

Sequence-based Prediction of the Cellular Toxicity Associated with Amyloid Aggregation within Protein Condensates

Attila Horvath, Michele Vendruscolo, and Monika Fuxreiter*



Cite This: *Biochemistry* 2022, 61, 2461–2469



Read Online

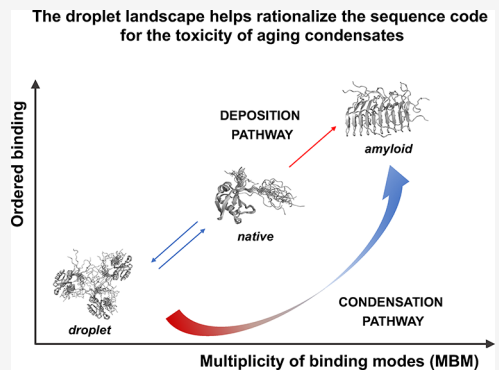
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Various neurological dysfunctions are associated with cytotoxic amyloid-containing aggregates formed through the irreversible maturation of protein condensates generated by phase separation. Here, we investigate the amino acid code for this cytotoxicity using TDP-43 deep-sequencing data. Within the droplet landscape framework, we analyze the impact of mutations in the amyloid core, aggregation hot-spot, and droplet-promoting residues on TDP-43 cytotoxicity. Our analysis suggests that TDP-43 mutations associated with low cytotoxicity moderately decrease the probability of droplet formation while increasing the probability of multimodal binding. These mutations promote both ordered and disordered binding modes, thus facilitating the conversion between the droplet and amyloid states. Based on this understanding, we develop an extension of the FuzDrop method for the sequence-based prediction of the cytotoxicity of aging condensates and test it over 20,000 TDP-43 variants. Our analysis provides insight into the amino acid code that regulates the cytotoxicity associated with the maturation of liquid-like condensates into amyloid-containing aggregates, suggesting that, at least in the case of TDP-43, mutations that promote aggregation tend to decrease cytotoxicity, while those that promote droplet formation tend to increase cytotoxicity.



INTRODUCTION

Increasing evidence demonstrates that proteins can populate three fundamental states in the cellular environment. In addition to the native and amyloid states,^{1,2} proteins can sample a dense, liquid-like state, which is reversibly formed from the native state through a phase separation process and is prone to irreversibly mature into the amyloid state.^{3–5} This liquid-like state, also known as the droplet state, seems to be generally accessible to proteins⁶ and associated with a wide range of physiological processes that involve the formation of multicomponent functional assemblies referred to as membraneless organelles.^{7,8}

The droplet state is tightly regulated by the protein homeostasis system,⁹ including molecular chaperones,¹⁰ ubiquitin ligases, and the autophagy system.¹¹ Upon dysregulation of the balance between the native state and the droplet state, however, the latter can evolve into the amyloid state (Figure 1).¹² During this process of amyloid formation, known as the condensation pathway, cytotoxic intermediates can be generated.^{13–17} Cytotoxicity can be caused by a wide variety of molecular mechanisms, including protein mislocalization, a lack of availability of functional partners, or a presence of nonphysiological partners, and by changes in the protein structure, leading to oligomerization. In addition, a delayed reconversion to the native state of proteins trapped in a gel-like form can be due to recruitment of other cellular components.¹⁸

We previously investigated the amino acid code that determines the condensation pathway to amyloid formation.²¹ We suggested that the conversion of the droplet state to the amyloid state can be described by a droplet landscape based on the observation that the droplet and amyloid states are stabilized by different binding modes.^{21,22} The droplet state is mostly stabilized by disordered interactions, comprising heterogeneous binding patterns among the same residues (Figure 1A). The amyloid state, in contrast, is stabilized by ordered interactions, which are formed by well-defined contacts between residues (Figure 1A).

An important role in this discussion is played by protein regions that can sample both ordered and disordered binding modes. Regions that sample a multiplicity of binding modes (MBM), including both disordered and ordered interactions (Figure 1B), can be identified based on their sequences.²² Our analysis suggests that regions that change their binding modes upon alterations in the cellular environment overlap with regions becoming the amyloid cores of the condensates and

