A machine learning framework for predicting and modulating condition-dependent protein phase separation

Jangwon Bae[1†], Minjun Kang[2†], Donghyuk Lee[1], Kuk-Jin Yoon[3*], and Yongwon Jung[1*]

1. Department of Chemistry, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea

2. The Robotics Program, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea

3. Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea

[†]These authors contributed equally to this work.

[*]Correspondence:

Yongwon Jung, E-mail: ywjung@kaist.ac.kr

Kuk-Jin Yoon, E-mail: kjyoon@kaist.ac.kr

Abstract

Protein phase separation is a fundamental process in organizing membraneless organelles and is implicated in a wide range of pathological conditions. Crucially, rather than being an intrinsic feature of specific proteins, phase separation is a condition-dependent phenomenon governed by environmental parameters, including protein concentration, temperature, and solvent composition. However, most existing machine-learning models infer phase separation propensity from amino-acid sequences alone, failing to capture these context-dependent behaviors. Here, we present LLPSense, a machine-learning framework that integrates pre-trained protein language model embeddings with environmental parameters for substantially more accurate, condition-aware prediction of protein phase separation. We demonstrate LLPSense's predictive power and utility through three key experimental experimental demonstrations. First, the model revealed that SGTA, previously unrecognized as a phase-separating protein, exhibits complex, temperature-dependent reentrant phase behavior. Second, LLPSense accurately predicted mutations in α-synuclein that either enhance or suppress phase separation, enabling systematic mapping of residues potentially relevant to Parkinson's disease. Third, using model-guided mutagenesis, we inverted the phase behavior of UBQLN4 from high-temperature to low-temperature separation. Collectively, LLPSense provides a robust computational tool for interrogating the condition-dependent landscape of protein phase separation, enabling mechanistic studies of disease-associated phase separation as well as the rational design of programmable condensates.

Introduction

Biomolecular phase separation of proteins and nucleic acids drives the formation of diverse membraneless organelles and contributes to the spatial organization of biochemical processes within cells[1]. Dysregulation of protein phase behavior has been increasingly implicated in the pathogenesis of various diseases, including neurodegenerative disorders[2-4]. Many proteins have been shown to undergo phase separation, thereby forming membraneless reaction hubs that regulate diverse cellular processes[5]. Despite substantial progress, experimentally identifying phase-separating proteins and characterizing their phase behavior remain labor-intensive and limited in throughput. To overcome these limitations, numerous machine-learning (ML) approaches have been introduced. In particular, several ML predictors[6-9] have been trained to estimate phase separation propensity directly from protein sequences, achieving superior performance compared with earlier *in silico* screening methods[10]. Among them, the recently developed PSPire model[11] exhibits remarkable predictive accuracy, even for proteins lacking intrinsically disordered regions (IDRs). Another ML model, PSPHunter[12], identified key residues that drive phase separation and demonstrated that truncating only six residues in GATA3 was sufficient to inhibit its phase separation behavior.

Previous sequence-based ML prediction models have relied on static binary classification (i.e., whether a protein is capable of phase separation or not). However, natural liquid-liquid phase separation (LLPS) of

proteins is a dynamic, stimulus-responsive phenomenon governed by complex physicochemical conditions (e.g., pH, temperature)[13]. Because conventional methods rely exclusively on sequence feature similarity, they inherently treat phase separation as a fixed property, failing to account for the fact that the same protein's behavior varies with conditions. This reliance restricts their predictive resolution and limits the precise estimation of LLPS propensity. In line with this static perspective, the validation of existing tools like PSPHunter focused primarily on variants that suppress LLPS, whereas natural mutants exhibit both suppressing and enhancing phase separation. Therefore, capturing the full spectrum of such modulation necessitates bidirectional scales that enable the precise comparison of phase separation probabilities, supported by a comprehensive dynamic range covering the global condition space. Droppler[14] represents a pioneering effort to address this limitation by employing a neural network to estimate LLPS probability under various conditions. Despite this advancement, its predictive performance remains limited, thereby restricting broader applicability. The authors of Droppler acknowledged that modeling condition-dependent LLPS is substantially more challenging than sequence-only classification. Consequently, developing a robust predictor that integrates both intrinsic protein sequences and extrinsic conditional variables is essential for the scalable computational analysis of LLPS behavior across large protein datasets. Such a condition-aware framework would provide a powerful platform to investigate LLPS biology and enable the *in silico* design of programmable biomolecular condensates.

In this study, we present LLPSense, a condition-aware ML framework designed to predict protein phase behavior across diverse physicochemical environments. LLPSense accurately estimates the phase separation probability of a given protein sequence under a wide range of parameters, including protein concentration, temperature, pH, crowding agents, salt concentration, and glycerol content. We thoroughly validated the predictive power and experimental utility of LLPSense through multiple analyses. Initially, we challenged the model to screen for overlooked phase-separating proteins within datasets previously classified as negative. LLPSense flagged 16.4% of these sequences as potential phase-separating candidates, and this predicted subset included proteins already reported in the literature to undergo phase separation. Notably, this screen identified SGTA as a previously unrecognized LLPS protein, whose complex reentrant, temperature-dependent phase behavior was accurately predicted and subsequently confirmed experimentally. LLPSense also pinpointed specific residues in Parkinson's disease-associated α-synuclein, predicting substitutions that either enhance or suppress phase separation. Comparative analysis with a sequence-only predictor confirmed that the condition-aware framework facilitates mutational analysis at single-residue resolution and bidirectional prediction. We additionally demonstrated that the conditional dependence of LLPS can be modulated by altering intrinsic sequence properties. Specifically, mutational engineering of the lower critical solution temperature (LCST)-type protein UBQLN4 shifted its phase behavior toward an upper critical solution temperature (UCST)-like regime, highlighting that LLPSense can guide the rational design of sequence modifications for tuning condition-specific phase behavior.

Results

Protein language model embeddings facilitate accurate prediction of phase separation

Motivated by the recent success of applying language-based encoding methods, such as Word2Vec, to LLPS prediction[7,12], we hypothesized that leveraging advanced pretrained protein language models (pLMs) could provide a more accurate and efficient framework. Among available pLMs, we selected ProtT5[15] to extract high-quality sequence embeddings for both training and inference, as prior work (Seq2Phase)[16] has demonstrated that pLM-derived embeddings enable accurate prediction of proteins recruited to biomolecular condensates. Before integrating environmental variables, we validate this pLM-based strategy using the conventional sequence-based, binary LLPS prediction task. We selected the previously developed PSPire as a comparative reference because, in addition to its high predictive performance, it demonstrated improved identification of phase-separating proteins with low intrinsically disordered region (IDR) content. This distinction is crucial for analyzing whether pLMs can provide generalized protein representations, capturing phase separation properties across the full structural spectrum. Accordingly, we trained and evaluated models using the dataset used in PSPire, which classifies human proteins into three categories: intrinsically disordered phase-separating proteins (ID-PSPs), structured phase-separating proteins (noID-PSPs), and non-phase-separating proteins (non-PSPs). ID-PSPs and noID-PSPs were treated as positive samples, and non-PSPs as negatives.

We developed a gradient-boosted decision tree classifier based on XGBoost[17]. A previous study demonstrated that tree-based classifiers, such as XGBoost and Random Forest, outperform neural-network architectures and other ML approaches for LLPS prediction[12]. We first obtained pLM-based embeddings using ProtT5, which encodes an amino acid sequence of length N into an N × 1,024 embedding matrix. Mean pooling was then applied across the sequence dimension to obtain a single 1,024-dimensional vector representation. These averaged embeddings were subsequently used as input features for model training and evaluation, resulting in a model we termed LLPSeq (Fig. 1a). The resulting model demonstrated remarkable predictive performance. In both the ID and noID test sets, the area under the receiver operating characteristic curve (AUROC) was comparable to values reported for PSPire, while the area under the precision-recall curve (AUPRC) surpassed it (Fig. 1b–e). To benchmark protein language embeddings against conventional feature-based models with a similar architecture, we constructed a baseline model trained on 55 knowledge-based engineered features (e.g., molecular weight, net charge, hydrophilicity, and low-complexity region content) widely used in previous studies[6,7]. The overall performance of this baseline model was inferior to that of PSPire and LLPSeq, especially in the IDR-lacking (noID) test set (Fig. 1d, e). These findings indicate that, even without handcrafted feature engineering, embeddings derived directly from pretrained protein language models provide a more effective and generalizable representation for LLPS prediction.

