Leveraging a large language model to predict protein phase transition: a physical, multiscale and interpretable approach

Mor Frank[a], **Pengyu Ni**[a], **Matthew Jensen**[a], **and Mark B Gerstein**[a,b,c,d]

[a]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06510, USA; [b]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA; [c]Department of Computer Science, Yale University, New Haven, CT 06511, USA; [d]Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA

This manuscript was compiled on November 21, 2023

Abstract

Protein phase transitions (PPTs) from the soluble state to a dense liquid phase (forming droplets via liquid-liquid phase separation) or solid aggregates (such as amyloid) play key roles in pathological processes associated with age-related diseases such as Alzheimer's disease (AD). Several computational frameworks are capable of separately predicting the formation of protein droplets or amyloid aggregates based on protein sequences, yet none have tackled the prediction of both within a unified framework. Recently, large language models (LLMs) have exhibited great success in protein structure prediction; however, they have not yet been used for PPTs. Here, we fine-tune a LLM for predicting PPTs and demonstrate its superior performance compared to suitable classical benchmarks. Due to the "black-box" nature of the LLM, we also employ a classical random forest model along with biophysical features to facilitate interpretation. Finally, focusing on AD-related proteins, we demonstrate that greater aggregation is associated with reduced gene expression in AD, suggesting a natural defense mechanism.

protein phase transition | computational models | gene expression

DRAFT

Protein phase transitions (PPTs) play an important role in both physiology and disease. Proteins typically exist in a soluble form, but they have the capacity to transition into droplets (i.e., undergo liquid-liquid phase separation [LLPS]) and solid condensates. Numerous studies indicate that proteins undergoing LLPS are involved in various cellular processes, such as transcription regulation, postsynaptic signaling, and genome organization (1–3). In addition to the physiological importance of LLPS, emerging evidence indicates that, in neurodegenerative diseases, proteins can induce a liquid-to-solid transition to form amyloid aggregates (4, 5). Unlike LLPS, which is largely reversible, the liquid-to-solid transition is often irreversible (6).

Although it is well established that external conditions affect protein misfolding and PPT (7), understanding the fundamental factors inherently encoded by the protein sequence that drive PPT becomes an additional objective. In this regard, recent studies have shown the association between LLPS and the biophysical properties encoded by the amino acid sequence, utilizing machine learning algorithms and language models (8–11), and demonstrated that at least 40% of the human proteins, named "droplet-drivers," undergo spontaneous LLPS under physiological conditions (12). However, previous modeling methods predominantly concentrate on predicting LLPS or aggregation as distinct objectives rather than describing a unified process of PPT from the soluble native state to droplet and amyloid formation. In addition, the usage of state-of-the-art techniques, such as attention-based language models, has not been fully leveraged to predict the physical mechanisms underlying diseases.

Under cellular conditions, LLPS is meta-stable for most proteins, as it is kinetically trapped by a free energy barrier that inhibits the transition to a solid state (e.g., amyloid aggregation) (13). Yet, it remains unclear whether the ability to maintain this liquid meta-stable state is inherently encoded in the amino acid sequence and which biophysical properties can drive amyloid formation. Various sequence-related features affecting the formation of protein condensates have been reported, such as hydrophobic and disordered interactions (14). Here, we investigate the hypothesis that PPT is encoded by the amino acid sequence and can be directly predicted from it. Previous research using a biophysical quantity called "binding mode entropy" (BME) demonstrated that droplet-forming proteins exhibit low BME values, indicating a strong preference for disordered binding. In contrast, amyloid-forming proteins exhibit high BME values, as they can engage in both disordered and ordered interactions (15).

Extending this approach, we calculated multiple knowledge-based features directly from the amino acid sequence and incorporated them into classical machine learning models. We also fine-tuned a large language model (LLM) called ESMFold (16) to produce contextu-

Significance Statement

The protein phase transition is a physical mechanism associated with both physiological processes and age-related diseases. Here, we present a modeling approach for predicting a specific protein sequence's propensity to undergo phase transitions directly from its sequence. Our methodology involves utilizing a large language model to analyze the likelihood of a given protein sequence existing in a particular material state. Additionally, for enhanced interpretability, we incorporate a classical knowledge-based model. Our results suggest the potential for accurately predicting the propensity to form either liquid or solid condensates. Furthermore, our findings indicate the potential regulation of this propensity by gene expression under pathological conditions to prevent aggregation.

Dr. Mor Frank - end-to-end modeling framework from concept to actual methodology of both the machine learning and language models. Code writing covers all the pipeline from feature engineering, and modeling approaches to analyzing modeling results as well as writing the manuscript. Dr. Pengyu Ni - uncovering gene-to-protein relations in AMP-AD dataset, statistical analyses of gene expressions as well as helping in reviewing the paper. Dr. Matthew Jensen - preprocessed the AMP-AD dataset and obtained the gene expression matrix, Prof. Mark B Gerstein - contributed as an expert in the area of biophysics-related computational models through scientific and professional discussions as well as reviewing the manuscript.

No competing interests.

[a]E-mail: pi@gersteinlab.org

alized embedding representations of the protein sequences. Each of the models predicted the propensity of a given protein sequence to undergo a phase transition, forming droplets and amyloid aggregates in particular (Fig. 1).

We further applied our modeling framework to unlabeled protein sequences extracted from the Accelerating Medicines Partnership program for Alzheimer's disease (AD) (AMP-AD) knowledge portal (17) to predict their propensity to form liquid and solid condensates. Subsequently, we examined the abundance of the predicted droplets versus proteins prone to amyloid formation.

Finally, in an effort to establish a link between PPT propensity and gene expression, we compared the gene expression of amyloid beta precursor (APP) and tau proteins, both known to contribute pathologically to AD, to better understand how PPT is regulated under pathological conditions. Our results suggest the possibility that the tendency to form amyloid aggregates, as well as liquid droplets, is encoded by the amino acid sequence. Furthermore, this process appears to be self-regulated by gene expression, thereby controlling protein aggregation, which is associated with pathological conditions such as AD.

