Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS

Qidong Wan (darrenwanxx@gmail.com)
eternbio https://orcid.org/0000-0003-4583-3105
Hao He
eternbio
Jidong Zhu
eternbio

Article

Keywords:

Posted Date: February 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2571470/v1

Abstract

Liquid-liquid phase separation (LLPS) is a process that underpins the formation of membrane compartments and regulates various biological processes in cells. Intrinsically disordered proteins and regions (IDPs/IDRs) play a significant role in LLPS as they are a class of proteins that undergo monomeric and heterotypic interactions, driving phase separation. Although many computational methods are available to study the sequences that determine phase separation, the quantitative amino-acid (AA) contribution remains poorly understood. To address this issue, we have developed BERTIG, a novel, interpretable deep learning framework that predicts the LLPS capability of IDRs with a high level of accuracy. The framework utilizes the Integrated Gradients (IG) algorithm and Bayesian optimization, while incorporating prediction probability (Proba), attribution score (AS), and model score (MS) to produce quantitative interpretations of both wild and mutated forms of IDPs. BERTIG has been shown to accurately identify and validate key AAs and motifs responsible for LLPS in disordered proteins, with performance comparable to experimental results and superior to other methods. Thus, BERTIG is a versatile, powerful, and interpretable model that will greatly enhance characteristics understanding of the increasing number of proteins, including prion-like proteins.

Introduction

The advancement of computational methods for interpreting phase separation of proteins from protein sequences has proceeded along two distinct yet complementary approaches. One approach focuses on the stickers-and-spacers framework[1–4], which takes into account the multivalence of interactions. The other approach focuses on the conformational entropy framework[5], which considers the free and binding states.

The stickers-and-spacers model, based on the a priori identification of stickers and spacers, can be utilized to describe branched or linear associative polymers, such as multivalent protein and RNA molecules. In the context of intrinsically disordered regions (IDRs), stickers are believed to be short linear motifs (SLiMs) of 1-10 residues, while spacers are the intervening residues in the IDR[2]. The open-source computational engine LASSI can be utilized to calculate full phase diagrams for coarse-grained representations of multivalent proteins on simple cubic lattices[1].

The stickers-and-spacers framework incorporates our understanding of molecular driving forces into thermodynamic or kinetic simulations of protein physics or statistical approximations thereof. However, this framework is limited to prion-like low-complexity domains (PLCDs) and it can be challenging to identify the stickers versus spacers, determine the strengths of different types of stickers, and assess the effects of spacers[6].

The conformational entropy framework, a complementary approach, is based on the idea that the droplet state is stabilized by the large conformational entropy resulting from nonspecific side-chain interactions upon binding[5,7,8]. This entropy can be predicted from the amino acid sequences. A tool named FuzDrop[5] was developed to predict the droplet-promoting propensity of proteins and their droplet-promoting profiles using the conformational entropy of their free states (probabilities of pD) and binding states (probabilities of pDD).

While the conformational entropy framework has a strong theoretical foundation, it has limitations in practice. Specifically, it fails to produce a quantitative value for the contribution of amino acids to liquid-liquid phase separation (LLPS), which restricts its utility for further biological applications. Additionally, both the stickers-and-spacers framework and the conformational entropy framework are unable to account for the effects of genetic variations on proteins.

Previous studies have generated several important questions that require further investigation: (1) How can the contribution of amino-acid-resolution to phase separation in disordered and ordered regions be evaluated? (2) What are the key regions and motifs that most significantly contribute to phase separation? (3) How can the effects of genetic variations on phase separation be assessed?

To address these questions, we have developed a new and innovative deep learning framework called BERTIG. BERTIG consists of two main components: BERT and IG (Integrated Gradients). BERT is an LLPS prediction model that is constructed using a pretrained-finetuned approach, while IG is a powerful quantitative interpretation model based on the integrated gradients algorithm. A key feature of BERTIG is the integration of prediction probability direction (Proba), attribution score (AS), and model score (MS), which makes the integrated gradients algorithm accurate and reliable.

We demonstrate the capabilities of BERTIG on two related tasks: the LLPS prediction of IDR proteins and the amino-acid-resolution interpretation of proteins. BERTIG predicts the LLPS of IDRs with higher AUC and AP compared to existing computational methods. When applied to datasets of transcription factor (TF) proteins, BERTIG is able to identify the key amino acids and motifs driving phase separation. We also apply BERTIG to mutated proteins, specifically TDP-43 and FUS, to derive total attribute scores for phase separation that are in close agreement with experimental results from published studies.

Results

The BERTIG framework

The results of the BERTIG framework demonstrate significant improvement in the accuracy of LLPS prediction and amino-acid-resolution attribution interpretation. This is achieved by incorporating the state-of-the-art pretrained-finetuned language model and the IG method, which is restricted by the direction of Proba, AS, and MS.

Furthermore, the results highlight the importance of properly exploring the parameters of steps and batch size in the approximation of the integral, which is crucial for accurate amino-acid-resolution attribution interpretation. This aspect has received limited attention in the literature.

The BERTIG framework predicts both LLPS and attribution for a protein using the primary amino acid sequence as input. The complete architecture and training procedure are outlined in Fig. 1a.

The BERTIG framework consists of two key components. Firstly, a pre-trained BERT model (ProtBert-BFD[10] with 30 blocks) is fine-tuned (Fig. 1a) using protein sequences from four public datasets (LLPSdb[11], PhaSePro[12], FuzDrop[5], OpenCell[13]) (training data in Supplementary Data 1) as inputs. This stage outputs a prediction probability for the likelihood of liquid-liquid phase separation (LLPS). In the second stage, Bayesian optimization is used to determine the optimal parameters for the Integrated Gradients (IG) algorithm, which calculates an amino-acid-resolved attribution score for the LLPS prediction. The results are expressed as the loss score (denoted as error_score_neg_reciprocal) and delta (the difference between the model score and attribution score).

Key innovations in the framework include well-defined loss function restricted by the direction of 3 metrics (Proba, AS, MS), the usage of Bayesian optimization to explore the best step and batch size, and the SOTA paradigm of pretrained-finetuned language model. We reinforce the notion of iterative parameters refinement of steps and batch size that contributes markedly to accuracy of AA attribution and the integration of 3 metrics to the reliability of the interpretation.

Bayesian Optimization Of Ig

The central component of the BERTIG framework is the integration of three indicators: Proba, AS, and MS. To improve the accuracy and reliability of interpretation, these indicators are binarized to either 1 or -1. The binarization process is performed as follows: if Proba is greater than 0.5, it will be binarized to 1; otherwise, -1. If AS or MS is over 0, it is assigned the value of 1, denoted as ASb or MSb; otherwise, -1. The direction coef (DC) is then determined by checking if Proba, ASb, and MSb have the same direction. DC is defined as 1 if Proba = ASb = MSb and − 1 otherwise.

The Integrated Gradients (IG) approach, which is employed in the BERTIG framework, is defined as the path integral of gradients along a straight-line path from the baseline x' to the input x (Fig. 1b; see Methods for details). This path integral can

be approximated using Riemann Sum or Gauss Legendre quadrature rule. To optimize the steps and batch size in the framework, a well-defined loss function is employed, which is defined as

$$L = DC \times \frac{-1}{|AS - MS|}$$

where the parameters are described as above. As a guideline, we suggest a cut-off value of L = -10 for mutated proteins and L = -5 for natural proteins.

To understand how BERTIG interpreted the proteins, we estimate the relative importance of key components by evaluating several ablation models on 21 mutated proteins (Supplementary Data 2; Supplementary Figs. 2–4) in a fp16 half-precision setting with the label-agnostic fashion. Our results, presented in Fig. 1c, indicated that a combination of multiple factors contribute to the accuracy of the model. Our analysis of 12 mutated TDP-43 proteins demonstrated that accuracy improved progressively from the baseline model to the Step-Batch-AS_MS-Proba model, emphasizing the need to explore both step size and batch size along with AS, MS, and Proba. For the 9 mutated FUS proteins, maximum accuracy was achieved by the Step-AS_MS and Step-AS_MS-Proba models, highlighting the significance of step size and suggesting that the default batch size may be suitable in some cases. The comparison of Step-AS_MS with Step-AS_MS-Proba showed an increase in accuracy from 0.17 to 0.25 for TDP-43 proteins, indicating the importance of the direction of Proba. The comparison of Step-AS_MS-Proba with Step-Batch-AS_MS-Proba demonstrated an increase in accuracy from 0.25 to 0.50, which emphasized the significance of batch size. For the FUS proteins, the comparison of Step-AS_MS_size with Step-AS_MS showed an increase in accuracy from 0.22 to 0.78, revealing the crucial role played by the direction of AS and MS. In conclusion, the direction of Proba, AS, and MS, as well as step size and batch size, play a crucial role in the interpretation of proteins by the BERTIG framework. Further details on each ablation model are provided in the Methods section.

High Accuracy Of Llps Prediction

To construct a Liquid-Liquid Phase Separation (LLPS) model, we assembled a total of 4,583 Intrinsically Disordered Proteins (IDPs) from four datasets, LLPSdb, PhaSePro, FuzDrop, and OpenCell (see distribution in Supplementary Fig. 1). The proportion of IDPs undergoing LLPS was 1211 to 3372 (LLPS: non LLPS = 1211: 3372). A small curation was performed to determine if a protein could undergo phase separation independently (see Methods for details of datasets process; see Supplementary Fig. 1 for the datasets visualization).

