35637), the gene (FUS), or by using the "RNA-binding" search key (Last accessed on May 30 2022). The summary page contains a high-quality image of the experimentally confirmed cellular location (left panel in Figure 3). Due to its multifunctionality in RNA processing, FUS is mostly observed in the nucleus [38]. In physiological conditions, the low levels of the protein are distributed in the cytoplasm [39], where FUS transports and manages RNA through the dynamic liquid-like subcellular compartments, such as ribonucleoprotein or stress granules [40]. However, the cytoplasmic concentration of FUS significantly increases when noxious mutations lead to aggregation [41].

This progressively aberrant process is manifested by neurodegenerative diseases in humans [41]. Although plenty of accumulated evidence points to the influence of distinct factors on the cellular behavior of FUS, its primary sequence still holds many cues. To frame the physicochemical properties of full-length FUS, we used the analytical approach offered by the SingleSEQ module of BIAPSS.

The average metrics, available in the *Summary* of the SingleSEQ module, indicate that the 526-residue-long sequence of FUS contains over 80% disorder and only 8% order. The solvent-accessibility predictions show the same aspect ratio between exposure and burial. The contents of aromatic, hydrophobic, polar, and charged residues are 10%, 42%, 40%, and 17%, respectively, with a slight excess of positive charge. Such a rough overview described by a set of averages gives some general insight into the protein properties, but it conceals some local distributions that are important for the identification of the preferential interactions.



REGION		NCPR	κ
PLD	(1-212)	-0.024	0.217
GAR-1	(213-274)	0.145	0.135
RRM	(275-369)	-0.053	0.208
GAR-2	(370-422)	0.151	0.134
ZnF	(423-454)	0.000	0.162
GAR-3	(455-505)	0.098	0.144

Figure 3. The **left** panel shows the cellular location of FUS (image source: BIAPSS). The protein is predominantly located in the nucleus. The physiological low levels of FUS found in the cytoplasm typically self-organize to membraneless compartments, such as stress granule or ribonucleoprotein granule. The aberrant disease-related aggregates are mostly localized in the cytoplasm. The methylation (purple stars) of C-terminal arginines (green tail) in the wild-type FUS strongly promotes phase separation and gelation. The phosphorylation (magenta stars) of serine and threonine in the N-terminus (blue tail) dissolve liquid-liquid droplets. The tyrosine-to-phenylalanine mutants (yellow stars) in the N-terminus and hypomethylation of arginines in the C-terminus increase aggregation. The **right** panel shows classification of intrinsically disordered ensemble regions (CIDER) [42] for the FUS sequence split into functional segments.

Therefore, we conduct a detailed analysis of the composition and complexity of the FUS sequence, and we present the resulting patterns in Figure 4. Compared to any reference set of proteins (use *Composition and Complexity* app), this one is extremely enriched in glycine, which makes up nearly 1/3 of the full sequence. Another 20% of the amino acid content consists primarily of serine and glutamine. Although the dominant content of these three amino acids suggests the generally low complexity of the sequence, their distribution along the sequence is strongly heterogeneous. Indeed, the calculated low information content of the sequence is mainly localized around protein terminals and clearly corresponds to three fragments with high glycine concentrations (LCR2: residues 164–267; LCR3: residues 370–420; LCR4: residues 454–507). These regions also exclusively accumulate total arginines, which, together with glycine, form a series of RGG repeating motifs that are known to bind RNA specifically [32]. Both serine and glutamine are mostly localized at the N-terminus, being more clearly clustered within LCR1 (1–163). LCR1 additionally gathers 24/35 available tyrosines, and, thus, it has visibly distinct enrichment (SQYG) that is known to occur in prion-like domains (PLD) [43]. By using the *Domains*, *Motifs*, *Repeats* application, we also found that the remaining compositionally more complex regions of the C-terminus (I287-L365 and R422-D453) match the PF00076 and PF00641 Pfam domains (i.e., the RNA recognition motif (RRM) and RNA-binding zinc finger (ZnF), respectively). The robust predictions (for details, see Methods) unanimously show that RRM is a well-folded FUS domain, while the other fragments remain disordered.