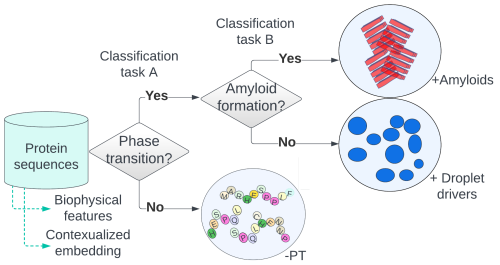

Fig. 1. Modeling approach for PPT. In classification task A proteins exhibiting experimental evidence of undergoing a phase transition, forming either droplets or amyloids, are consolidated into a single dataset (+Droplet drivers and +Amyloids). Phase transition propensity is predicted versus the preference to maintain the native soluble state (-PT). In classification task B the unified dataset is utilized to predict the propensity to form droplets versus amyloid aggregates.

Results and discussion

Leveraging ESMFold, a large language model, for predicting protein phase transitions.

Recently, pre-trained language models employing deep bidirectional attention have proven to be useful in embedding biological sequences (16, 18). Here, we utilize the evolutionary scale ESMFold model, which was trained on a large database of evolutionarily diverse protein sequences (19). Unlike the AlphaFold model (20), which employs a deep neural network architecture to facilitate information exchange between protein sequences (i.e., an Evoformer block), the ESMFold utilizes a masked transformer to analyze a single sequence. This distinction makes ESMFold an order of magnitude faster than AlphaFold in predicting protein structures. Leveraging this advantage, we fine-tuned the ESMFold for predicting PPT in the above-mentioned classification tasks. To accomplish this, we used the pre-trained ESM-2, a transformer-based language model that is part of the ESMFold architecture. Additionally, we incorporated a linear layer on top the model to handle the classification task instead of using the folding module (Fig. 2).

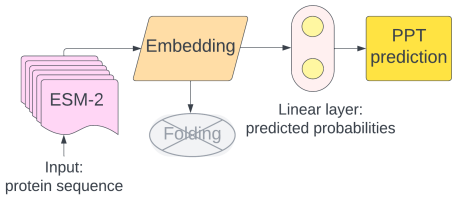

Fig. 2. ESM-2, a pre-trained LLM via masking, is utilized to predict PPT. The folding module is removed from the original architecture of the ESMFold model, while a linear head is added on top for employing a classification task.

In the following sections, we demonstrate that this modeling approach for predicting PPT exhibits better performance than classical model benchmarks (Supplementary Tables S1 and S2). The discriminative power of the LLM versus the random forest model is demonstrated through principal component analysis (PCA) of the embedding vectors extracted from the last layer of the LLM versus the PCA of the biophysical features (Fig. 3).

Database construction of the biophysical features.

We used two databases containing full protein sequences that represent the propensity to form droplets (*+Droplet drivers*) and amyloid aggregates (*+Amyloids*), following the process in (15) with no further modifications. It should be noted that our analyses excluded proteins observed to exhibit both droplet and amyloid formation. This decision was driven by the limited availability of data, with only 20 proteins reported to possess such dual capabilities. However, our model captures the degree to which proteins are prone to undergo LLPS during the process of amyloid formation. In addition, we restricted our analyses to droplet drivers to specifically focus on the biophysical properties of proteins that do not require an interacting partner to undergo LLPS. Moreover, we constructed a negative database of well-structured proteins randomly chosen from the database used in (8). These proteins are highly unlikely to undergo LLPS and are not reported to form amyloid aggregates, thus demonstrating a minimal propensity for undergoing a phase transition (*-PT*) (Materials and methods).

We computed the following biophysical features considering multiple protein scales (Fig. 4 (A-E), and Table 1). First, we identified

Table 1. Biophysical features.

Feature	Reference
1. Fraction of disordered regions	(21)
2. Gyration radius	(22)
3. Instability index	(23)
4. Probability of solubility	(24)
5. Avg. hydrophobicity	(25)

disordered residues (Materials and methods) by using the iupred3 algorithm (21). In accordance with (15), our analyses revealed that +Droplet drivers had a higher fraction of disordered regions compared to +Amyloids, which preserved a more ordered structure as demonstrated in Fig. 4A. Second, to account for the 3D structure of the proteins, we used the pdb files extracted from the AlphaFold (26) database or used the ESMFold model (19) and calculated the gyration radius of each protein sequence (Materials and methods). We concluded that +Amyloids tend to have a more compact structure, consistent with their tendency to have more ordered interactions

DRAFT

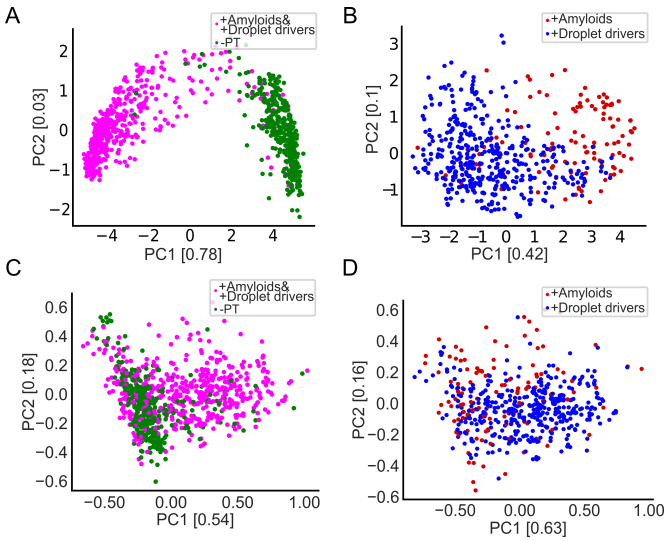

Fig. 3. PCA of the embedding vectors, extracted from the LLM, and the biophysical features for both classification tasks. Each principle direction is presented along with its explained variance. The embedding vector of each protein sequence is the mean of all the 320-dimensional embedding vectors representing the amino acids in the sequence and was extracted from the last hidden layer. (A) and (B) The PCA of the embedding vectors are presented for classification tasks A (+Droplet drivers and +Amyloids versus -PT) and B (+Droplet drivers versus +Amyloids), correspondingly. (C) and (D) The PCA of the biophysical features of the two classification tasks, A and B, are shown.