Special Issue: Protein Condensates

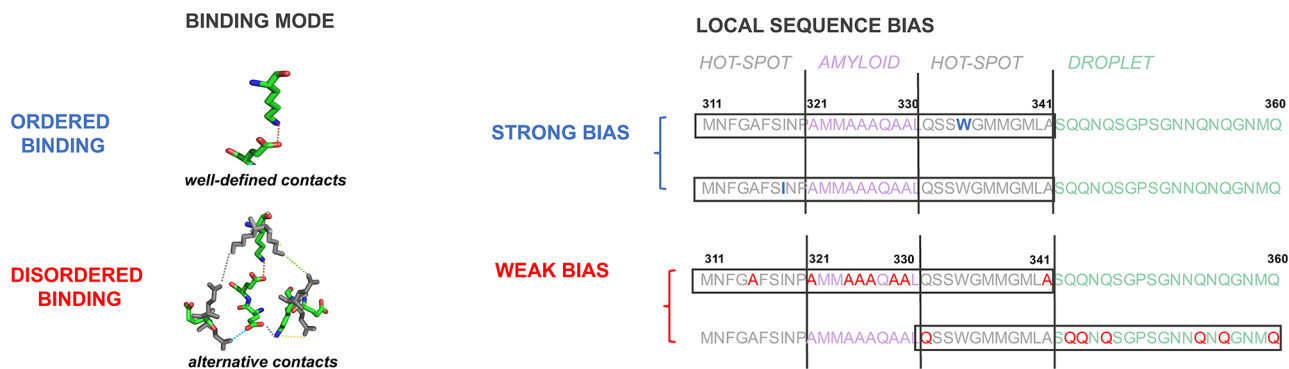
Received: August 29, 2022

Revised: October 20, 2022

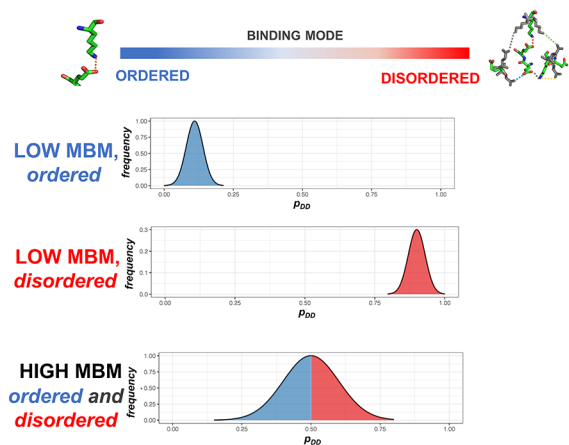
Published: November 7, 2022



A BINDING MODE OF INTERACTION



B MULTIPLICITY OF BINDING MODES



C DROPLET LANDSCAPE

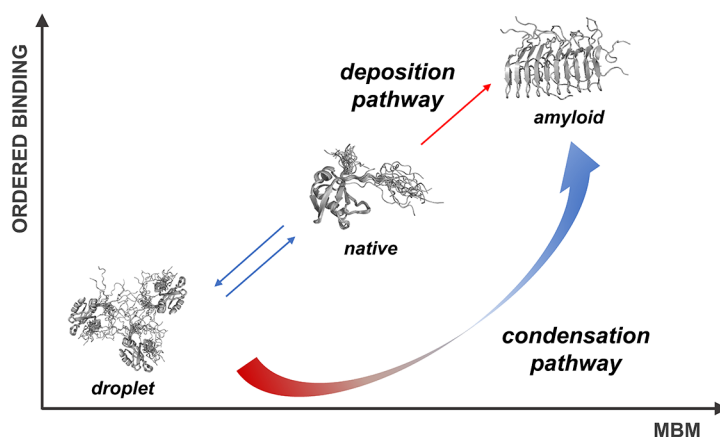


Figure 1. Interaction modes in the droplet and amyloid states of proteins. (A) The local sequence bias of a protein region determines its binding mode. Illustration of residues in regions with strong (blue) and weak (red) local sequence bias in the droplet (green), amyloid (purple), and hot-spot (gray) regions of TDP-43.¹⁹ The local sequence bias of a given region is defined based on the difference between the amino acid composition of the region itself and its flanking sequences.²⁰ A strong bias leads to ordered binding modes with a well-defined contact pattern (orange dotted line). A weak bias leads to alternative contact patterns between similar sites of the same set of residues (colored dotted lines). (B) The multiplicity of binding modes (MBM) quantifies the spectrum of interaction modes sampled in the bound state. Protein interactions sample a wide range of binding modes from ordered (blue) to disordered (red). The MBM is low when only one of these modes is sampled. This means that in the bound state, the protein exhibits well-defined contact patterns (ordered binding, blue) or variable, heterogeneous binding patterns (disordered binding, red), but it is unlikely to switch between these scenarios. In the case of high MBM, a protein region can exhibit both types of interactions and can switch between ordered and disordered binding modes, depending on its partner, post-translational modification, or other cellular conditions. (C) Landscape of binding modes of the three main cellular states of proteins. The *x* axis reports the multiplicity of binding modes (MBM), and the *y* axis reports the ordered binding modes. The droplet state is characterized by disordered interactions (low degree of ordered binding) as well as low MBM. That is, residues promoting the droplet state unlikely change disordered to ordered interactions. In contrast, residues that form the amyloid state tend to exhibit ordered binding mode and high MBM, since they can change between disordered and ordered binding. The amyloid state can be formed through the deposition pathway (diagonal) through unfolding of the native state or through the condensation pathway (along the arrow passing through the lower-right corner) of irreversible maturation of the droplet state.

thus can serve as aggregation hot-spots.²¹ Based on this insight, we could discriminate between FUS mutations associated with amyotrophic lateral sclerosis (ALS) and those not associated with the disease.^{23,24} In particular, our analysis indicates that disease-associated mutations increased the multiplicity of binding modes, i.e., the probability of sampling both ordered and disordered interactions, while similar mutations not affecting this property tended to be not associated with ALS.²¹

Here, we investigated the amino acid code that determines cytotoxicity of the aging condensates using deep-sequencing data of >20,000 mutants of the TAR binding protein 43 (TDP-43).¹⁹ The presence of neuronal aggregates of mutant TDP-43 is a molecular hallmark of ALS.^{17,25,26} The cytotoxicity of TDP-43 mutants measured in yeast indicated that mutations

that enhanced amyloid formation decreased cytotoxicity,¹⁹ while mutations that affected the secondary structure increased cytotoxicity. By extending previous results on ALS-associated mutations,^{23,24} our analysis here indicates that TDP-43 mutants increase the probability of sampling both ordered and disordered interactions and in particular increase the multiplicity of binding modes of droplet-promoting residues. Based on this understanding, we report an extension of the FuzDrop method⁶ for the sequence-based prediction of the cytotoxicity of aging condensates.

RESULTS

Aggregation Is Induced by Mutations Affecting the Local Sequence Bias. We previously reported the prediction

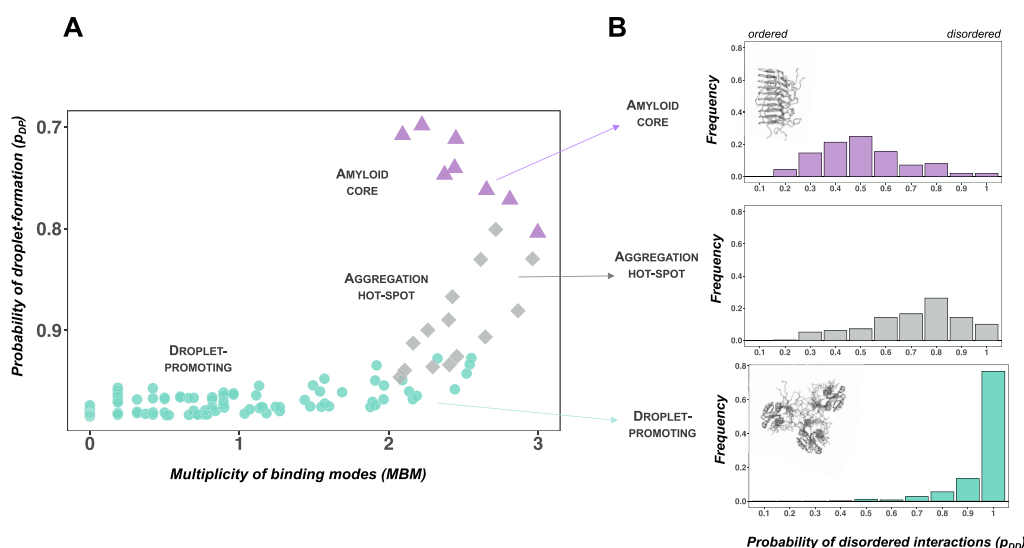


Figure 2. Droplet landscape representation of the LC region of TDP-43. (A) Droplet landscape of the LC region of TDP-43 (residues 262–414). The x axis of the droplet landscape corresponds to the multiplicity of binding modes (MBM).²² The y axis of the droplet landscape is defined by the droplet-promoting propensity (p_{DP}).⁶ The droplet landscape of the wild-type TDP-43 LC domain (residues 262–414) illustrates droplet-promoting residues (residues 262–311 and 342–414, green circles) in the lower-left section of the droplet landscape with high p_{DP} values and low MBM values. In contrast, residues forming the amyloid core (residues 321–330, purple triangles), exhibiting low p_{DP} values and high MBM values. Regions that readily convert into amyloids (i.e., “aggregation hot-spots”, residues 312–320 and 331–341, gray diamonds¹⁹) are in the lower cross section, with high p_{DP} values and high MBM values. These residues prefer disordered binding configurations, as well as also sample ordered states, as reflected by high interaction multimodality. (B) Frequencies of different binding modes in the TDP-43 LC domain. The frequencies of different binding modes from ordered to disordered interactions are shown for the amyloid core (purple), aggregation hot-spot (gray), and droplet-promoting region (green). While the amyloid core and the aggregation hot-spot exhibit wide distributions by sampling both the ordered and disordered interactions, leading to high MBM, the droplet region mostly samples disordered (unimodal) interactions, leading to low MBM.