The BERT architecture was trained with supervised learning on the LLPS dataset and its performance was evaluated using AUC for hyperparameter exploration. The best model achieved an AUC of 0.95 (as shown in Fig. 2a1) and an average precision (AP) of 0.86 on the test set (20% of the LLPS dataset) (as shown in Fig. 2a2).

In comparison with other Phase-Separation Predictors, such as DeePhase[14], PSAP[15], LLPhyScore[16], and PScore[17], the performance of our BERTIG model was found to be superior on two evaluation sets, LLPSDB-v2 (Figs. 2b1 and 2b2) and PhaSePro (Figs. 2c1 and 2c2). Notably, the first-generation predictor, PScore, which is based solely on planar pi–pi interactions, showed impressive performance, highlighting the crucial role of pi–pi interactions in IDRs phase separation. Our results also indicated that DeePhase, which combines physical features and word2vec features and is ranked second, showed comparable AUC and AP values with BERTIG, as a phase-separation predictor based on protein sequence embeddings. DeePhase was evaluated using a pre-trained word2vec model on the Swiss-Prot database, creating 200-dimensional embedding vectors for every 3-gram. The comparable results of BERTIG and DeePhase suggest the effectiveness of protein sequence embeddings for predicting LLPS. However, BERTIG outperformed Dee Phase, highlighting the advantage of using a massive pre-trained language model. Additionally, the other two models, which rely on traditional methods of extracting physiochemical features, also showed good performance.

In conclusion, the comparison of models highlights that phase separation can be predicted based purely on protein sequence embeddings, and the pretrained-finetuned language model could be served as a general framework.

Aa Interpretation

The performance of BERTIG in interpreting the contribution of each amino acid (AA) to liquid-liquid phase separation (LLPS) was tested using two well-studied proteins, TDP-43 and FUS.

TDP-43 is comprised of an N-terminal domain (NTD; residues 1-103), two RNA recognition motifs (RRM1 and RRM2; residues 104–176 and residues 191–262), and a C-terminal domain (CTD; residues 274–414) [18] (Fig. 3a1). The NTD contains a ubiquitin-like fold with one alpha-helix and six beta-sheets that promotes TDP-43 self-oligomerization in a concentration-dependent manner[19,20]. The C-terminus of TDP-43 has prion-like domains that are known to be involved in TDP-43 phase separation and aggregation[21–23].

The results of the BERTIG analysis showed that two major regions (1-150 and 250–400) of TDP-43 played a positive role in LLPS (Fig. 3b1). These regions align with the main regions of linear interacting peptides (LIP) as identified in the Mobidb database[24] (positions 80–102 and 263–414) (Fig. 3c1), and the main regions of binding modes (positions 376–395 and 403–414) (Fig. 3d1). This supports the conclusion that both the NTD and CTD of TDP-43 contribute positively to LLPS and is consistent with experimental results[25–28].

Note

LIP (Residues interacting with another molecule that preserve structural linearity in the bound state. Also called short linear motifs (SLiMs), molecular recognition features (MoRFs), protean segments (ProS)).

FUS is composed of 526 amino acids and several conserved domains, including the prion-like domain (PrLD), an RNA-recognition motif (RRM), two arginine-glycine rich regions (RGGs), a zinc finger (ZnF) domain, and a PY-nuclear localization signal[18] (PY-NLS) (Fig. 3a2). These domains govern the assembly of FUS through inter- and intramolecular interactions, with the PrLD forming homotypic cross-β interactions[30–32], the R residues in RGG domains forming intramolecular interactions[33], and the Y residues in the PrLD forming intermolecular interactions[34]. The PY-NLS interacts with importins, which regulate FUS condensation[35,36].

According to the BERTIG interpretation, two major regions (140–270 and 420–500) were identified as making important contributions to LLPS (Fig. 3b2). This conclusion is supported by consensus from the mobidb database, which identifies the 4 main regions in linear interacting peptide (LIP) as positions 34–150, 285–370, 432–443, and 507–526 (Fig. 3c2), and the two main regions in binding mode as positions 1-277 and 391–506 (Fig. 3d2). The results of the BERTIG analysis are highly consistent with the experimental results[40–42] reported and the consensus from the mobidb database. Particularly, the AAs at positions 150–250 in the pi-pi interaction region showed greater contributions to LLPS[37] (Fig. 3b2).

Key Aas For Llps

To determine if the crucial amino acids (AAs) responsible for phase separation could be identified, we analyzed AA sequences of motifs (sequences of more than 2 consecutive AAs with either positive or negative attributes) in different proteins. Our first examination focused on the well-studied PLCDs of 112 proteins. We analyzed the AA compositions of paired groups with opposite attributions and applied loss threshold filtering, which resulted in a remaining 90 proteins (Supplementary Data 3; Supplementary Figs. 6, 10, 11). In the PLCD regions (Fig. 4a, shown in green), the most influential AAs for LLPS were YHWNG, with YWNG being validated in previous literature[6,33]. In contrast, for the entire protein (Fig. 4a, shown in orange), the top 5 AAs were YWFGV, with a slight variation where V replaced H.

Then, we analyzed 308 TF proteins (Fig. 4a blue color) assembled from 3 sources (see details of data process in Methods; Supplementary Data 4; Supplementary Figs. 5–6, 12–13): human transcription factor effector domain[41] (Supplementary Data

5; Supplementary Figs. 8–9), Low-complexity Aromatic-Rich Kinked Segments (LARKS)[30] (Supplementary Data 6; Supplementary Figs. 8–9) and the eukaryotic linear motif (ELM)[42] (Supplementary Data 7; Supplementary Figs. 8–9).

The top 5 AAs were found to be YWKNC (Fig. 4a blue color) in TF proteins, which differed significantly from the results for the prion-like proteins. Y was, however, always the most important AA for LLPS.

In comparison to the paired groups (positive score vs negative score) (Fig. 4b) in the PLCDs region, the top 5 AAs found to be favorable for phase separation were GQSNA in terms of total number and GQYAS in terms of total score. V was found to be the most unfavorable for phase separation ($p < 0.05$). Meanwhile, for the entire set of 90 prion-like proteins, the top 5 significant AAs were GSPQT in total number and GSPTQ in total score (Extended Data Fig. 1a). When comparing the 90 prion-like proteins with the 308 TF proteins (Extended Data Fig. 1b), the top 5 significant AAs in TF proteins were found to be PASGT in both total number and score, displaying a high level of consistency (common in PSGT).

Our findings provide important insights into the AAs driving LLPS across a diverse range of protein classes.

Motif Discovery

To further evaluate the important functional regions for liquid-liquid phase separation (LLPS), we employed BERTIG to analyze 308 transcription factor (TF) proteins (Supplementary Data 4; Supplementary Fig. 7c) with confirmed motifs. The results showed that a remarkable 85% of the motifs were consistent. Our analysis revealed MEF2D to be the top 1 gene in terms of common motifs (Fig. 5a).

In the case of MEF2D, the newly predicted motifs were the following (Fig. 5b): {'19–21': 'VTF', '26–30': 'FGLMK', '32–35': 'AYEL', '37–38': 'VL', '42–50': 'EIALIIFNH', '52–55': 'NKLF', '57–58': 'YA'}. As Fig. 5c showed, the regions 38–49 was LIP that could interact with another molecule predicted by ANCHOR[43], which was also in our newly predicted motifs. The region 10–57, highly consistent with our newly predicted motifs regions, was the domain of “transcription factor, MADS-box” (Fig. 5d), that may play an import role in LLPS.

In the case of the ARX gene, our analysis revealed that the region 60–350 made a great contribution to the phase separation (Extended Data Fig. 2a), which was in consistency with LIP (Extended Data Fig. 2b).

Overall, these results demonstrate the accuracy of BERTIG in predicting new motifs and important functional regions for LLPS in TF proteins.

Additionally, we identified several LLPS-favorable motifs that were statistically significant, such as PP, AA, SS, GG, KK, QQ, RR, TT, AAAA, PPPP, EEEE, QQQQ, among others. The top 10 significant ($p < 0.05$) motifs of different lengths for the 308 TF proteins are shown in Fig. 5e:

For 2 amino-acid lengths: PP, AA, AP, SS, PA, SP, PG, PS, GG, TS;

For 3 amino-acid lengths: PPP, AAA, SSS, PAA, PPA, AAG, AAP, APP, PSS, PAP;

For 4 amino-acid lengths: AAAA, PPPP, CGKA, GGAG, AAAG, AAGG, APPA, APPP, PPPA, QNRR.

In the case of the 90 PLCDs proteins, the top 10 significant motifs, when only the PLCDs regions were considered, are listed in Extended Data Fig. 3a:

2 amino-acid lengths: YN, GN, NG, QP, TA, NY, FG, PS, AS, MN;

3 amino-acid lengths: GGY.

However, the top 10 significant motifs when considering the whole PLCDs proteins are different, as shown in Extended Data Fig. 3b:

2 amino-acid lengths: GG, SS, TS, AS, AP, GP, GA, PP, PG, SP;

3 amino-acid lengths: GGG, GGY, GGR, AAA, NSS, PPG, PSS, PTT, QQA, TGS.

It can be observed that only AS and GGY were the same significant motifs when comparing PLCDs regions to PLCDs proteins. On the other hand, there were 9 common significant motifs between TF proteins and PLCDs proteins: GG, SS, TS, AP, PP, PG, SP, AAA, PSS.