compared to +Droplet driver proteins, which primarily demonstrate disordered interactions (Fig. 4B). This compact structure maximizes the interactions, potentially encoded in the amino acid sequence, and promotes aggregation. Third, we calculated the instability index, II, of a given protein sequence as shown in Fig. 4C, (Materials and methods). +Droplet drivers are associated with high II in accordance with their unstable disordered structure under physiological conditions (27) as opposed to -PT, which shows high stability. The +Amyloids showed a relatively low II, aligning with their propensity to undergo misfolding while maintaining a stable structure of beta sheet-rich assemblies (28). This is opposed to +Droplet drivers, which are associated with a high fraction of disordered regions and a lack of a stable 3D structure (29). We next estimated the solubility probability for all protein sequences using the solubility weighted index, SWI (Materials and methods). We found that +Droplet drivers and -PT exhibited a higher probability of solubility than +Amyloids and the unlabeled sequences extracted from the AMP-AD dataset (Fig. 4D). This characteristic could be linked to their important cellular-level functions while evading aggregation. Finally, we calculated the mean hydrophobicity using the Kyte and Doolittle hydropathy scale for each protein sequence (Fig. 4E). We found that +Droplet drivers showed the lowest values, suggesting an essential role of hydrophilic interactions in the LLPS process. These interactions allow the chains to remain disordered, maintaining the condensates in a liquid state rather than a solid one (14). In contrast, +Amyloids exhibited higher values associated with hydrophobic interactions, which promote aggregation.

Prediction of phase transition.

We developed a two-step modeling framework. First, we predicted whether a given protein would undergo a phase transition or prefer to remain in the native state. For this, we considered the +Droplet drivers and +Amyloids as a unified dataset, comparing them against the -PT dataset (Fig. 1, classification task A). The protein sequences within the unified class of +Droplet drivers and +Amyloids were further used to predict their propensity to form droplets or amyloid aggregates (Fig. 1, classification task B). Both classification tasks were divided into training and test sets in an 80:20 ratio.

We calculated the contribution of each biophysical feature using the Gini score (31), as presented in Fig. 5. Notably, the order of the features was preserved, suggesting that both classification tasks are driven by the same physical principles. Specifically, the instability index was identified as the most important feature, followed by the fraction of disordered regions, both of which are known to drive protein PPT (12).

For both classification tasks, two clusters could be distinguished through PCA performed on either the embedding vectors or the biophysical features. Remarkably, for both classification tasks, a clearer distinction between the two classes was achieved when using the LLM embedding vectors (Figs. 3A and 3B). This result emphasizes the importance of latent relationships derived from the protein sequence, which affect the likelihood of PPT occurrence. A notable separation was achieved between the unified dataset of +Droplet drivers and +Amyloids versus -PT (Figs. 3A and 3C). Additionally, we observed a relatively clear separation between +Droplet drivers and +Amyloids (Figs. 3B and 3D). This difference supports the idea that the trade-off between the proper functionality of droplets and the risk of forming solid aggregates, i.e., amyloid, is governed by the biophysical properties encoded in the amino acid sequence. Nonetheless, distinguishing between +Droplet drivers and +Amyloids was more challenging as droplet-promoting proteins can transition into an amyloid state under certain conditions, as demonstrated in a recent study of LLPS-prone proteins (32).

To evaluate the performance of the two models, we performed five-fold cross-validation in a stratified manner. We used accuracy, arithmetic mean of recall and precision, and the area under the receiver-operator-characteristic (AUROC) to evaluate the random forest classifier and the LLM as shown in Fig. 6. The two models successfully distinguished between proteins that have a high propensity to undergo a phase transition, whether droplets or amyloid aggregates, as opposed to those that tend to remain in their native state (Fig. 6A). The distinction between +Droplet drivers and +Amyloids was fairly notable as well (Fig. 6B). This classification task is much more challenging, as the +Amyloids can have both disordered and ordered binding modes (i.e., interactions) that are associated with both liquid and solid con-

DRAFT

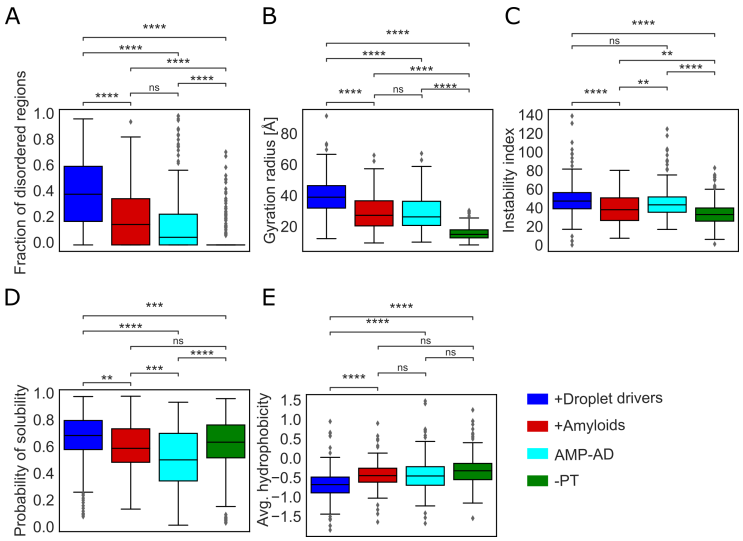

Fig. 4. Wide range of biophysical features demonstrating the differences between the four datasets. (A) Fraction of disordered regions, (B) Gyration radius, (C) Instability index, (D) Probability of solubility, and (E) Average Hydrophobicity. A significant difference was found between +Droplet drivers and +Amyloids. When comparing -PT to +Droplet drivers or +Amyloids, a significant difference emerged in most cases. As expected, +Droplet drivers show a more disordered structure compared to +Amyloid, while -PT shows almost no disordered regions (30). +Amyloids have a similar gyration radius as the AMP-AD dataset, which are both lower compared to +Droplet drivers. +Amyloids show reduced instability index values compared to +Droplet drivers. Overall, the solubility probability of +Droplet drivers and -PT is higher compared to +Amyloids and AMP-AD datasets. +Droplet drivers show the lowest average hydrophobicity, and -PT shows the highest. Significance was tested using the Welch statistical t-test. **** p <= 1.00e-04, *** 1.00e-04 < p <= 1.00e-03, ** 1.00e-03 < p <= 1.00e-02, and ns is not significant.

