# Learning the molecular grammar of protein condensates from sequence determinants and embeddings

Kadi L. Saar[a,b,1], Alexey S. Morgunov[a,1,2], Runzhang Qi[a], William E. Arter[a], Georg Krainer[a], Alpha A. Lee[b], and Tuomas P. J. Knowles[a,b,3]

[a]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; and [b]Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, United Kingdom

Edited by William A. Eaton, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, and approved February 17, 2021 (received for review September 13, 2020)

**Intracellular phase separation of proteins into biomolecular condensates is increasingly recognized as a process with a key role in cellular compartmentalization and regulation. Different hypotheses about the parameters that determine the tendency of proteins to form condensates have been proposed, with some of them probed experimentally through the use of constructs generated by sequence alterations. To broaden the scope of these observations, we established an in silico strategy for understanding on a global level the associations between protein sequence and phase behavior and further constructed machine-learning models for predicting protein liquid–liquid phase separation (LLPS). Our analysis highlighted that LLPS-prone proteins are more disordered, less hydrophobic, and of lower Shannon entropy than sequences in the Protein Data Bank or the Swiss-Prot database and that they show a fine balance in their relative content of polar and hydrophobic residues. To further learn in a hypothesis-free manner the sequence features underpinning LLPS, we trained a neural network-based language model and found that a classifier constructed on such embeddings learned the underlying principles of phase behavior at a comparable accuracy to a classifier that used knowledge-based features. By combining knowledge-based features with unsupervised embeddings, we generated an integrated model that distinguished LLPS-prone sequences both from structured proteins and from unstructured proteins with a lower LLPS propensity and further identified such sequences from the human proteome at a high accuracy. These results provide a platform rooted in molecular principles for understanding protein phase behavior. The predictor, termed DeePhase, is accessible from https://deephase.ch.cam.ac.uk/.**

liquid–liquid phase separation | biomolecular condensates | protein biophysics | machine learning | language models

Liquid–liquid phase separation (LLPS) is a widely occurring biomolecular process that underpins the formation of membraneless organelles within living cells (1–4). This phenomenon and the resulting condensate bodies are increasingly recognized to play important roles in a wide range of biological processes, including the onset and development of metabolic diseases and cancer (5–11). Understanding the factors that drive the formation of protein-rich biomolecular condensates has thus become an important objective and been the focus of a large number of studies, which have collectively yielded valuable information about the factors that govern protein phase behavior (3, 4, 12, 13).

While changes in extrinsic conditions, such as temperature, ionic strength, or the level of molecular crowding, can strongly modulate LLPS (14–17), of fundamental importance to condensate formation is the linear amino acid sequence of a protein, its primary structure. A range of sequence-specific factors governing the formation of protein condensates have been postulated with electrostatic interactions, $\pi$–$\pi$ and cation–$\pi$ contacts, and hydrophobic interactions and the valency and patterning of the low-complexity regions (LCRs) in particular brought forward as central features (12, 13, 18–22). The predictive power of some of these hypotheses has been recently reviewed (23). In parallel, studies examining the relationship between protein phase behavior and its sequence alterations through deletion, truncation, and site-specific mutation events have determined various sequence-specific features to be important in modulating the protein phase separation of specific proteins, such as the high abundance of arginine and tyrosine residues in the context of the fused in sarcoma (FUS)-family proteins (22), the positioning of tryptophan and other aromatic amino acid residues in TAR DNA-binding protein 43 (TDP-43) (24), arginine- and glycine-rich disordered domains in LAF-1 protein (25), and multivalent interactions for the UBQLN2 protein (26).

## Significance

The tendency of many cellular proteins to form protein-rich biomolecular condensates underlies the formation of subcellular compartments and has been linked to various physiological functions. Understanding the molecular basis of this fundamental process and predicting protein phase behavior have therefore become important objectives. To develop a global understanding of how protein sequence determines its phase behavior, we constructed bespoke datasets of proteins of varying phase separation propensity and identified explicit biophysical and sequence-specific features common to phase-separating proteins. Moreover, by combining this insight with neural network-based sequence embeddings, we trained machine-learning classifiers that identified phase-separating sequences with high accuracy, including from independent external test data.

Author contributions: K.L.S., A.S.M., and T.P.J.K. designed research; K.L.S. and A.S.M. performed research; K.L.S., A.S.M., and R.Q. contributed new reagents/analytic tools; K.L.S. and A.S.M. analyzed data; K.L.S., A.S.M., and T.P.J.K. wrote the paper; and W.E.A., G.K., and A.A.L. discussed the results and reviewed the manuscript.

Competing interest statement: Parts of this work have been the subject of a patent application filed by Cambridge Enterprise Limited, a fully owned subsidiary of the University of Cambridge.

This article is a PNAS Direct Submission.

[1] K.L.S. and A.S.M. contributed equally to this work.

[2] Present address: Fluidic Analytics, Unit A, The Paddocks Business Centre, Cambridge CB1 8DH, United Kingdom.

[3] To whom correspondence may be addressed. Email: tpjk2@cam.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2019053118/-/DCSupplemental.

Published April 7, 2021.

To broaden the scope of these observations and understand on a global level the associations between the primary structure of a protein and its tendency to form condensates, here, we developed an in silico strategy for analyzing the associations between LLPS propensity of a protein and its amino acid sequence and used this information to construct machine-learning classifiers for predicting LLPS propensity from the amino acid sequence (Fig. 1). Specifically, by starting with a previously published LLPSDB database collating information on protein phase behavior under different environmental conditions (27) and by analyzing the concentration under which LLPS had been observed to take place in these experiments, we constructed two datasets including sequences of different LLPS propensity and compared them to fully ordered structures from the Protein Data Bank (PDB) (29) as well as the Swiss-Prot (30) database. We observed phase-separating proteins to be more hydrophobic, more disordered, and of lower Shannon entropy and have their low-complexity regions enriched in polar residues. Moreover, high LLPS propensity correlated with high abundance of polar residues yet the lowest saturation concentrations were reached when their abundance was balanced with a sufficiently high hydrophobic content.

Moreover, we used the outlined sequence-specific features as well as implicit protein sequence embeddings generated using a neural network-derived word2vec model and trained classifiers for predicting the propensity of unseen proteins to phase separate. We showed that even though the latter strategy required no specific feature engineering, it allowed constructing classifiers that were comparably effective at identifying LLPS-prone sequences as the model that used knowledge-based features, demonstrating that language models can learn the molecular grammar of phase separation. Our final model, combining knowledge-based features with unsupervised embeddings, showed a high performance both when distinguishing LLPS-prone proteins from structured ones and when identifying them within the human proteome. Overall, our results shed light onto the physicochemical factors modulating protein condensate formation and provide a platform rooted in molecular principles for the prediction of protein phase behavior.

## Results and Discussion

**Construction of Datasets and Their Global Sequence Comparison.** To link the amino acid sequence of a protein to its tendency to form biomolecular condensates, we collated data from two publicly available datasets, the LLPSDB (27) and the PDB (29), and constructed three bespoke datasets—LLPS$^+$ and LLPS$^-$ comprising intrinsically disordered proteins with high and low LLPS propensity, respectively, and PDB$^*$ consisting of sequences that were very unlikely to phase separate (see below). To create the first two datasets, the LLPSDB dataset was reduced to naturally occurring sequence constructs with no posttranslational modifications and further filtered for sequences that were observed to undergo phase separation on their own (i.e., homotypically), in isolation from other components, such as DNA, RNA, or an additional protein (*Materials and Methods*). In the resulting dataset, the mean concentration at which each construct had been observed to phase separate was estimated and a threshold of 100 μM was used to divide the sequences according to their high propensity (LLPS$^+$; 137 constructs from 77 unique UniProt IDs; Dataset S1) or low propensity to phase separate (Fig. 1*B*). The latter set (25 constructs) was then combined with the sequences from the LLPSDB that had not been observed to phase separate homotypically (72 constructs) and this created the LLPS$^-$ dataset (84 constructs from 52 unique UniProt