To summarize, the comparison of 308 TF proteins and 90 PLCDs revealed that a high variability in the distribution of motifs may be a hallmark of the diversity of proteins.

Effects Of Mutations On Llps

In this study, we focused on the examination of two well-studied proteins, TDP-43 and FUS. TDP-43 fragments containing the C-terminal PLCDs are known to be a pathological hallmark of ALS and FTD, with the majority of TDP-43 mutations associated with ALS located within the PLCDs [44,45]. Mutations in the nuclear localization sequence (NLS) in the NTD have also been shown to cause cytoplasmic localization and aggregation of TDP43[46,47]. Similarly, mutations in FUS have been linked to both sporadic and familial cases of ALS and are associated with the accumulation of FUS-positive inclusions in the cytoplasm of degenerating neurons and glia, as well as decreased nuclear FUS[48–50].

In order to predict the phase behavior of these mutated proteins, a set of validated mutations were assembled, curated, and analyzed. For TDP-43, 12 mutated proteins were obtained from the literature: G294V[51], G298S[52], N319G[53], P320G[53], A321G[22], A321V[22], A324G[53], A326P[54], Q327G[53], G335A[54], M337P[54], M337V[55], with only 3 of these mutations (A321G, A326P, and M337P) being weaker than the natural protein according to the published literature. For FUS, 9 mutated proteins were tested: G156E[59], R216C[52], G230C[60], R244C[59], G399V[61], R521C[52], R521G[61], R521H[52], P525L[62], with only 1 (G399V) being weaker than the natural protein.

The ability of LLPS to accurately predict the phase behavior of these mutated proteins was assessed based on the attribution score (AS) in a fp32 setting with double-float precision. Results showed that 11 of the 12 mutated TDP-43 proteins were correctly interpreted (A326P misinterpreted), with an accuracy of 92% (Fig. 6a1). Similarly, 7 of the 9 mutated FUS proteins were correctly interpreted (G156E and R521C misinterpreted), with an accuracy of 78% (Fig. 6a2).

The assessment of the effect of mutated proteins on LLPS requires a comprehensive evaluation of the entire protein and its functional regions, rather than focusing solely on the mutation loci. In the case of TDP-43, the impact of the A321 residue was thoroughly analyzed (Fig .6b1). The 321–340 region is a crucial helical structure that contributes to LLPS through self-interaction. Research has shown that the A321G mutation decreases phase separation, resulting in decreased helical population and shorter helical segments[22]. Our results, as shown in Fig. 6c1, confirm this conclusion. The A321G mutated TDP-43 (AS = 1.5590) was weaker than the natural TDP-43 protein (AS = 1.9313), and the surrounding AA scores were mostly negative, indicating a reduction in LLPS. On the other hand, the A321V mutation (AS = 2.6680) was found to enhance phase separation, with bigger AA scores than the natural TDP-43 (AS = 1.9313), which indicates increased hydrophobicity[22].

Similarly, the impact of the R521 residue in the PY-NLS of the FUS protein was studied. The R521 residue is crucial in the binding of FUS to the chaperone protein Kapβ2, which facilitates its localization within the nucleus[56]. Mutations R521G and R521H result in altered conformation, decreased binding to RNA, and formation of large condensates, which can lead to aberrant trafficking and cytoplasmic retention. Our results, as shown in Fig. 6c2, reveal that both R521G (AS = 0.6485) and R521H (AS = 1.0399) mutated FUS proteins were stronger than the natural FUS protein (AS = 0.6306), but with different

patterns. In R521G mutated FUS, AA scores were generally small, with more positive values emerging, whereas R521H had more big positive values.

These results indicate that BERTIG can effectively predict the phase separation ability of mutated proteins.

Interpretations Of Different Models

To compare the ability of different models in interpreting of mutated proteins, we selected the M337P-mutated TDP-43 and the G230C-mutated FUS. These mutations were chosen to demonstrate the differing effects they have on the phase behavior of the proteins.

The M337P mutation in TDP-43 weakens its ability to undergo liquid-liquid phase separation due to disruption of the CR helix-helix interactions and the extension of the helical region. This results in a nearly complete loss of TDP-43's autoregulatory capacity[54].

In contrast, the G230C mutation in FUS protein improves its ability to undergo liquid-liquid phase separation. This is because glycine, which is classified as a "spacer," plays an important role in controlling the fluidity of FUS condensates[33]. The G230C mutant displays a range of interactions with RNA and exhibits significantly more dynamic Förster resonance energy transfer (FRET) fluctuations[57].

We then compared the predictions of BERTIG with those of other models, namely FuzDrop[5], LLPhyScore[16], and PScore[17].

As depicted in Fig. 7, BERTIG demonstrated the greatest differentiation power and model interpretation ability. The 321–340 region of TDP-43 is a crucial helical structure for LLPS and the M337P mutation caused most of the scores in the region predicted by BERTIG to shift from positive to negative (Fig. 7a1). This shift was not observed in the other models (FuzDrop in Fig. 7b1, LLPhyScore in Extended Data Fig. 4a1, and PScore in Extended Data Fig. 4b1).

Similarly, the G230C mutation in FUS enhanced its LLPS ability, as BERTIG showed the PLCDs (residues 1-239) to have more positive scores compared to the natural FUS, particularly in the main PLCDs regions with residues 1 to 100 (Fig. 7a2). However, no such improvement was observed in the other models (FuzDrop in Fig. 7b2, LLPhyScore in Extended Data Fig. 4a2, and PScore in Extended Data Fig. 4b2).

While the sticker-spacer framework is also useful for interpreting IDRs, it is only applicable to PLCDs and it is challenging to identify stickers and spacers, as well as their strengths and effects. Thus, no comparison to BERTIG was performed in this study.

In conclusion, BERTIG is the most superior tool currently available for the distinguishability of quantitative AA interpretation.

Discussion

Intrinsically disordered proteins (IDPs) play a crucial role in biomolecular condensation[7, 58–60]. However, detecting IDPs through comparative modeling is challenging due to the lack of knowledge of their invariants, or the set of function-preserving sequence perturbations. To address this challenge, we combined a prediction model based on BERT with a novel interpretation method, direction-restricted IG, to uncover the quantitative AA contribution and key motifs involved in phase separation. While IG has been used in the interpretation of protein-DNA/RNA binding sites[61], protein-ligand binding[62] and protein sequence–function relationships[63], no optimization was conducted before, thus our work represents the first optimization of the IG algorithm for interpretable protein predictions.

We optimized the IG algorithm by introducing a direction-restriction method to ensure that the interpretation results are meaningful. The interpretation would not be valid if the attribution score of a protein is positive while the model score is negative, regardless of the small delta value. Likewise, if the prediction probability is over 0.5, but the attribution score is

negative or the model score is negative, the interpretation would also be invalid. To address these issues, we utilized Bayesian optimization to find the best parameters for the steps and batch size for a more accurate and reliable interpretation.

Our results showed that BERTIG outperforms feature-based methods and word2vector-based deep learning models in predicting the liquid-liquid phase separation (LLPS) of proteins. This is demonstrated through a comprehensive comparison of different models on well-established LLPS datasets. Overall, our findings highlight the importance of optimizing interpretable algorithms, such as IG, for accurate and reliable predictions of IDPs. Hence, the optimization of IG is undoubtedly necessary. Generally, the optimization of any of the interpretable algorithms that returns delta is important and inevitable, such as DeepLift, GradientShap, DeepLiftShap and so on.

In terms of interpretation, our results showed, through a thorough comparison of various models on mutated TDP-43 and FUS proteins, that BERTIG was the only tool capable of deciphering the impact of mutations on the liquid-liquid phase separation at the amino acid level. The existing sticker-spacer framework is limited to prion-like proteins[3,6,64] only, while BERTIG has the potential to be applied to a wider range of intrinsically disordered proteins and facilitate the understanding of their phase separation behavior. Furthermore, BERTIG offers a more insightful interpretation of mutated proteins compared to the sticker-spacer framework as it considers the interactions between stickers in a more nuanced manner, whereas the latter treats such interactions in a general sense.

Our findings highlight the diversity of phase separation mechanisms among different types of proteins and demonstrate the versatility of BERTIG as a tool for providing insights into a diverse set of proteins. This will significantly accelerate the study of liquid-liquid phase separation in an increasing number of proteins.

Methods

BERTIG finetuned dataset for LLPS prediction

All data in the present study were downloaded from public datasets. There were totally 5777 proteins (LLPS : non LLPS = 1333: 4444) (Dataset S1) assembled from 4 sources separately: 109 proteins in the PhaSePro database (https://phasepro.elte.hu), 589 proteins in the LLPSDB-v2 dataset (http://bio-comp.org.cn/llpsdb), 882 proteins labeled manually in the OpenCell dataset[13] (https://opencell.czbiohub.org/), 4197 proteins in the FuzDrop dataset from the paper[5].