of whether a protein region forms ordered or disordered interactions based on its local sequence bias²⁰ (Figure 1A). The local sequence bias of a region arises from a difference in the amino acid composition between the region itself and its flanking sequences. When the bias is large, a well-defined interaction tends to be established through an ordered binding mode²⁰ (Figure 1A,B). This type of interaction is usually established in specific complexes of proteins with sequences of high complexity. In contrast, when the bias is small, disordered interactions are promoted due to competing binding sites with similar properties²⁰ (Figure 1A). In this case, a variety of alternative interaction patterns can be established among the same set of residues, resulting in a heterogeneous bound state (Figure 1A,B). Disordered binding modes are usually linked with low-complexity (LC) sequences and/or structural disorder.²⁷ We previously reported that the droplet state is driven by such disordered interactions,⁶ which can be formed through a wide range of sequence motifs⁵ of aromatic, charged, and hydrophobic contacts.²⁸

Thus, based on the local sequence bias, sequence elements promoting droplet formation and amyloid formation can be identified. Based on these insights, one can predict that aggregation could be initiated by sequence elements where mutations, post-translational modifications, or interactions with other cellular factors considerably increase the sequence bias and promote ordered interactions (Figure 1B,C).

Droplet Landscape Representation of Aggregation within Protein Condensates. As noted above (Figure 1C), to describe the transition from the droplet to the amyloid state, one can use a droplet landscape (Figure 2A). This landscape helps understand how the local sequence bias can be modulated by the cellular conditions, which cause a change

in binding modes, thus leading to the conversion between the droplet and amyloid states.²¹ The x axis of the droplet landscape corresponds to the multiplicity of binding modes (MBM), which is computed based on the Shannon entropy of binding modes (S_{bind}), as defined in the FuzPred method²² (Figure 2A). The y axis of the droplet landscape is defined by the residue-specific droplet-promoting propensity (p_{DP}) of the FuzDrop method, which characterizes the likelihood for spontaneous phase separation⁶ (Figure 2A).

In the droplet landscape representation, droplet-promoting residues are found in the lower-left section of the droplet landscape (the “droplet region”), as they have a high probability to form the droplet state mostly by disordered interactions. Thus, the multiplicity of binding modes of the droplet regions is low (Figure 2A,B). By contrast, the amyloid-promoting residues are in the upper-right section of the droplet landscape (the “amyloid region”), as they have lower probability to undergo phase separation, by simultaneously sampling both ordered and disordered interactions. Thus, residues in the amyloid regions have a high multiplicity of binding modes²⁹ (Figure 2A,B). Residues in the lower-right section of the droplet landscape have a high probability to form the droplet state and sample both ordered and disordered interactions. Therefore, these residues can initiate aggregation in the protein condensates (“aggregation hot-spot”) (Figure 2A). Hot-spot residues have a high multiplicity of binding modes, as they sample both ordered and disordered interactions (Figure 2B).

The methods to compute the values of the multiplicity of binding modes (MBM) and residue-specific droplet-promoting propensity (p_{DP}) have been published previously^{6,22} and are publicly available.³⁰ The p_{DP} values are computed using a