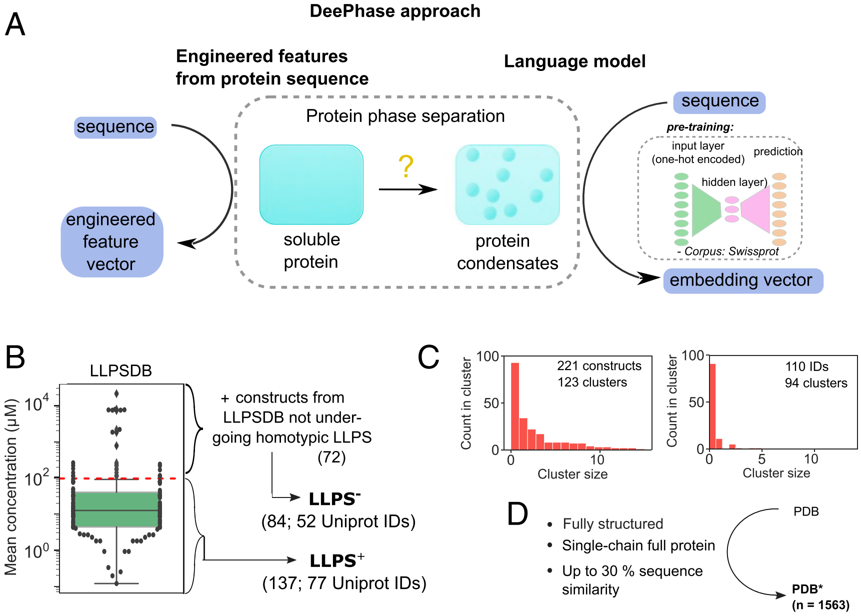

**Fig. 1.** (*A*) DeePhase predicts the propensity of proteins to undergo phase separation by combining engineered features computed directly from protein sequences with protein sequence embedding vectors generated using a pretrained language model. The DeePhase model was trained using three datasets, namely two classes of intrinsically disordered proteins with a different LLPS propensity (LLPS$^+$ and LLPS$^-$) and a set of structured sequences (PDB$^*$). (*B*) To generate the LLPS$^+$ and LLPS$^-$ datasets, the entries in the LLPSDB database (27) were filtered for single-protein systems. The constructs that phase separated at an average concentration below $c = 100$ μM were classified as having a high LLPS propensity (LLPS$^+$; 137 constructs from 77 UniProt IDs) with the remaining 25 constructs together with constructs that had not been observed to phase separate homotypically classified as low-propensity dataset (LLPS$^-$; 84 constructs from 52 UniProt IDs). (*C*) The 221 sequences clustered into 123 different clusters [*Left*, CD-hit clustering algorithm (28) with the lowest threshold of 0.4]. (*Right*) The 110 parent sequences showed high diversity by forming 94 distinct clusters. (*D*) The PDB$^*$ dataset (1,563 constructs) was constructed by filtering the entries in the PDB (29) to fully structured full-protein single chains and clustering for sequence similarity with a single entry selected from each cluster.

IDs; Dataset S2). We note that the classification does not imply that the proteins in the LLPS$^-$ dataset cannot phase separate under any conditions, but it rather allowed us to learn information about sequence-specific features that correlate with enhanced LLPS propensity. To further characterize the diversity of the datasets, we clustered the sequences for similarity (*Materials and Methods*) and found the 221 constructs across the LLPS$^+$ and LLPS$^-$ datasets to form 123 unique clusters (Fig. 1*C*, *Left*). Moreover, when reducing the combined dataset of 221 sequences down to a single sequence from each UniProt ID by keeping the longest sequence for the cases where more than a single construct had been derived from the same parent sequence, we identified 94 different clusters (Fig. 1*B*, *Right*). Overall, these results indicate a noticeable amount of sequence diversity, which is essential for building models that generalize well.

An additional dataset, PDB$^*$, consisting of entries sampled from the PDB was constructed for it to serve as an alternative training set comprising sequences that are highly unlikely to phase separate. This dataset was constructed by including sequences from the PDB that did not include any disordered residues (112,572 chains) filtered for lengths above 50 amino acids with the selected sequences verified via mapping to their UniProt IDs (*Materials and Methods*). The remaining sequences (13,325) were clustered for sequence identity to retain no more than a single sequence from each cluster (*Materials and Methods*). This process reduced the original PDB dataset to a diverse set of 1,563 fully structured sequences (PDB$^*$; Dataset S3; Fig. 1*D*).

We compared the generated datasets across a range of global sequence-specific features with the aim to understand the factors that are linked with enhanced condensate formation propensity using the Swiss-Prot database (30) as a reference control (Fig. 2 *A*–*E*; full distributions are shown in *SI Appendix*, Fig. S1). From the analysis, we first concluded that the average construct in the LLPS$^+$ dataset (cyan) did not differ in its length from the average construct in Swiss-Prot (gray) but it was longer than the average construct in the PDB$^*$ (magenta; Fig. 2*A*) or the LLPS$^-$ (orange) datasets, which is consistent with theoretical expectations as increasing the protein length decreases the entropic cost per amino acid of confining a protein into a condensate. We next estimated the hydrophobicity of all of the constructs in the four datasets using the Kyte and Doolittle hydropathy scale (31) and concluded LLPS-prone constructs to be less hydrophobic than the sequences in any of the other three datasets (Fig. 2*B*). Finally, we noted that LLPS-prone sequences exhibited lower Shannon entropy than the sequences in the PDB$^*$ or the Swiss-Prot dataset (Fig.