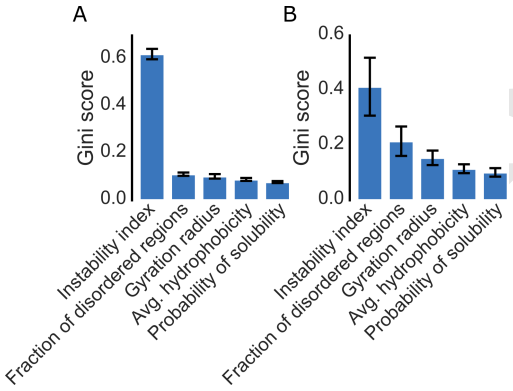

Fig. 5. Feature importance calculated for the two classification tasks by the mean Gini score derived from the 5-fold cross-validation along with the standard deviation. The order of the features (from the most important to the lowest) was preserved. (A) Classification task A: +Droplet drivers and +Amyloids versus -PT. The instability index shows a significant discrimination power while other features have similar contributions. (B) Classification task B: +Droplet drivers versus +Amyloids. The instability index and the fraction of disordered regions are the most important ones along with a gradual decrease in the Gini score of the other features, emphasizing their accumulated contribution to the prediction.

densates. Remarkably, all of the metric quantities were higher in the LLM than those of the random forest classifier for both classification tasks. The high predictive power of the LLM is in line with the PCA presented in Figs. 3A and 3B, where a clearer distinction between the two classes was achieved by the embedding vectors than the biophysical features. This indicates that the LLM successfully identifies regions that drive PPT, suggesting that various biophysical properties, beyond the ones presented in Fig. 4, are encoded by the amino acid sequence.

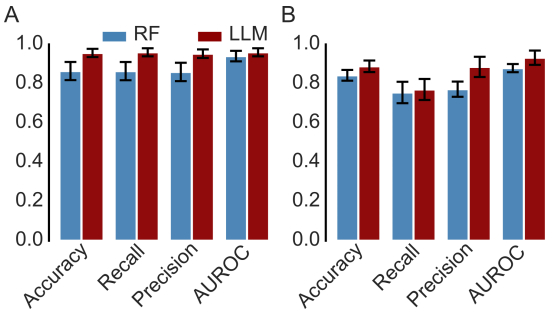

Fig. 6. Metrics evaluating the random forest (RF) classifier and the LLM by using 5-fold cross-validation. For each metric, the mean and standard deviation are presented for the two classification tasks. The LLM demonstrated the highest performance. (A) Classification task A: +Droplet drivers and +Amyloids versus -PT. Both models showed consistently high performance for all four metrics, which were all above 0.8 for RF and 0.9 for the LLM. (B) Classification task B: +Droplet drivers versus +Amyloids. Performances were still high, however, lower than classification task A due to the challenging distinction between the two classes.

Next, we created an ensemble of the random forest model and the LLM (ML_LM model) by averaging their predicted probabilities (Materials and methods). To evaluate the performance of the ML_LM model, we calculated the same metrics for the test set. In both classification tasks, the LLM demonstrated a better performance

than the random forest model in almost all of the metrics, while the ML_LM model achieved the highest AUROC (Fig. 7). Furthermore, the ML_LM model significantly improved the ability to identify protein sequences belonging to either +Droplet drivers or +Amyloids (Fig. 7B, Recall metric).

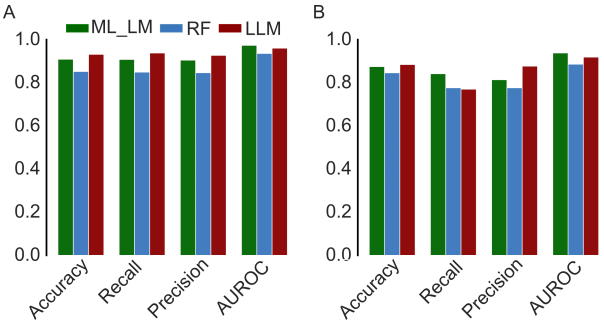

Fig. 7. Performance evaluation on the test sets by using the four metrics: Accuracy, arithmetic mean of recall, precision, and AUROC. (A) The performance of classification task A: +Droplet drivers and +Amyloids vs. -PT showed consistently high measures all above 0.9 for the LLM and the ML_LM. (B) For classification task B: +Droplet drivers vs. +Amyloids all four metrics were higher than 0.8 for the ML_LM.

To quantify a protein's inherent tendency to preferentially adopt a specific state, we introduced a scoring function called the *"Transition score,"* which is defined by a log-odds ratio using the probability of undergoing PPT (Materials and methods). First, we calculated the probability of transitioning from the native state, specifically into either droplets or amyloid states (classification task A). We applied Eq. 4 on the unified dataset of +Droplet drivers and +Amyloid versus -PT, where p is the probability of being in the unified +Droplet drivers and +Amyloids class. Our analysis resulted in a clear separation, with the largest densities located in negative and positive ranges, suggesting the predictability of PPT (Fig. 8A). Second, we used this measure to distinguish between +Droplet drivers and +Amyloids by using p in Eq. 4 as the probability of being in the +Amyloids class (classification task B). In this classification task, a lower transition score suggests that the protein samples more disordered interactions than ordered ones, potentially promoting LLPS prior to amyloid formation. We achieved a reasonable discrimination between these two classes (Fig. 8B), with some overlap between the two distribution scores. However, high densities were located within both negative and positive ranges.

Predictions of unlabeled protein sequences.