119 **LLPSense accurately predicts condition-dependent protein phase behavior**

120 Building on the promising performance of the pLM-based LLPSeq model, we extended this framework to
121 develop a comprehensive predictor that integrates both intrinsic sequence information and extrinsic
122 environmental conditions. A significant bottleneck in developing condition-aware models has been the limited
123 availability of large, standardized datasets suitable for machine learning. Although LLPSDB v2[18] provides a
124 broad collection of *in vitro* LLPS observations, substantial curation is required to render it amenable to model
125 training. The previously developed Droppler model[14] was trained on a filtered subset of LLPSDB v1[19],
126 comprising 896 entries across across 366 sequences. By contrast, we constructed a rigorously curated and
127 filtered dataset derived from LLPSDB v2, comprising 3,081 entries across 922 unique sequences, with a specific
128 focus on single-protein entries.

129 For each entry, we extracted and standardized a comprehensive set of 13 condition variables, including
130 protein concentration, temperature, pH, six representative categories of crowding agents, concentrations of three
131 representative salts (NaCl, KCl, and $MgCl_2$), and glycerol percentage. These 13 variables were sufficient to
132 capture the majority of experimentally annotated entries. In contrast, complex co-solutes such as heparin or
133 RNA were excluded due to data sparsity and mechanistic complexity. Collectively, these selected parameters
134 capture key physicochemical modulators of LLPS, encompassing solvent polarity, ionic strength, and
135 macromolecular crowding effects. Furthermore, to accommodate the continuous nature of experimental phase
136 diagrams, we implemented a systematic interval-partitioning strategy. This method subdivides reported ranges
137 into uniformly spaced samples, thereby augmenting the effective number of training instances while preserving
138 experimental validity (see Methods for details). All curated datasets and preprocessing scripts used in this study
139 are publicly available to facilitate future machine-learning research on condition-dependent protein phase
140 separation (see Data and Code Availability).

141 Next, sequence embeddings derived from ProtT5 were concatenated with the normalized condition variables
142 to construct an integrated feature matrix. This combined representation served as the input for an XGBoost-
143 based architecture, and we designated this predictive framework as LLPSense (Fig. 1a). This multimodal
144 integration enables the model to simultaneously capture intrinsic sequence determinants of LLPS and extrinsic
145 physicochemical factors that modulate condensate formation. We employed a 5-fold cluster-based cross-
146 validation (CV) strategy to assess the generalizability of LLPSense and identify optimal hyperparameters. To
147 prevent overestimation of performance due to sequence homology between training and test sets, proteins were
148 grouped into 234 distinct clusters using MMseqs2[20], ensuring that sequences in different clusters shared <20%
149 identity at 90% coverage. All proteins within the same cluster were assigned to the same CV fold, thereby
150 maintaining strict separation between training and validation sets. Under this cluster-based 5-fold CV, LLPSense
151 demonstrated robust predictive performance, achieving an average AUROC of 0.77 and an AUPRC of 0.74 (Fig.
152 1f–h). This performance significantly surpasses the AUROC value reported for Droppler (0.64). Furthermore,
153 even when Droppler was retrained on our expanded dataset, LLPSense continued to exhibit superior predictive

accuracy (Fig. 1h, Supplementary Fig. 1). Altogether, these results demonstrate that our integrated modeling strategy provides an effective and reliable framework for predicting condition-dependent phase separation. Given that modeling conditional dependencies is significantly more challenging than sequence-only prediction, the achieved performance marks a pivotal milestone in the field. We next investigated whether this *in silico* predictive accuracy translates into practical utility in real experimental settings.

Identification of potential phase-separating proteins in the previously classified negative dataset

We screened the human proteome dataset previously classified as non-PSP (used in the PSPire study) using LLPSense. Remarkably, a substantial fraction of proteins (16.4%, 1,682/10,252) were predicted to undergo phase separation within a temperature range of 10–50 °C under physiological conditions (100 μM protein concentration, pH 7.3, and 160 mM NaCl; Fig. 2a). Among these proteins, SS18[21], FOXP2[22], and Atrophin-1[23] (Fig. 2b) have been reported to exhibit phase-separating behavior, despite their inclusion in the non-PSP set. This discrepancy underscores a critical limitation in the existing negative datasets. Such issues arise primarily because a definitive list of proteins that are strictly incapable of LLPS has not been established. As a result, previous studies have resorted to heuristic negative sets derived from indirect criteria, such as subtracting known LLPS-positive proteins from the proteome. However, the boundary between "LLPS-competent" and "LLPS-incompetent" proteins is inherently fluid, varying across environmental conditions and experimental methods. This ambiguity complicates model training and further emphasizes the necessity for condition-aware strategies.

Among the candidates newly identified by LLPSense, SGTA stood out due to its predicted temperature-dependent reentrant phase behavior, which displays characteristics of both upper critical solution temperature (UCST) and lower critical solution temperature (LCST) regimes (Fig. 2c). To experimentally validate this prediction, SGTA was recombinantly prepared, and its phase behavior was assessed using solution turbidity measurements, a commonly used quantitative indicator of phase separation propensity. We observed a PEG concentration-dependent increase in SGTA turbidity (A_{350}), which confirms its intrinsic ability to phase separate (Fig. 2d). Consistent with the predicted reentrant profile, turbidity increased modestly at 30–35 °C but showed a much greater increase at both low (4 °C) and elevated (≥ 40 °C) temperatures (Fig. 2e). Microscopic images of SGTA condensates corroborated these findings, revealing that the total volume of condensates was larger at elevated temperatures (Fig. 2f). Collectively, these results demonstrate that LLPSense not only identifies previously unrecognized phase-separating candidates but also successfully captures their intricate, condition-dependent phase separation behavior.

To further assess the predictive applicability of LLPSense, we randomly selected three proteins (ASCL1, FBLL1, SEMG2) from candidates predicted to exhibit a phase separation threshold within the 10–50 °C range (Supplementary Fig. 2) and evaluated their behavior *in vitro*. ASCL1 and FBLL1 aggregated during dialysis against PBS (pH 7.3, 160 mM NaCl), resulting in final concentrations below 10 μM. FBLL1 is known to localize within the nucleolus[24], suggesting that it may participate in the phase behavior of this membraneless organelle.

189 SEMG2 is a semenogelin protein responsible for the gelation of seminal fluid[25]. Upon addition of PEG8000,
190 SEMG2 displayed marked turbidity (Supplementary Fig. 3), consistent with macromolecular condensation.

191 The aggregation of ASCL1 and FBLL1 may reflect their high intrinsic LLPS propensity, as phase-separating
192 proteins frequently exhibit marginal solubility and an increased tendency toward aggregation[26,27]. Furthermore,
193 we observed a concentration-dependent transition in material state: SGTA formed large liquid-like droplets at
194 160 μM (with 10% PEG8000), whereas small gel-like condensates were formed at a lower concentration (Fig.
195 2f). SEMG2 also formed interconnected gel-like condensates rather than canonical liquid droplets (Fig. 2g).
196 Despite extensive research, experimental criteria that clearly distinguish liquid-like condensates from percolated
197 gels and irreversible aggregates remain poorly defined[28]. Accordingly, while LLPSense effectively predicts
198 condensation propensity, it does not yet discriminate among distinct material states. With the expanding
199 understanding of biomolecular phase behavior, LLPS, gelation, aggregation, and fibrillization are viewed not
200 as discrete end states but as interconnected regimes within a shared energy landscape[29-31]. Our preliminary tests
201 with ASCL1, FBLL1, SEMG2, and SGTA illustrate this continuum, highlighting both the broad predictive reach
202 of LLPSense and the necessity for future data refinements that explicitly account for the full liquid–gel–solid
203 spectrum.