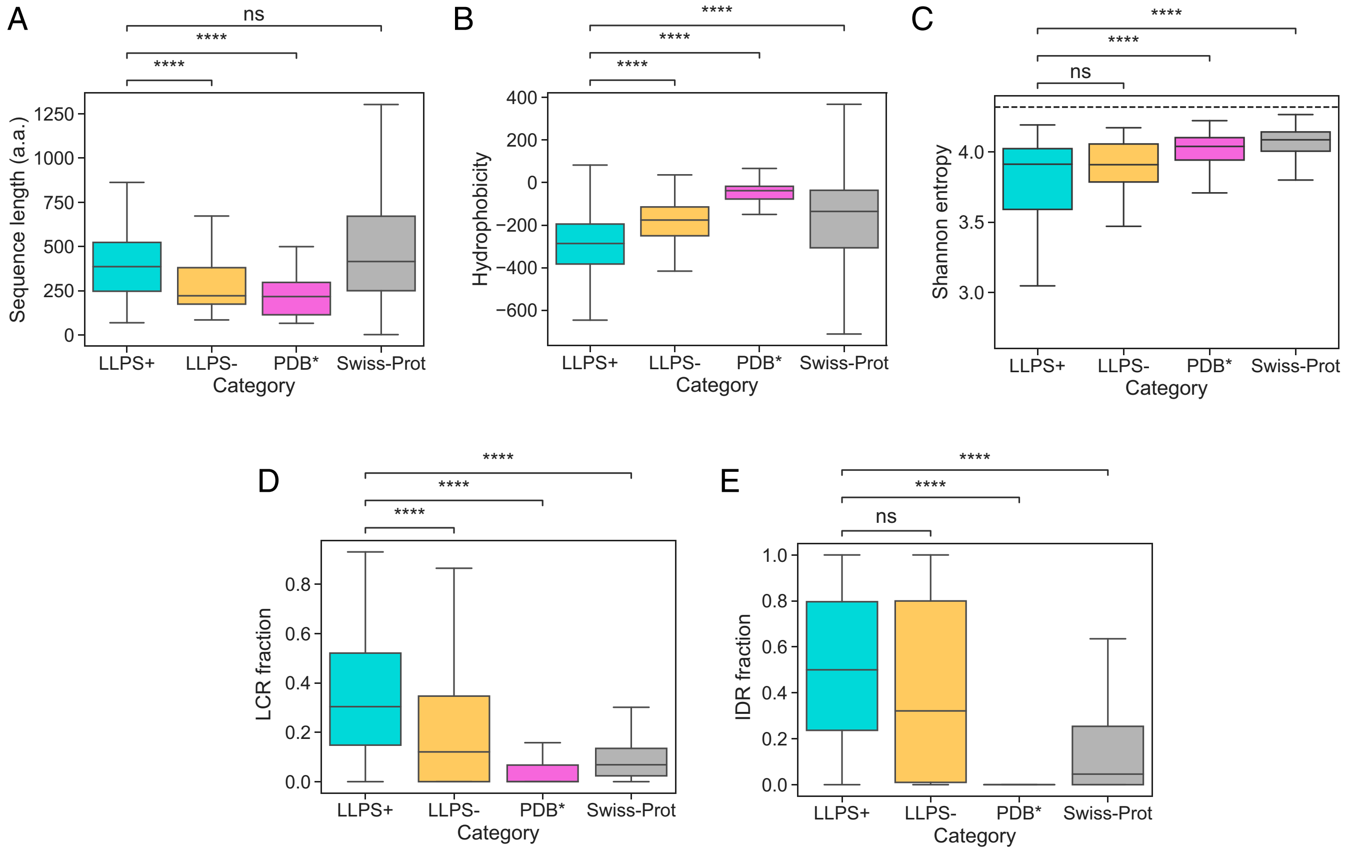

**Fig. 2.** (*A*–*E*) Comparison of the (*A*) sequence length (in amino acids, a.a.), (*B*) hydrophobicity, (*C*) Shannon entropy, the fraction of sequence that is part of (*D*) the low-complexity regions (LCRs) and (*E*) the intrinsically disordered regions (IDRs) for the three training datasets and the Swiss-Prot. Comparative analysis highlighted that the average construct in the LLPS$^+$ dataset (cyan) was longer than in the LLPS$^-$ (orange) and the PDB$^*$ (magenta) datasets and less hydrophobic and had a higher LCR fraction than sequences in the LLPS$^-$, the PDB$^*$, or the Swiss-Prot (gray) datasets. It also had a lower Shannon entropy and a higher IDR fraction than sequences in the PDB$^*$ or the Swiss-Prot datasets. The boxes bound data between the upper and the lower quartile, and the center lines indicates the mean value. The ends of the whiskers correspond to values that exceed the boundaries of the interquartile range by 1.5 times its size or to the most extreme value. Significance was tested with a Mann–Whitney test, $^{****}$ denotes a $P$ value below $10^{-4}$, and ns denotes no significance at $P \leq 0.01$. Full distributions are shown in *SI Appendix*, Fig. S1. The dashed line in *C* corresponds to the case when all amino acids are present at equal frequencies.

2*C*)—an effect that can be linked to their extended prion-like domains (21, 22).

To understand how sequence complexity and the extent of disorder were linked to the tendency of proteins to undergo phase separation, we employed the SEG algorithm (32) to extract LCRs for all of the sequences in the four datasets and the IUPred2 algorithm (33) to identify their disordered regions (*Materials and Methods*). This analysis revealed that constructs in the $\mathrm{LLPS}^{+}$ dataset had a larger fraction of the sequences that was part of the LCRs than of the sequences in any of the other three datasets (Fig. 2*D*) and a higher degree of disorder than sequences in the $\mathrm{PDB}^{*}$ or the Swiss-Prot dataset (Fig. 2*E*).

**Amino Acid Composition of the Constructs Undergoing LLPS.** Having ascertained the length of the low-complexity and intrinsically disordered regions as basic parameters that set the constructed datasets apart (Fig. 2), we next set out to analyze the amino acid composition of these regions. By classifying the amino acid residues into polar, hydrophobic, aromatic, cationic, and anionic categories (*Materials and Methods*), we observed that the propensity of proteins to undergo LLPS was associated with a higher relative content of polar (blue) and a reduced relative content of hydrophobic (orange) and anionic (purple) residues across the full amino acid sequence (Fig. 3*A*). The increased abundance of polar residues was particularly pronounced within the LCRs with aromatic (green) and cationic (red) residues also being overrepresented within these regions compared to the sequences in the Swiss-Prot database (Fig. 3*B*), consistent with previous observations and findings (18, 21). Moreover, we observed that not only are sequences with a high LLPS propensity enriched in polar

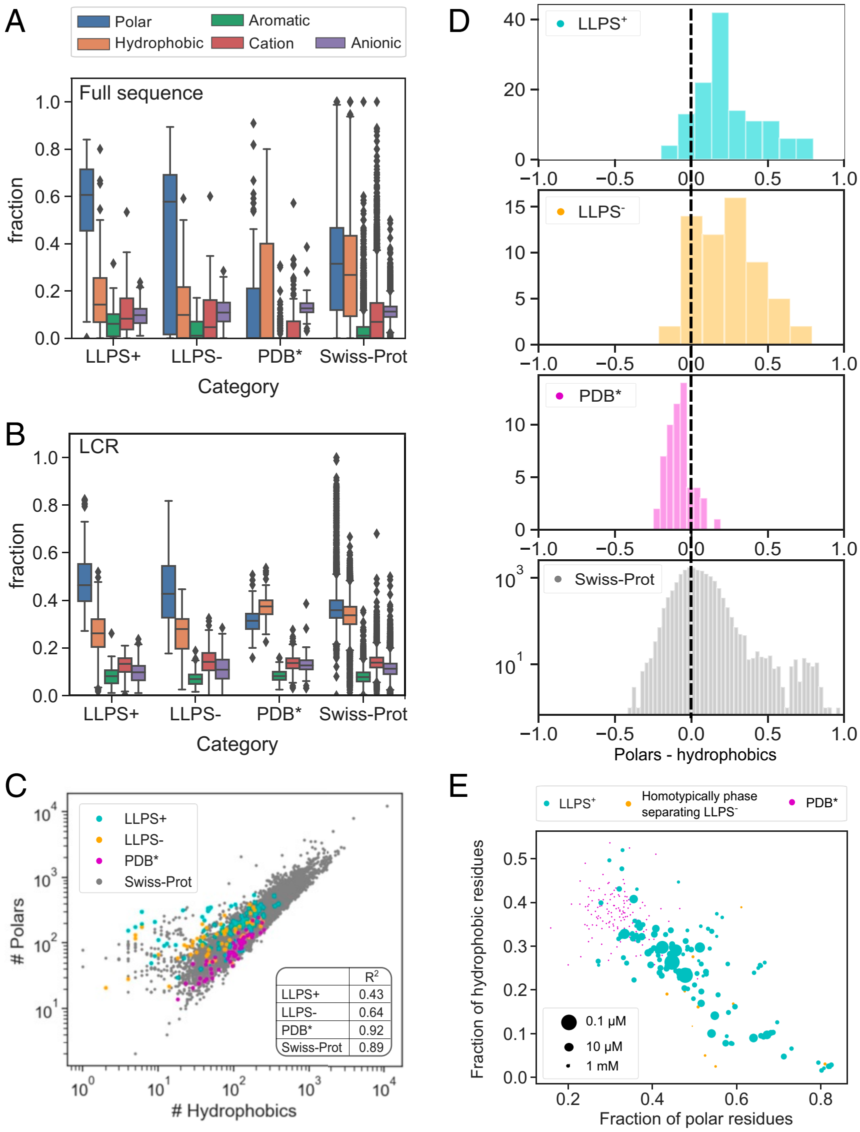

**Fig. 3.** Comparison of the amino acid composition of the sequences within the $\mathrm{LLPS}^{+}$, $\mathrm{LLPS}^{-}$, $\mathrm{PDB}^{*}$, and Swiss-Prot datasets. (*A*) LLPS-prone sequences were enriched in polar amino acid residues (blue) and depleted in hydrophobic (orange) and anionic (purple) residues. (*B*) The elevated abundance of polar residues was particularly pronounced within the LCR with the sequences in the $\mathrm{LLPS}^{+}$ also having their LCR regions enriched in aromatic (green) and cationic (red) residues compared to in the Swiss-Prot database. (*C*) The relationship between polar and hydrophobic residues was less tightly conserved among LLPS-prone sequences as indicated by smaller $R^2$ values (*Inset*). (*D*) Compared to the Swiss-Prot and the $\mathrm{PDB}^{*}$ datasets, LLPS-prone proteins showed an overabundance of polar residues relative to hydrophobic ones. (*E*) When further examining the relationship between the saturation concentration of a construct (indicated by the marker size; saturation concentration of 10 mM was assumed for $\mathrm{PDB}^{*}$) and its amino acid composition, a strong overabundance of polar residues was seen to lead to an increase in saturation concentration with most LLPS-prone sequences showing a tight balance between the abundances of polar and hydrophobic residues.

residues; they also showed a much less tightly conserved relationship between polar and hydrophobic residues (Fig. 3*C*). The high relative abundance of hydrophobic residues and their narrowly defined fraction are likely linked to the requirement of a hydrophobic core underpinning the more structured nature of the proteins in the PDB$^*$ dataset.