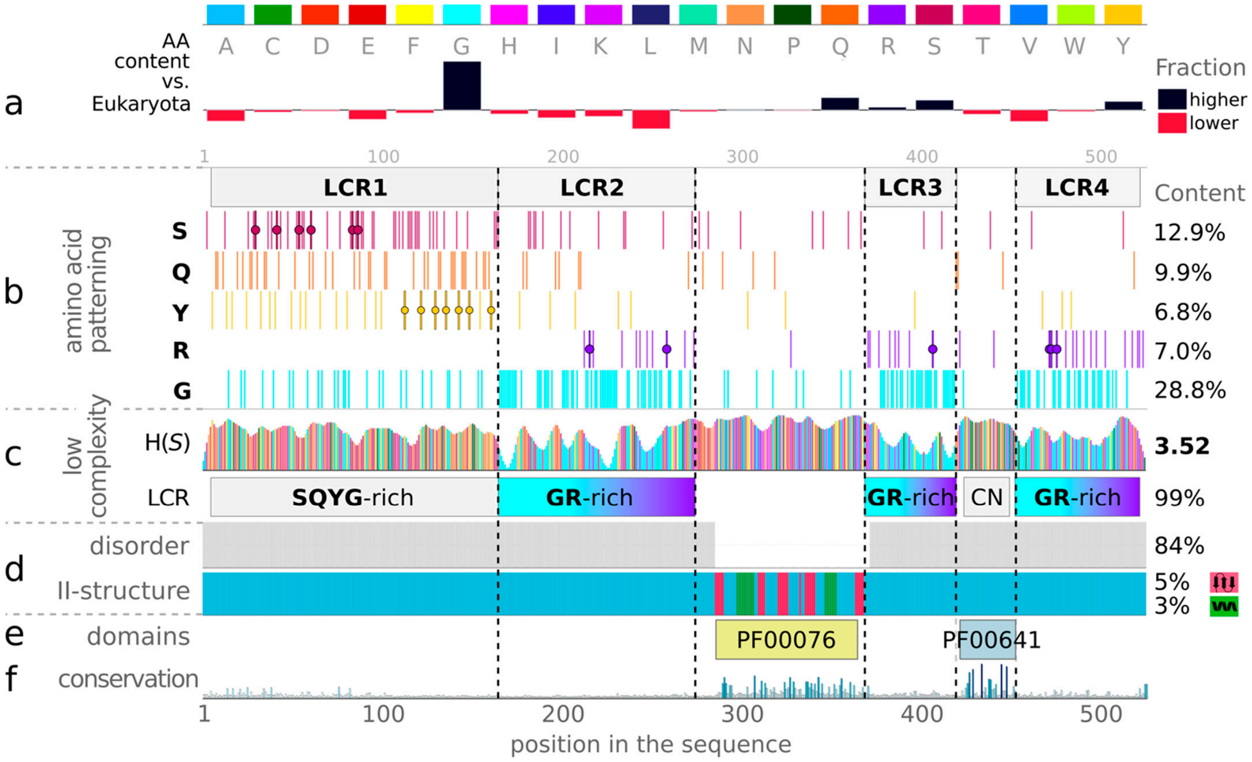


Figure 4. Sequence composition and complexity of FUS. The upper panel (**a**) shows the amino acid (AA) content of the query sequence compared to the Eukaryota dataset (black indicates higher and red the lower content than the reference). The bottom panel shows, the following information for a specific query sequence: (**b**) the patterning of enriched amino acids (S (magenta), Q (orange), Y (yellow), R (purple), G (cyan)); (**c**) low-complexity measures (a color scale corresponding to each amino acid) provided as regions of particular AA enrichments (LCR row), and the sequence information content (H(S) row, Shannon entropy); (**d**) consensus of predicted disorder regions (gray) and secondary-structure assignment (helix in green, strand in magenta, coil in light blue); (**e**) detected Pfam domains; and (**f**) evolutionary conservation derived from the multiple sequence alignment against UniRef50 (blue shades). The amino acid patterning section contains points corresponding to the locations of the most relevant serine phosphorylation sites (residues 30, 42, 54, 61, 84, 87) [37], arginine methylations (residues 216, 259, 407, 472, 473, 476) [44], and tyrosine mutations (residues 113, 122, 130, 136, 143, 149, 161) [44].

The seed MSAs prepared for FUS within the *Sequence Conservation* application further confirm that both domains are evolutionarily conserved members of Pfam families: RRM_1 and zf-RanBP, respectively (see bottom rows in Figure 4). The visual inspection of the amino acid content and the distribution of FUS allows us to identify and isolate specific regions in the protein (Figure 5). Furthermore, we have performed a physicochemical featurization of these segments, using the *Chemical Properties Patterns* app, which reveals the preferred interactions when coupled with biomolecular conditionals that are known from experiments. The recent experimental reports show that the isolated prion-like domain (PLD) (residues 1–214, or even residues 1–163) can undergo self-organization, forming liquid droplets when kept at high protein levels or high salt concentrations [44,45]. This N-terminal fragment is enriched in amino acids, whose side chains are multivalent, as shown in Figure 5. Thus, the dense pattern of polarity comes from the enrichment in S, Q, Y, where Y, Q, and G also provide π-electron centers for π-π-stacking. Most of them are also able to be both donors and acceptors of side-chain protons for hydrogen bonding (HB). In line with this, the intermolecular-interaction profiles derived from simulations of the 120–163 region indicate the most frequent contacts between QQ > QY > YY > SY and other pairs of enriched amino acids [46]. All of these observations suggest that the homotypic phase separation of wild-type PLD monomers is driven by balanced contributions from hydrogen bonding and π-stacking. Indeed, several mutagenesis studies show that Y→A substitution disrupts phase separation by the removal of both components of the interaction, while Y→F mutants are significantly more aggregation-prone, due to the strengthening of the binding via tighter hydrophobic F-π-stacking at the cost of losing HB contributions of polar tyrosine [44,46]. It is also worth noting that the PLD region is completely deficient of positive charge, with a minor net charge per residue of −0.01 (M1-S165: −0.012; and M1-G212: −0.024), which places it within the weak polyelectrolyte region on the CIDER diagram (right panel in Figure 3) [42]. However, an excess of serine and threonine in this region provides an ability to introduce a strongly negative charge through multiple phosphorylations. After phosphorylation, the dominant force becomes electrostatic repulsion, which is known to disrupt both phase separation and aggregation [37]. The central region of the PLD (residues 39–95) was proposed as the core of aberrant fibrils, which, in solid-state form, structured cross-β-sheets [37]. The same structural properties have not been unambiguously confirmed in the condensed phase of liquid–liquid mixing. Undoubtedly, however, our algorithms detected, along this region, structural motifs known as low-complexity amyloid-like reversible folded segments (LARKS) [33]. In our analysis, the most effective predictors of structural properties showed, for these motifs, some tendency towards an extended secondary structure and a slightly increased probability of burial (bottom panel in Figure 5; use *Secondary Structure*, *Solvent Accessibility* and *Structural Disorder* applications of the BIAPSS SingleSEQ module). Interestingly, the prediction in eight-letter notation detected a turn or bend within each of the structural motifs, which explains their flexible nature. These findings, together with the ambiguous experimental results, may suggest some variations in the structural state in the PLD core, and specifically the disorder-to-order transition driven by biomolecular conditionals.

The remaining part of the FUS sequence, referred to as the C-terminus, contains two well-known domains (RRM and ZnF) and three glycine–arginine-rich regions (GARs), which are detected using the *Domains, Motifs, Repeats* application. All components are significant players in binding RNA. Zinc finger supports only the recognition of the specific GGU motif, while the RRM domain and RGG repeats are universal towards a variety of RNAs [47]. Both folded domains of FUS are much less polar than the PLD, as seen from the BIAPSS-based physicochemical features in Figure 5. They also have a lower content of side chains that are able to engage in π-stacking or hydrogen bonding. However, the charged residues are pretty abundant in the composition of RRM and ZnF, which explains the functional role of electrostatic interactions towards the binding of nucleic acids or stabilizing folds via salt bridges [48,49].

All three GARs are the least polar regions of the protein (see Figure 5; use *Chemical Properties Patterns* app of BIAPSS SingleSEQ module). The dense patterning of hydrophobicity arises from glycine excess. The rich π-electron-containing systems, other than aromatic side chains, originate mainly from the abundance of the arginine's guanidino group. Arginine is also a source of excess positive charge at the C-terminus. The experimental studies consistently confirm that the isolated C-terminus does not undergo phase separation [44]. However, liquid–liquid droplets rapidly occur when mixed with N-terminal monomers [44]. Moreover, the LLPS of full-length wild-type FUS is more robust than the heterotypic mixing of the N- and C-terminals and the homotypic self-assembling of N-terminal monomers [44]. This suggests the higher priority of cation-π (R-Y) stacking over π-π (Y-Y) stacking, while both are reinforced by hydrogen bonds. Another experimental study showed that R→K mutants, who no longer have the ability of π-π-stacking but retain charge, can still undergo phase separation. In turn, R→A substitutions prevent phase separation because they lose the π-system, cation, and ability of side-chain hydrogen bonding. Interestingly, the recent report indicates that stacking interactions, including cation-π (e.g., RY, KF) and especially π-π (e.g., YY and RY, and even RQ), are most robust over a wide range of salt concentrations [45]. The hydrophobic contribution from π-electron-containing systems becomes the main force that strengthens the contact in high salt. In these conditions, the screening of usually dominant electrostatic contributions is significant. Surprisingly, changing the partitioning of the different forces makes the interaction of the two positively charged arginines attractive under these conditions [45]. The set of diverse chemical groups in arginine is a unique feature among the other amino acid side chains. With its high reactivity, the need for precise regulation comes, and so arginine can be tuned to a preferred state by posttranslational methylation.