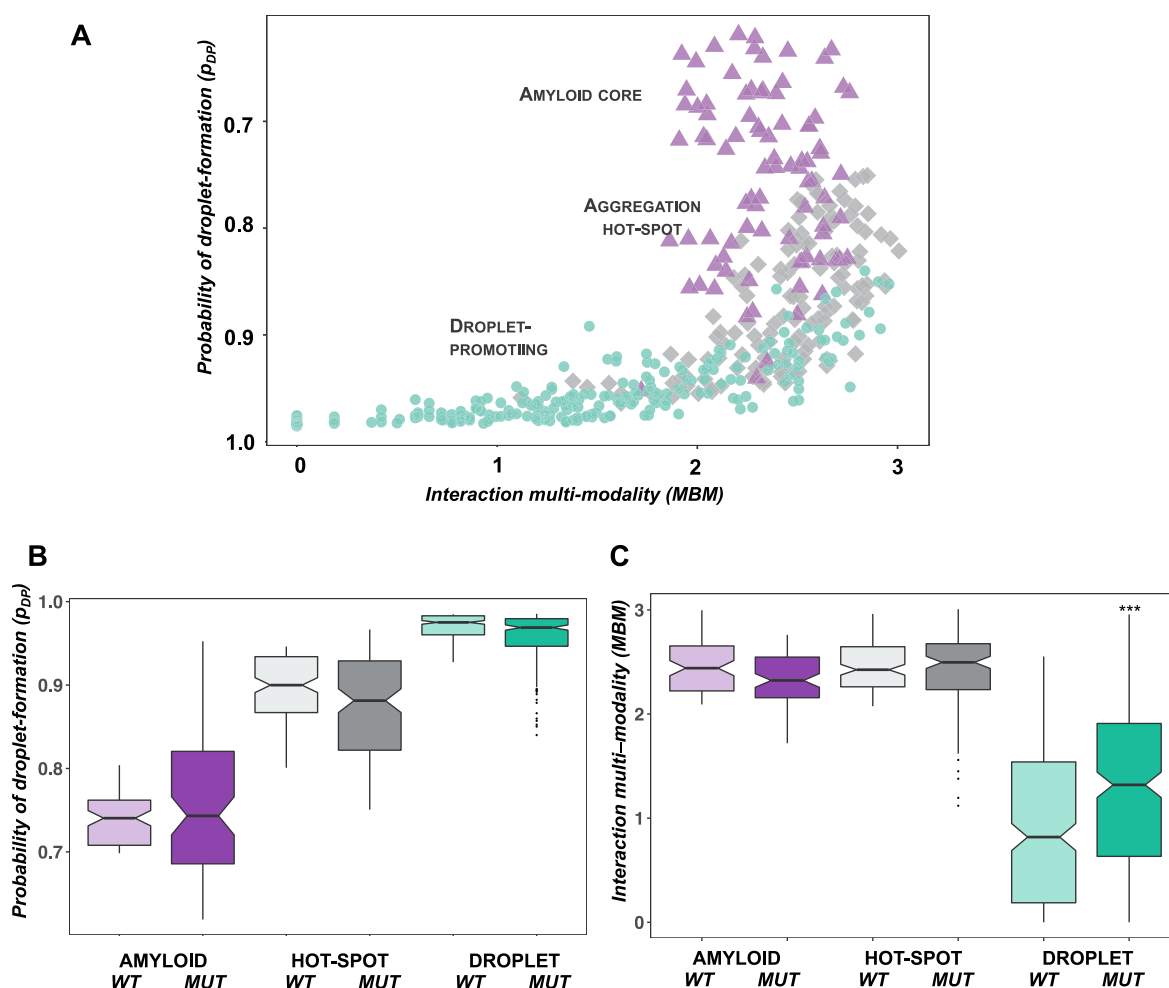


Figure 3. Single mutations increase the MBM of the TDP-43 LC domain. We analyzed 498 single mutations with a change in cytotoxicity ($\Delta e_{tox} > 3\sigma$).¹⁹ (A) Droplet landscape of TDP-43 single mutants. Amyloid core residues (residues 321–330, purple triangles) and droplet-promoting residues (residues 262–311 and 342–414, green circles) considerably overlap with the aggregation hot-spot region (residues 312–320 and 331–341, gray diamonds). This indicates a high probability to phase separate (p_{DP} values, y axes) and a high multiplicity of binding modes (MBM, x axis) that reflects sampling both disordered and ordered interactions. (B) Comparison of droplet propensities of wild-type and mutant TDP-43 residues. No significant change was calculated between the phase separation probability of wild-type (light) and mutant residues (dark) in the amyloid core (purple), aggregation hot-spot (gray), and droplet region (green). (C) Comparison of MBM of wild-type and mutant TDP-43 residues. Mutations in the droplet region (dark green) significantly ($p < 10^{-3}$) increase the MBM as compared to the wild-type values (light green), reflecting a shift in binding modes toward ordered interactions. The statistical significance was computed by the Mann–Whitney test of the R program.

binary logistic model (Methods, eq. 4) and a scoring function based on the conformational entropy of the free and bound states (Methods, eq. 5). This method was shown to be robust to identify droplet-promoting regions under physiological conditions.⁶

The droplet landscape of the wild-type TDP-43 LC domain (residues 262–414) illustrates distinct features of regions promoting droplet formation and amyloid formation and serves as hot-spots for aggregation (Figure 2A). Droplet-promoting residues (residues 262–311 and 342–414) are in the lower-left section of the droplet landscape and have high p_{DP} values and low MBM values. In contrast, residues forming the amyloid core (residues 321–330^{31,32}) are in the upper-right section, exhibiting low p_{DP} values and high MBM values. Regions that readily convert to amyloids (i.e., “aggregation hot-spots”; residues 312–320 and 331–341¹⁹) are in the lower cross section with high p_{DP} values and high MBM values. These residues prefer disordered binding configurations, as

well as also sample ordered states, as reflected by high MBM (Figure 2A,B).

TDP-43 LC Mutations Increase the MBM of Droplet-Promoting Residues. A deep mutagenesis approach was recently reported to generate >50,000 TDP-43 variants.¹⁹ The cytotoxicity of these variants was assessed by monitoring growth rates in *Saccharomyces cerevisiae*, and mutants decreasing the growth rate were considered cytotoxic.¹⁹ We analyzed the impact of these mutations on the droplet landscape of the droplet region, amyloid core, and aggregation hot-spot residues of 498 single mutants, leading to a large change in cytotoxicity ($\Delta e_{tox} > 3\sigma$, where σ is the average change over all the mutations;¹⁹ Figure 3). Both the amyloid core and the droplet region are shifted toward the aggregation hot-spot region, with high p_{DP} values (≥ 0.75) and high MBM values (> 2.25) (Figure 3A). That is, both the amyloid core and droplet-promoting residues could sample both ordered and disordered interactions, thus facilitating the conversion between these states. Along these lines, our analysis indicated