A small curation was performed to determine if a protein could undergo phase separation independently. For the PhaSePro dataset, only spontaneous phasing or DNA/RNA-dependent proteins were included in the positive set. In the LLPSdb-v2 set, only one protein system per item was included. The LLPSDB-v2 dataset was filtered based on both in vivo/in cell and one component experiments. The OpenCell dataset is a human protein localization resource generated from 1,311 CRISPR-edited cell lines harboring fluorescent tags. In accordance to the original annotation of “nuclear punctae”, all the 1311 raw cell images were analyzed with the software of CellProfiler[65], and labelled with “positive”, “negative” and “unknown”. Only 882 “positive” and “negative” proteins were retained. The FuzDrop dataset included the entire 445 LLPS and 3911 non-LLPS proteins. All the 4 datasets were merged and then de-duplicated according to the sequence based on the order of priority: PhaSePro > LLPSDB > FuzDrop > OpenCell. Finally, we constructed the most comprehensive protein phase-separation dataset from nearly all the main data sources of LLPS.

All the proteins were then filtered out by the cut-off 0.05 in IDR score that originated from the mobidb database[24] (https://mobidb.bio.unipd.it/) based on the order of priority: curated-disorder-priority > prediction-disorder-alphafold > prediction-disorder-mobidb_lite > 1 - homology-domain-pfam. Please take care that if the IDR score > 1 - homology-domain-pfam, the final IDR score will shift to the last one. We filled the IDR score with 1.0 for the mutated proteins that did not exist in the mobidb database. Finally, only 4583 proteins (LLPS: non-LLPS = 1211: 3372) were left to construct the models.

These 4583 training proteins were randomly split into training, validation, and testing sets in a ratio of 0.64:0.16:0.20 (Dataset S1).

Bertig Dataset For Interpretation

In this study, three datasets were utilized for interpretation purposes: (1) a collection of 12 mutated TDP-43 proteins and 9 mutated FUS proteins sourced from the literature (Dataset S2), (2) 112 prion-like proteins from the Iglesias database[66], and (3) a merged dataset of 385 TF proteins sourced from three different databases, including human transcription factor effector domains[41], Low-complexity Aromatic-Rich Kinked Segments (LARKS)[30] and the eukaryotic linear motif (ELM)[42] .

The 21 mutated proteins were carefully curated and verified through a manual process. The set of 112 prion-like proteins was sourced from Iglesias and filtered by a loss threshold of -5, resulting in 90 proteins (Dataset S3). The merged dataset of 385 TF proteins was filtered by a loss threshold of -5, resulting in 308 proteins (Dataset S4).

The human transcription factor effector domain database contains 924 effector domains across 594 human TFs, which are classified into activator domains (ADs), repressor domains (RDs), and bifunctional (Bif) domains. The database was filtered by activity and confidence level, resulting in 219 proteins, which were further filtered by a loss threshold of -5, resulting in 208 proteins (Dataset S5). The Low-complexity Aromatic-Rich Kinked Segments dataset consists of 400 functional regions characterized by the formation of kinked β-sheets. The dataset was filtered by TF, resulting in 59 proteins, which were further filtered by a loss threshold of -5, resulting in 56 proteins (Dataset S6). The Eukaryotic Linear Motif database provides a comprehensive repository of manually curated and experimentally validated short linear motifs. The latest version of the database consists of 2390 proteins and was filtered by TF, resulting in 188 proteins, which were further filtered by a loss threshold of -5, resulting in 172 proteins (Dataset S7).

Bertig Framework

We developed a finetuned model for the prediction of Liquid-Liquid Phase Separation (LLPS) from the pre-trained BERT model named ProtBert-BFD[10] with 30 blocks using the four public datasets (LLPSdb[11], PhaSePro[12], FuzDrop[5], OpenCell[13]). This model takes protein sequences as input. The HuggingFace package (version 4.20.1) was utilized to finetune the model, with parameters set to a maximum sequence length of 512 amino acids, a learning rate of 5e-5, a batch size of 8, gradient accumulation steps of 20, and a weight decay of 0.01. The model was trained on a single Nvidia Tesla V100 GPU with 32 GB graphics memory, using an early stop strategy where the training would terminate if the validation loss did not decrease for 10 consecutive epochs. The performance of the model was evaluated using the Area Under the Curve (AUC) metric, computed using the scikit-learn package (version 1.1.1).

To perform amino-acid-wise interpretation, we adopted the Integrated Gradients (IG) algorithm. IG is an axiomatically justified model that assigns an importance score to each input feature by approximating the integral of gradients of the model's output. The method calculates the path integral of the gradients along the straight-line path from the baseline x' to the input x:

$$\text{IntegratedGrads}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F(x) along the i^{th} dimension, and α is the scaling coefficient. x_i indicates the input protein sequence, $x_i\prime$ indicates the baseline filled with zero for the corresponding protein sequence. The integral of integrated gradients can be efficiently approximated via a summation using a Riemann Sum or Gauss Legendre quadrature rule.

To ensure accurate and reliable interpretation, the direction of prediction probability (proba), attribution score (AS), and model score (MS) were included in the IG algorithm. The total approximated integrated gradients are denoted as AS, while the total true integrated gradients are denoted as MS. The direction coefficient (DC) was defined as 1 if the direction of Proba, AS, MS is the same, and − 1 otherwise. The binarized process was performed such that when the proba was greater than 0.5, it was

binarized as 1, and − 1 otherwise. The same was applied to AS and MS, denoted as ASb and MSb, respectively. The final result was defined as DC = 1 if Proba = ASb = MSb, and − 1 otherwise. The loss function (L) for optimization of steps and batch size was defined as

$$L = DC \times \frac{-1}{|AS - MS|}$$

The point to note in the formula is that AS and MS are both the original scores, not the binarized values.

In practice, we used the Captum package (version 0.5.0) on a single Nvidia Tesla A100 GPU of 80 GB graphics memory capacity to make the interpretations of amino acids. Our results showed that steps between 50 and 2000 and batch size between 2 and 50 were sufficient to approximate the integral. Based on our experience, we recommend a cut-off value of -10 for the mutated proteins and − 5 for the natural proteins. We also used the HyperOpt package (version 0.2.7) to automate the search for optimal hyperparameter configuration using Bayesian Optimization and the Sequential Model-Based Global Optimization (SMBO) methodology. The HyperOpt package is an open-source python package that uses the Tree-based Parzen Estimators (TPE) algorithm to select the optimal hyperparameters that maximize a user-defined objective function. By defining the functional form and bounds of each hyperparameter, the TPE algorithm efficiently searches through the complex hyperspace to find the optimums.

Ablation Test

To understand the workings of BERTIG, we trained and evaluated various ablation models of 21 mutated proteins (Supplementary Data 2) in a label-agnostic manner on an Nvidia A100 GPU with 80 GB of graphics memory. The following ablation models were used to estimate the relative importance of key components of the architecture:

1. Baseline Model: This model uses the default step size of 50 and batch size equal to the protein length as specified in the Captum package. All other ablation models are compared to this baseline.
2. Step-AS_MS-Size Model: This model was trained with step sizes ranging from 50 to 2000, without considering the direction of AS and MS, and using the default batch size of the protein length.
3. Step-AS_MS Model: This model was trained with step sizes ranging from 50 to 2000 and includes the direction of AS and MS in the loss function. The batch size used is the default value.
4. Step-AS_MS-Proba Model: This model was trained with step sizes ranging from 50 to 2000 and includes the direction of AS, MS, and Proba in the loss function. The batch size used is the default value.
5. Step-Batch-AS_MS-Proba Model: This model was trained with step sizes ranging from 50 to 2000 and batch sizes ranging from 5 to 50, and includes the direction of AS, MS, and Proba in the loss function. Both the step size and batch size were explored in this model.

Comparison Between Bertig And Other Models On Llps Prediction And Interpretation

The PhaSePro and LLPSDB-v2 datasets were employed to evaluate various models for their capacity in predicting liquid-liquid phase separation (LLPS) of proteins. The PScore model, which was developed with the aim of exploring the relationship between planar protein-protein contacts and phase separation, can only be applied to protein sequences that are longer than 140 amino acids. As a result, only 109 proteins in PhaSePro (Dataset S8) and 498 proteins in LLPSDB-v2 (Dataset S9) that met the length requirement were considered.

The PScore[17] model: It provides a single score per protein sequence and has been validated to accurately differentiate between known phase separating proteins and other proteomic sequences. The authors trained a statistical potential for predicting pi-contact frequency in proteins using the non-redundant crystal structure subset of the PDB (80% of 17388 proteins used for

training and 20% for testing). The final predictor operates by averaging frequencies of sp2 groups in specific sequence contexts and comparing them to distributions of sp2 groups with the same sequence identity. The authors also developed a predictor for phase separation propensity of a protein sequence based on the pi-contact predictor and a set of 11 proteins known to phase separate in vitro. The phase separation predictor was trained using a stochastic optimization process to maximize the score difference between the lowest scoring member of the 11-member training set and the highest scoring 1% of the PDB training set. The AUC values for the predictor were estimated using bootstrap with 10,000 iterations against the test and human sets.

The DeePhase[14] model: It is a more comprehensive approach, which predicts LLPS by combining physical features and word2vec features, resulting in 200-dimensional embedding vectors for every 3-gram. The pre-training was performed on the Swiss-Prot database using 3-grams as words and a context window size of 25. The skip-gram pretraining procedure was used with the gensim library and created 200-dimensional embedding vectors for every 3-gram. The final 200-dimensional protein embeddings were obtained by summing all the constituent 3-gram embeddings. The classifiers were built using the Python scikit-learn package with default parameters and no hyperparameter tuning was performed. The dataset was split into a training and validation set in a 1:4 ratio and 25-fold cross-validation was used to estimate the performance of the model. The final prediction score was calculated as the sum of the probability of the sequence belonging to the LLPS + and half of the probability of it belonging to the LLPS − dataset.