204

205 ***In silico* mapping of the α-synuclein mutational landscape at single-residue resolution**

206 Many diseases are associated with mutations that induce aberrant enhancement[32-35] or suppression[36,37] of phase
207 separation. However, the previously reported mutation-prediction framework[12] was primarily optimized to
208 identify substitutions or truncations that suppress LLPS. This restriction arises because models relying solely
209 on sequence information are prone to overfitting existing positive datasets. In contrast, the LLPSense framework
210 estimates LLPS propensity with higher resolution across diverse physicochemical conditions, enabling a
211 nuanced, bidirectional assessment of both enhancing and inhibitory mutations. To illustrate this capability, we
212 examined α-synuclein, whose LLPS behavior is increasingly recognized as an intermediate state preceding
213 pathological fibrillization[38]. By systematically substituting each of the 140 residues of α-synuclein and
214 computing the resulting change in LLPS probability, we aimed to predict specific mutations that modulate phase
215 separation.

216 LLPSense enabled rapid probability screening across the entire mutational landscape, comprising all possible
217 single-point substitutions (140 residues × 19 variants; Fig. 3a). To experimentally validate the predicted effects,
218 we selected two mutations projected to exert the strongest influence on phase separation (Fig. 3b). Specifically,
219 the substitution of glutamate with methionine at position 28 (E28M) within the amphipathic N-terminal domain
220 yielded the highest predicted increase in LLPS probability; we termed this the 'E1 variant'. Conversely, the
221 substitution of valine with aspartate at position 77 (V77D) within the hydrophobic non-amyloid component
222 (NAC) domain caused the most significant reduction, designated as the 'S1 variant'. In contrast, mutations within
223 the acidic C-terminal tail elicited relatively minor fluctuations in predicted phase behavior.

Using the E1 and S1 variants as scaffolds, we performed iterative rounds of prediction to identify additional mutations that confer additive effects while retaining previously mutated positions. Through this recursive procedure, we constructed a series of variants with multiple mutations, designating the phase separation enhanced set as E2, E4, and E8, and the suppressed set as S2 (Supplementary Table 1). For these variants, the numerical suffix indicates the cumulative number of incorporated mutations. In addition, to verify that LLPSense accurately predicted E1 based on the specific position (E28) and residue (E-to-M), two control mutants (E28P and E130M) were selected for their lowest predicted LLPS probabilities among possible combinations (Fig. 3c). E28P represents a different substitution at the same position, while E130M represents the identical E-to-M substitution at a different position. Both control mutants exhibit predicted probabilities significantly lower than the E1 variant and slightly below the WT baseline (Fig. 3d).

We expressed and purified all α-synuclein variants under identical conditions and assessed their phase separation behavior by measuring turbidity across nine combinations of PEG8000 and protein concentrations. Heatmap visualization of ΔTurbidity, calculated by subtracting the WT turbidity from that of each variant, revealed a stark contrast between the enhanced and suppressed variant sets (Fig. 3e, Supplementary Fig. 4). Summing the ΔTurbidity values across all nine conditions yielded the following order: E2 > E1 > WT (baseline) > E28P > E130M > S1 > S2 (Fig. 3f). This ranking was in excellent agreement with LLPSense predictions (Fig. 3d), with the only minor deviation occurring between E130M and E28P, where predicted differences were marginal. During PBS dialysis, the hyper-enhanced variants E4 and E8 exhibited pronounced aggregation and structural instability (Supplementary Fig. 5a, b). This observation corroborates our earlier discussion that condensation-prone sequences often exhibit a concomitant increase in aggregation propensity. Nonetheless, both variants still showed turbidity levels higher than WT (Supplementary Fig. 5c). Taken together, these results demonstrate that LLPSense accurately identifies mutations that modulate phase separation and can reliably pinpoint sequence alterations that shift α-synuclein toward LLPS-prone states.

To examine the advantages of the condition-aware framework, we compared the predictive landscape of LLPSense with that of LLPSeq, a sequence-based predictor. Despite the exclusion of α-synuclein from its training dataset, LLPSeq predicted a high probability of 0.78 for the WT, a result that likely stems from overfitting to the homologous β-synuclein sequence. Consequently, LLPSeq classified most single-point mutations as suppressive (90.2%; 2,039/2,260). In contrast, even with the target protein included in its training, LLPSense identified a substantial fraction of mutations as enhancing (24.0%; 543/2,260) (Supplementary Fig. 6a, b), demonstrating a broad bidirectional prediction range. In addition, LLPSense can circumvent this saturation issue through its environmental sensitivity. By modulating parameters such as protein concentration or crowding agent levels, the model can shift the baseline into a neutral regime, near a predicted probability of 0.5, even if a sequence exhibits high intrinsic probability. This capability ensures a sufficient dynamic range to resolve subtle mutational effects that are otherwise masked in static, sequence-only models. LLPSeq also predicted a monotonic decrease in phase separation potential for enhanced mutants (E1–E8), showing an increasingly pronounced discrepancy from the intended cumulative modulation (Supplementary Fig. 6c).

260 We also benchmarked LLPSense on α-synuclein variants previously reported as pathogenic and characterized
261 for LLPS propensity using turbidity and droplet-size measurements[38-41] (Supplementary Fig. 7). Based on
262 experimental observations from prior literature, mutations were categorized into four distinct groups consisting
263 of significantly enhanced, mildly enhanced, comparable to WT, and suppressed variants. LLPSense
264 demonstrated high qualitative consistency with these experimental groupings by accurately reflecting the
265 relative rankings and propensity shifts across all tested substitutions. The absolute probability for the WT
266 sequence may be subject to moderate overestimation due to its inclusion in the training dataset. However, the
267 relative comparisons and predicted trends across the various mutants remain highly valid, as the model
268 accurately reflects the impact of each substitution at single-residue resolution.

269

270 **LLPSense-guided mutations reprogram UBQLN4 from LCST to UCST phase behavior**

271 Most phase-separating proteins exhibit upper critical solution temperature (UCST) behavior, where phase
272 separation is favored at lower temperatures due to reduced entropic costs. In contrast, certain families, such as
273 elastin-like polypeptides (ELPs)[42] and the UBQLN family[43], display LCST properties, undergoing phase
274 separation as the temperature increases. Our curated dataset included ELPs and UBQLN2, but excluded
275 UBQLN1 and UBQLN4. Notably, while UBQLN1 shares high sequence homology with UBQLN2 (76.21%
276 identity, 82.79% similarity), UBQLN4 is more distantly related (58.72% identity, 70.19% similarity). When the
277 Droppler model was retrained on our dataset, it correctly classified the highly homologous UBQLN1 as an
278 LCST-type protein but failed to identify the more divergent UBQLN4 (Supplementary Fig. 8a). By contrast,
279 LLPSense accurately predicted the LCST behavior of both UBQLN1 and UBQLN4 (Fig. 5a), demonstrating
280 improved generalizability and sensitivity to the sequence determinants underlying temperature-dependent phase
281 behavior.