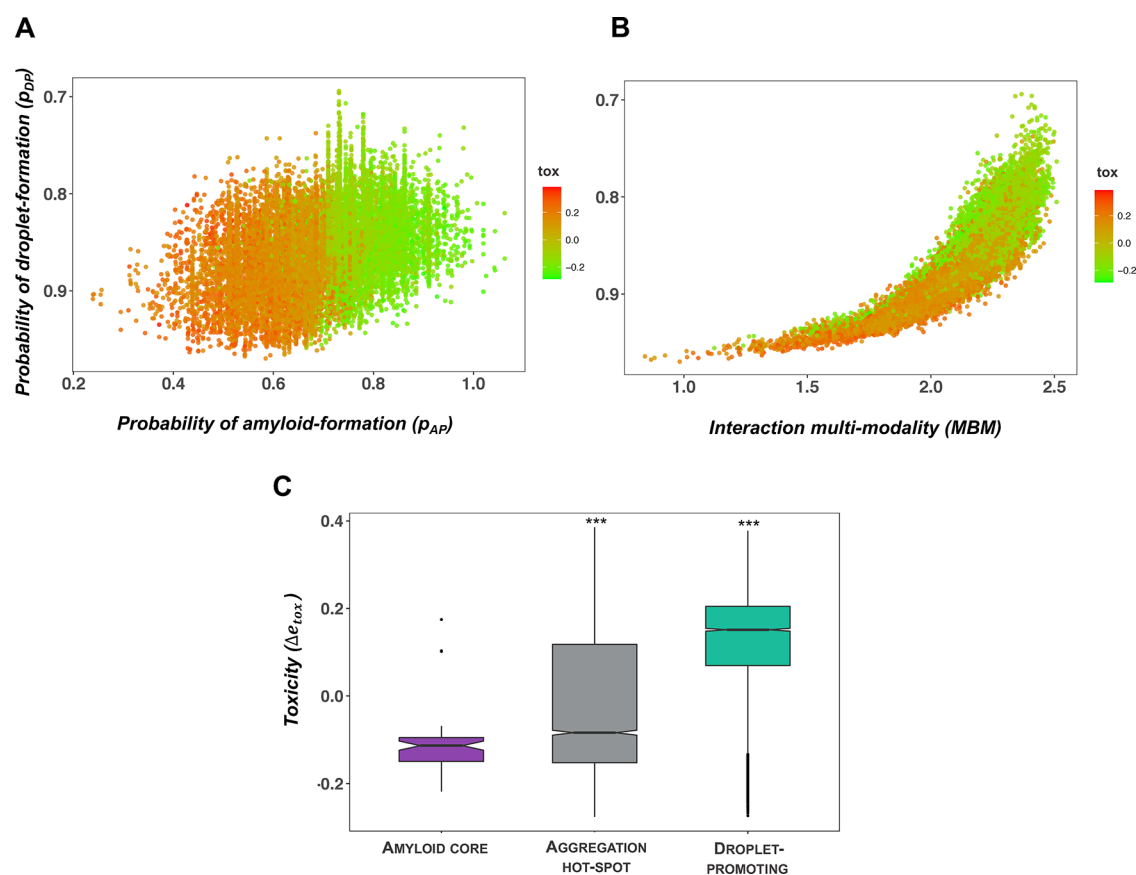


Figure 4. Investigation of the amino acid code of the cytotoxicity of TDP-43 LC double mutants. (A) Amyloid formation propensity is linked with a reduction of the cytotoxicity of TDP-43 LC double mutants. Variants with increased droplet-forming probabilities are more toxic than variants that tend to form amyloids. The cytotoxicity scale ranges from green (not cytotoxic) to red (cytotoxic) and is shown on the right. (B) Droplet formation propensity is linked with the cytotoxicity of TDP-43 LC double mutants. Droplet-promoting residues located at the left bottom of the droplet landscape (high p_{DP} and low MBM values) exhibit higher cytotoxicity than amyloid-promoting residues at the top right with lower p_{DP} and high MBM values. The cytotoxicity scale, which ranges from green (not toxic) to red (toxic), is shown on the right. (C) Cytotoxicity is linked with high droplet formation propensity. Variants were classified based on their position on the droplet landscape: droplet-promoting ($p_{DP} > 0.85$; MBM < 2.0 ; green), amyloid-promoting ($p_{DP} < 0.75$; MBM > 2.25 ; purple), and aggregation hot-spot ($0.75 < p_{DP} < 0.85$; MBM > 2.25 ; gray). The variants promoting droplet formation are significantly ($p < 10^{-6}$) more toxic than amyloid-promoting variants and aggregation hot-spots. Cytotoxicity values were taken from ref 19.

that while the p_{DP} values did not significantly change for the droplet, amyloid, and aggregation hot-spot regions (Figure 3B), the MBM values significantly increased for the droplet region (Figure 3C).

Cytotoxicity of TDP-43 Mutants Is Linked with Droplet Formation. Then, we analyzed how molecular determinants of the condensation pathway (i.e., the conversion from droplet to the amyloid state²¹) are related to cytotoxicity (Tables S1 and S2). TDP-43 variants with increased aggregation propensities did not exhibit higher cytotoxicity (Figure 4A), in accord with previous results.¹⁹ Consistently, double mutants located in the droplet region of the landscape ($p_{DP} > 0.85$; MBM < 2.0 ; Methods) have increased cytotoxicity as compared to variants located in the amyloid region ($p_{DP} < 0.75$; MBM > 2.25) (Figure 4B). Then, we compared the cytotoxicity of the amyloid, hot-spot, and droplet regions using the classification based on the droplet landscape (see Methods), obtaining results that suggest that double mutations affecting droplet formation exhibit significantly higher cytotoxicity than those promoting amyloid formation or serve as aggregation hot-spots (Figure 4C). Our analysis also indicated that mutations increasing glycine (G) or proline (P) content also increase cytotoxicity, while

those depleting these residues decrease cytotoxicity (Figure S1). This is in accord with previous results that G and P facilitate the self-organization of elastomeric sequences.³³

These results are consistent with the conclusion that TDP-43 aggregation-promoting mutants may not provide major contributions to cytotoxicity. Instead, mutations affecting the droplet state and perturbing disordered interactions increase cytotoxicity.

Extension of the FuzDrop Method to Predict the Cytotoxicity of Aging Condensates. The analysis reported above indicates that the molecular determinants of the condensation pathway offer insight into the cytotoxicity of the mutants. Thus, we probed whether we can quantitatively estimate the change in experimental cytotoxicity¹⁹ based on these quantities: the mutation-induced changes in droplet-forming (Δp_{DP}) and amyloid-forming (Δp_{AP}) probabilities as well as the change in MBM (ΔMBM). Droplet-promoting propensities (p_{DP}) were computed by the FuzDrop program,⁶ and amyloid-promoting propensities (p_{AP}) were obtained from the solubility scores by the CamSol program,³⁴ and MBM values were derived from the FuzPred program.²² We determined the differences in these quantities for the mutant and the wild-type sequences (Methods).

We used optimized random forest approaches with the out-of-bag (OOB) validation technique (Methods) on Δp_{DP} , Δp_{AP} , and ΔMBM parameters used for TDP-43 single and double mutants with $\Delta e_{\text{tox}} \geq 3\sigma$,¹⁹ respectively. We obtained Pearson's correlation coefficient between the experimental (Δe_{tox}) and predicted cytotoxicity (Δp_{tox}) values of $r = 0.975$ for 498 single mutants and $r = 0.983$ for 23,802 double mutants (Table S3). Then, we developed a combined model for both single and double TDP-43 mutants (Methods). The combined model also gave a good performance, with $r > 0.90$ between the experimental and predicted cytotoxicity values of the different data sets (Table 1 and Figure 5). In particular, it

Table 1. Correlation between Experimental (Δe_{tox}) and Predicted (Δp_{tox}) Changes in Cytotoxicity^a

mutations	data set	<i>N</i>	<i>R</i>
double	all	23,802	0.911
	droplet region	7,296	0.903
	ALS-associated	2,430	0.885
single	all	498	0.933
	droplet region	271	0.880

^aRandom forest models were developed on a combined set of single and double mutants, respectively. *N* is the size of the data set. Pearson's correlation coefficients were computed by the R program.

also exhibited a comparable performance on the data set of droplet region mutations (Table 1), which provide a major contribution to cytotoxicity. Then, we applied the model to ALS-associated mutants³⁵ (Table S4). Using 2430 double mutants, where at least one of the mutations was associated with ALS, we obtained Pearson's correlation coefficient of $r = 0.89$ (Table 1). Although cytotoxicity of TDP-43 was assessed in a model organism,¹⁹ this analysis suggests the presence of general molecular mechanisms.

Our analysis indicates that changes in cytotoxicity during droplet maturation can be predicted from the protein sequence based on the change in droplet-forming probability (Δp_{DP}), amyloid-forming probability (Δp_{AP}), and change in multiplicity of binding modes (ΔMBM).

DISCUSSION AND CONCLUSIONS

The possibility for proteins of populating different states creates a challenge for the protein homeostasis system, since dysregulated transitions into nonfunctional assemblies can generate pathological processes.^{12,16} In particular, aging condensates often appear to cause cytotoxicity and to be associated with neurological disorders.^{25,36,37}

In this study, we have investigated the amino acid code of the cytotoxicity of aging protein condensates. Our approach is based on the analysis of the binding modes in the droplet and amyloid states.^{21,38} Since changes in the multiplicity of binding modes due to sequence modifications (e.g., post-translational modifications) or cellular properties (e.g., localization) may enable proteins to switch between the different states,⁵ we reasoned that changes in binding modes may also affect interactions with cellular partners contributing to promiscuity.