The PSAP[15] model: The model used literature curation to determine a total of 90 high-confident phase separating proteins, which were then subjected to various sequence analysis techniques to extract various features. These features were then used to train a Random Forest Classifier to predict the score of each protein in the proteome. The performance of the classifier was evaluated using various metrics, including the area under curve (AUC) of the receiving operating characteristics (ROC) and precision and recall. The results were based on 10-fold cross-validation and the final model was trained on the full training set.

The LLPhyScore[16] model: The model, a novel predictor of IDR-driven phase separation, is based on a comprehensive set of physical interaction features and is trained on a curated set of phase-separation-driving proteins with various negative training sets, including the PDB and human proteome. The features included Pi-Pi contacts, hydrogen bonding terms, water/carbon contact counts, secondary structure, disorder, charge, cation-pi, and kinked beta. This model outputs residue-level scores in a protein sequence, which are smoothed over a window of 50 neighboring residues. Further details can be found in the original paper.

The FuzDrop[5] model: This model is designed to predict droplet-promoting regions and proteins that undergo spontaneous phase separation, based on the idea that the droplet state is stabilized by the large conformational entropy resulting from nonspecific side-chain interactions. FuzDrop outputs the probability of each residue being involved in spontaneous phase separation using a binary logistic model. Further details can be found in the original paper.

Statistical analysis

In the analysis of key amino acids and motifs discovery, we established a criterion for determining an appropriate region of a protein by requiring that it contain more than two consecutive amino acid scores greater than 0.02 or less than − 0.025, which is the default value in the Captum package (Captum version 0.5).

The Scipy (version 1.9.0) was utilized to compare changes in each residue and motif in two paired groups (plus and minus) for statistical analysis. The plus group represents the attribution score of each amino acid that is greater than 0.02, and the minus group represents the attribution score of each amino acid that is less than − 0.025. The scores for each of the 20 amino acids were tallied separately, and the scores were summed in the corresponding regions of the protein, both in the plus and minus groups. If the normal distribution was confirmed (as determined by the Shapiro-Wilk test), a two-tailed, paired T-test was performed, otherwise, the Wilcoxon signed-rank test was used. $P \leq 0.05$ was considered statistically significant.

Declarations

Acknowledgements

We thank G. Zhu, J. Xie, Y. Yan, Z. Gao., H. Yan for their feedback and operational support at various stages of this project. Thanks also to the SHENZHEN BKUNYUN Cloud team and the Ali Cloud team for their help with computing resources.

Author contributions

Q.W., J.Z., and H.H. conceived the project. Q.W. developed the BERTIG framework. Q.W. performed most of the data analysis. H.H. performed the OpenCell dataset analysis. Q.W. wrote the manuscript.

Competing interests

J.Z. is a co-founder of Etern Biopharma. Q.W. and H.H. are employees of Etern Biopharma.

References

1. Choi, J.-M., Dar, F. & Pappu, R. V. LASSI: A lattice model for simulating phase transitions of multivalent proteins. PLOS Comput. Biol. **15**, e1007028 (2019).
2. Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. 27 (2020) doi:10.1146/annurev-biophys-121219-081629.
3. Alshareedah, I., Moosa, M. M., Pham, M., Potoyan, D. A. & Banerjee, P. R. Programmable viscoelasticity in protein-RNA condensates with disordered sticker-spacer polypeptides. Nat. Commun. **12**, 1–14 (2021).
4. Ranganathan, S. & Shakhnovich, E. I. Dynamic metastable long-living droplets formed by sticker-spacer proteins. eLife **9**, e56159 (2020).
5. Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M. & Vendruscolo, M. Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci.* **117**, 33254–33262 (2020).
6. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nat. Chem. **14**, 196–207 (2022).
7. Vendruscolo, M. & Fuxreiter, M. Sequence Determinants of the Aggregation of Proteins Within Condensates Generated by Liquid-liquid Phase Separation. J. Mol. Biol. **434**, 167201 (2022).
8. Mullick, P. & Trovato, A. *Sequence based prediction of protein phase separation into disordered condensates using machine learning*. http://biorxiv.org/lookup/doi/10.1101/2021.12.13.472521 (2021) doi:10.1101/2021.12.13.472521.
9. Kokhlikyan, N. *et al.* Captum: A unified and generic model interpretability library for PyTorch. Preprint at https://doi.org/10.48550/arXiv.2009.07896 (2020).
10. Elnaggar, A. *et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning*. http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199554 (2020) doi:10.1101/2020.07.12.199554.
11. Wang, X. *et al.* LLPSDB v2.0: an updated database of proteins undergoing liquid–liquid phase separation *in vitro*. Bioinformatics **38**, 2010–2014 (2022).
12. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res.* gkz848 (2019) doi:10.1093/nar/gkz848.
13. Cho, N. H. *et al.* OpenCell: Endogenous tagging for the cartography of human cellular organization. Science **375**, eabi6983 (2022).
14. Saar, K. L. *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci.* **118**, e2019053118 (2021).
15. van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. Cell Rep. **34**, 108705 (2021).
16. Cai, H., Vernon, R. M. & Forman-Kay, J. D. An Interpretable Machine-Learning Algorithm to Predict Disordered Protein Phase Separation Based on Biophysical Interactions. Biomolecules **12**, 1131 (2022).

17. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. eLife **7**, e31486.
18. Portz, B., Lee, B. L. & Shorter, J. FUS and TDP-43 Phases in Health and Disease. Trends Biochem. Sci. **46**, 550–563 (2021).
19. Chang, C. *et al.* The N-terminus of TDP-43 promotes its oligomerization and enhances DNA binding affinity. Biochem. Biophys. Res. Commun. **425**, 219–224 (2012).
20. Mompeán, M. *et al.* The TDP-43 N-terminal domain structure at high resolution. FEBS J. **283**, 1242–1260 (2016).
21. Conicella, A. E. *et al.* TDP-43 α-helical structure tunes liquid–liquid phase separation and function. *Proc. Natl. Acad. Sci.* **117**, 5883–5894 (2020).
22. Conicella, A. E., Zerze, G. H., Mittal, J. & Fawzi, N. L. ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. Struct. Lond. Engl. 1993 **24**, 1537–1549 (2016).
23. Mompeán, M. *et al.* "Structural characterization of the minimal segment of TDP-43 competent for aggregation". Arch. Biochem. Biophys. **545**, 53–62 (2014).
24. Piovesan, D. *et al.* MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res. **49**, D361–D367 (2020).
25. Wang, A. *et al.* A single N-terminal phosphomimic disrupts TDP-43 polymerization, phase separation, and RNA splicing. EMBO J. **37**, e97452 (2018).
26. Carter, G. C., Hsiung, C.-H., Simpson, L., Yang, H. & Zhang, X. N-terminal Domain of TDP43 Enhances Liquid-Liquid Phase Separation of Globular Proteins. J. Mol. Biol. **433**, 166948 (2021).
27. Wang, L., Kang, J., Lim, L., Wei, Y. & Song, J. TDP-43 NTD can be induced while CTD is significantly enhanced by ssDNA to undergo liquid-liquid phase separation. Biochem. Biophys. Res. Commun. **499**, 189–195 (2018).
28. Jiang, L.-L. *et al.* The N-terminal dimerization is required for TDP-43 splicing activity. Sci. Rep. **7**, 6196 (2017).
29. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021).
30. Hughes, M. P. *et al.* Atomic structures of low-complexity protein segments reveal kinked β-sheets that assemble networks. 10 (2018) doi:10.1126/science.aan6398.
31. Murray, D. T. *et al.* Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. Cell **171**, 615–627.e16 (2017).
32. Luo, F. *et al.* Atomic structures of FUS LC domain segments reveal bases for reversible amyloid fibril formation. Nat. Struct. Mol. Biol. **25**, 341–346 (2018).
33. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell **174**, 688–699.e16 (2018).
34. Harrison, A. F. & Shorter, J. RNA-binding proteins with prion-like domains in health and disease. Biochem. J. **474**, 1417–1438 (2017).
35. Qamar, S. *et al.* FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. Cell **173**, 720–734.e15 (2018).
36. Yoshizawa, T. *et al.* Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. Cell **173**, 693–705.e22 (2018).
37. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. Cell **176**, 419–434 (2019).
38. Hsu, H. & Lachenbruch, P. A. Paired *t* Test. in *Wiley StatsRef: Statistics Reference Online* (eds. Balakrishnan, N. et al.) (Wiley, 2014). doi:10.1002/9781118445112.stat05929.
39. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). Biometrika **52**, 591 (1965).
40. Pratt, J. W. Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. J. Am. Stat. Assoc. **54**, 655–667 (1959).
41. Soto, L. F. *et al.* Compendium of human transcription factor effector domains. Mol. Cell **82**, 514–526 (2022).
42. Kumar, M. & Michael, S. The Eukaryotic Linear Motif resource: 2022 release. 12 doi:10.1093/nar/gkab975.
43. Dosztanyi, Z., Meszaros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics **25**, 2745–2746 (2009).