Since the high abundance of polar residues relative to hydrophobic ones clearly correlated with elevated LLPS propensity (Fig. 3*D*), we next aimed to explore whether a very high content of polar residues affects protein phase behavior. This analysis was motivated by associative polymer theory and the "spacers and sticker" framework (21) whereby the formation of intermolecular interactions and the onset of protein phase separation are facilitated by an interplay between "spacer" and "sticker" regions. To this effect, we first evaluated the saturation concentrations of all of the 149 constructs that had been seen to undergo homotypic phase separation (Fig. 1*B*) as the lowest concentration at which the particular construct had been seen to phase separate. We then used these estimates to examine how the saturation concentration varied with the amino acid composition of the protein (Fig. 3*E*). We also included proteins in the PDB$^*$ dataset and they were visualized such that their marker size would correspond to a saturation concentration of 10 mM. This analysis suggested that low abundance of polar residues leads to proteins establishing a structured confirmation, which in turn lowers their propensity to undergo phase separation. By contrast, after the relative abundance of polar residues over hydrophobic ones reached a critical value of about 2, an increase in the saturation concentration was again observed. The latter effect likely originated from insufficient sticker regions that would facilitate an effective phase-separation process. Taken together, these results illustrate the importance of both disordered spacer and hydrophobic sticker regions for a high phase separation propensity (21).

**Model for Classifying the Propensity of Unseen Sequences to Phase Separate.** We next developed machine-learning classifiers that could predict the propensity of proteins to undergo phase separation using the constructed datasets (LLPS$^+$, LLPS$^-$, and PDB$^*$) for training the models. To convert the sequences into feature vectors we used the aforementioned engineered features (EFs) in combination with distributional semantics-based sequence embeddings. Specifically, to generate such embeddings, we pretrained a word2vec model using the Swiss-Prot database as the corpus and overlapping 3-grams as words. This pretrained model was then used to convert protein sequences into 200-dimensional embedding vectors (Fig. 4*A* and *Materials and Methods*). Such an approach—exploiting the availability of large unlabeled datasets to learn meaningful low-dimensional sequence representations—has been previously shown to serve as an effective transfer learning strategy for predicting the properties of proteins (34). Crucially, when we mapped the embedding vectors for each 3-gram on a two-dimensional (2D) plane [Multicore-TSNE library (35); axes have no particular meaning], we saw that the 3-grams clustered according to their estimated biophysical properties, such as hydrophobicity (Fig. 4*B*) and isoelectric point (Fig. 4*C*), illustrating the capability of the pretrained language model (LM) to capture biophysically relevant information.

We used a dimensionality reduction approach (35) to visualize the feature vectors of all of the data points in the training data on a 2D plane both for the case when EFs (sequence length, hydrophobicity, Shannon entropy, the fraction of the sequence identified to be part of the LCRs and IDRs [Fig. 2] and the fraction of polar, aromatic, and cationic amino acid residues within the LCRs [Fig. 3]) and the word2vec-based embeddings were used (Fig. 4 *D–E*). This process revealed a notable degree of separation between the LLPS$^+$ (cyan) and the PDB$^*$ (magenta) datasets with the sequences in the LLPS$^-$ dataset spread between them. The clustering between the LLPS$^+$ and the PDB$^*$ datasets was clearer for the engineered features, probably because some of the dimensions, such as the degree of disorder, set the two classes apart very distinctly (Fig. 2*E*). This is in contrast to LM-based embedding vectors where such differences are not confined to a single dimension. This mapping of multidimensional vectors into a lower-dimensional manifold served a visual purpose and did not aim to give direct quantitative insight into how effectively classifiers that use these featurization strategies could perform.

We next set out to gain an insight into how well these two feature types could distinguish between the three classes of proteins. Specifically, we trained random forest classifiers for each of the three pairs of data and estimated their performance using a 25-fold cross-validation test with 20% of the data left out for validation each time (*Materials and Methods*). For the cases when PDB$^*$ was used as a training set, a random subset sized equally to the LLPS$^+$ dataset was sampled to ensure that the model encountered a comparable number of sequences from each class during the training process. Moreover, to ensure generalizability, the data were split into training and validation sets in a stratified manner, so that sequences with the same UniProt ID would belong to the same set. Using accuracy, precision, and the area under the receiver-operator-characteristic (ROC) curve (AUROC) as performance metrices, we saw that both EF- and LM-based features allowed efficient distinction between LLPS$^+$ and PDB$^*$ as well as between LLPS$^-$ and PDB$^*$ with distinction between the LLPS$^+$ and the LLPS$^-$ datasets being the most challenging. These performance metrices were robust to the test–train split that was used (*SI Appendix*, Fig. S2). For all proceeding analysis we used the following four models: models EF-1 and LM-1 that had been trained on LLPS$^+$ and the PDB$^*$ and, additionally, models EF-multi and LM-multi that were constructed as multiclass classifiers trained to simultaneously distinguish between all of the three datasets (*Materials and Methods*).

**Performance of the Models on an External Dataset.** Having established the high cross-validation performance of the models within our generated datasets, we set out to test the models on external test data. Specifically, we evaluated their capability on two tasks: 1) distinguishing sequences with a high LLPS propensity from sequences very unlikely to undergo phase separation and 2) identifying LLPS-prone sequences from within the human proteome.

First, to construct an external set of LLPS-prone proteins, we used the PhaSepDB (36) database. After removing from this database the UniProt IDs that overlapped with any of the LLPSDB entries and hence with our training data, we obtained a set of 196 LLPS-prone human sequences (Dataset S5). A further examination of these 196 sequences highlighted that 35 of them included no intrinsically disorder regions. In general, it is known that while fully structured proteins can in principle undergo phase separation, they usually do so at high concentrations and would hence not normally be regarded as LLPS prone (37). To validate this trend, we examined the LLPSDB database (used for constructing the training datasets) where the phase behavior of all of the protein constructs was listed together with environmental conditions. This analysis revealed that all of the experiments in which a fully structured protein had been observed to be phase separated were performed under an extensive amount of molecular crowding (e.g., dextran, ficoll, polyethylene glycol) or in a nonhomotypical environment (e.g., in the presence of lipids). It is thus likely that the fully ordered sequences within the PhaSepDB that were identified as phase separating in a keyword-based literature search similarly phase separated under high concentrations,