282 To experimentally validate these predictions, we examined UBQLN4 *in vitro* and observed a clear increase
283 in turbidity and condensate formation with rising temperature, which confirms its LCST nature (Fig. 5b, c). We
284 next investigated whether LLPSense could guide the rational reprogramming of this LCST profile toward a
285 UCST-like regime. Adapting a strategy analogous to the α-synuclein mutational screen, we prioritized mutations
286 based on the slope of the difference in predicted LLPS probability between the variant and wild-type (ΔProb =
287 $Prob_{var} - Prob_{WT}$) across the temperature gradient. We selected the mutations exhibiting the most negative ΔProb
288 slopes, which indicated a shift toward UCST behavior, and incorporated them into the sequence (Fig. 5d). To
289 ensure a sufficient cumulative impact to invert the temperature dependence while safeguarding against the loss
290 of phase-separation propensity, mutations were introduced across three iterative rounds with ten substitutions
291 accumulated per round (Supplementary Table 2). This stepwise engineering yielded a UBQLN4-UCST variant
292 that formed condensates resembling percolated clusters (Supplementary Fig. 8b). The variant successfully
293 displayed the targeted inversion in temperature dependence, exhibiting markedly higher turbidity at lower
294 temperatures compared to elevated temperatures (Fig. 5c). These results demonstrate that LLPSense serves not

only as a predictor of condition-dependent phase behavior but also as a tool for rationally tuning it through targeted sequence modifications.

Discussion

In this study, we presented LLPSense, a unified machine-learning framework that bridges the gap between intrinsic sequence determinants and extrinsic environmental factors for accurate prediction of protein phase separation. By harnessing high-dimensional embeddings from pre-trained protein language models, LLPSense overcomes the limitations of previous sequence-only predictors, thereby enabling high-resolution characterization of phase behavior across diverse physicochemical landscapes.

Our comprehensive validation highlights the transformative potential of this condition-aware approach. First, LLPSense identified proteins exhibiting atypical phase separation behaviors, such as SGTA, within datasets previously classified as non-phase-separating. This finding reveals latent inaccuracies in current negative datasets and demonstrates the model's ability to capture complex, reentrant phase behaviors. Second, by mapping the mutational landscape of α-synuclein, we demonstrated that the model can bidirectionally predict mutations that modulate phase separation propensity. Importantly, this capability facilitates the rapid screening of residues harboring pathogenic potential, streamlining the identification and prioritization of high-risk variants implicated in LLPS-associated pathologies. Lastly, we successfully reprogrammed the thermodynamic profile of UBQLN4 from an LCST to a UCST regime through targeted sequence modification. This achievement serves as a proof-of-concept that LLPSense can guide the rational design of proteins with tunable, environment-responsive phase properties.

Furthermore, the modular architecture of LLPSense offers significant potential for future scalability. By leveraging embeddings as the core feature set, our framework is designed to be adaptable to the rapid advancements in deep learning. As more sophisticated protein language models (pLMs) emerge, LLPSense allows for the seamless integration of these next-generation representations. Concurrently, the predictive power of LLPSense is expected to grow with the continued accumulation of high-quality experimental data. This data-driven evolution will extend the framework's applicability beyond binary classification, paving the way for more granular characterizations, such as distinguishing among diverse condensate material states.

Ultimately, LLPSense establishes a foundation for programmable biomolecular condensation. By providing a reliable platform for predicting and manipulating condition-dependent phase transitions, this work paves the way for novel applications in synthetic biology, the development of condensate-modulating therapeutics, and a deeper physical understanding of cellular organization.

Methods

Sampling of experimental condition data

Entries from LLPSDB v2 were processed into protein names, amino acid sequences, and condition-related variables, including protein concentration, temperature, pH, crowding agents, three types of salts (NaCl, KCl, and $MgCl_2$), and glycerol percentage. Crowding agents were classified into six representative categories: PEG 300–1k, PEG 3k–6k, PEG 8k–20k, ficoll, dextran MW ≤ 40 kDa, and dextran MW ≥ 70 kDa. PEG and dextran with unspecified molecular weights were assigned to the PEG 3k–6k and Dextran MW ≤ 40 kDa categories, respectively. Experimental conditions were parsed from a structured Excel dataset into numerical feature vectors. Fields specified as fixed values or ranges ('a-b', 'a-', and '-b') were interpreted according to predefined feature-specific limits, sampling densities, and sampling intervals (Supplementary Table 3). Closed ranges were uniformly sampled either as continuous intervals or as discrete unit-based grids, depending on the feature type. Open-ended ranges were expanded 10-fold above or 0.1-fold below the specified boundary, and subsequently clipped to the global minimum or maximum allowed for each feature. A Cartesian product of all condition values was generated to enumerate every valid experimental configuration. If the resulting combinatorial expansion exceeded a predefined limit ($n > 100$), a random subset of configurations was sampled. Finally, all features were normalized to the range [0, 1] using feature-specific maximum values.

Dataset preparation for LLPSeq training and evaluation

We trained and evaluated LLPSeq using the dataset and strategy employed in PSPire. The positive protein lists (ID-PSPs and noID-PSPs) were identical. However, a minor fraction of the non-PSPs negative set was excluded because sequence updates in the UniProt database altered their lengths, causing them to fall outside the original PSPire length filtering criteria (30/8,323 in the training set and 2/1,961 in the testing set). A total of 55 engineered features were extracted in accordance with the methodology described in our previous study[7] to train a baseline model for comparison against the pLM-based LLPSeq.

Model training and hyperparameter optimization

The LLPSense model was constructed using the eXtreme Gradient Boosting (XGBoost) framework, a scalable implementation of gradient-boosted decision trees. The model was trained with the binary:logistic objective function to predict the probability of phase separation, minimizing the logarithmic loss (logloss) metric. To ensure robust generalization, we optimized key hyperparameters governing the tree architecture and learning process—specifically n_estimators, max_depth, learning_rate, min_child_weight, subsample, colsample_bytree, gamma, and regularization terms (reg_alpha, reg_lambda). Hyperparameters for both LLPSeq and LLPSense were optimized using the Optuna framework. For LLPSense, model selection was conducted using a 5-fold cluster-based cross-validation scheme. The final optimized hyperparameter configurations for each model are summarized in Supplementary Table 4.

364 **Standard prediction condition and candidate screening**

365 Unless otherwise specified, phase-separation probabilities were computed under standard conditions: 100 μM
366 protein concentration, pH 7.3, 160 mM NaCl, and 25 ºC. Specifically for candidate screening, we employed a
367 temperature-based filtering approach where predicted probabilities were smoothed using a centered moving
368 average (CMA) with a window size of 9. Candidates were identified as proteins exhibiting a probability
369 exceeding 0.5 within the 10–50 °C range. For experimental validation, we selected SGTA alongside three
370 randomly drawn proteins (ASCL1, FBLL1, and SEMG2) from a pool of sequences shorter than 1,000 amino
371 acids, specifically those exhibiting a predicted phase transition threshold within the 10–50 °C screening window.
372 The AlphaFold database structures[44] of the proteins used in the experiments are shown in Supplementary Fig.
373 9.

374

375 **Model-guided mutagenesis for design of UBQLN4-UCST**

376 We employed an iterative model-guided mutagenesis strategy under standard prediction conditions in the
377 absence of crowding agents. For each round, temperature-dependent probability curves were smoothed using a
378 CMA with a window size of 15. The temperature sensitivity of each single-point mutant was quantified by
379 calculating the slope of the difference in predicted LLPS probability between the variant and wild-type (ΔProb
380 = $Prob_{var}$–$Prob_{WT}$) across the temperature gradient via linear regression. The comprehensive list of these
381 mutations and their calculated ΔProb slopes for each round is provided in Supplementary Table 2. We selected
382 the top 10 mutations exhibiting the most negative ΔProb slopes, which indicated a shift from LCST toward
383 UCST behavior, and incorporated them into the sequence. This procedure was repeated for three consecutive
384 rounds, resulting in a final UBQLN4-UCST variant harboring 30 cumulative mutations. Note that for
385 visualization in Fig. 4d, the probability curve of UBQLN4-UCST was plotted in the presence of 5% PEG 8k-
386 20k using a CMA window size of 15 for clarity. The AlphaFold database structure of UBQLN4 WT and the
387 predicted structure[45] of UBQLN4-UCST are shown in Supplementary Fig. 9. Finally, although these variants
388 were initially identified using a version of LLPSense prior to its final refinement, the overall predicted
389 temperature-dependent trends and selection logic remain consistent despite minor numerical discrepancies
390 compared to the latest model version (Supplementary Fig. 10).