Our analysis indicates that mutations promoting well-defined, ordered interactions and aggregation decrease cytotoxicity, in agreement with previous observations.¹⁹ In contrast, mutations that perturb disordered interactions and TDP-43 droplets tend to increase cytotoxicity. These observations are in accord with previous results that structurally labile regions (LARKs) are associated with TDP-43-linked pathologies.^{31,32} Earlier results also suggested that protein hydrogels may contain amyloid-like structures.^{39,40}

In conclusion, our analysis suggests that the amino acid code for the cytotoxicity of aging droplets may be similar to that for the condensation pathway from the droplet to the amyloid states and that amyloid aggregation within condensates may have a partially protective role against cytotoxicity. It will be interesting to investigate whether these conclusions will extend beyond the case of TDP-43 investigated here.

METHODS

Probability of Disordered Interactions. The probability of disordered interactions, p_{DD} , is estimated for each amino A_i as²⁰

$$p_{\text{DD}}(A_i) = \text{median}[\pi_{\text{DD}}(R_i)]_N \quad (1)$$

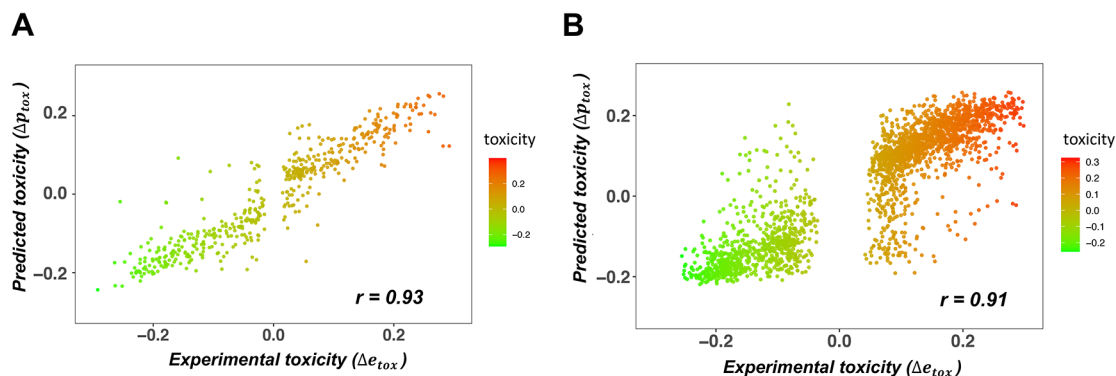


Figure 5. Correlation between experimental (Δe_{tox}) and predicted (Δp_{tox}) changes in cytotoxicity upon TDP-43 single (A) and double (B) missense mutants. Predictions were performed with the extended FuzDrop method based on three parameters: the change in droplet-promoting probability (Δp_{DP}), amyloid-promoting probability (Δp_{AP}), and change in multiplicity of binding modes (ΔMBM). (A) Prediction of the cytotoxicity of single mutants. Application of random forest models (Methods) on 498 single variants. (B) Prediction of the cytotoxicity of double mutants. Application of random forest models (Methods) on 23,802 double variants, where the parameters were averaged for the region of residues 312–341. Only 10% of the data is shown for clarity, and the *R* value is computed for the whole data set (Table 1). Pearson's correlation coefficients were calculated in R. The cytotoxicity scale ranges from green (not toxic) to red (toxic) and is shown on the right panels. Cytotoxicity values were taken from ref 19.

where $\pi_{DD}(R_i)$ is the probability of disordered binding mode of R_i , a region of 5–9 residues around A_i , and N is the number of possible regions R_i . We refer to $\pi_{DD}(R_i)$ as the “binding mode probability” of region R_i because a value of 0 indicates a binding from fully disordered to fully ordered states and a value of 1 indicates a binding from fully disordered to fully disordered states. p_{DD} is computed using the FuzPred program.²⁰

Residue-Specific Multiplicity of Binding Modes (MBM). The MBM was derived from the Shannon entropy of binding modes (S_{bind}),²² which quantifies the variability of binding modes at the amino acid level. To define S_{bind} , we start by defining the frequency f of different possible binding modes for an amino acid A_i :

$$f[\pi_{DD}(R_i)] = \frac{n_R[\pi_{DD}(R_i)]}{N} \quad (2)$$

To calculate f , the binding modes are divided into discrete bins (usually 10), and n_R is the number of binding regions within a given bin. The Shannon entropy of binding modes (S_{bind}) is then defined as the entropy of the frequencies of f :²²

$$S_{bind}(A_i) = -\sum f[\pi_{DD}(R_i)] \log_2 f[\pi_{DD}(R_i)] \quad (3)$$

where the summation is over the $[\pi_{DD}(R_i)]$ bins.

Residue-Specific Droplet-Promoting Probability. The droplet-promoting propensity profile p_{DP} quantifies the probability of spontaneous phase separation⁶

$$p_{DP}(A_i) = \frac{\exp F_S(A_i)}{1 + \exp F_S(A_i)} \quad (4)$$

where $F_S(A_i)$ is a scoring function for residue A_i

$$F_S(A_i) = \lambda_1 p_D(A_i) + \lambda_2 p_{DD}(A_i) + \gamma \quad (5)$$

Here, $p_D(A_i)$ is the probability of disorder in the free state, and $p_{DD}(A_i)$ is the probability of disordered binding. $p_D(A_i)$ approximates the conformational entropy in the unbound form, while $p_{DD}(A_i)$ estimates the conformational entropy of binding. λ_1 and λ_2 are the linear coefficients of the predictor variables and γ is a scalar constant (intercept), which were determined using the binary logistic model.⁶ p_D was derived from the disorder score as computed using the ESpritz NMR algorithm.⁴¹ The p_{DD} values were predicted by the FuzPred method.²⁰ $p_{DP} = 0.60$ is the threshold used to predict whether a residue is readily involved in spontaneous phase separation.⁶

Residue-Specific Amyloid-Promoting Probability. The amyloid-promoting propensity profile p_{AP} of a protein expresses its sequence-dependent probability to aggregate. Amyloid-promoting propensity profiles were obtained by the solubility profiles obtained by the CamSol program (p_{CS}).³⁴ $p_{AP} = -p_{CS}$ and $p_{AP} = 0.90$ is the threshold above which a protein is predicted to readily aggregate.⁴²

Analysis of TDP-43 Deep-Sequencing Data. We analyzed 498 single and 23,802 double TDP-43 missense mutants with $\Delta e_{tox} \geq 3\sigma$.¹⁹ Mutations were assigned to amyloid core (321–330 residues), droplet (262–311; 342–414), and “aggregation hot-spot” regions (312–320; 331–341) based on experimental data σ .¹⁹ For classification of double mutations, we used only both mutations in the same region. Droplet-promoting probabilities (p_{DP}) were computed by the FuzDrop program,⁶ and amyloid-promoting propensities (p_{AP}) were obtained from the solubility scores by the CamSol program.³⁴ In the case of single mutants, we

determined the differences in these quantities computed for the mutant and the wild-type (UniProt Q13148) residue. In the case of double mutants, we averaged the p_{DP} , p_{AP} , and MBM values for the 312–341 residue region and computed the difference between the average values of the mutant and wild-type sequence.