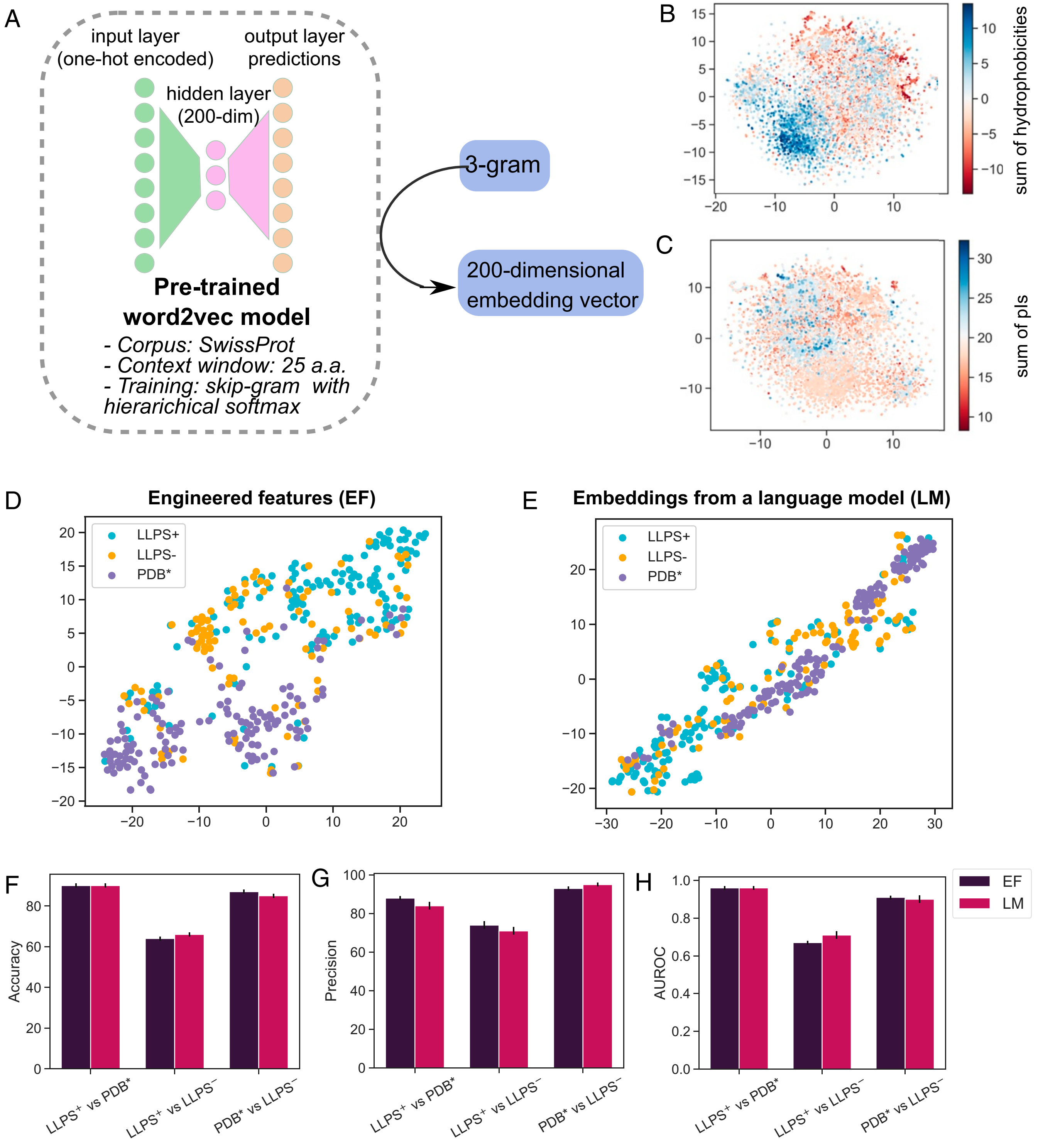

**Fig. 4.** (*A*) A single-hidden-layer language model (LM) was pretrained to learn embedding vectors for each amino acid 3-gram (*Materials and Methods*). The generated embedding vectors clustered 3-grams according to their (*B*) hydrophobicity (evaluated as the sum of the Kyte and Doolittle hydrophobicity values of the individual amino acids in the 3-gram) and (*C*) isoelectric point (pI; sum of the pI values of the individual amino acids). Dimensionality reduction from the 200-dimensional vectors to the 2D plane was performed with the Multicore-TSNE library (35). Visualizing the similarity of the sequences in the LLPS$^+$ (cyan), the LLPS$^-$ (orange), and the PDB$^*$ (magenta) datasets in a 2D plane revealed noticeable clustering between the LLPS$^+$ and the PDB$^*$ datasets both for (*D*) EF vectors and (*E*) embedding vectors generated by the LM. (*F* and *G*) Accuracy (*F*) and precision (*G*) of the featurization strategies when distinguishing between the three pairs of the protein classes using a random forest classifier (performance shown for 25-fold cross-validation). (*H*) Both approaches were highly effective at distinguishing between the LLPS$^+$ and the PDB$^*$ datasets (AUROC of 0.96 ± 0.01 for both). When discriminating between the more similar LLPS$^+$ and LLPS$^-$ datasets, the AUROC scores were lower with the LM-based features potentially reaching a slightly higher score than EFs (AUROCs of 0.67 ± 0.01 and 0.71 ± 0.01, respectively).

nonhomotypical conditions, or a notable level of molecular crowding, and their phase transition cannot be directly linked to the protein sequence as it was not triggered exclusively by homotypic interactions between protein molecules. Motivated by this argument, we eliminated fully structured sequences, which yielded a list of 161 sequences serving as our external test data (Dataset S6). The set of proteins highly unlikely to undergo phase separation (Dataset S7) was created by random sampling an equal number (161) of sequences from the PDB* dataset (Dataset S3) after having removed from it all of the sequences that were used in the training process (Dataset S4). The predictions made by the four models on the sequences that were part of the external test data are highlighted in Fig. 5 *A* and *C* (circles, LLPS-prone sequences; triangles, nonphase-separating sequences) with the shaded regions corresponding to the predictions the models made on the human Swiss-Prot (20,365 entries). Using these data, we constructed the ROC curves for this task (Fig. 5 *B* and *D*, dashed line) and concluded that all of the four models were able to effectively distinguish between the two types of proteins with their AUROCs ranging between 0.98 and 0.99 (Fig. 5*E*).

Additionally, we set out to gain an insight into the capability of the models to identify LLPS-prone sequences from the human proteome. As the phase behavior of many proteins remains unstudied, it is challenging to generate a dataset of nonstructured proteins that do not undergo phase separation. To obtain an estimate of the rate of false negative predictions of the models when identifying LLPS-prone sequences, we relied on the exhaustive keyword-based literature search performed as part of the construction of the PhaSepDB database (36). This database was curated by extracting all publications from the NCBI PubMed database that included phase separation-related keywords in their abstracts. The resulting 2,763 papers were manually rechecked to obtain publications that described membraneless organelles and related proteins, and they were further manually filtered to these proteins that had been observed to undergo phase separation experimentally either in vitro or in vivo. Making a conservative assumption that all of the proteins that were not identified as LLPS positive through this PubMed search are nonphase separating, we could estimate an upper bound for the false positive rate of each of the models in identifying LLPS-prone sequences. The approximated ROC curves with respect to (w.r.t.) proteome are shown in Fig. 5 *B* and *D*, solid lines. All of the four models showed a notable predictive performance with the AUROCs w.r.t. proteome varying between 0.74 and 0.81 (Fig. 5*E*). This performance is in contrast to control