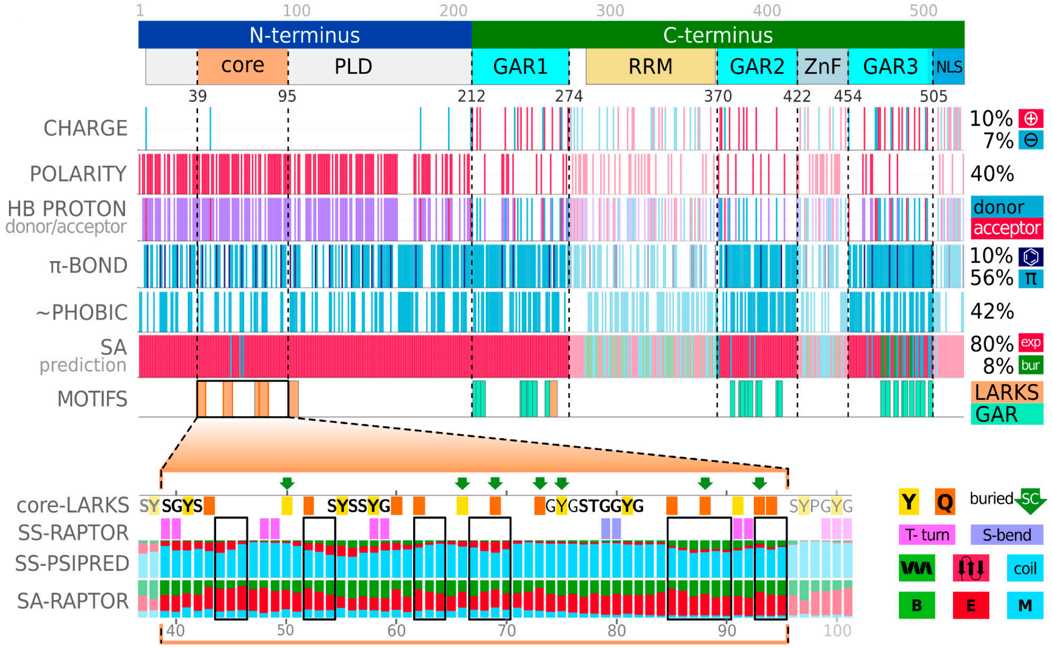


Figure 5. Physicochemical and structural properties of FUS. The various characteristics are shown along the protein-sequence split on the N-terminus (blue) and C-terminus (green). The full-length N-terminus corresponds to a highly polar prion-like domain (PLD in gray). The PLD contains a core region (residues 39–95 in orange), in which multiple LARKS motifs were detected (orange bars in the MOTIFS row) and evidenced to form fibrils in a solid state. The C-terminus contains two well-folded domains detected by Pfam search (RRM in yellow, ZnF in gray), and three glycine–arginine-rich (GAR in cyan) regions. All components of the C-terminus are known to have a functional role in RNA binding. The following rows show the physicochemical patterning along the full-length FUS sequence, including charge (positive (magenta), negative (blue)), polarity, donor (blue)/acceptor (magenta)/both (purple) of side-chain proton for hydrogen bonding, π-electron-containing systems (blue), with separation of aromatic ones (dark blue), hydrophobicity (blue), predicted solvent accessibility (SA) in 3-letter notation (exposed (red), buried (green), medium (blue)). The zoom of the PLD core is shown at the bottom panel, where the SS and SA rows contain the predicted probabilities of secondary structure (helix (green), strand (red), coil (blue)) and solvent accessibility. The green arrows indicate the side chains buried in the fibril core, while the black frames highlight segments that form strands of a cross-β motif [37].

Thus, under physiological conditions, FUS is highly methylated [44]. This limits self-assembly via interactions with tyrosine and promotes a functional role of intermolecular interactions with other proteins and nucleic acids. Therefore, phase separation and the gelation of FUS can increase by the hypomethylation of arginines within RGG-rich regions or the insertion of additional ones into the C-terminus [44]. All of these findings come together to demonstrate the significant role of the arginine side chain in phase separation. Tyrosine and glutamine are similarly relevant.

2.3. FUS, LLPS Regulated in the Context-Dependent Tuning of Preferred Forces

Note that the original results generated for FUS (identifier: P35637) using the BIAPSS web platform are shown in Supplementary Materials. The findings described in the previous section are briefly summarized below. FUS is a predominantly disordered protein (80%), composed of two functional regions: the prion-like domain (residues 1–212) and the RNA-binding C-terminal fragment (residues 213–526). Surprisingly, although the sequence is more than 65% composed of only five amino acids (G >> S >Q > R~Y), their distribution is highly variable. In particular, low sequence complexity occurs mainly in three structurally flexible glycine–arginine-rich regions. They provide hydrophobicity, π-electron-containing centers, and a positive charge of the C-terminus. In contrast, the enrichment of the N-terminus in serine, glutamine, and tyrosine makes it strongly polar but negatively charged (due to a few aspartic acids), with numerous aromatic centers and side chains capable of hydrogen bonding. These very different physicochemical properties of the protein terminals are sensitive to environmental changes and allow for the context-dependent regulation of the protein's cellular behavior. In particular, arginine can be tuned to a preferred state by posttranslational methylation [44], while serine/threonine phosphorylation introduces a highly negative charge [37]. Both modifications prevent the formation of aberrant self-assembly. In the first case, the methyl groups hinder cation-π stacking between arginine and tyrosine, limiting the phase separation driven by contacts between the N- and C-terminals. In the second case, phosphorylation introduces strong electrostatic repulsion between the N-termini. It inhibits homotypic phase separation, driven by the π-stacking of aromatic tyrosines. However, both mechanisms have no effect under high salinity due to the significant screening of electrostatics. In such conditions, even interactions of arginines become attractive due to hydrophobic contributions and π-π stacking [45]. Therefore, these observations demonstrate that the peculiar physicochemical properties of amino acid residues play a significant role in phase separation. The multifunctional chemical groups of amino acids make them reactive and multivalent. These features aid in the context-dependent tuning between preferred modes of interactions. They can work synergistically or alternatively, and their regulation depends on the environmental conditions, the state of posttranslational modifications, and the presence of binding partners.