Next, we applied our modeling framework to 189 unlabeled protein sequences extracted from the AMP-AD database (17) (dataset S4 and Fig. S4), with predictions derived from the ML_LM model. The transition score of +Droplet drivers and +Amyloids versus -PT revealed two distributions. Most of the protein sequences were predicted to have a high propensity to undergo a phase transition, forming either droplets or amyloid aggregates, which is in line with previous findings (12) (Figs. 9A and 9C and dataset dataset S5). Next, we further evaluated the propensity of these proteins to form either +Droplet drivers or +Amyloids. Our analyses demonstrated that most proteins exhibited a high propensity to undergo a phase transition, mainly forming droplets (Figs. 9B and 9D and dataset dataset S6) emphasizing the physiological relevance of LLPS (33–35). Our predictions for two proteins known to be pathological in AD, APP and tau, revealed their propensity to undergo phase transitions (Fig. 10). Specifically, APP

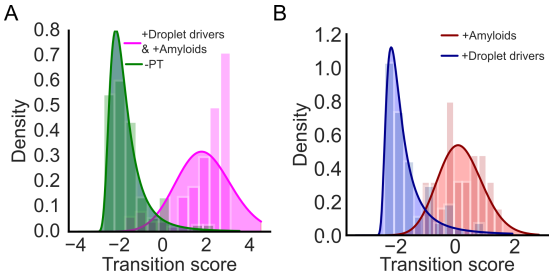

Fig. 8. Probability density of the transition scores along with the fitted alpha distributions (solid line) for the two test sets using ML_LM. (A) Transition scores of classification task A: +Droplet drivers and +Amyloids versus -PT. A clear distinction was achieved between the distributions of the two classes by being located separately in the negative and positive ranges. (B) Transition scores of classification task B: +Amyloids are associated with higher transition scores, which are mostly in the positive range.

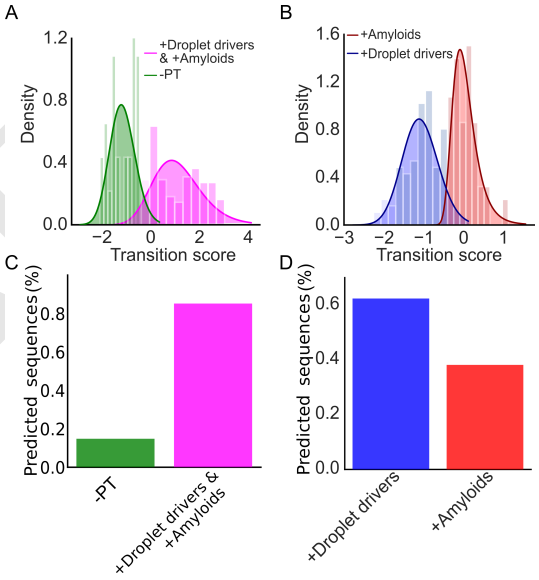

Fig. 9. Probability density of the transition scores along with the fitted alpha distributions (solid line) calculated for the unlabeled protein sequences along the percentage of the predicted sequences for each classification task. (A) +Droplet drivers and +Amyloids are shifted to positive scores compared to -PT, which is concentrated in the negative range. (B) +Droplet drivers and +Amyloids are mostly located in the negative and positive ranges, correspondingly. (C) The majority of the unlabeled proteins were predicted as +Droplet drivers and +Amyloids. (D) Most of the unlabeled protein sequences were predicted as having high droplet propensity.

was predicted as having a high propensity to form amyloid aggregates along with a medium transition score (i.e., persisting in the droplet state to a medium extent). This is in line with the nature of this protein to form amyloid peptides upon proteolytic cleavage, such as beta-amyloid 42, which is considered to be very toxic and could potentially serve as an early biomarker for AD (36, 37). This indicates that the products of APP might have a lower free energy barrier for amyloid formation and exhibit fewer disordered regions typically associated with LLPS. In contrast, the protein tau was predicted as a +Droplet driver with a lower transition score, suggesting a preference for the

DRAFT

droplet state. This aligns with tau's lack of a well-defined structure and its enrichment with hydrophilic amino acids, which can facilitate LLPS. This is in accordance with the observation that phase-separated tau droplets can both serve physiological functions (e.g., microtubule assembly) and act as an intermediate toward aggregation (38). Hence, LLPS might be considered a significant initiating factor in the onset of age-related diseases. External factors, such as phosphorylation – a potent regulator of tau (39–41) – can further contribute to the conversion of LLPS into aggregation. Taken together, these predictions support the idea that amyloid beta plaques produced from APP can promote the formation of tau entanglements arising from aged droplets, which in turn damage neurons (42).

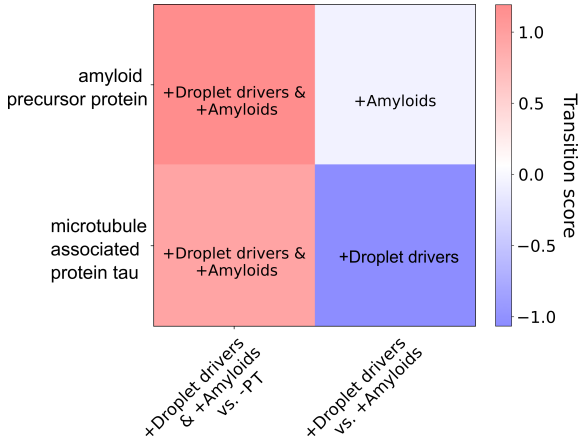

Fig. 10. PPT prediction for APP and tau, two AD-related proteins. Transition scores along with the predicted class were calculated by the ML_LM model. Both proteins exhibited a strong propensity to undergo PPT (left panel). APP was predicted as having a high propensity to form amyloid aggregates while tau was predicted as having a high propensity to form droplets (right panel).

To gain a deeper understanding of how the LLM makes PPT predictions, we extracted self-attention maps for APP and tau (Fig. 11). In both classification tasks, the LLM identified patterns within the protein sequences, indicating that certain combinations of amino acids exhibit greater significance than others. Specifically, for classification task A, local attention was useful in predicting PPT (Figs. 11A and 11C). However, to distinguish between droplet and amyloid-forming proteins, a broader contextualized interaction was needed. This suggests that the model seeks additional information to make predictions (Figs. 11B and 11D), likely due to the complexity of this classification task.