Extending the FuzDrop Method to Predicting Cytotoxicity of Protein Droplets upon Mutations. We defined Δp_{tox} using a machine learning method with three input parameters: the difference between the mutant and wild-type protein in residue-specific droplet-promoting probability (Δp_{DP}) as computed by the original FuzDrop method,⁶ the change in residue-specific amyloid-promoting probability (Δp_{AP}) obtained as the negative of the solubility score of the CamSol program,³⁴ and the change in multiplicity of binding modes (ΔMBM) obtained as the differences between the S_{bind} by the FuzPred method.²²

The models were built using the random forest method with the OOB validation technique where two-thirds of the original data set is used for training and validation is performed on the remaining part. Random forest models with the highest Pearson’s correlation coefficients were inferred using grid optimization on the parameters of the number of individual decision trees (*n_{tree}*) and the number of variables used at each split (*m_{try}*) with the randomForest package using the R program. Models were developed on single and double mutation data sets, respectively (Table S3), as well as using a combined data set of 498 single mutants and 23802 double mutants. In-house R scripts used to generate data and figures as well as the serialized random forest models can be downloaded from the GitHub repository (https://github.com/ahorvath/Biochemistry_2022.git).

The models were tested on all mutants with $\Delta e_{tox} \geq 3\sigma$, as well as on mutations of the droplet region (Table 1 and Table S3). In addition, the models were tested on mutations, where at least one of the mutations was ALS-associated³⁵ (Table S4).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.biochem.2c00499>.

Characterization of the single and double TDP-43 mutants and ALS-associated mutants and the performance of the model (PDF)

Molecular determinants of the condensation pathway (XLSX)

Molecular determinants of the condensation pathway (XLSX)

Single and double mutation data sets (XLSX)

ALS-associated mutants (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Monika Fuxreiter – Department of Biomedical Sciences and Department of Physics and Astronomy, University of Padova, Padova, PD 35131, Italy; orcid.org/0000-0002-4463-6727; Email: monika.fuxreiter@unipd.it

Authors

Attila Horvath – John Curtin School of Medical Research, The Australian National University, Canberra 2600, Australia

Michele Vendruscolo – Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK; orcid.org/0000-0002-3616-1610

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.biochem.2c00499>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

M.F. would like to thank the financial support of the AIRC Foundation for Cancer Research I.G. 26229.