391

392 **Expression and purification of proteins**

393 All proteins used in this study were cloned into the pET-21a(+) vector using the *NdeI* and *XhoI* restriction sites,
394 allowing for expression with the vector-encoded C-terminal His-tag. Complete protein sequences are provided
395 in Supplementary Table 5. The recombinant plasmids were transformed into *Escherichia coli* BL21(DE3) cells
396 and cultured in LB medium supplemented with 100 μg/mL ampicillin at 37 °C. When the OD_{600} reached 0.4–
397 0.7, protein overexpression was induced by adding 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG),
398 followed by overnight incubation at 20 °C. Cells were harvested by centrifugation at 6000 rpm (~6400 × g) and
399 washed once with 1×PBS (160 mM NaCl, pH 7.3). The cell pellets were resuspended in Ni-IDA binding buffer
400 (20 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole, pH 8.0) and lysed by sonication. Lysates were clarified by

401 centrifugation at 12,000 rpm (~16,000 × g) for 15 min, and the supernatants were collected. The supernatant
402 was loaded onto a Ni-IDA Excellose (Takara) resin column, washed with wash buffer (20 mM Tris-HCl, 150
403 mM NaCl, 50 mM imidazole, pH 8.0), and eluted with elution buffer (20 mM Tris-HCl, 150 mM NaCl, 300
404 mM imidazole, pH 8.0). For α-synuclein variants, the eluate was subjected to heat shock at 65 °C, followed by
405 centrifugation at 13,000 rpm (~16,000 × g) and filtration using a 0.2 μm syringe filter (ADVANTEC). Protein
406 solutions were dialyzed against 1×PBS at 25 °C, followed by filtration through a 0.2 μm syringe filter. Protein
407 concentrations were determined by measuring A_{280}. Final protein solutions were aliquoted and flash-frozen in
408 liquid nitrogen and stored at −80 °C for long-term use.
409
410 **Protein phase separation assay**
411 Protein phase separation was assessed in 96-well black plates with clear bottoms (Thermo Fisher). For each
412 condition, 100 μL of protein solution was loaded per well. Turbidity was quantified by measuring absorbance
413 at 350 nm using a spectral scanning multimode reader (VarioSkan Flash, Thermo Fisher) 30 min after the
414 initiation of phase separation. Baseline correction was performed by subtracting the absorbance of PBS control
415 wells containing the corresponding concentration of PEG8000. The reported turbidity values represent the
416 average of ten independent measurements. For microscopic analysis, confocal imaging was performed using a
417 laser-scanning confocal microscope (A1R HD25 with Eclipse Ti2, Nikon) equipped with a 60×/1.40 NA
418 apochromatic oil-immersion objective. All images were acquired after allowing the samples to settle for 60 min
419 in BSA-passivated 96-well plates.
420
421 **Implementation details**
422 All computational experiments were conducted using a single NVIDIA RTX A6000 GPU. Feature extraction
423 utilizing the ProtT5 model required approximately 131 ms per protein on average (benchmarked on the human
424 proteome restricted to lengths under 4,000 residues). For inference, LLPSense processed a single protein-
425 condition pair in approximately 34 ms. Notably, inference consumed less than 8 GB of VRAM, demonstrating
426 the framework's capability for computationally efficient large-scale screening.
427

428 ## Data and code availability

429 The curated dataset, trained models, and associated code in this study will be made publicly available at
430 https://github.com/NearNiah/LLPSense.git upon submission of the manuscript.

431

432 ## Author contributions

433 J.B. and M.K. conceived the study, developed the model, and wrote the code. D.L. curated the datasets under
434 the guidance of J.B. J.B. performed the experiments with support from D.L. M.K. performed the computational

analysis and evaluated the model performance. J.B. conducted the data analysis, visualized the results and schemes, and drafted the manuscript. Y.J. and M.K. reviewed and edited the manuscript. K.-J.Y. and Y.J. supervised the project.

Competing interests

Y.J., J.B., M.K., and K.-J.Y. are inventors on a patent application filed by Korea Advanced Institute of Science and Technology related to the LLPSense technology (Application No. KR 10-2025-0208052). The other author declares no competing interests.

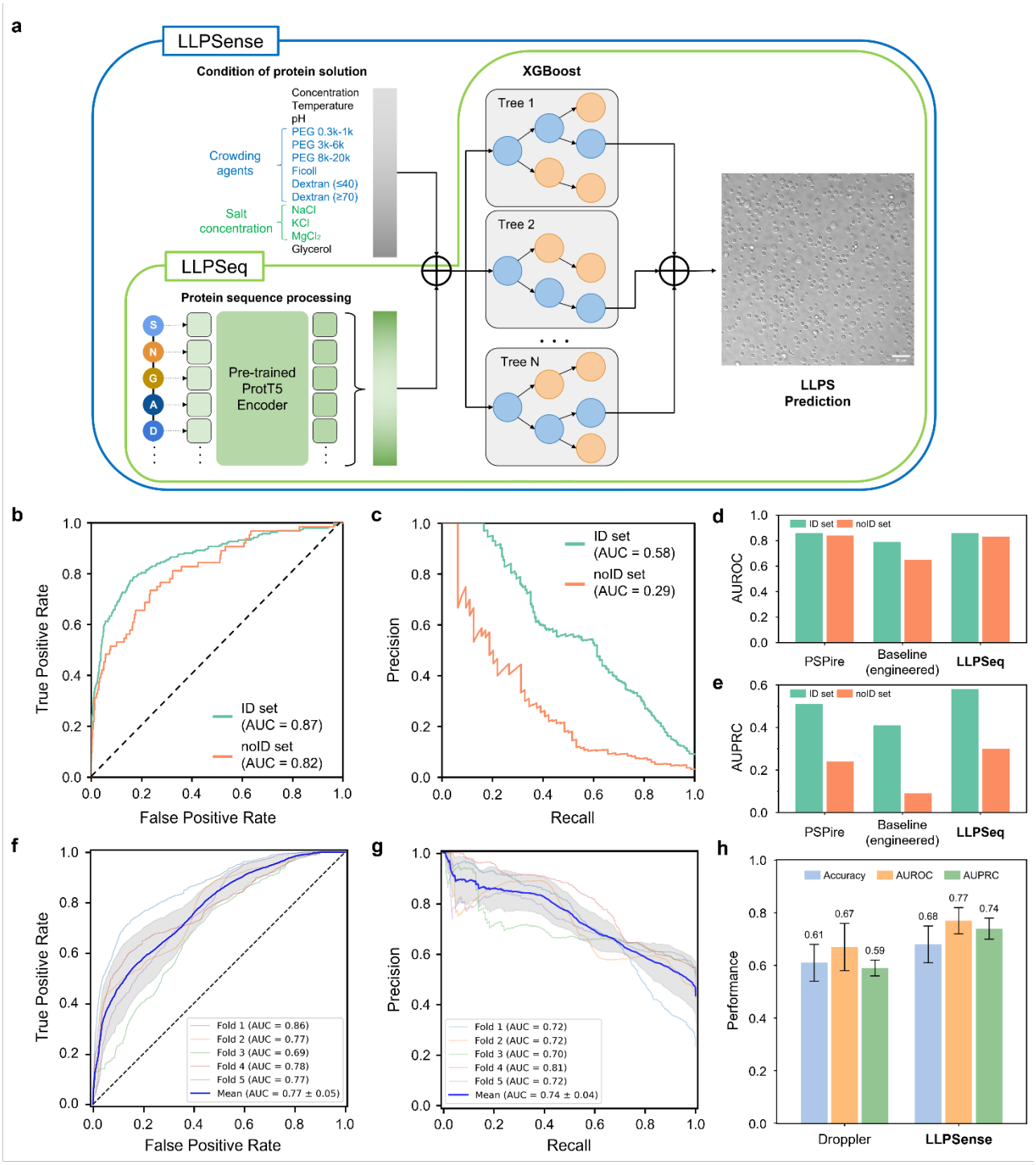

Figure 1. Architectural framework and performance benchmarking of LLPSeq and LLPSense. (a) Schematic illustration of the LLPSeq and LLPSense frameworks. **(b, c)** Predictive performance of the pLM (protein language model)-based LLPSeq model, shown as (b) receiver operating characteristic (ROC) curves and (c) precision-recall (PR) curves. Summary of **(d)** AUROC and **(e)** AUPRC scores of LLPSeq against PSPire and a baseline model (trained on 55 engineered features). **(f, g)** Evaluation of LLPSense via 5-fold cluster-based cross-validation, displaying (f) ROC and (g) PR curves. Shaded regions represent ± 1 s.d. from the mean. **(h)** Performance comparison on the sequence-condition integrated task against the Droppler model, which was retrained on the same expanded dataset.