3. Materials and Methods

3.1. Sequence Complexity and Physicochemical Decoration

3.1.1. Sequence Complexity

Low-complexity regions (LCRs) in proteins are compositionally biased fragments of sequences that often have low amino acid diversity and repeats of short motifs of the sequential or structural kinds. Many reports point to their functional or regulatory roles, frequently also associated with subcellular phase separation [19]. The LCRs of LLPS proteins have been detected by using several state-of-the-art tools, such as SIMPLE [50], CAST [51], fLPS [52], and SEG [53]. The original hits were parsed by in-house algorithms to merge overlapping regions enriched in different amino acids, and only the integrated and unified results have been kept.

Shannon Entropy describes the information content held in data and it is a frequently used measure of protein-sequence complexity. We implemented a module for the on-the-fly calculation of it within BIAPSS services. The typical window length for compositional effects is between 5 and 20. The results can be displayed in:

- Residue-resolution mode (residue option; smoother output):

$$S_i = \frac{1}{N}\sum_{j=1}^{N} S_{(j,N)}$$

where the Shannon entropy ($S_{(i)}$) at sequence position i is a sum of entropies at all windows containing this position, normalized by the window length (N);

- Window-resolution mode (block option):

$$S_{(j,N)} = -\sum_{aa=1}^{AA=20} f_{aa} log_2(f_{aa})$$

where the Shannon entropy ($S_{(j,N)}$) at the j-th sequence window of the length (N) is summed over the fractions (f_{aa}) of 20 biogenic amino acids. The value is assigned to the center position within the window. The $S_{(j,N)}$ ranges from 0 (where only one residue is present within the sequence window) to $\log_2(N)$ (all positions are different). Therefore, the lower the Shannon entropy, the less complex the sequence is.

3.1.2. Physicochemical Decoration

To examine the physicochemical properties of LLPS-driving proteins, we identified, along each sequence, the patterns of polarity (Ser, Thr, Tyr, Gln, Asn, Cys, Met), hydrophobicity (Gly, Ala, Val, Ile, Leu, Pro, Phe), and detected π-stacking centers (Arg, Asn, Asp, Gln, Glu, Gly (note that, due to the lack of the side chain, glycine can stack via π-electrons from a peptide bond and hydrogen bonding via backbone carbonyl or amide), including those within aromatic rings (Phe, Tyr, Trp, His). We also provided the charge-distribution split between positively (His, Lys, Arg) and negatively (Glu, Asp) charged residues. For each feature, both the arrangement along the sequence and the fraction of residues are provided.

3.1.3. Electrostatics

It is well established that the electrostatic interactions often affect the solubility and stabilize the binding interface in the liquid–liquid demixing of biomolecules. The recently proposed charge-decoration parameters emerged as a measure of charge distribution along the protein sequence. In addition to the overall charge content, these descriptors are seen as important factors that shape the protein conformations, especially within low-complexity regions [54]. Following these discoveries, we calculated and compared the charge-decoration parameters; namely:

- SCD (sequence charge decoration) is implemented following the formulation by Sawle and Ghosh [55];
- OCS (overall charge symmetry) is implemented following the formulation by Das and Pappu [56];
- FCR (fraction of charged residues) is defined as a sum of the fractions of positive and negative charges.

3.2. HMMER-Based Sequence Conservation and Functional-Domain Detection

The multiple sequence alignment (MSA) and consensus profile were prepared using an efficient HMMER method (*phmmer* + *hmmalign* and *hmmbuild*, respectively), which employs a probabilistic hidden Markov model (HMM) [35], and are significantly more accurate compared to BLAST-based searches. Because some of the LLPS sequences are highly unique (detection of the remote homologs is needed), and because the MSA is reliable if at least several dozen homologous sequences are available, we used sequences selected from various UniProt subsets. Specifically, SwissProt, UniRef50, and UniRef90 differ in the size and increasing sequence identity of entries [57–59]. To identify sequence regions with significant evolutionary conservation, we derived three additional MSA-based parameters: strength, diversity, and character. The MSA strength of the sequence conservation informs on how much the specific position is held by evolution. This measure normalizes results from the *hmmlogo* tool to a discrete range from 0 (poorly conserved) to 5 (highly conserved). The *hmmlogo* computes letter heights along the sequence, depending on the information content of the position. The MSA diversity defines the number of different amino acids detected at a given position in the MSA, and is provided in discrete scale from 0 (highly conserved) to 5 (poorly conserved) (0—one, 1—two, 2—three, 3—four, 4—five or six, 5–7 and more amino acids at the aligned position). The MSA character describes the chemical nature of the most common amino acids at a given position in the multiple sequence alignment. We distinguished the following attributes: polar, charge, aromatic, another π-system, hydrophobic, and other (G or P).

Some LLPS proteins are composed of one or more well-known domains. The identification of these functional regions alongside regions of low complexity or disorder can provide additional insights into the regulatory role of phase separation. Therefore, we have performed a Pfam search for all LLPS proteins, reporting the detected domains and incorporating the original Pfam seed-MSAs for corresponding regions of LLPS sequence (instead of full-length ones) to derive more reliable evolutionary conservation descriptors.

3.3. Short Sequential and Structural Motifs Specific for LLPS Sequences

Short linear motifs (SLiMs) are short fragments along the sequence, often situated in the intrinsically disordered regions, generally showing high structural flexibility and evolutionary conservation. We systematically detected various short sequential and structural motifs. The implemented algorithms used the list of grouped motifs' instances, defined by regular expressions, as the keys to search protein sequences prone to phase separation. Among motifs known from the literature as relevant for phase behavior, our analysis includes short structural stretches of protein sequence, such as LARKS [33] and steric zippers [60]; glycine–arginine-rich regions (GARs) [32]; and new sequential repetitive n-mers.

3.4. Structural Properties Derived from Sequence-Based Predictions

Bearing in mind the predictive nature of sequence-based methods and, hence, their limited accuracy, comparing several of them and choosing the final consensus has proven to be successful in many approaches. In our study, we comprised predictions from at least three to six widely used tools for each biomolecular characteristic. While almost every method is available as a web server, due to the size and complexity of our analyses, we employed standalone versions. The raw data derived from these standalone tools during the high-performance computing was initially parsed, filtered, and simplified to a uniform CSV format, and deposited in our online repository at https://biapss.chem.iastate.edu/download.html, accessed on 1 April 2022.