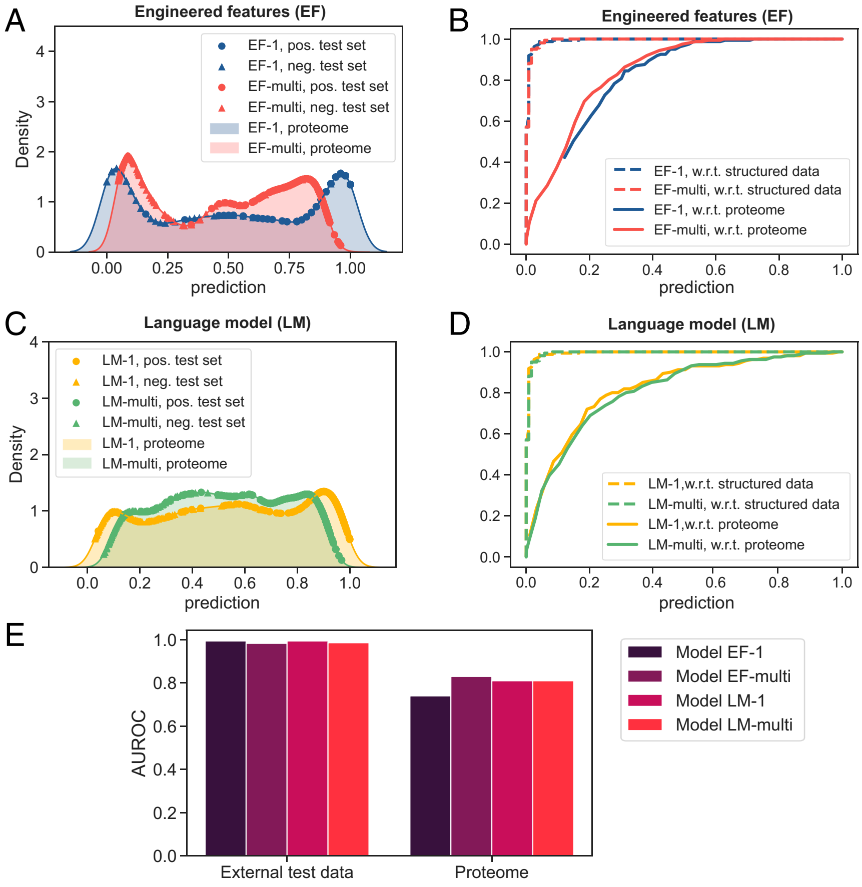

**Fig. 5.** Performance of the models on external data when 1) discriminating between LLPS-prone sequences and structured proteins and 2) identifying LLPS-prone proteins from the human proteome. (*A*) The prediction profiles of model EF-1 (trained on LLPS$^+$ and PDB*) and model EF-multi (trained on all three protein classes) on external test data comprising 161 LLPS-prone sequences (pos.; colored circles) and 161 sequences highly unlikely to undergo phase separation (neg.; colored triangles) and the human proteome (colored region; 20,291 proteins). The positive part of the external dataset was constructed based on the PhaSepDB database with sequences that had their Uniprot IDs overlapping with the training data excluded. The negative half was based on the PDB*. (*B*) ROC curves of the models 1) on the external test data (dashed line) and 2) when identifying LLPS-prone sequences from the human proteome by regarding all proteins that had not been reported to phase separate as nonphase separating (lower bound for the false positive rate; solid line). (*C* and *D*) Same data for models LM-1 and LM-multi where 200-dimensional representations learned from a pretrained word2vec model were used for featurization. (*E*) Comparison of the AUROC values for ROC curves shown in *B* and *D* for the two tasks.

experiments where, prior to training the models, the labels of the sequences were randomly reshuffled (*SI Appendix*, Fig. S3) and the actual sequence compositions were replaced by randomly sampling amino acids from the Swiss-Prot database (*SI Appendix*, Fig. S4). All in all, the results indicate that our models can distinguish between LLPS-prone and structured ones and they can also identify LLPS-prone proteins from within the human proteome. These results also highlight that w2v-based featurization creates meaningful low-dimensional representations that can be used for building classifiers for downstream tasks, in this case, for the prediction of protein phase behavior, without requiring prior insight into the features that govern the process.

**Comparison of Explicitly Engineered Features and Learned Embeddings.** The use of two distinct featurization approaches—one that used only knowledge-based features and another one that relied on hypothesis-free embedding vectors—provided us with the opportunity to investigate whether, in addition to being able to predict LLPS propensity, language models can also learn the underlying features of protein condensate formation. The prediction profiles of the different models shown in Fig. 5 suggest that the difference between the two featurization approaches was most pronounced when only $\mathrm{LLPS}^{+}$ and $\mathrm{PDB}^{*}$ datasets were used for training (models EF-1 and LM-1). We thus focused the exploratory analysis on these two models involving the same training data but a different featurization strategy.

First, we noticed that when the predictions of the models were binned by intrinsic disorder, the predictions correlated with the degree of disorder for both models (Fig. 6 *A* and *B*; data across the full human proteome). While this correlation was not unexpected for model EF-1 for which disorder was an explicit input feature, the presence of such a correlation in the case of LM-1 suggested that not only can language models predict LLPS propensity, they also can capture information about the biophysical features underpinning this process. Second, we hypothesized that a key difference between models EF-1 and LM-1—whether or not disorder was used as an explicit input feature—may equip LM-1 with an enhanced capability to discriminate between disordered sequences of varying LLPS propensity. We tested this hypothesis by examining the predictions of the two models on highly disordered sequences (IDR fraction above 0.5) from the low LLPS-propensity dataset, $\mathrm{LLPS}^{-}$. As the models were trained on the $\mathrm{LLPS}^{+}$ and $\mathrm{PDB}^{*}$ datasets, none of the sequences from $\mathrm{LLPS}^{-}$ were part of the training data for these two models. Our analysis (Fig. 6*C*) revealed that while both models could distinguish the low LLPS-propensity datasets from structured proteins, LM-1 was also able to discriminate between two classes of disordered proteins of varying LLPS propensity (orange and red lines) in contrast to EF-1 for which the predictions for these two classes of disordered proteins with varying LLPS propensity were nearly identical (green and blue lines).

**DeePhase Model.** Finally, with models EF-multi and LM-multi using different input features but still demonstrating comparably good performance, we created our final model, termed DeePhase, where the prediction on every sequence was set to be the average prediction made by the two models. As expected, the models could effectively distinguish between the LLPS-prone and the structured proteins in the external test dataset (Fig. 7*A*; cyan and pink regions; AUROC of 0.99). When identifying LLPS-prone sequences from the human proteome (cyan and gray regions) as outlined earlier, AUROC of 0.84 was reached, which was comparable to or slightly exceeded what models EF-multi (0.83) and LM-multi (0.81) achieved on their own.

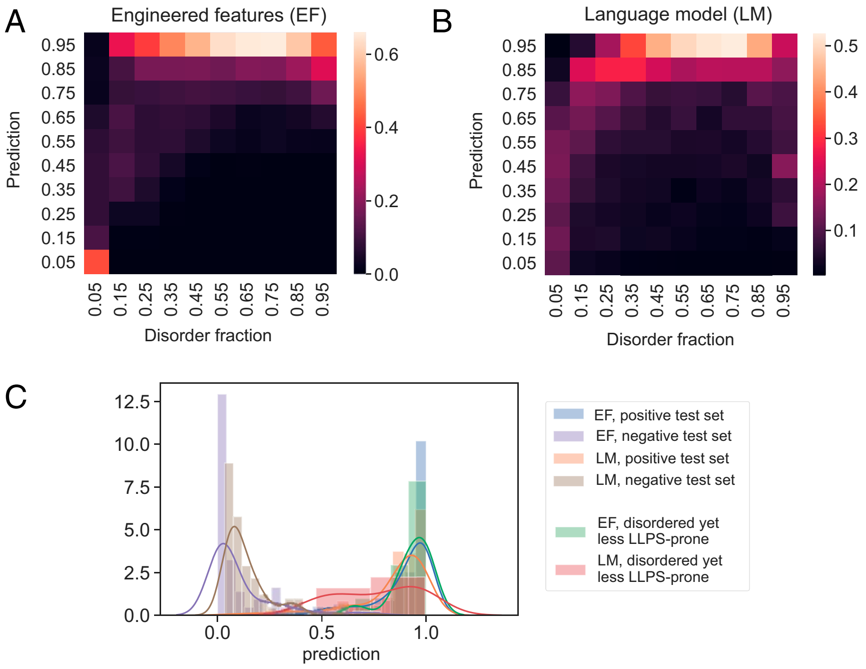

**Fig. 6.** Comparison of the models constructed using EFs and embedding vectors extracted from a LM. (*A* and *B*) The predicted LLPS-propensity score correlated with the disorder content when both EF- and LM-based embeddings were used. This trend indicated the language model was able to learn a key underlying feature associated with a high LLPS propensity. (*C*) The prediction profiles of models EF-1 and LM-1 on LLPS-prone sequences (external positive test set), structured sequences (external negative test set), and disordered sequences with a low LLPS propensity (sequences with IDR fraction above 0.5 that were part of the $\mathrm{LLPS}^{-}$ dataset and underwent homotypic phase separation). As these models were trained on the $\mathrm{LLPS}^{+}$ and $\mathrm{PDB}^{*}$ datasets, they did not use $\mathrm{LLPS}^{-}$ as training data. Model LM-1 stood out for its capability to distinguish between disordered proteins with different phase-separation propensities (orange and red lines) in contrast to EF-1 that made nearly identical predictions on these two classes of proteins (blue and green lines).