44. Gao, J., Wang, L., Yan, T., Perry, G. & Wang, X. TDP-43 proteinopathy and mitochondrial abnormalities in neurodegeneration. Mol. Cell. Neurosci. **100**, 103396 (2019).
45. Chia, R., Chiò, A. & Traynor, B. J. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. **17**, 94–102 (2018).
46. Barmada, S. J. *et al.* Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. J. Neurosci. **30**, 639–649 (2010).
47. Tziortzouda, P., Van Den Bosch, L. & Hirth, F. Triad of TDP43 control in neurodegeneration: autoregulation, localization and aggregation. Nat. Rev. Neurosci. **22**, 197–208 (2021).
48. Vance, C. *et al.* Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science **323**, 1208–1211 (2009).
49. Deng, H., Gao, K. & Jankovic, J. The role of FUS gene variants in neurodegenerative diseases. Nat. Rev. Neurol. **10**, 337–348 (2014).
50. Loughlin, F. E. *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 **73**, 490–504.e6 (2019).
51. Kreiter, N. *et al.* Age-dependent neurodegeneration and organelle transport deficiencies in mutant TDP43 patient-derived neurons are independent of TDP43 aggregation. Neurobiol. Dis. **115**, 167–181 (2018).
52. Mann, J. R. *et al.* RNA Binding Antagonizes Neurotoxic Phase Transitions of TDP-43. Neuron **102**, 321–338.e8 (2019).
53. Zhou, X. *et al.* Mutations linked to neurological disease enhance self-association of low-complexity protein sequences. Science **377**, eabn5582 (2022).
54. Hallegger, M. *et al.* TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell **184**, 4680–4696.e22 (2021).
55. Ling, S.-C. *et al.* ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 13318–13323 (2010).
56. Swetha, R. G., Ramaiah, S. & Anbarasu, A. R521C and R521H mutations in FUS result in weak binding with Karyopherinβ2 leading to Amyotrophic lateral sclerosis: a molecular docking and dynamics study. J. Biomol. Struct. Dyn. **35**, 2169–2185 (2017).
57. Niaki, A. G. *et al.* Loss of Dynamic RNA Interaction and Aberrant Phase Separation Induced by Two Distinct Types of ALS/FTD-Linked FUS Mutations. Mol. Cell **77**, 82–94.e4 (2020).
58. Holehouse, A. S., Ginell, G. M., Griffith, D. & Böke, E. Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates: Published as part of the *Biochemistry* virtual special issue “Protein Condensates”. Biochemistry **60**, 3566–3581 (2021).
59. Muñoz-Gil, G. *et al. Phase separation of tunable biomolecular condensates predicted by an interacting particle model*. http://biorxiv.org/lookup/doi/10.1101/2020.09.09.289876 (2020) doi:10.1101/2020.09.09.289876.
60. Schuster, B. S. *et al.* Biomolecular Condensates: Sequence Determinants of Phase Separation, Microstructural Organization, Enzymatic Activity, and Material Properties. J. Phys. Chem. B **125**, 3441–3451 (2021).
61. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. **30**, 214–226 (2020).
62. McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci.* **116**, 11624–11629 (2019).
63. Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proc. Natl. Acad. Sci.* **118**, e2104878118 (2021).
64. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science **367**, 694–699 (2020).
65. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**, R100 (2006).

66. Iglesias, V. *et al.* In silico Characterization of Human Prion-Like Proteins: Beyond Neurological Diseases. Front. Physiol. **10**, 314 (2019).

Figures

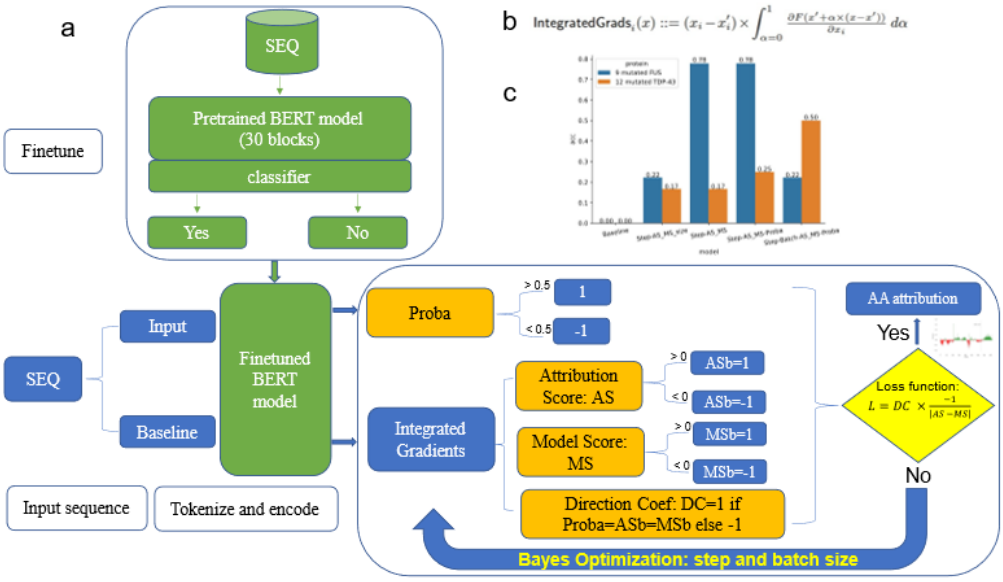

Fig. 1| BERTIG framework and ablation test.

a, BERTIG framework. The framework consists of two parts: the pretrained BERT model that is finetuned (represented by the green color), and the interpretation optimization using the Integrated Gradients algorithm (represented by the blue color). Arrows show the information flow among the various components. AS (Attribution Score): indicates the total approximated integrated gradients. ASb: AS binarization. MS (Model Score): indicates the total true integrated gradients. MSb: MS binarization. Proba: the prediction probability of LLPS. DC (Direction Coef): DC was derived from Proba, AS and MS scores which were binarized to 1 or -1.

b, The Integrated gradients formula. The method is defined as the path integral of the gradients along the straight-line path from the baseline x' to the input x. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F (x) along the i^{th} dimension, and α is the scaling coefficient. In terms of the protein, x_i indicates the input protein sequence, x_i' indicates the corresponding baseline filled with zero.

c, Ablation test. For the purpose of quantifying the effects of mutations, 9 FUS and 12 TDP-43 proteins with variations were modeled and optimized using a half-precision setting (fp16) in a label-agnostic fashion. The X-axis indicates the model under different conditions, while the Y-axis displays the accuracy of the LLPS interpretation as a percentage comparison to the wild-type protein. Different models notes: The Baseline model indicated that the default value of step size was 50 and batch size was the protein length as described in the Captum package[9]. The Step-AS_MS_size model indicated that the step size was explored, the direction of AS and MS were not considered, and the batch size was the default value. The Step-AS_MS model indicated the same as the Step-AS_MS_size model except that the direction of AS and MS was considered. The Step-AS_MS-Proba model indicated that in addition to the configuration of the Step-AS_MS model, the prediction probability was also considered. The Step-Batch-AS_MS-Proba model indicated that in addition to the configuration of the Step-AS_MS-Proba model, the batch size was also explored.

Figure 1

See image above for figure legend.