We aimed to further understand the relationship between the propensity of proteins to form droplets and amyloid aggregates along with their gene expression. To accomplish this, we analyzed the gene expression of APP and tau (17). Our analyses revealed an overall decrease in gene expression in AD samples compared to controls (Fig. 12). Additionally, we found that their transition scores correlate with the change in gene expression levels, indicating that a larger transition score is associated with a more significant decrease in gene expression ($p<0.005$ versus $p<0.02$). This suggests the existence of a negative feedback loop between protein levels and gene expression, tightly regulating protein aggregation in AD (and other neurodegenerative diseases), in accordance with (43, 44).

Conclusions

To test the hypothesis that PPT is encoded by the amino acid sequence, we simulated the entire process from the soluble native state to the formation of droplets and amyloid aggregates. To this end, we fine-tuned a LLM and computed biophysical features, which were then used as input for classical machine learning models (e.g., random forest classifier) to predict PPT. We highlighted the significance of self-attention-based language models in the prediction of PPT. Specifically, the LLM exhibited remarkably higher performance compared to the classical model benchmark. This emphasizes the capability of contextualized language models to decipher the functionality encoded

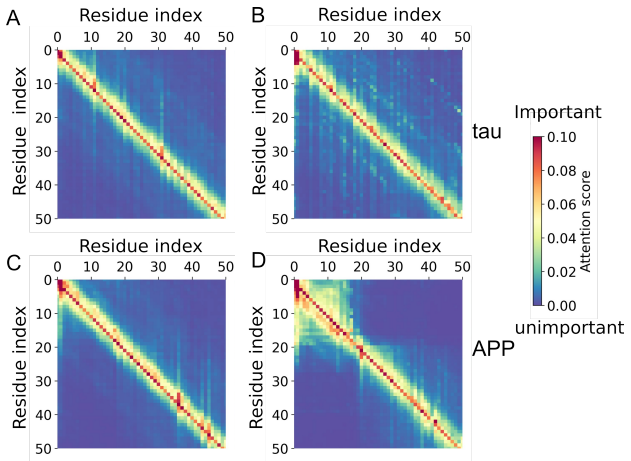

Fig. 11. Self-attention maps for the first 50 amino acids in sequence for tau and APP proteins extracted from the LLM (Left panel - classification task A. Right panel - classification task B). Attention scores referring to each amino acid (key tokens) are presented on the x-axis. The y-axis represents the target amino acids (query tokens) for which an attention score vector is computed. The attention scores were computed as the average across the 20 attention heads. These scores correspond to the level of attention that the query should have for each key. (A) and (C) exhibit strong local patterns where attention scores are close to being a band matrix with interest concentrated in small regions near the diagonal, implying very local self-attention contextualized interactions. In (B) and (D) attention scores are more sparse, implying broader contextualized interactions.

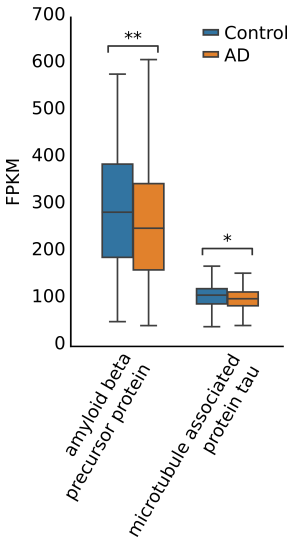

Fig. 12. Significant differences in gene expression of both APP (left) and tau (right) proteins were observed between AD samples and controls. The samples' age range spans from 67 to 90 years. ** p value < 0.005 and * p-value < 0.02. Significance was tested by using the Welch statistical t-test.

DRAFT

by the amino acid sequence. In addition, we demonstrated that PPT is dictated by biophysical properties that can be directly computed from the protein sequence. In this regard, we showed that phase-separated proteins are less stable and more disordered. Furthermore, proteins associated with reduced solubility, high compactness, and increased hydrophobicity might be more prone to form amyloid aggregates. Our ML_LM model also showed high performance, suggesting that a combination of knowledge-based and contextualized models can serve as an effective predictor for PPT. While implementing our ML_LM model on unlabeled protein sequences and calculating their transition scores, we identified proteins potentially inclined to undergo a phase transition, forming droplets in particular. We also applied our modeling approach to tau and APP proteins. Both of these proteins are thought to contribute to AD, but there are different hypotheses as to which aggregates first. Our findings indicate that in terms of protein sequence, the tau protein exhibits a higher propensity to undergo LLPS compared to APP. This suggests that tau tangles might predominantly manifest in the advanced stages of AD, with APP plaques being more prevalent in the early stages of the disease. Moreover, we demonstrated that in AD, the phase transition propensity of these two proteins can be modulated by gene expression. Taken together, our modeling approach can identify potential phase-separated proteins, including those with a high propensity to be trapped in the liquid state, which are of great interest as they can serve as early biomarkers and drug targets (45) to slow the progression of age-related diseases. In this regard, time-dependent PPT experiments are still needed to leverage the usage of LLPS as a stand-alone state and to discern its relationship with physiological and pathological conditions.

Materials and methods

Database construction of +Droplet drivers and +Amyloids.

The *+Droplet drivers* and *+Amyloids* datasets were constructed by using previously published databases with no further modifications (15). These two databases contain proteins with UniProt accession identifiers, where each protein was labeled as droplet driver or amyloid based on the evidence to form droplets or amyloid aggregates under various experimental conditions.

The droplet-forming proteins database was constructed by 404 droplet drivers proteins that undergo spontaneous LLPS (without the aid of a partner) and are not reported to form amyloid aggregates. The database was assembled from three public databases: PhaSepDB (http:/db.phasep.pro) (46), PhaSePro (https://phasepro.elte.hu) (47) and the LLPSDB database (http://bio-comp.org.cn/llpsdb) (38). The material state that was experimentally observed for such proteins was liquid or gel. These proteins tend to have a high free energy barrier to move forward to amyloid formation (12). We refer to these proteins as having a high propensity to form liquid condensates and label them as *+Droplet drivers* (dataset S1 and Fig. S1).