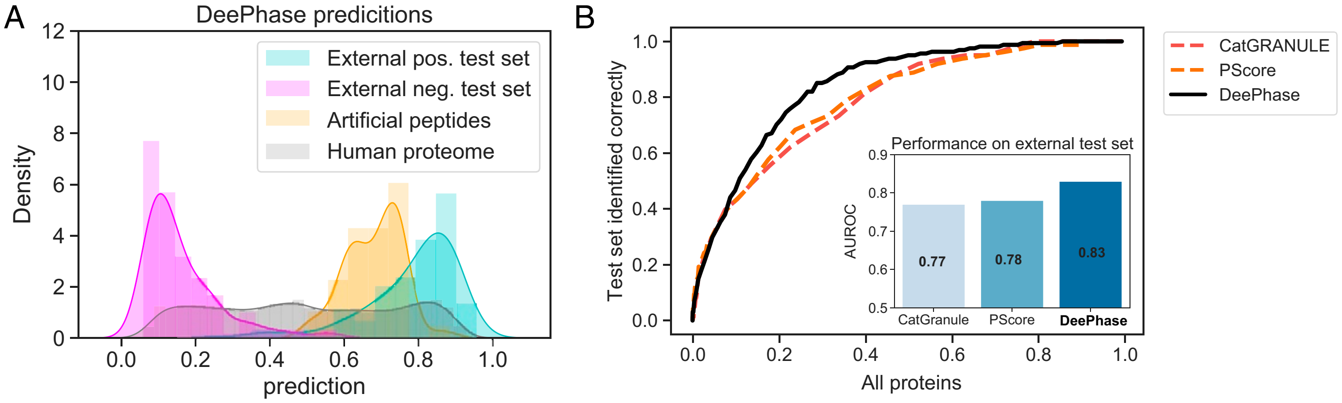

**Fig. 7.** Generalizability of DeePhase to evolutionarily nonrelated sequences and comparison to previously developed LLPS predictors. (*A*) DeePhase prediction profile on the human proteome (gray), on the external test data (161 LLPS-prone [cyan] and 161 structured proteins [pink]), and on a set of 73 artificial peptides that have been experimentally validated to be phase separate (38) (yellow). DeePhase allocated a high LLPS-propensity score for the latter dataset, indicating that its capability to evaluate phase behavior extends to evolutionary nonrelated sequences. (*B*) Comparison of DeePhase to CatGRANULE and PScore, the algorithms that were recently found to be the best performing for LLPS prediction (23), when identifying LLPS-prone sequences from the human proteome. For a reliable comparison, sequences were filtered for a length of 140 residues or above as this is the lowest threshold at which the PScore can be evaluated.

To further investigate the generalizability of DeePhase, we analyzed its performance after reducing the external dataset only to sequences that showed low similarity with the training data. Specifically, by clustering the external test and training data together [CD-hit algorithm (28), the lowest threshold of 0.4] and retaining only these test sequences that did not cocluster with any of the constructs in the training set, the external dataset was reduced from 161 sequences down to 109. With this reduction, AUROC w.r.t. the proteome dropped from 0.84 to 0.83, illustrating that the performance of DeePhase generalizes with regard to the sequences that do not share high sequence similarity with the training set. To test the limits of the DeePhase model further still, we also evaluated the LLPS-propensity score of a set of 73 artificial proteins that had experimentally been observed to phase separate in an earlier study (Dataset S8) (38). These constructs were not evolutionarily related to the sequences in our training set, yet DeePhase allocated a high LLPS propensity to them all (Fig. 7*A*, yellow).

To conclude, we compared the performance of DeePhase to two previously developed algorithms, PScore and CatGranule that had recently been identified as the best-performing algorithms for evaluating LLPS propensity of proteins in a comparative study (23). As the use of the PScore algorithm is limited to sequences that are longer than 140 residues we removed sequences shorter than this threshold value, which reduced the size of the proteome down to 18,473 sequences. On this dataset, the AUROC of DeePhase w.r.t. the proteome was 0.83, exceeding by over 10% what was achieved by the CatGRANULE and PScore models (Fig. 7*B*). We note that the comparison was constructed in a manner where it was ensured it would not favor the DeePhase model as any LLPS-prone sequences that DeePhase encountered during the training process had been excluded.

## Conclusion

To understand how protein sequence governs its phase behavior and build an algorithm for predicting LLPS-prone sequences, we constructed datasets of proteins of varying LLPS propensity. The analysis of the curated datasets highlighted that LLPS-prone sequences were less hydrophobic and had a higher degree of disorder and a lower Shannon entropy than an average protein in the Swiss-Prot database. Furthermore, our analysis of the amino acid compositions indicated that while LLPS-prone sequences were enriched in polar residues, the lowest saturation concentrations were reached when their abundance was balanced by hydrophobic residues. Relying on the generated datasets, we used the identified features as well as hypothesis-free embedding vectors generated by a language model to construct machine-learning classifiers for predicting protein phase behavior. We observed that the model built on unsupervised embedding vectors was able to predict LLPS propensity at a comparable accuracy to a model that relied on knowledge-based features, demonstrating the capability of language models to learn the molecular grammar of phase protein phase behavior. DeePhase, our final model that combined engineered features with unsupervised embeddings, showed a high performance both when distinguishing LLPS-prone proteins from structured ones and when identifying them within the human proteome, establishing a framework rooted in molecular principles for predicting of protein phase behavior.

## Materials and Methods

**Construction of the LLPS$^+$ and LLPS$^-$ Datasets.** The LLPS$^+$ and LLPS$^-$ datasets were constructed using the previously published LLPSDB database (accessed on 20 May 2020) (27). Specifically, the "LLPS_Natural_protein" repository from "Datasets classified by protein name" was used, which documented a total of 2,143 entries of proteins and their constructs examined for the occurrence of LLPS under various experimental conditions. The 2,143 entries were filtered down to systems that included only a single naturally occurring protein with no posttranslational modifications or repeat or single-site mutations and to experiments where the examined protein sequence was longer than 50 amino acids. This procedure resulted in a dataset with 769 experimental entries including a total of 231 unique constructs from 120 different UniProt IDs.

For each of the constructs, the experiments where the construct had been observed to phase separate were combined and the average concentration at which these positive experiments were performed was evaluated. When the latter concentration was below 100 μM, the construct was regarded as LLPS prone and it was included in the LLPS$^+$ dataset (Dataset S1). This process resulted in a total of 137 of such constructs from 77 unique UniProt IDs. The constructs that had been observed to phase separate at a concentration higher than 100 μM were combined with the constructs that had not been observed to phase separate to generate the LLPS$^-$ dataset. The latter dataset included a total of 94 entries from 61 unique UniProt IDs. The clustering was performed using CD-hit (28) using its lowest similarity threshold, 0.4.

**Construction of the PDB* Dataset.** Entries in the PDB (29) were used to generate a diverse set of proteins highly unlikely to undergo LLPS. Specifically, first, amino acid chains that were fully structured (i.e., did not include any disordered residues) were extracted, which resulted in a total of 112,572 chains. PDB chains were matched to their corresponding UniProt IDs using Structure Integration with Function, Taxonomy and Sequence service by the European Bioinformatics Institute, and entries where sequence length did not match were discarded. Duplicate entries were removed and the remaining 13,325 chains were clustered for their sequence identity using a conservative cutoff of 30%. One sequence from each cluster was selected, resulting in the final dataset of 1,563 sequences.

**Estimation of Physical Features from the Sequences.** A range of explicit physicochemical features was extracted for all of the sequences in the four datasets from their amino acid sequences (LLPS$^+$, LLPS$^-$, PDB$^*$, and Swiss-Prot). Specifically, the molecular weight of each sequence and its amino acid composition were calculated using the Python package BioPython. The hydrophobicity of each sequence was evaluated by summing the individual hydrophobicity values of the amino acids in the sequences using the Kyte and Doolittle hydropathy scale (31). The Shannon entropy of each sequence was estimated from the formula

$$H(X) = -\sum_{i=1}^{N=20} p_i \log_2 p_i, \qquad [1]$$

where $P$ corresponds to the frequency of each of the naturally occurring 20 amino acids in the sequence. The LCRs for each of the sequences were estimated using the SEG algorithm (32) with standard parameters. The disordered region was predicted with IUPred2a (33) that estimated the probability of disorder for each of the individual amino acid residues in the sequence. The disorder fraction of a sequence was calculated as the fraction of residues in the total sequence that were considered disordered where a specific residue was classified as disordered when the disorder probability stayed above 0.5 for at least 20 consecutive residues.