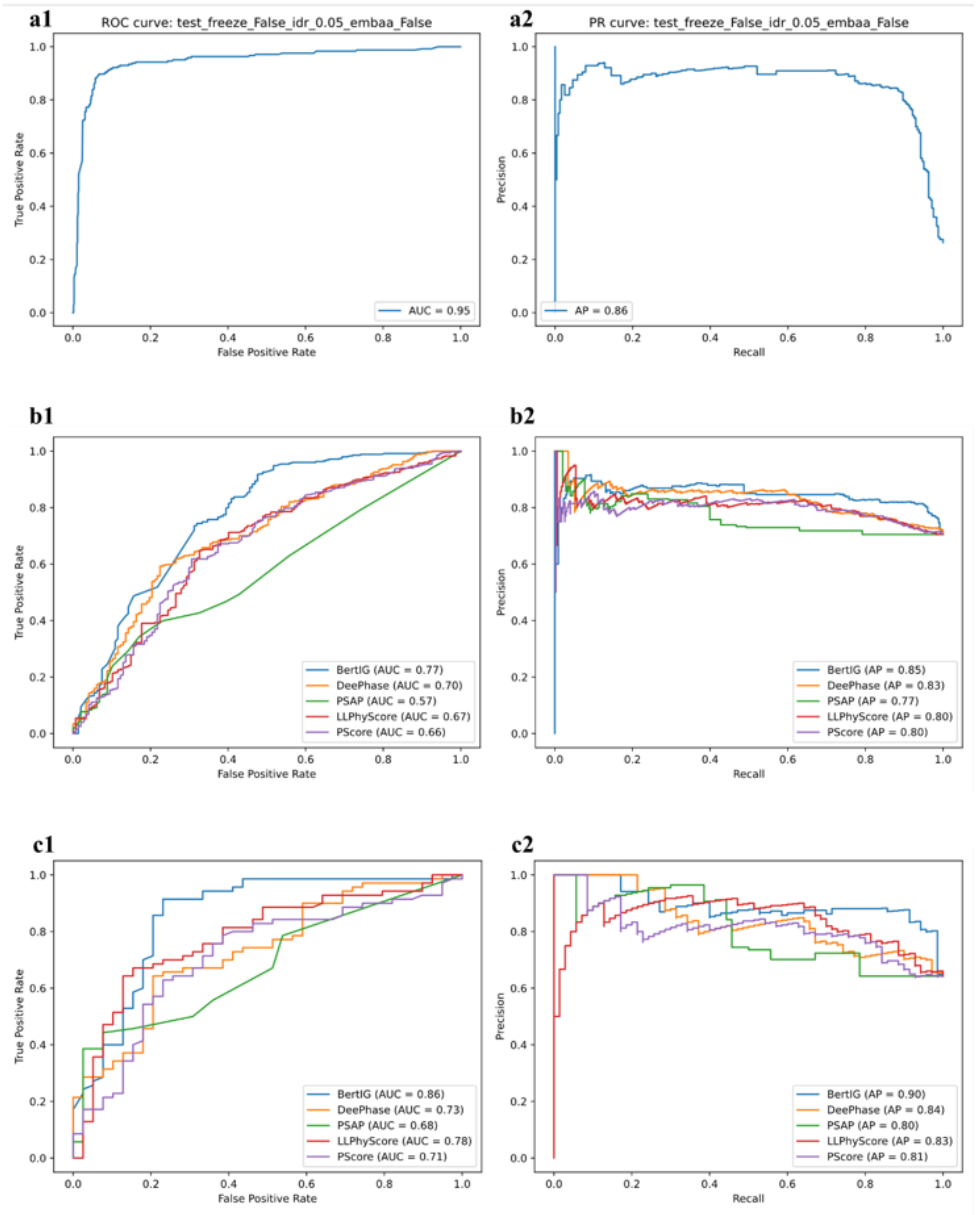

Figure 2

Model performance. ROC: receiver operating characteristic curves plotting the true positive rate and the false positive rate for a predictive model using different probability thresholds. AUC: the area under the ROC curve. PR (Precision-Recall): PR curves plotting the true positive rate and the positive predictive value for a predictive model using different probability thresholds. AP (Average Precision): summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

a1-a2, ROC and PR curves of BERTIG on the test dataset. **b1-b2,** ROC and PR curves of BERTIG and other methods on the LLPSDB-v2 dataset. **c1-c2,** ROC and PR curves of BERTIG and other methods on the PhaSePro dataset.

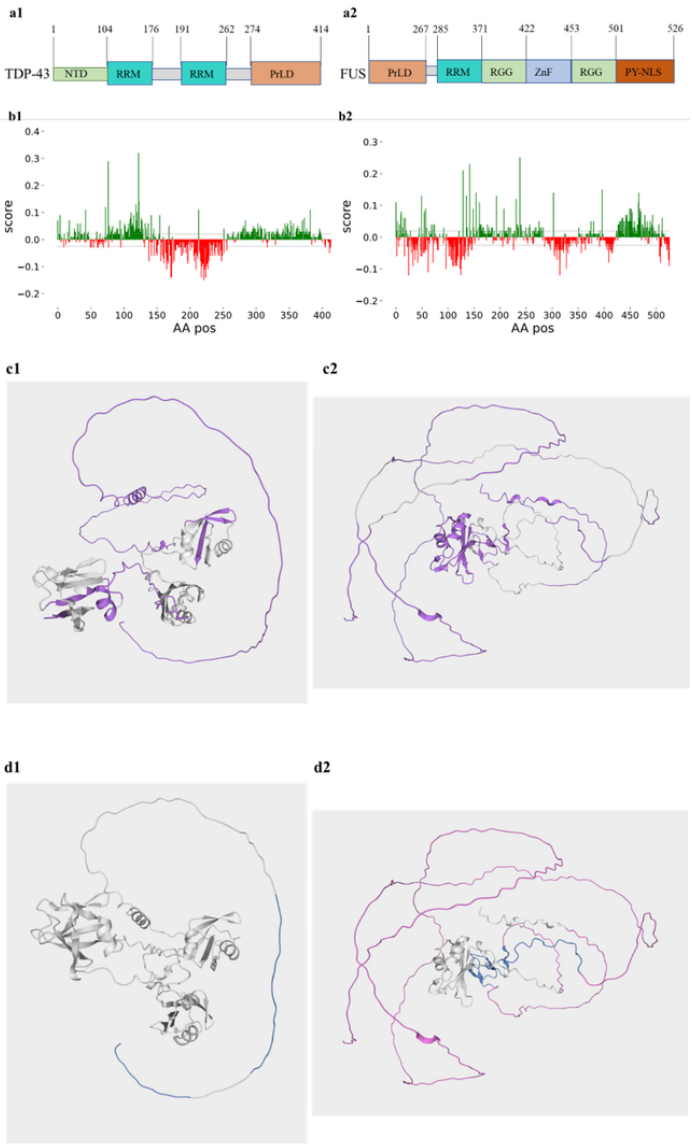

Figure 3

BERTIG produces highly accurate AAs attributions. a1-a2, schematic representation of the structure in TDP-43 and FUS. **b1-b2,** AAs attributions by BERTIG (green: promote phase separation, red: inhibit phase separation) in TDP-43 and FUS. **c1-c2,** linear interacting peptide (LIP) colored in purple from Alphafold2[29] and mobidb[24] in TDP-43 and FUS. **d1-d2,** binding mode (red: disorder-to-disorder binding mode (DD), blue: disorder-to-order binding mode (DO)) from Alphafold2[29] and mobidb[24] in TDP-43 and FUS.

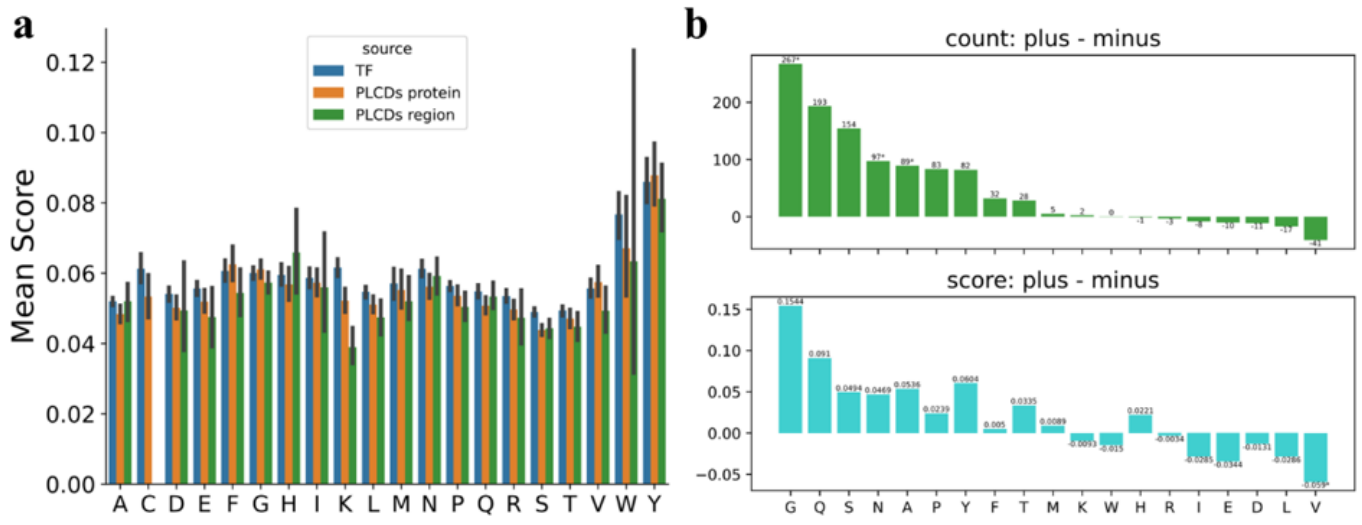

Figure 4

key AAs for LLPS. a, Mean score in each type of 20 AAs from 308 TF proteins, 90 PLCDs proteins and PLCDs regions. If an AA is in the region where more than 2 consecutive AAs scores are separately bigger than 0.02 (default positive score in the Captum package), mean score in each type of 20 AAs is derived.

b, AA comparison in 90 PLCDs regions. count plus: If an AA is in the region where more than 2 consecutive AA scores are separately bigger than 0.02, sum of counts in each type of 20 AAs is derived. count minus: If an AA is in the region where more than 2 consecutive AA scores are separately smaller than -0.025 (default minus score in the Captum package), sum of counts in each type of 20 AAs is derived. score sum plus: If an AA is in the region where more than 2 consecutive AA scores are separately bigger than 0.02, sum of attribution scores in each type of 20 AAs is derived. score sum minus: If an AA is in the region where more than 2 consecutive AA scores are separately smaller than -0.025, sum of attribution scores in each type of 20 AAs is derived. Significance was tested with paired T-test[38] if the normal distribution is satisfied (Shapiro-Wilk test[39]); Otherwise, Wilcoxon signed-rank test[40]. Reported p values are for the paired plus-minus AA score comparisons in each protein (*$p < 0.05$).