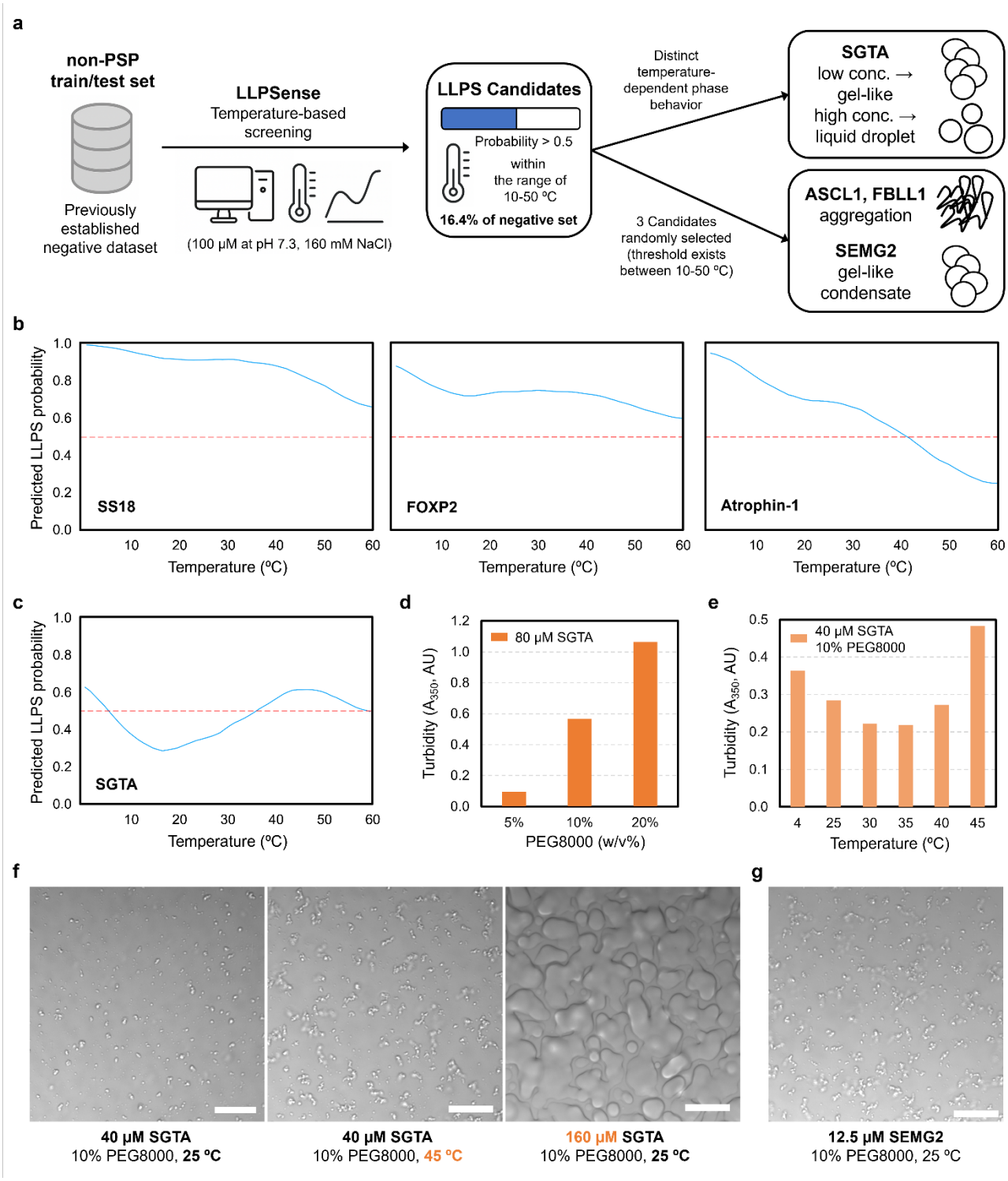

Figure 2. Experimental validation of LLPS candidates identified by LLPSense from a previously defined negative dataset. (a) Schematic overview of the screening and validation workflow. LLPSense re-evaluated a previously established non-PSP dataset, identifying 16.4% of entries as potential phase separation candidates (probability > 0.5 within 10–50 °C). **(b, c)** Predicted temperature profiles of LLPS candidates. The curve was smoothed using a centered moving average (window size = 15). (b) Literature-validated phase separating proteins previously misclassified in the non-PSP set. (c) The profile of SGTA protein exhibits distinctive reentrant phase behavior, characterized by mixed UCST/LCST regimes. **(d)** Turbidity measurements (A_{350}) for SGTA under crowding conditions. **(e)** Temperature-dependent turbidity validating the predicted reentrant behavior of SGTA. **(f)** Representative microscopy images showing the condition-dependent morphology of SGTA condensates. **(g)** SEMG2 condensates exhibiting interconnected gel-like clusters. (f, g) Scale bars, 20 μm.