Finally, the amino acid sequence and the LCRs were described for their amino acid content by allocating the residues to the following groups: amino acids with polar residues (serine, glutamine, asparagine, glycine, cysteine, threonine, proline), with hydrophobic residues (alanine, isoleucine, leucine, methionine, phenylalanine, valine), with aromatic residues (tryptophan, tyrosine, phenylalanine), with cationic residues (lysine, arginine, histidine), and with anionic residues (aspartic acid, glutamic acid).

**Protein Sequence Embeddings.** Protein sequence embeddings were evaluated using a pretrained word2vec model. Specifically, the pretraining was performed on the full Swiss-Prot database (accessed on 26 Jun 2020) using 3-grams as words and a context window size of 25—parameters that have been previously shown to work effectively when predicting protein properties via transfer learning (34). The skip-gram pretraining procedure with negative sampling was used and implemented using the Python gensim library (39) with its default settings. This pretraining process created 200-dimensional embedding vectors for every 3-gram. To evaluate the embedding vectors for the protein, each protein sequence was broken into 3-grams using all three possible reading frames and the final 200-dimensional protein embeddings were obtained by summing all of the constituent 3-gram embeddings.

**Machine-Learning Classifier Training and Performance Estimation.** All classifiers were built using the Python scikit-learn package (40) with default parameters. No hyperparameter tuning was performed for any of the models—while such a tuning step may have given an improvement in accuracy, it can also lead to an overfitted model that does not generalize well to unseen data. For performing the training, the dataset was split into a train and a validation test in a 1 : 4 ratio and 25-fold cross-validation was used to estimate the performance of the model. For each fold, the split was performed in a stratified manner, such that entries from the same UniProt ID were always grouped together either to the training set or to the validation set. A random seed of 42 was used throughout. Multiclass classifiers were trained using the multiclass classification module of the Python scikit-learn package (40). The final prediction score was calculated as the sum of the probability of the sequence belonging to the LLPS$^+$ and half of the probability of it belonging to the LLPS$^-$ dataset.

**Data Availability.** The data and code are available from GitHub, https://github.com/kadiliissaar/deephase. All the data is also included in Datasets S1–S8

**ACKNOWLEDGMENTS.** The research leading to these results has received funding from the Schmidt Science Fellows program in partnership with the Rhodes Trust (K.L.S.), a St John's College Junior Research Fellowship (to K.L.S.), a Trinity College Krishnan-Ang Studentship (to R.Q.) and the Honorary Trinity-Henry Barlow Scholarship (to R.Q.), the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in NanoScience and Nanotechnology (NanoDTC) (EP/L015978/1 to W.E.A.), the EPSRC Impact Acceleration Program (W.E.A., G.K., T.P.J.K.), the European Research Council under the European Union's Horizon 2020 Framework Program through the Marie Sklodowska-Curie Grant MicroSPARK (Agreement 841466 to G.K.), the Herchel Smith Fund of the University of Cambridge (to G.K.), a Wolfson College Junior Research Fellowship (to G.K.), the European Research Council under the European Union's Seventh Framework Program (FP7/2007–2013) through the European Research Council Grant PhysProt (Agreement 337969 to T.P.J.K.), and the Newman Foundation (T.P.J.K.).

1. C. P. Brangwynne *et al.*, Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
2. A. A. Hyman, C. P. Brangwynne, Beyond stereospecificity: Liquids and mesoscale organization of cytoplasm. *Dev. Cell* **21**, 14–16 (2011).
3. A. A. Hyman, C. A. Weber, F. Jülicher, Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014).
4. S. Boeynaems *et al.*, Protein phase separation: A new phase in cell biology. *Trends Cell Biol.* **28**, 420–435 (2018).
5. S. Alberti, A. Gladfelter, T. Mittag, Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
6. A. Aguzzi, M. Altmeyer, Phase separation: Linking cellular compartmentalization to disease. *Trends Cell Biol.* **26**, 547–558 (2016).
7. P. Li *et al.*, Phase transitions in the assembly of multivalent signaling proteins. *Nature* **483**, 336–340 (2012).
8. E. Sokolova *et al.*, Enhanced transcription rates in membrane-free protocells formed by coacervation of cell lysate. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11692–11697 (2013).
9. J. Berry, S. C. Weber, N. Vaidya, M. Haataja, C. P. Brangwynne, RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5237–E5245 (2015).
10. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
11. J. A. Riback *et al.*, Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* **168**, 1028–1040 (2017).
12. C. P. Brangwynne, P. Tompa, R. V. Pappu, Polymer physics of intracellular phase transitions. *Nat. Phys.* **11**, 899–904 (2015).
13. E. Gomes, J. Shorter, The molecular language of membraneless organelles. *J. Biol. Chem.* **294**, 7115–7127 (2019).
14. A. S. Raut, D. S. Kalonia, Effect of excipients on liquid–liquid phase separation and aggregation in dual variable domain immunoglobulin protein solutions. *Mol. Pharm.* **13**, 774–783 (2016).
15. K. Julius *et al.*, Impact of macromolecular crowding and compression on protein–protein interactions and liquid–liquid phase separation phenomena. *Macromolecules* **52**, 1772–1784 (2019).
16. L. Lemetti *et al.*, Molecular crowding facilitates assembly of spidroin-like proteins through phase separation. *Eur. Polym. J.* **112**, 539–546 (2019).
17. G. Krainer *et al.*, Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nat. Commun.* **12**, 1085 (2021).
18. E. W. Martin, T. Mittag, Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry* **57**, 2478–2487 (2018).
19. R. M. Vernon *et al.*, Pi-pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, e31486 (2018).
20. G. L. Dignon, R. B. Best, J. Mittal, Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
21. E. W. Martin *et al.*, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).
22. J. Wang *et al.*, A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699 (2018).
23. R. M. Vernon, J. D. Forman-Kay, First-generation predictors of biological protein phase separation. *Curr. Opin. Struct. Biol.* **58**, 88–96 (2019).
24. H. R. Li, W. C. Chiang, P. C. Chou, W. J. Wang, J. R. Huang, TAR DNA-binding protein 43 (TDP-43) liquid–liquid phase separation is mediated by just a few aromatic residues. *J. Biol. Chem.* **293**, 6090–6098 (2018).
25. S. Elbaum-Garfinkle *et al.*, The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7189–7194 (2015).
26. T. P. Dao *et al.*, Ubiquitin modulates liquid-liquid phase separation of UBQLN2 via disruption of multivalent interactions. *Mol. Cell* **69**, 965–978 (2018).
27. Q. Li *et al.*, LLPSDB: A database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic Acids Res.* **48**, D320–D327 (2020).
28. W. Li, A. Godzik, CD-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
29. H. M. Berman *et al.*, The Protein Data Bank (accessed on 26 June 2020). *Nucleic Acids Res.* **28**, 235–242 (2000).

30. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
31. J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
32. J. C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163 (1993).
33. 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. Mol. Biol.* **347**, 827–839 (2005).
34. E. Asgari, M. R. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS One* **10**, e0141287 (2015).
35. D. Ulyanov, Multicore-TSNE (2016). https://github.com/DmitryUlyanov/Multicore-TSNE. Accessed 15 January 2021.
36. K. You *et al.*, PhaSepDB: A database of liquid–liquid phase separation related proteins. *Nucleic Acids Res.* **48**, D354–D359 (2020).
37. H. X. Zhou, V. Nguemaha, K. Mazarakos, S. Qin, Why do disordered and structured proteins behave differently in phase separation? *Trends Biochem. Sci.* **43**, 499–516 (2018).
38. M. Dzuricky, B. A. Rogers, A. Shahid, P. S. Cremer, A. Chilkoti, De novo engineering of intracellular condensates using artificial disordered proteins. *Nat. Chem.* **12**, 814–825 (2020).
39. R. Řehůřek, P. Sojka, "Software framework for topic modeling with large corpora" in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, 2010), pp. 45–50.
40. F. Pedregosa *et al.*, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).