3.4.1. Secondary Structure

Protein secondary structure is a regular three-dimensional organization of local fragments along a polypeptide chain. The two most common secondary structural elements are alpha helices and beta sheets. Most of the predictors provide the secondary-structure assignment in 3-letter notation (ss3): H—helix, E—strand, C—coil, while the advanced ones (RaptorX and PORTER-5) also deliver more detailed 8-letter notation (ss8): H—α-helix; G—3_{10}-helix; I—π-helix; E—β-strand; B—β-bridge; T—HB-turn; S—bend; C–loop. In our benchmark study, we employed five well-established secondary-structure predictors: PSIPRED [61], RaptorX-SS8 [62], PORTER-5 [63], SPIDER-3-Single [64], and FELLS [65].

3.4.2. Solvent Accessibility

Solvent accessibility gives some insight into protein structural flexibility, indicating the exposed patches on the protein surface available for interactions with the solvent molecules. Some surface sites have high evolutionary conservation, which is suggestive of functional or structural importance. Since not many structures of phase-separating IDPs are known, the robust prediction of solvent accessibility can help to identify flexible regions prone to conformational changes upon binding. The assignment of solvent accessibility is usually provided in the 3-letter code: B—buried, E—exposed, M—medium. In our benchmark study, we employed three well-established solvent-accessibility predictors: RaptorX-Property [66], PaleAle 5.0 [67], SPOT-1D [68].

3.4.3. Structural Disorder

The sequence-based predictions indicate regions of increased structural flexibility, usually estimating the disorder probability at a given position in the sequence. Detecting highly flexible regions may support the identification of short sequence stretches of multivalent interactions that can be relevant to phase separation. In our benchmark study, we employed seven well-established predictors of structural disorder: RaptorX-Property [66], IUPred2A [69], SPOT-Disorder [70], DISOPRED (v2 and v3) [71], and PONDR® (FIT, VLXT, VSL2) [72]. Most of these methods return the probability of disorder for each position in the sequence. Usually, the residue is considered as ordered when the score is below 0.5. The protein-binding regions in disordered fragments were estimated using the ANCHOR method [73].

3.4.4. Contact Map

Contact-map application provides a more reduced representation of a protein structure using a binary two-dimensional matrix of distances between all possible amino acid-residue pairs. The commonly used definition assumes the threshold 6–10 Å as the distance between the pair of two Cα or Cβ atoms being in contact. The contact number of protein residues limits the number of possible protein conformations and helps encode a three-dimensional structure. In our benchmark study, we employed three state-of-the-art predictors of intramolecular contacts: RaptorX-Contact [74], ResPRE [75], SPOT-Contact [68,76].

3.5. Data Availability

The UniProt IDs of LLPS sequences were collected as a joint superset of deposits from primary LLPS databases (i.e., PhaSePro (https://phasepro.elte.hu/, accessed on 1 April 2022), PhaSepDB.v1 (http://db.phasep.pro/, accessed on 1 April 2022), LLPSDB (http://bio-comp.org.cn/llpsdb/, accessed on 1 April 2022)). Then, protein sequences were taken from the UniProt database, available at https://www.uniprot.org/, accessed on 1 April 2022. The cellular location of the protein was derived via web scraping of primary LLPS databases, UniProt and COMPARTMENTS (https://compartments.jensenlab.org/, accessed on 1 April 2022). The following resources were reviewed for the corresponding entries of experimental or predicted three-dimensional structures: PDBe (https://pdbe-kb.org/, accessed on 1 April 2022), Swiss-Model Repository (https://swissmodel.expasy.org/repository/, accessed on 1 April 2022), ModBase (http://salilab.org/modbase-cgi/, accessed on 1 April 2022), and AlphaFold DB (https://alphafold.ebi.ac.uk/, accessed on 1 April 2022).

The results of our comprehensive analysis, performed on 501 proteins, are available at https://biapss.chem.iastate.edu/download.html, accessed on 1 April 2022. For each file, the details of its content and methods used are comprehensively described. These files are used directly as the input for the web applications of the SingleSEQ and MultiSEQ modules in the BIAPSS platform. The results of the analysis can be explored interactively online, saved as high-quality PNG images, and used directly as figures in the publication.

3.6. Code Availability

3.6.1. Phase-Separation Predictors:

1. PSPredictor, web server version available at http://www.pkumdl.cn:8000/PSPredictor, accessed on 1 April 2022.

3.6.2. Low-Complexity-Region (LCR) Predictors (Sequence-Based):

1. SEG, standalone (1999), available at https://ftp.ncbi.nih.gov/pub/seg/seg/, accessed on 1 April 2022;

2. fLPS, standalone Sep 2017, available at https://github.com/pmharrison/flps, accessed on 1 April 2022;

3. SIMPLE, standalone V6-6.1, available at https://github.com/john-hancock/SIMPLE-V6, accessed on 1 April 2022;

4. CAST2, web server version available at http://structure.biol.ucy.ac.cy/CAST2/index.html, accessed on 1 April 2022.

3.6.3. Multiple Sequence Alignment/Build a Profile/Conservation Logo:

1. HMMER (phmmer, hmmalign, hmmbuild, hmmlogo), standalone 3.3, available at http://hmmer.org/download.html, accessed on 1 April 2022;
2. Pfam, the database search was used to detect functional domains, available at http://pfam.xfam.org/, accessed on 1 April 2022.

3.6.4. Secondary-Structure Prediction (Sequence-Based):

1. PSIPRED, standalone 4.02, available at https://github.com/psipred/psipred, accessed on 1 April 2022;
2. RAPTOR-X, standalone Version ID: Rev: 37223, available upon request at http://raptorx.uchicago.edu/download/, accessed on 1 April 2022;
3. PORTER, standalone v5, available at https://github.com/mircare/Porter5/, accessed on 1 April 2022;
4. SPIDER, standalone v3, available upon request at https://sparks-lab.org/downloads/, accessed on 1 April 2022;
5. FESS, standalone 2.0 (November 2016), available upon request at http://old.protein.bio.unipd.it/download/, accessed on 1 April 2022.

3.6.5. Solvent-Accessibility Prediction (Sequence-Based):

1. RAPTOR-X property, standalone v1.01 (October 2018), available upon request at http://raptorx.uchicago.edu/StructurePropertyPred/predict/, accessed on 1 April 2022;
2. PaleAle, standalone 5.0 (December 2019), available at https://github.com/mircare/Brewery, accessed on 1 April 2022;
3. SPOT-1D, standalone (July 2019), available upon request at https://servers.sparks-lab.org/downloads/, accessed on 1 April 2022.