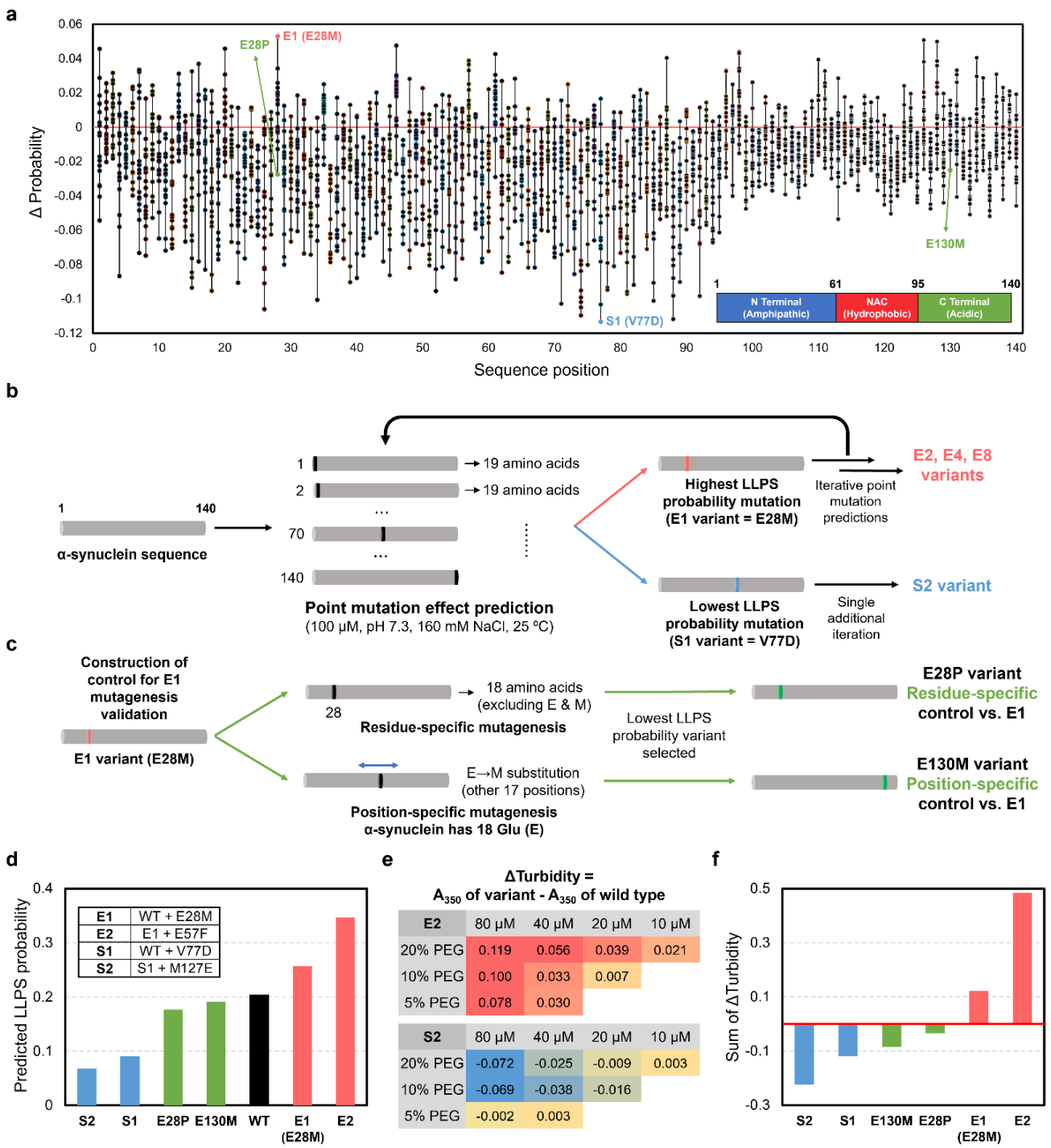

Figure 3. LLPSense-guided residue-level mapping and rational design of α-synuclein variants. (a) Predicted mutational landscape showing the shift in probability (ΔProbability) for all point mutants relative to the WT (red line). Key variants (E1, S1) and controls (E28P, E130M) are highlighted. The domain architecture of α-synuclein is annotated at the bottom. **(b)** Schematic workflow of the *in silico* saturation mutagenesis. E1 (E28M) and S1 (V77D) were identified as the strongest modulators, serving as scaffolds for recursive multi-mutant design (E2–E8, S2). **(c)** Schematic of control variant construction to verify the specificity of LLPSense. Controls were designed to decouple position and residue effects: E28P (same position, different residue) and E130M (different position, same residue). **(d)** Predicted probabilities for designed variants. The inset details the specific mutations. **(e)** Experimental ΔTurbidity heatmaps showing condition-dependent phase behavior relative to WT. Warm colors indicate enhancement; cool colors indicate suppression. **(f)** Summed ΔTurbidity across all conditions. The experimental ranking mirrors the predictions in (d), verifying the distinct separation of enhanced and suppressed variants relative to the WT baseline (red line).

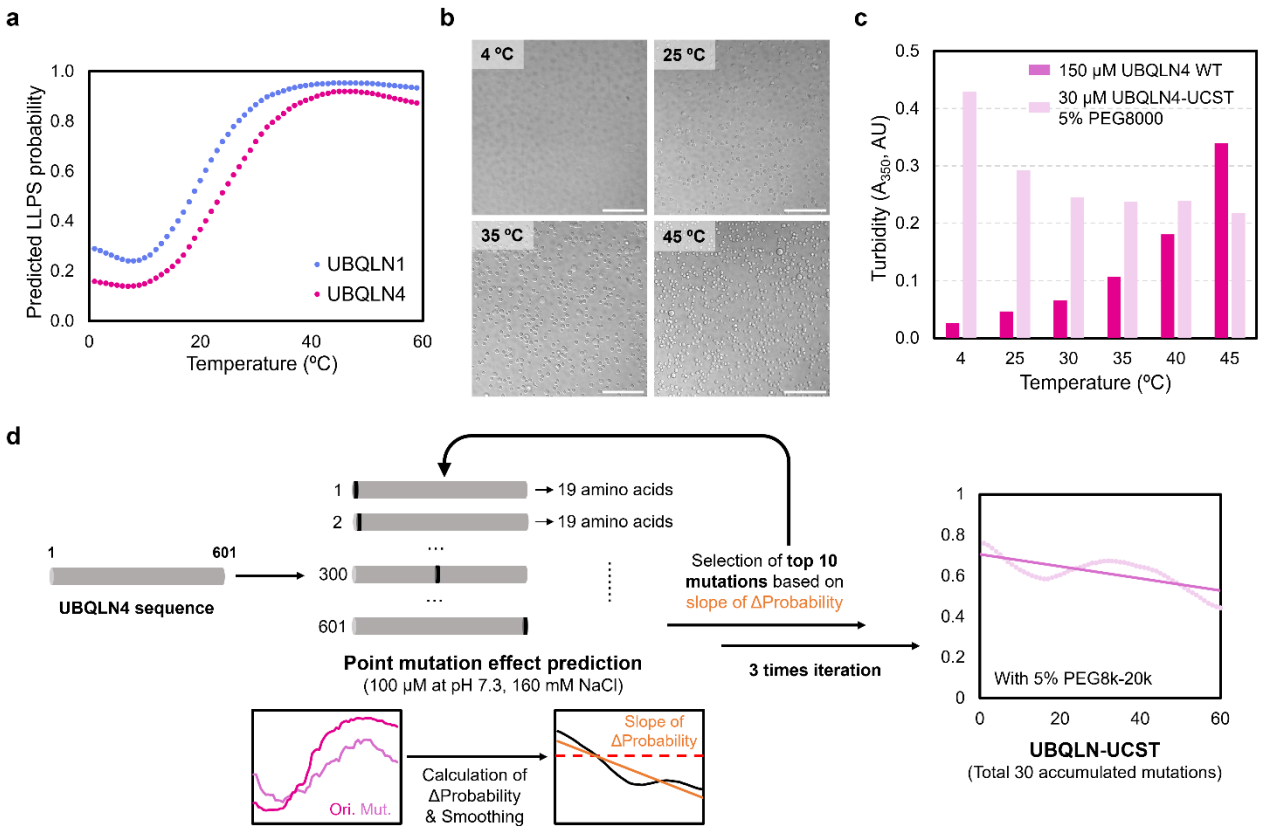

Figure 4. LLPSense-guided identification and rational reprogramming of UBQLN4 phase behavior. (a) Temperature-dependent LLPS probability profiles predicted by LLPSense for UBQLN1 and UBQLN4. The model accurately captures the LCST characteristics of both proteins. Predicted points were smoothed using a centered moving average (CMA, window size = 15). **(b)** Representative microscopy images of UBQLN4 WT condensates formed across a temperature gradient. Scale bars, 50 μm. **(c)** Experimental turbidity measurements (A_{350}) comparing the thermal response of UBQLN4 WT and the UBQLN4-UCST variant. While the WT (dark magenta) exhibits distinct LCST behavior, the UBQLN4-UCST (light pink) displays a UCST-like profile. **(d)** Schematic workflow and computational strategy for the rational reprogramming of phase behavior. The design strategy prioritized the top *N=10* mutations in each round exhibiting the largest slope of ΔProbability in the desired direction (sign) to invert the condition-dependent profile. While the selection is based on individual single-mutation evaluations, the representative plot displays the simultaneous introduction of the top 10 mutations to more effectively visualize the predicted shift in phase behavior.