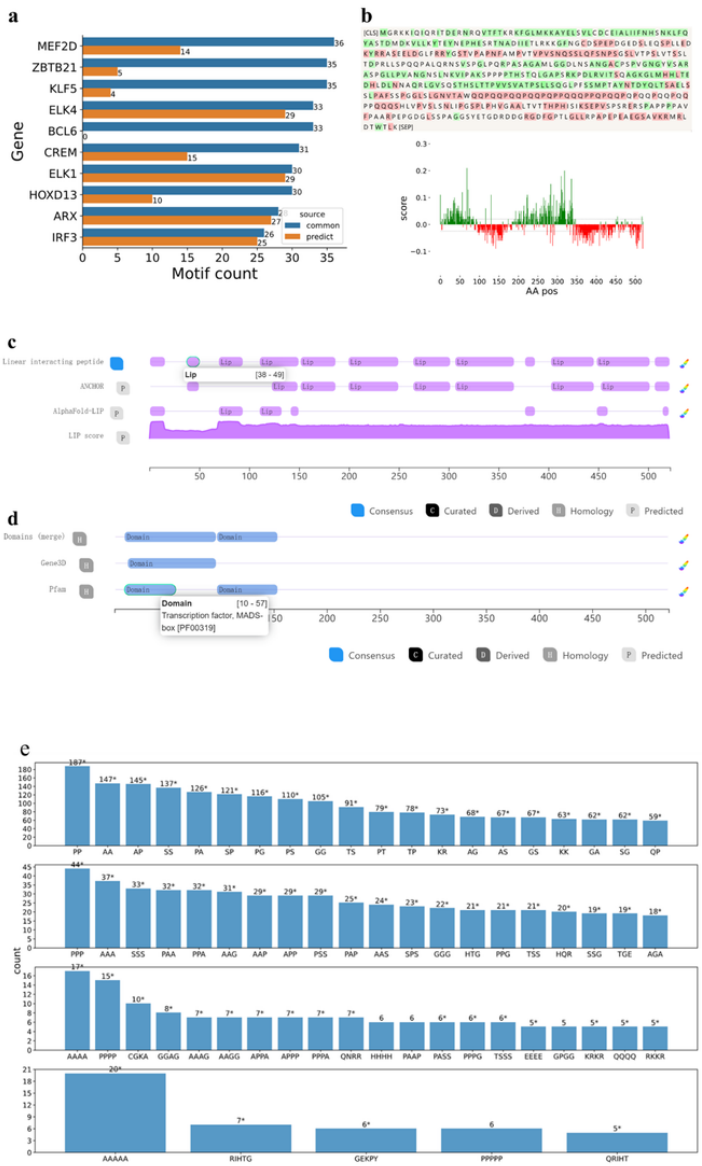

Figure 5

Motifs comparisons. If there were more than 2 consecutive AAs in the region where their scores were separately bigger than 0.02 (default plus score in the Captum package), the region was defined as motif. The predicted motifs were counted for different length. **a,** Top 10 genes in 308 TF proteins based on all motifs. **b,** AAs attributions by BERTIG (green: promote phase separation, red: inhibit phase separation) in MEF2D. **c,** Linear interacting peptide (LIP) residues interacting with another molecule in MEF2D from the mobidb database. **d,** Structured residues in MEF2D from the mobidb database. **e,** Motifs with the positive score in 308 TF proteins.

Significance was tested with paired T-test[38] if the normal distribution is satisfied (Shapiro-Wilk test[39]); Otherwise, Wilcoxon signed-rank test[40]. Reported p values are for the paired plus-minus AA score comparisons in each protein (*p < 0.05).

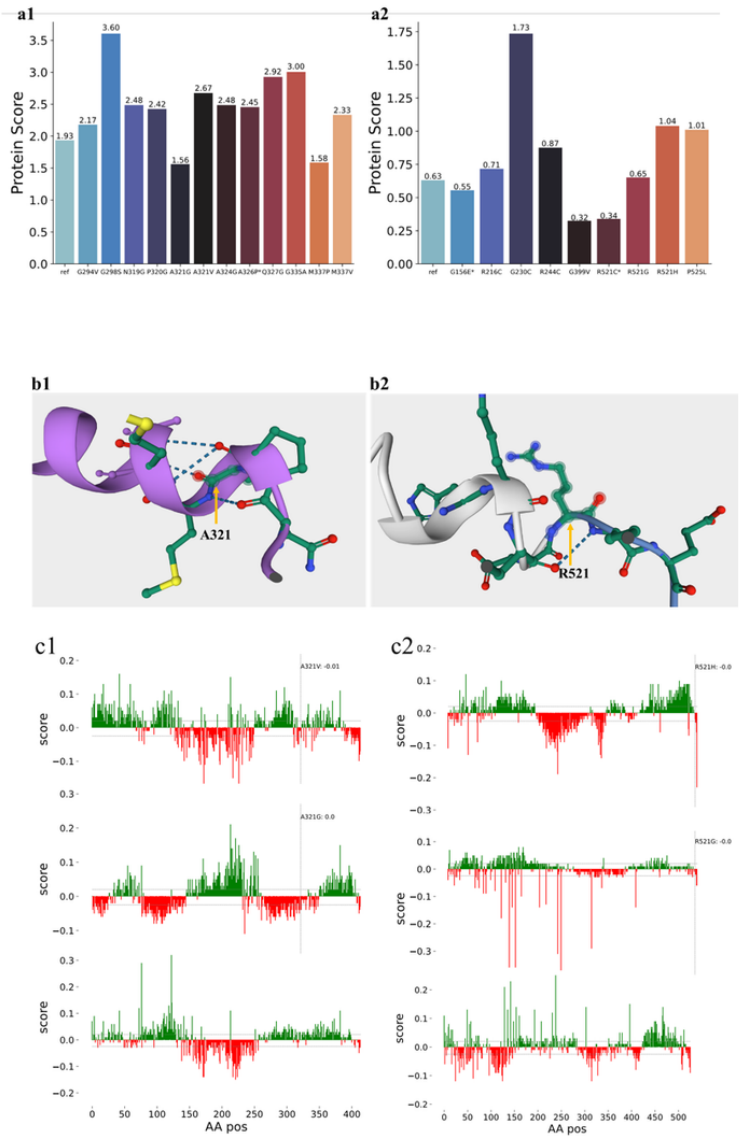

Figure 6

Effect of mutated proteins on LLPS. a1-a2, AS based LLPS ability of mutated TDP-43 and FUS proteins. Mutations with the "*" were misinterpreted, such as: A326P in TDP-43; and G156E, R521C in FUS. **b1-b2,**Residue A in the position 321 in the helical segment of TDP-43 and R in the position 521 of FUS protein. **c1-c2,** The amino-acid-resolution contribution plots of mutated TDP-43 and FUS proteins.

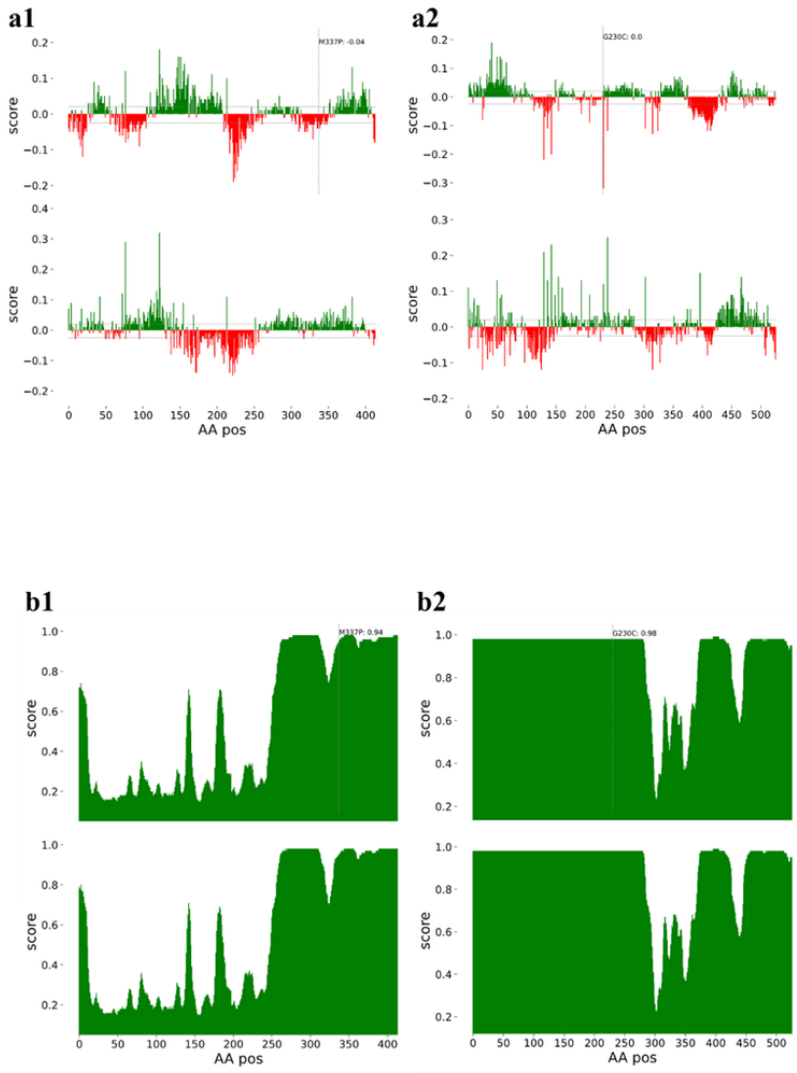

Figure 7

Interpretations of different models on the M337P mutated TDP-43 (left) and G230C mutated FUS (right). a1-a2, BERTIG. **b1-b2,** FuzDrop. Note: bottom: wild protein, top: mutated protein.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AccurateandefficientinterpretationofquantitativeaminoacidattributionfordisorderedproteinsundergoingLLPSV2.1supp.docx
- ExtendedDataFigs.docx