3.6.6. Structural-Disorder Prediction (Sequence-Based):

1. RAPTOR-X property, standalone v1.01 (October 2018), available upon request at http://raptorx.uchicago.edu/StructurePropertyPred/predict/, accessed on 1 April 2022;
2. UPred2A, standalone (November 2019), available upon request at https://iupred2a.elte.hu/download_new, accessed on 1 April 2022;
3. DISOPRED, standalone v2 and v3.1, available at https://github.com/psipred/disopred, accessed on 1 April 2022;
4. SPOT-Disorder2, standalone (February 2019), available upon request at https://sparks-lab.org/downloads/, accessed on 1 April 2022;
5. VSL2, standalone (November 2019), downloaded from http://www.dabi.temple.edu/disprot/download/VSL2.tar.gz (not available now), accessed on 1 April 2022;
6. PONDR-FIT, web server available at http://original.disprot.org/pondr-fit.php, accessed on 1 April 2022;
7. PONDR-VLXT, web server available at http://www.pondr.com/, accessed on 1 April 2022.

3.6.7. Contact-Map Prediction (Sequence-Based):

1. RAPTOR-X Contact, web server available at http://raptorx.uchicago.edu/ContactMap/, accessed on 1 April 2022;
2. ResPRE, standalone (November 2019), available at https://zhanglab.ccmb.med.umich.edu/ResPRE/download/ResPRE.zip, accessed on 1 April 2022;
3. SPOT-Contact, standalone v3 (June 2007), available upon request at https://sparks-lab.org/downloads/, accessed on 1 April 2022.

The raw data collected from the third-party software were parsed utilizing custom python/bash algorithms to provide the unified format and derive the consensus properties.

The output files are available at https://biapss.chem.iastate.edu/download.html, accessed on 1 April 2022.

4. Conclusions

In conclusion, many proteins undergo liquid–liquid phase separation (LLPS), which drives the biogenesis of various membraneless organelles. The interplay between the protein sequence and the LLPS potential is poorly understood. The BIAPSS web platform, which provides the means for the analysis, visualization, and interpretation of data for LLPS proteins, is designed to uncover the sequence-encoded signals of LLPS proteins. This BIAPSS platform stands out as an efficient and user-friendly visualization framework that facilitates the integration and comparison of the physicochemical and structural features of the vast majority of known phase-separating proteins. With the rapid growth of experimental data on a single-case basis, we expect the increased need for computational infrastructure that consolidates some generalized insights. Hence, we have also developed a feature-rich module for analyzing multiple protein sequences. The interactive interface, with content-rich labels and tooltips, makes data exploration and interpretation easy. Both the web applications and raw datasets are broadly accessible on multiple operating systems and popular browsers. The presented case study of FUS shows that the BIAPSS-inferred biophysical regularities accurately identify regions prone to phase separation and facilitate the design of precise sequence modifications for various applications.

While the current version of BIAPSS enables the convenient and insightful analysis of large nonredundant and high-quality LLPS protein supersets, there are several directions through which our platform could expand to discover unknown LLPS proteins. These include analyzing the physicochemical and structural properties for customized protein sequences and introducing several LLPS indicators trained with machine learning to reveal coupled effects. The new functionality will find applications for flexible sequence redesign to introduce or modulate phase separation. Among many beneficial uses, it could tailor the properties of modern biomaterials or open up new directions in the development of medical therapies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23116204/s1.