References

1 Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017). https://doi.org/10.1038/nrm.2017.7

2 Niu, X. *et al.* Biomolecular condensates: Formation mechanisms, biological functions, and therapeutic targets. *MedComm* **4**, e223 (2023). https://doi.org/10.1002/mco2.223

3 Wang, B. *et al.* Liquid-liquid phase separation in human health and diseases. *Signal Transduct. Target. Ther.* **6**, 290 (2021). https://doi.org/10.1038/s41392-021-00678-1

4 Ding, M., Xu, W., Pei, G. & Li, P. Long way up: rethink diseases in light of phase separation and phase transition. *Protein Cell* **15**, 475–492 (2024). https://doi.org/10.1093/procel/pwad057

5 Smokers, I. B. A., Visser, B. S., Slootbeek, A. D., Huck, W. T. S. & Spruijt, E. How Droplets Can Accelerate Reactions-Coacervate Protocells as Catalytic Microcompartments. *Acc. Chem. Res.* **57**, 1885–1895 (2024). https://doi.org/10.1021/acs.accounts.4c00114

6 van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep.* **34**, 108705 (2021). https://doi.org/10.1016/j.celrep.2021.108705

7 Saar, K. L. *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci. USA* **118** (2021). https://doi.org/10.1073/pnas.2019053118

8 Chu, X. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**, 72 (2022). https://doi.org/10.1186/s12859-022-04599-w

9 Chen, Z. *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc. Natl. Acad. Sci. USA* **119**, e2115369119 (2022). https://doi.org/10.1073/pnas.2115369119

10 Vernon, R. M. & Forman-Kay, J. D. First-generation predictors of biological protein phase separation. *Curr. Opin. Struct. Biol.* **58**, 88–96 (2019). https://doi.org/10.1016/j.sbi.2019.05.016

11 Hou, S. *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. *Nat. Commun.* **15**, 2147 (2024). https://doi.org/10.1038/s41467-024-46445-y

12 Sun, J. *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat. Commun.* **15**, 2662 (2024). https://doi.org/10.1038/s41467-024-46901-9

13 Poudyal, M. *et al.* Intermolecular interactions underlie protein/peptide phase separation irrespective of sequence and structure at crowded milieu. *Nat. Commun.* **14**, 6199 (2023). https://doi.org/10.1038/s41467-023-41864-9

14 Raimondi, D. *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). https://doi.org/10.1093/bioinformatics/btab350

15 Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022). https://doi.org/10.1109/TPAMI.2021.3095381

16 Miyata, K. & Iwasaki, W. Seq2Phase: language model-based accurate prediction of client proteins in liquid-liquid phase separation. *Bioinform. Adv.* **4**, vbad189 (2024). https://doi.org/10.1093/bioadv/vbad189

17 Chen, T. Q. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 785–794 (2016). https://doi.org/10.1145/2939672.2939785

18 Wang, X. *et al.* LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* **38**, 2010–2014 (2022). https://doi.org/10.1093/bioinformatics/btac026

19 Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res.* **48**, D320–D327 (2020). https://doi.org/10.1093/nar/gkz778

20 Steinegger, M. & Soding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018). https://doi.org/10.1038/s41467-018-04964-5

21 Kuang, J. *et al.* SS18 regulates pluripotent-somatic transition through phase separation. *Nat. Commun.* **12**, 4090 (2021). https://doi.org/10.1038/s41467-021-24373-5

22 Vaglietti, S. *et al.* PolyQ length-based molecular encoding of vocalization frequency in FOXP2. *iScience* **26**, 108036 (2023). https://doi.org/10.1016/j.isci.2023.108036

23 Hinz, J., Lehnhardt, L., Zakrzewski, S., Zhang, G. & Ignatova, Z. Polyglutamine expansion alters the dynamics and molecular architecture of aggregates in dentatorubropallidoluysian atrophy. *J. Biol. Chem.* **287**, 2068–2078 (2012). https://doi.org/10.1074/jbc.M111.318915

24 Zhang, D. *et al.* Identification of FBLL1 as a neuron-specific RNA 2'-O-methyltransferase mediating neuronal differentiation. *Proc. Natl. Acad. Sci. USA* **121**, e2406961121 (2024). https://doi.org/10.1073/pnas.2406961121

25 Malm, J., Hellman, J., Magnusson, H., Laurell, C. B. & Lilja, H. Isolation and characterization of the major gel proteins in human semen, semenogelin I and semenogelin II. *Eur. J. Biochem.* **238**, 48–53 (1996). https://doi.org/10.1111/j.1432-1033.1996.0048q.x

26 Visser, B. S., Lipinski, W. P. & Spruijt, E. The role of biomolecular condensates in protein aggregation. *Nat Rev Chem* **8**, 686–700 (2024). https://doi.org/10.1038/s41570-024-00635-w

27 Franzmann, T. M. & Alberti, S. Protein Phase Separation as a Stress Survival Strategy. *Cold Spring Harb. Perspect. Biol.* **11** (2019). https://doi.org/10.1101/cshperspect.a034058

28 Alberti, S. & Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nat. Rev. Mol. Cell Biol.* **22**, 196–213 (2021). https://doi.org/10.1038/s41580-020-00326-6

29 Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017). https://doi.org/10.1126/science.aaf4382

30 Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019). https://doi.org/10.1016/j.cell.2018.12.035

31 Vendruscolo, M. & Fuxreiter, M. Protein condensation diseases: therapeutic opportunities. *Nat. Commun.* **13**, 5550 (2022). https://doi.org/10.1038/s41467-022-32940-7

32 Zhu, G. *et al.* Phase Separation of Disease-Associated SHP2 Mutants Underlies MAPK Hyperactivation. *Cell* **183**, 490–502 e418 (2020). https://doi.org/10.1016/j.cell.2020.09.002

33 Wegmann, S. *et al.* Tau protein liquid-liquid phase separation can initiate tau aggregation. *EMBO J.* **37**, EMBJ201798049 (2018). https://doi.org/10.15252/embj.201798049

34 Mackenzie, I. R. *et al.* TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron* **95**, 808–816 e809 (2017). https://doi.org/10.1016/j.neuron.2017.07.025

35 Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015). https://doi.org/10.1016/j.cell.2015.09.015

36 Li, C. H. *et al.* MeCP2 links heterochromatin condensates and neurodevelopmental disease. *Nature* **586**, 440–444 (2020). https://doi.org/10.1038/s41586-020-2574-4

37 Bouchard, J. J. *et al.* Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compartments. *Mol. Cell* **72**, 19–36 e18 (2018). https://doi.org/10.1016/j.molcel.2018.08.027

38 Ray, S. *et al.* alpha-Synuclein aggregation nucleates through liquid-liquid phase separation. *Nat. Chem.* **12**, 705–716 (2020). https://doi.org/10.1038/s41557-020-0465-9

39 Xu, B. *et al.* Distinct Effects of Familial Parkinson's Disease-Associated Mutations on alpha-Synuclein Phase Separation and Amyloid Aggregation. *Biomolecules* **13** (2023). https://doi.org/10.3390/biom13050726

40 Chandran, A. *et al.* Dual effect of α-synuclein disease variants on condensate formation. *bioRxiv*, 2025.2006.2006.657340 (2025). https://doi.org/10.1101/2025.06.06.657340

41 Ghosh, D. *et al.* The Parkinson's disease-associated H50Q mutation accelerates alpha-Synuclein aggregation in vitro. *Biochemistry* **52**, 6925–6927 (2013). https://doi.org/10.1021/bi400999d

42 Tanaka, N., Suyama, K., Tomohara, K. & Nose, T. Exploring LCST- and UCST-like Behavior of Branched Molecules Bearing Repeat Units of Elastin-like Peptides as Side Components. *Biomacromolecules* **25**, 7156–7166 (2024). https://doi.org/10.1021/acs.biomac.4c00751

43 Dao, T. P., Rajendran, A., Galagedera, S. K. K., Haws, W. & Castaneda, C. A. Short disordered termini and proline-rich domain are major regulators of UBQLN1/2/4 phase separation. *Biophys. J.* **123**, 1449–1457 (2024). https://doi.org/10.1016/j.bpj.2023.11.3401

44 Bertoni, D. *et al.* AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. *Nucleic Acids Res.* (2025). https://doi.org/10.1093/nar/gkaf1226

45 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w