The amyloid-forming proteins database was constructed by 113 proteins consisting of core amyloid regions. The amyloids can be either functional or pathological (http://amypro.net) (48). This dataset is enriched with experimentally validated amyloidogenic sequence regions. For these proteins, LLPS was not observed, suggesting a low free energy barrier to rapidly form amyloid aggregates from liquid condensates, which cannot be readily observed. We refer to the proteins in this database as having a high propensity to form solid aggregates and label them as *+Amyloids* (dataset S2 and Fig. S2).

Database construction of -PT.

We randomly selected 360 sequences from the PDB* dataset constructed in (8). We intentionally did not use a larger number of sequences to keep the overall dataset relatively balanced. These sequences are folded into a well-defined structure with high stability and are not enriched in any disordered residues (as reflected by the low fraction of disordered regions and instability index in Figs. 4A, 4C, dataset S3 and Fig. S3). The sequences were extracted from the protein data bank (49).

Computing biophysical features from protein sequences.

In order to examine whether biophysical properties are encoded in the amino acid sequence and can drive PPT, we calculated the following quantities to be fed into a classical machine learning model benchmark (e.g., random forest classifier). First, we evaluated disordered regions by using the IUPred3 algorithm, which predicts the probability for each amino acid in the sequence to be disordered by using an energy estimation method (50). Next, we followed a similar process described in (8). Specifically, in each protein sequence, we identified regions with at least 20 consecutive amino acids with a probability above 0.5. These regions were considered disordered regions. Finally, we calculated the fraction of these regions. Second, in order to account for the 3D structure of the protein and its compactness in the 3D space, we calculated the gyration radius (26) as follows (19, 22):

$$R_g = \sqrt{\sum_{i=1}^{n} m^i (r^i - R_c)^2 / M}, \qquad [1]$$

where m^i is the mass of the i-th atom, which is located in the 3D space by the r^i coordinates, R_c is the coordinates of the center of mass, and M is the mass of all the atoms in the protein. This calculation was performed by using pdb files, which contain information about each of the protein atoms coordinates in the 3D space. The files were extracted from the AlphaFold database for the +Droplet drives and the +Amyloids datasets, as their UniProt accession identifier was provided. In case a pdb file was missing, it was constructed by ESMFold (v1) (19). In addition, we took advantage of the fast computing of protein structures by ESMFold and used this model to compute the pdb files of the -PT proteins, as their UniProt accession identifiers were not provided. Third, we calculated the instability index, II, a statistical measure of a given protein sequence, which is calculated by a weighted sum of dipeptides based on their differing occurrence in stable and unstable proteins as follows (23):

$$II = (10/L) \sum_{i=1}^{L-1} DIWV(x_i, y_{i+1}), \qquad [2]$$

where L is the length of the protein sequence, x_i, y_{i+1} is a dipeptide, and $DIWV$ is the dipeptide instability weight value. A II below 40 is considered a stable protein. We next estimated the probability of solubility of all the protein sequences by using the solubility weighted index, SWI, calculated by the weighted average of the normalized B-factor of amino acid residues. Using the SWI the probability of solubility is calculated as follows (24):

DRAFT

$$Probability\ of\ solubility = \frac{1}{1 + exp(-(aSWI + b))}, \quad [3]$$

where a and b are fitting parameters. Finally, we calculated the mean hydrophobicity of each protein sequence using the Kyte and Doolittle hydropathy scale, where each amino acid is represented by a hydrophobicity score between -4.6 and 4.6. A score of 4.6 is the most hydrophobic while a score of -4.6 is the most hydrophilic (25).

ESM-2 - Transformer protein language model.

We used the transfer learning approach by fine-tuning the attention-based ESM-2 model transformer with a linear layer on top for sequence classification task (19). The model is composed of 6 layers and 8M parameters. We found this approach useful while dealing with relatively small datasets as we have here. Specifically, we trained the model on our datasets (by using a ratio of 80:20 train-test sets). The model was trained with a batch size of 8 and 2 epochs. In this LLM, each amino acid is represented by a 320-dimensional embedding vector. We repeated this process separately for each of the two classification tasks described in Fig. 1.

Random forest - Machine learning classifier.

First we normalized our features by performing min-max normalization. Next, we trained a random forest classifier using the Python scikit-learn package (51) with class weight "balanced" and set the max depth of the trees to 5 to avoid overfitting. The model followed the same workflow as the LLM as described in Fig. 1.

Combined model: Random forest and LLM.

For each classification task, the combined probability of being in a certain class was calculated by averaging the probabilities of the RF classifier and the LLM. To label the sequences, we set a minimum threshold of 0.6 to be classified as -PT and +Droplet drivers for classification tasks A and B, correspondingly.

Transition score calculation.

$$Transition\ score = \log \frac{p}{(1-p)}, \quad [4]$$

where p is the average probability to undergo PPT, which is calculated by using the random forest and LLM predictions.

Data availability. The data used or generated in this work is provided by the following files.

dataset S1

dataset S2

dataset S3

dataset S4

dataset S5

dataset S6

ACKNOWLEDGMENTS. We would like to thank Prof. Andrew Miranker from the Department of Molecular Biophysics and Biochemistry, Yale University, for fruitful discussions in relation to protein phase transitions from an experimental viewpoint. His expertise was helpful in shaping the direction of this work.