Author Contributions: Conceptualization, A.E.B.-D.; methodology, A.E.B.-D.; software, A.E.B.-D.; validation, A.E.B.-D., D.A.P. and V.N.U.; formal analysis, A.E.B.-D., D.A.P., V.N.U.; investigation, A.E.B.-D., D.A.P. and V.N.U.; writing—original draft preparation, A.E.B.-D., D.A.P. and V.N.U.; writing—review and editing, A.E.B.-D., D.A.P. and V.N.U.; visualization, A.E.B.-D.; funding acquisition, A.E.B.-D. and D.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Roy J. Carver Charitable Trust through the Iowa State University Bioscience Innovation Postdoctoral Fellowship (to A.E.B.-D.) and the National Institute of General Medical Sciences of the National Institutes of Health [R35GM138243 to D.A.P.].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in this article and in Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Brangwynne, C.P.; Eckmann, C.R.; Courson, D.S.; Rybarska, A.; Hoege, C.; Gharakhani, J.; Julicher, F.; Hyman, A.A. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **2009**, *324*, 1729–1732. [CrossRef] [PubMed]
2. Brangwynne, C.P.; Mitchison, T.J.; Hyman, A.A. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4334–4339. [CrossRef] [PubMed]
3. Li, P.; Banjade, S.; Cheng, H.C.; Kim, S.; Chen, B.; Guo, L.; Llaguno, M.; Hollingsworth, J.V.; King, D.S.; Banani, S.F.; et al. Phase transitions in the assembly of multivalent signalling proteins. *Nature* **2012**, *483*, 336–340. [CrossRef] [PubMed]
4. Banjade, S.; Rosen, M.K. Phase transitions of multivalent proteins can promote clustering of membrane receptors. *eLife* **2014**, *3*, e04123. [CrossRef]
5. Banjade, S.; Wu, Q.; Mittal, A.; Peeples, W.B.; Pappu, R.V.; Rosen, M.K. Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E6426–E6435. [CrossRef]
6. Uversky, V.N. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **2017**, *44*, 18–30. [CrossRef]
7. Uversky, V.N. Protein intrinsic disorder-based liquid-liquid phase transitions in biological systems: Complex coacervates and membrane-less organelles. *Adv. Colloid. Interface Sci.* **2017**, *239*, 97–114. [CrossRef]
8. Darling, A.L.; Zaslavsky, B.Y.; Uversky, V.N. Intrinsic disorder-based emergence in cellular biology: Physiological and pathological liquid-liquid phase transitions in cells. *Polymers* **2019**, *11*, 990. [CrossRef]
9. Sawyer, I.A.; Bartek, J.; Dundr, M. Phase separated microenvironments inside the cell nucleus are linked to disease and regulate epigenetic state, transcription and RNA processing. *Semin. Cell Dev. Biol.* **2019**, *90*, 94–103. [CrossRef]
10. Turoverov, K.K.; Kuznetsova, I.M.; Fonin, A.V.; Darling, A.L.; Zaslavsky, B.Y.; Uversky, V.N. Stochasticity of biological soft matter: Emerging concepts in intrinsically disordered proteins and biological phase separation. *Trends Biochem. Sci.* **2019**, *44*, 716–728. [CrossRef]
11. Uversky, V.N.; Finkelstein, A.V. Life in phases: Intra- and inter- molecular phase transitions in protein solutions. *Biomolecules* **2019**, *9*, 842. [CrossRef] [PubMed]
12. Nesterov, S.V.; Ilyinsky, N.S.; Uversky, V.N. Liquid-liquid phase separation as a common organizing principle of intracellular space and biomembranes providing dynamic adaptive responses. *Biochim. Biophys. Acta Mol. Cell. Res.* **2021**, *1868*, 119102. [CrossRef] [PubMed]
13. Uversky, V.N. Recent developments in the field of intrinsically disordered proteins: Intrinsic disorder-based emergence in cellular biology in light of the physiological and pathological liquid-liquid phase transitions. *Annu. Rev. Biophys.* **2021**, *50*, 135–156. [CrossRef] [PubMed]
14. Antifeeva, I.A.; Fonin, A.V.; Fefilova, A.S.; Stepanenko, O.V.; Povarova, O.I.; Silonov, S.A.; Kuznetsova, I.M.; Uversky, V.N.; Turoverov, K.K. Liquid-liquid phase separation as an organizing principle of intracellular space: Overview of the evolution of the cell compartmentalization concept. *Cell. Mol. Life Sci.* **2022**, *79*, 251. [CrossRef]
15. Dignon, G.L.; Best, R.B.; Mittal, J. Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.* **2020**, *71*, 53–75. [CrossRef] [PubMed]
16. Choi, J.M.; Holehouse, A.S.; Pappu, R.V. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **2020**, *49*, 107–133. [CrossRef]
17. Wang, J.; Choi, J.M.; Holehouse, A.S.; Lee, H.O.; Zhang, X.; Jahnel, M.; Maharana, S.; Lemaitre, R.; Pozniakovsky, A.; Drechsel, D.; et al. A molecular grammar governing the driving forces for phase separation of prion-like rna binding proteins. *Cell* **2018**, *174*, 688–699.e16. [CrossRef]
18. Uversky, V.N.; Kuznetsova, I.M.; Turoverov, K.K.; Zaslavsky, B. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Lett.* **2015**, *589*, 15–22. [CrossRef]
19. Banani, S.F.; Lee, H.O.; Hyman, A.A.; Rosen, M.K. Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell. Biol.* **2017**, *18*, 285–298. [CrossRef]
20. Savojardo, C.; Martelli, P.L.; Casadio, R. Protein-protein interaction methods and protein phase separation. *Annu. Rev. Biomed. Data Sci.* **2020**, *3*, 89–112. [CrossRef]
21. Borcherds, W.; Bremer, A.; Borgia, M.B.; Mittag, T. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr. Opin. Struct. Biol.* **2021**, *67*, 41–50. [CrossRef] [PubMed]
22. Zaslavsky, B.Y.; Ferreira, L.A.; Uversky, V.N. Driving forces of liquid-liquid phase separation in biological systems. *Biomolecules* **2019**, *9*, 473. [CrossRef] [PubMed]
23. Tsang, B.; Pritisanac, I.; Scherer, S.W.; Moses, A.M.; Forman-Kay, J.D. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* **2020**, *183*, 1742–1756. [CrossRef] [PubMed]
24. Saar, K.L.; Morgunov, A.S.; Qi, R.; Arter, W.E.; Krainer, G.; Knowles, T.P. Machine learning models for predicting protein condensate formation from sequence determinants and embeddings. *bioRxiv* **2020**. [CrossRef]
25. You, K.; Huang, Q.; Yu, C.; Shen, B.; Sevilla, C.; Shi, M.; Hermjakob, H.; Chen, Y.; Li, T. PhaSepDB: A database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* **2020**, *48*, D354–D359. [CrossRef]
26. Meszaros, B.; Erdos, G.; Szabo, B.; Schad, E.; Tantos, A.; Abukhairan, R.; Horvath, T.; Murvai, N.; Kovacs, O.P.; Kovacs, M.; et al. PhaSePro: The database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **2020**, *48*, D360–D367. [CrossRef]