REFERENCES

- (1) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 384–396.
- (2) Michaels, T. C. T.; Šarić, A.; Curk, S.; Bernfur, K.; Arosio, P.; Meisl, G.; Dear, A. J.; Cohen, S. I. A.; Dobson, C. M.; Vendruscolo, M.; Linse, S.; Knowles, T. P. J. Dynamics of oligomer populations formed during the aggregation of Alzheimer's Aβ42 peptide. *Nat. Chem.* **2020**, *12*, 445–451.
- (3) 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.
- (4) Hyman, A. A.; Weber, C. A.; Jülicher, F. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **2014**, *30*, 39–58.
- (5) Fuxreiter, M.; Vendruscolo, M. Generic nature of the condensed states of proteins. *Nat. Cell Biol.* **2021**, *23*, 587–594.
- (6) 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. U. S. A.* **2020**, *117*, 33254–33262.
- (7) Boeynaems, S.; Alberti, S.; Fawzi, N. L.; Mittag, T.; Polymenidou, M.; Rousseau, F.; Schymkowitz, J.; Shorter, J.; Wolozin, B.; Van Den Bosch, L.; Tompa, P.; Fuxreiter, M. Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.* **2018**, *28*, 420–435.
- (8) Lyon, A. S.; Peeples, W. B.; Rosen, M. K. A framework for understanding the functions of biomolecular condensates across scales. *Nat Rev Mol Cell Biol.* **2021**, *22*, 215–235.
- (9) Alberti, S.; Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nat Rev Mol Cell Biol.* **2021**, *22*, 196–213.
- (10) Walters, R. W.; Muhrlad, D.; Garcia, J.; Parker, R. Differential effects of Ydj1 and Sis1 on Hsp70-mediated clearance of stress granules in *Saccharomyces cerevisiae*. *RNA.* **2015**, *21*, 1660–1671.
- (11) Buchan, J. R.; Kolaitis, R. M.; Taylor, J. P.; Parker, R. Eukaryotic stress granules are cleared by autophagy and Cdc48/VCP function. *Cell.* **2013**, *153*, 1461–1474.
- (12) Vendruscolo, M.; Fuxreiter, M. Protein condensation diseases: therapeutic opportunities. *Nat. Commun.* **2022**, *13*, 5550.
- (13) Schmidt, H. B.; Görlich, D. Nup98 FG domains from diverse species spontaneously phase-separate into particles with nuclear pore-like permselectivity. *eLife.* **2015**, *4*, No. e04251.
- (14) Murakami, T.; Qamar, S.; Lin, J. Q.; Schierle, G. S.; Rees, E.; Miyashita, A.; Costa, A. R.; Dodd, R. B.; Chan, F. T.; Michel, C. H.; Kronenberg-Versteeg, D.; Li, Y.; Yang, S. P.; Wakutani, Y.; Meadows, W.; Ferry, R. R.; Dong, L.; Tartaglia, G. G.; Favrin, G.; Lin, W. L.; Dickson, D. W.; Zhen, M.; Ron, D.; Schmitt-Ulms, G.; Fraser, P. E.; Shneider, N. A.; Holt, C.; Vendruscolo, M.; Kaminski, C. F.; St George-Hyslop, P. ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron.* **2015**, *88*, 678–690.
- (15) Patel, A.; Lee, H. O.; Jawerth, L.; Maharana, S.; Jahnke, M.; Hein, M. Y.; Stoyanov, S.; Mahamid, J.; Saha, S.; Franzmann, T. M.; Pozniakovski, A.; Poser, I.; Maghelli, N.; Royer, L. A.; Weigert, M.; Myers, E. W.; Grill, S.; Drechsel, D.; Hyman, A. A.; Alberti, S. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell.* **2015**, *162*, 1066–1077.
- (16) Alberti, S.; Dormann, D. Liquid-Liquid Phase Separation in Disease. *Annu. Rev. Genet.* **2019**, *53*, 171–194.
- (17) Mathieu, C.; Pappu, R. V.; Taylor, J. P. Beyond aggregation: Pathological phase transitions in neurodegenerative disease. *Science.* **2020**, *370*, 56–60.
- (18) Tan, P. S.; Aramburu, I. V.; Mercadante, D.; Tyagi, S.; Chowdhury, A.; Spitz, D.; Shammas, S. L.; Gräter, F.; Lemke, E. A. Two Differential Binding Mechanisms of FG-Nucleoporins and Nuclear Transport Receptors. *Cell Rep.* **2018**, *22*, 3660–3671.
- (19) Bolognesi, B.; Faure, A. J.; Seuma, M.; Schmedel, J. M.; Tartaglia, G. G.; Lehner, B. The mutational landscape of a prion-like domain. *Nat. Commun.* **2019**, *10*, 4162.
- (20) Miskei, M.; Horvath, A.; Vendruscolo, M.; Fuxreiter, M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.* **2020**, *432*, 2289–2303.
- (21) Vendruscolo, M.; Fuxreiter, M. Sequence determinants of the aggregation of proteins within condensates generated by liquid-liquid phase separation. *J. Mol. Biol.* **2021**, 167201.
- (22) Horvath, A.; Miskei, M.; Ambrus, V.; Vendruscolo, M.; Fuxreiter, M. Sequence-based prediction of protein binding mode landscapes. *PLoS Comput. Biol.* **2020**, *16*, No. e1007864.
- (23) Niaki, A. G.; Sarkar, J.; Cai, X.; Rhine, K.; Vidaurre, V.; Guy, B.; Hurst, M.; Lee, J. C.; Koh, H. R.; Guo, L.; Fare, C. M.; Shorter, J.; Myong, S. Loss of Dynamic RNA Interaction and Aberrant Phase Separation Induced by Two Distinct Types of ALS/FTD-Linked FUS Mutations. *Mol. Cell.* **2020**, *77*, 82.
- (24) Rhine, K.; Makurath, M. A.; Liu, J.; Skanchy, S.; Lopez, C.; Catalan, K. F.; Ma, Y.; Fare, C. M.; Shorter, J.; Ha, T.; Chemla, Y. R.; Myong, S. ALS/FTLD-Linked Mutations in FUS Glycine Residues Cause Accelerated Gelation and Reduced Interactions with Wild-Type FUS. *Mol. Cell.* **2020**, *80*, 666.
- (25) Gopal, P. P.; Nirschl, J. J.; Klinman, E.; Holzbaur, E. L. Amyotrophic lateral sclerosis-linked mutations increase the viscosity of liquid-like TDP-43 RNP granules in neurons. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E2466–E2475.
- (26) Li, Y. R.; King, O. D.; Shorter, J.; Gitler, A. D. Stress granules as crucibles of ALS pathogenesis. *J. Cell Biol.* **2013**, *201*, 361–372.
- (27) Gianni, S.; Freiburger, M. I.; Jemth, P.; Ferreira, D. U.; Wolynes, P. G.; Fuxreiter, M. Fuzziness and Frustration in the Energy Landscape of Protein Folding Function, and Assembly. *Acc. Chem. Res.* **2021**, *54*, 1251–1259.
- (28) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife.* **2018**, *7*, No. e31486.
- (29) Cao, Q.; Boyer, D. R.; Sawaya, M. R.; Ge, P.; Eisenberg, D. S. Cryo-EM structures of four polymorphic TDP-43 amyloid cores. *Nat. Struct. Mol. Biol.* **2019**, *26*, 619–627.
- (30) Hatos, A.; Tosatto, S. C. E.; Vendruscolo, M.; Fuxreiter, M. FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid-liquid phase separation and aggregation of proteins. *Nucleic Acids Res.* **2022**, *50*, W337–W344.
- (31) Conicella, A. E.; Zerze, G. H.; Mittal, J.; Fawzi, N. L. ALS Mutations Disrupt Phase Separation Mediated by alpha-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure.* **2016**, *24*, 1537–1549.
- (32) Guenther, E. L.; Cao, Q.; Trinh, H.; Lu, J.; Sawaya, M. R.; Cascio, D.; Boyer, D. R.; Rodriguez, J. A.; Hughes, M. P.; Eisenberg, D. S. Atomic structures of TDP-43 LCD segments and insights into reversible or pathogenic aggregation. *Nat. Struct. Mol. Biol.* **2018**, *25*, 463–471.

- (33) Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomes, R. Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. *Structure* **2006**, *14*, 1667–1676.
- (34) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **2015**, *427*, 478–490.
- (35) Kapeli, K.; Martinez, F. J.; Yeo, G. W. Genetic mutations in RNA-binding proteins and their roles in ALS. *Hum. Genet.* **2017**, *136*, 1193–1214.
- (36) 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.
- (37) Mackenzie, I. R.; Nicholson, A. M.; Sarkar, M.; Messing, J.; Purice, M. D.; Pottier, C.; Annu, K.; Baker, M.; Perkerson, R. B.; Kurti, A.; Matchett, B. J.; Mittag, T.; Temirov, J.; Hsiung, G. R.; Krieger, C.; Murray, M. E.; Kato, M.; Fryer, J. D.; Petrucelli, L.; Zinman, L.; Weintraub, S.; Mesulam, M.; Keith, J.; Zivkovic, S. A.; Hirsch-Reinshagen, V.; Roos, R. P.; Zuchner, S.; Graff-Radford, N. R.; Petersen, R. C.; Caselli, R. J.; Wszolek, Z. K.; Finger, E.; Lippa, C.; Lacomis, D.; Stewart, H.; Dickson, D. W.; Kim, H. J.; Rogaeva, E.; Bigio, E.; Boylan, K. B.; Taylor, J. P.; Rademakers, R. TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron* **2017**, *95*, 809–816.
- (38) Fuxreiter, M. Classifying the Binding Modes of Disordered Proteins. *Int. J. Mol. Sci.* **2020**, *21*, 8615.
- (39) Kato, M.; Han, T. W.; Xie, S.; Shi, K.; Du, X.; Wu, L. C.; Mirzaei, H.; Goldsmith, E. J.; Longgood, J.; Pei, J.; Grishin, N. V.; Frantz, D. E.; Schneider, J. W.; Chen, S.; Li, L.; Sawaya, M. R.; Eisenberg, D.; Tycko, R.; McKnight, S. L. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **2012**, *149*, 753–767.
- (40) Kato, M.; McKnight, S. L. A Solid-State Conceptualization of Information Transfer from Gene to Message to Protein. *Annu. Rev. Biochem.* **2018**, *87*, 351–390.
- (41) Walsh, I.; Martin, A. J.; Di Domenico, T.; Tosatto, S. C. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **2012**, *28*, 503–509.
- (42) Sormanni, P.; Vendruscolo, M. Protein Solubility Predictions Using the CamSol Method in the Study of Protein Homeostasis. *Cold Spring Harbor Perspect. Biol.* **2019**, *11*, a033845.