1. BR Sabari, et al., Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).
2. S Chong, et al., Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, eaar2555 (2018).
3. BA Gibson, et al., Organization of chromatin by intrinsic and regulated phase separation. *Cell* **179**, 470–484 (2019).
4. RM Nisbet, JC Polanco, LM Ittner, J Götz, Tau aggregation and its interplay with amyloid-β. *Acta neuropathologica* **129**, 207–220 (2015).
5. S Ambadipudi, J Biernat, D Riedel, E Mandelkow, M Zweckstetter, Liquid–liquid phase separation of the microtubule-binding repeats of the alzheimer-related protein tau. *Nat. communications* **8**, 275 (2017).
6. A Zbinden, M Pérez-Berlanga, P De Rossi, M Polymenidou, Phase separation and neurodegenerative diseases: a disturbance in the force. *Dev. cell* **55**, 45–68 (2020).
7. TJ Nott, et al., Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. cell* **57**, 936–947 (2015).
8. KL Saar, et al., Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci.* **118**, e2019053118 (2021).
9. G van Mierlo, et al., Predicting protein condensate formation using machine learning. *Cell reports* **34**, 108705 (2021).
10. A Wei, L Wang, Deep sequence representation learning for predicting human proteins with liquid-liquid phase separation propensity and synaptic functions in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pp. 1–8 (2022).
11. X Chu, et al., Prediction of liquid–liquid phase separating proteins using machine learning. *BMC bioinformatics* **23**, 1–13 (2022).
12. M Hardenberg, A Horvath, V Ambrus, M Fuxreiter, M Vendruscolo, Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci.* **117**, 33254–33262 (2020).
13. G Vecchi, et al., Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc. Natl. Acad. Sci.* **117**, 1015–1020 (2020).
14. GL Dignon, RB Best, J Mittal, Biomolecular phase separation: from molecular driving forces to macroscopic properties. *Annu. review physical chemistry* **71**, 53–75 (2020).
15. M Vendruscolo, M Fuxreiter, Sequence determinants of the aggregation of proteins within condensates generated by liquid-liquid phase separation. *J. Mol. Biol.* **434**, 167201 (2022).
16. Z Lin, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
17. AK Greenwood, et al., The ad knowledge portal: A repository for multi-omic data on alzheimer's disease and aging. *Curr. protocols human genetics* **108**, e105 (2020).
18. N Brandes, D Ofer, Y Peleg, N Rappoport, M Linial, Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
19. Z Lin, et al., Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* (2022).
20. J Jumper, et al., Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
21. G Erdős, M Pajkos, Z Dosztányi, Iupred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic acids research* **49**, W297–W303 (2021).
22. MY Lobanov, N Bogatyreva, O Galzitskaya, Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **42**, 623–628 (2008).
23. K Guruprasad, BB Reddy, MW Pandit, Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng. Des. Sel.* **4**, 155–161 (1990).
24. BK Bhandari, PP Gardner, CS Lim, Solubility-weighted index: fast and accurate prediction of protein solubility. *Bioinformatics* **36**, 4691–4698 (2020).
25. J Kyte, RF Doolittle, A simple method for displaying the hydropathic character of a protein. *J. molecular biology* **157**, 105–132 (1982).
26. M Varadi, et al., Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439–D444 (2022).
27. R Pancsa, W Vranken, B Mészáros, Computational resources for identifying and describing proteins driving liquid–liquid phase separation. *Briefings Bioinforma.* **22**, bbaa408 (2021).
28. C Soto, L Estrada, J Castilla, Amyloids, prions and the inherent infectious nature of misfolded protein aggregates. *Trends biochemical sciences* **31**, 150–155 (2006).

DRAFT

29. VN Uversky, Paradoxes and wonders of intrinsic disorder: Stability of instability. *Intrinsically disordered proteins* **5**, e1327757 (2017).
30. S Brocca, R Grandori, S Longhi, V Uversky, Liquid–liquid phase separation by intrinsically disordered protein regions of viruses: Roles in viral life cycle and control of virus–host interactions. *Int. J. Mol. Sci.* **21**, 9045 (2020).
31. BH Menze, et al., A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics* **10**, 1–16 (2009).
32. D Raimondi, et al., In silico prediction of in vitro protein liquid–liquid phase separation experiments outcomes with multi-head neural attention. *Bioinformatics* **37**, 3473–3479 (2021).
33. A Boija, et al., Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855 (2018).
34. H Lu, et al., Phase-separation mechanism for c-terminal hyperphosphorylation of rna polymerase ii. *Nature* **558**, 318–323 (2018).
35. SJ Nair, et al., Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat. structural & molecular biology* **26**, 193–203 (2019).
36. N Andreasen, et al., Cerebrospinal fluid β-amyloid (1-42) in alzheimer disease: differences between early-and late-onset alzheimer disease and stability during the course of disease. *Arch. neurology* **56**, 673–680 (1999).
37. V Rudajev, J Novotny, Cholesterol as a key player in amyloid β-mediated toxicity in alzheimer's disease. *Insights Brain Dis. Mech. 2021* p. 232 (2022).
38. Q Li, et al., Llpsdb: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic acids research* **48**, D320–D327 (2020).
39. S Wegmann, et al., Tau protein liquid–liquid phase separation can initiate tau aggregation. *The EMBO journal* **37**, e98049 (2018).
40. AF Arnsten, D Datta, K Del Tredici, H Braak, Hypothesis: Tau pathology is an initiating factor in sporadic alzheimer's disease. *Alzheimer's & Dementia* **17**, 115–124 (2021).
41. M Islam, et al., Tau liquid–liquid phase separation: at the crossroads of tau physiology and tauopathy. *J. Cell. Physiol.* (2022).
42. S Makin, The amyloid hypothesis on trial. *Nature* **559**, S4–S4 (2018).
43. YM Ayala, et al., Tdp-43 regulates its mrna levels through a negative feedback loop. *The EMBO journal* **30**, 277–288 (2011).
44. E Duranti, C Villa, Molecular investigations of protein aggregation in the pathogenesis of amyotrophic lateral sclerosis. *Int. J. Mol. Sci.* **24**, 704 (2022).
45. K Kamagata, et al., Characterization of design grammar of peptides for regulating liquid droplets and aggregates of fus. *Sci. reports* **11**, 1–13 (2021).
46. K You, et al., Phasepdb: a database of liquid–liquid phase separation related proteins. *Nucleic acids research* **48**, D354–D359 (2020).
47. B Mészáros, et al., Phasepro: the database of proteins driving liquid–liquid phase separation. *Nucleic acids research* **48**, D360–D367 (2020).
48. M Varadi, G De Baets, WF Vranken, P Tompa, R Pancsa, Amypro: a database of proteins with validated amyloidogenic regions. *Nucleic acids research* **46**, D387–D392 (2018).
49. HM Berman, et al., The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
50. Z Dosztanyi, V Csizmok, P Tompa, I Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. molecular biology* **347**, 827–839 (2005).
51. F Pedregosa, et al., Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

DRAFT