27. Li, Q.; Peng, X.; Li, Y.; Tang, W.; Zhu, J.; Huang, J.; Qi, Y.; Zhang, Z. LLPSDB: A database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res.* **2020**, *48*, D320–D327. [CrossRef]
28. Wang, X.; Zhou, X.; Yan, Q.; Liao, S.; Tang, W.; Xu, P.; Gao, Y.; Li, Q.; Dou, Z.; Yang, W.; et al. LLPSDB v2.0: An updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* **2022**, *38*, 2010–2014. [CrossRef]
29. Ning, W.; Guo, Y.; Lin, S.; Mei, B.; Wu, Y.; Jiang, P.; Tan, X.; Zhang, W.; Chen, G.; Peng, D.; et al. DrLLPS: A data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res.* **2020**, *48*, D288–D295. [CrossRef]
30. Li, Q.; Wang, X.; Dou, Z.; Yang, W.; Huang, B.; Lou, J.; Zhang, Z. Protein databases related to liquid-liquid phase separation. *Int. J. Mol. Sci.* **2020**, *21*, 6796. [CrossRef]
31. Alberti, S.; Gladfelter, A.; Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **2019**, *176*, 419–434. [CrossRef] [PubMed]
32. Chong, P.A.; Vernon, R.M.; Forman-Kay, J.D. RGG/RG motif regions in RNA binding and phase separation. *J. Mol. Biol.* **2018**, *430*, 4650–4665. [CrossRef] [PubMed]
33. Hughes, M.P.; Sawaya, M.R.; Boyer, D.R.; Goldschmidt, L.; Rodriguez, J.A.; Cascio, D.; Chong, L.; Gonen, T.; Eisenberg, D.S. Atomic structures of low-complexity protein segments reveal kinked beta sheets that assemble networks. *Science* **2018**, *359*, 698–701. [CrossRef] [PubMed]
34. Do Amaral, M.J.; Araujo, T.S.; Diaz, N.C.; Accornero, F.; Polycarpo, C.R.; Cordeiro, Y.; Cabral, K.M.S.; Almeida, M.S. Phase separation and disorder-to-order transition of human brain expressed X-linked 3 (hBEX3) in the presence of small fragments of tRNA. *J. Mol. Biol.* **2020**, *432*, 2319–2348. [CrossRef] [PubMed]
35. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **2013**, *41*, e121. [CrossRef] [PubMed]
36. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412–D419. [CrossRef]
37. Murray, D.T.; Kato, M.; Lin, Y.; Thurber, K.R.; Hung, I.; McKnight, S.L.; Tycko, R. Structure of FUS protein fibrils and its relevance to self-Assembly and phase separation of low-complexity domains. *Cell* **2017**, *171*, 615–627.e16. [CrossRef]
38. Birsa, N.; Ule, A.M.; Garone, M.G.; Tsang, B.; Mattedi, F.; Chong, P.A.; Humphrey, J.; Jarvis, S.; Pisiren, M.; Wilkins, O.G.; et al. FUS-ALS mutants alter FMRP phase separation equilibrium and impair protein translation. *Sci. Adv.* **2021**, *7*, eabf8660. [CrossRef]
39. Birsa, N.; Bentham, M.P.; Fratta, P. Cytoplasmic functions of TDP-43 and FUS and their role in ALS. *Semin. Cell Dev. Biol.* **2020**, *99*, 193–201. [CrossRef]
40. Protter, D.S.W.; Parker, R. Principles and properties of stress granules. *Trends Cell Biol.* **2016**, *26*, 668–679. [CrossRef]
41. Emmanouilidis, L.; Esteban-Hofer, L.; Damberger, F.F.; de Vries, T.; Nguyen, C.K.X.; Ibanez, L.F.; Mergenthal, S.; Klotzsch, E.; Yulikov, M.; Jeschke, G.; et al. NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation. *Nat. Chem. Biol.* **2021**, *17*, 608–614. [CrossRef] [PubMed]
42. Holehouse, A.S.; Das, R.K.; Ahad, J.N.; Richardson, M.O.; Pappu, R.V. CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **2017**, *112*, 16–21. [CrossRef] [PubMed]
43. King, O.D.; Gitler, A.D.; Shorter, J. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Res.* **2012**, *1462*, 61–80. [CrossRef]
44. Qamar, S.; Wang, G.; Randle, S.J.; Ruggeri, F.S.; Varela, J.A.; Lin, J.Q.; Phillips, E.C.; Miyashita, A.; Williams, D.; Strohl, F.; et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-pi interactions. *Cell* **2018**, *173*, 720–734. [CrossRef] [PubMed]
45. Krainer, G.; Welsh, T.J.; Joseph, J.A.; Espinosa, J.R.; Wittmann, S.; de Csillery, E.; Sridhar, A.; Toprakcioglu, Z.; Gudiskyte, G.; Czekalska, M.A.; et al. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nat. Commun.* **2021**, *12*, 1085. [CrossRef]
46. Murthy, A.C.; Dignon, G.L.; Kan, Y.; Zerze, G.H.; Parekh, S.H.; Mittal, J.; Fawzi, N.L. Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.* **2019**, *26*, 637–648. [CrossRef]
47. Loughlin, F.E.; Lukavsky, P.J.; Kazeeva, T.; Reber, S.; Hock, E.M.; Colombo, M.; Von Schroetter, C.; Pauli, P.; Clery, A.; Muhlemann, O.; et al. The solution structure of FUS bound to RNA reveals a bipartite mode of RNA recognition with both sequence and shape specificity. *Mol. Cell* **2019**, *73*, 490–504.e496. [CrossRef]
48. Maris, C.; Dominguez, C.; Allain, F.H. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **2005**, *272*, 2118–2131. [CrossRef]
49. Shazman, S.; Mandel-Gutfreund, Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **2008**, *4*, e1000146. [CrossRef]
50. Alba, M.M.; Laskowski, R.A.; Hancock, J.M. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **2002**, *18*, 672–678. [CrossRef]
51. Promponas, V.J.; Enright, A.J.; Tsoka, S.; Kreil, D.P.; Leroy, C.; Hamodrakas, S.; Sander, C.; Ouzounis, C.A. CAST: An iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* **2000**, *16*, 915–922. [CrossRef] [PubMed]
52. Harrison, P.M. fLPS: Fast discovery of compositional biases for the protein universe. *BMC Bioinform.* **2017**, *18*, 476. [CrossRef]
53. Wootton, J.C.; Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **1993**, *17*, 149–163. [CrossRef]

54. Bianchi, G.; Longhi, S.; Grandori, R.; Brocca, S. Relevance of electrostatic charges in compactness, aggregation, and phase separation of intrinsically disordered proteins. *Int. J. Mol. Sci.* **2020**, *21*, 6208. [CrossRef] [PubMed]
55. Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **2015**, *143*, 085101. [CrossRef]
56. Das, R.K.; Pappu, R.V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13392–13397. [CrossRef]
57. UniProt, C. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
58. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169. [CrossRef]
59. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
60. Riek, R. The three-dimensional structures of amyloids. *Cold Spring Harb. Perspect. Biol.* **2017**, *9*, a023572. [CrossRef]
61. Buchan, D.W.A.; Jones, D.T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **2019**, *47*, W402–W407. [CrossRef]
62. Wang, Z.; Zhao, F.; Peng, J.; Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **2011**, *11*, 3786–3792. [CrossRef] [PubMed]
63. Torrisi, M.; Kaleel, M.; Pollastri, G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci. Rep.* **2019**, *9*, 12374. [CrossRef] [PubMed]
64. Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.* **2018**, *39*, 2210–2216. [CrossRef] [PubMed]
65. Piovesan, D.; Walsh, I.; Minervini, G.; Tosatto, S.C.E. FELLS: Fast estimator of latent local structure. *Bioinformatics* **2017**, *33*, 1889–1891. [CrossRef]
66. Wang, S.; Li, W.; Liu, S.; Xu, J. RaptorX-Property: A web server for protein structure property prediction. *Nucleic Acids Res.* **2016**, *44*, W430–W435. [CrossRef] [PubMed]
67. Kaleel, M.; Torrisi, M.; Mooney, C.; Pollastri, G. PaleAle 5.0: Prediction of protein relative solvent accessibility by deep learning. *Amino Acids* **2019**, *51*, 1289–1296. [CrossRef]
68. Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics* **2019**, *35*, 2403–2410. [CrossRef]
69. Meszaros, B.; Erdos, G.; Dosztanyi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. [CrossRef]
70. Hanson, J.; Paliwal, K.K.; Litfin, T.; Zhou, Y. SPOT-Disorder2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genom. Proteom. Bioinform.* **2019**, *17*, 645–656. [CrossRef]
71. Jones, D.T.; Cozzetto, D. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **2015**, *31*, 857–863. [CrossRef] [PubMed]
72. Xue, B.; Dunbrack, R.L.; Williams, R.W.; Dunker, A.K.; Uversky, V.N. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **2010**, *1804*, 996–1010. [CrossRef] [PubMed]
73. Dosztanyi, Z.; Meszaros, B.; Simon, I. ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **2009**, *25*, 2745–2746. [CrossRef] [PubMed]
74. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **2017**, *13*, e1005324. [CrossRef]
75. Li, Y.; Hu, J.; Zhang, C.; Yu, D.J.; Zhang, Y. ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **2019**, *35*, 4647–4655. [CrossRef]
76. Singh, J.; Litfin, T.; Singh, J.; Paliwal, K.; Zhou, Y. SPOT-Contact-LM: Improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **2022**, *38*, 1888–1894. [CrossRef]