RESEARCH ARTICLE

Pervasive, conserved secondary structure in highly charged protein regions

Catherine G. Triandafillou[1⊙], **Rosalind Wenshan Pan**[2⊙], **Aaron R. Dinner**[3], **D. Allan Drummond**[4,5]*

1 Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, United States of America, **3** Department of Chemistry, University of Chicago, Chicago, Illinois, United States of America, **4** Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois, United States of America, **5** Department of Medicine, Section of Genetic Medicine, The University of Chicago, Chicago, Illinois, United States of America

⊙ These authors contributed equally to this work.
* dadrummond@uchicago.edu

Abstract

Understanding how protein sequences confer function remains a defining challenge in molecular biology. Two approaches have yielded enormous insight yet are often pursued separately: structure-based, where sequence-encoded structures mediate function, and disorder-based, where sequences dictate physicochemical and dynamical properties which determine function in the absence of stable structure. Here we study highly charged protein regions (>40% charged residues), which are routinely presumed to be disordered. Using recent advances in structure prediction and experimental structures, we show that roughly 40% of these regions form well-structured helices. Features often used to predict disorder—high charge density, low hydrophobicity, low sequence complexity, and evolutionarily varying length—are also compatible with solvated, variable-length helices. We show that a simple composition classifier predicts the existence of structure far better than well-established heuristics based on charge and hydropathy. We show that helical structure is more prevalent than previously appreciated in highly charged regions of diverse proteomes and characterize the conservation of highly charged regions. Our results underscore the importance of integrating, rather than choosing between, structure- and disorder-based approaches.

Author summary

Understanding how the functions of protein molecules are encoded in their amino-acid sequence remains one of the defining challenges in molecular biology. Considerable insight has been gained by understanding how sequence encodes three-dimensional structure—particularly with the recent emergence of high-quality structure prediction. Yet it has also become clear that many protein behaviors arise from disorder, without stable structure, and tools to predict disorder from protein sequences have similarly proliferated. We have found that a large and important class of protein sequences which is commonly presumed and predicted to be disordered—highly charged regions—surprisingly form

well-structured helices in roughly 40% of cases, far higher than predicted by both classic
and cutting-edge disorder predictors. By studying the evolution of these sequences, we
show how many of the intuitive methods for identifying disorder are in fact compatible
with structured regions. By illuminating this blind spot in detail, and showing that a sim-
ple method can reliably identify these structured regions, we highlight the surprising chal-
lenges that still remain in developing both conceptual intuition and detailed predictive
capability when both structure and disorder play important biological roles.

Introduction

In the overarching quest to understand how genotype shapes phenotype, the question of how
protein sequence encodes protein function has proved a rich and enduring challenge. An early
and still pervasive conceptual framework in which stable sequence-encoded protein structures
confer biological function has been met by a newer (yet by now firmly established) companion
approach born from the recognition that many functions—binding, selective recruitment, for-
mation of large-scale structures, and more—can be achieved by sequences which do not adopt
a stable conformation (intrinsically disordered regions, or IDRs). Although neither approach
is exclusive of the other, and indeed they anchor a continuum [1], for historical and methodo-
logical reasons many analyses adopt one approach or the other based on various heuristics [2–
7]. Such heuristics have had an outsized impact on how sequence-function maps are explored.

In one early and influential study of IDRs, Uversky and colleagues discovered that plotting
mean net charge against mean hydropathy (hydrophobicity) permits a dividing line to be
drawn separating folded from disordered proteins [2,8]. In these analyses, highly charged,
weakly hydrophobic sequences have a strong tendency to be disordered. More recently devel-
oped heuristics go beyond composition: simulation studies suggest that the degree of mixing
of opposite charges within a highly charged, nearly net-neutral (polyampholyte) sequence is a
predictor for the biophysical properties of such polypeptides, specifically whether they form
expanded or compact structures in solution [3,9]. These studies assume that the sequences in
question do not take on well-defined structures, largely based on the observation that many
disordered proteins are polyampholytes (~75% of known IDRs have a fraction of charged resi-
dues (FCR) > 0.35 [10]). These findings, although based on a few hand-picked sequences from
different organisms and proteins, have broadly informed the analysis of many other sequences
[11–13].

Other analyses, while still converging on the general finding that disorder is associated
most strongly with a bias toward charged residues and away from hydrophobic residues, have
emphasized extreme compositional biases themselves as strong predictors of disorder [4,5].
The most common quantitative description of sequence compositional bias is the Shannon
entropy, often referred to as complexity [6,14]. Complexity here has a statistical, not biological,
interpretation; "simple" sequences such as homopolymers or sequences composed of only a
few types of amino acids have low sequence entropy and thus low complexity.

Low-complexity regions (LCRs) have in the past decade experienced a surge of attention,
driven by the observation that they are associated with mesoscale organization in cells: clusters,
granules, hydrogels [15–17], membraneless organelles, and a host of related structures now
referred to as biomolecular condensates [18]. Charged LCRs in particular play a crucial role in
biomolecular condensation in highly influential model systems, mediating complex coacerva-
tion [7] and phase separation [19–22]. Particular sequence features such as enrichment with
positively charged residues like arginine and with conformationally flexible glycine, most

memorably in the RGG motif [23], appear often in RNA binding proteins that are known to condense; interactions with cationic residues can be modulated by negatively charged regions, leading to the proposal of a molecular grammar for such interactions [24]. While further work is needed to understand when and how these interactions drive condensation, it is also important to develop a better understanding of their biophysical properties, and especially how the regions behave in isolation.

Together, these lines of inquiry both reflect and create conditions in which highly charged, low-hydrophobicity LCRs may be studied nearly exclusively through the lens of disorder [3,10,25]. Because of the historical roots of the disorder presumption—particularly that many of the paradigm-shaping observations were made as databases of sequences and structures were in their infancy—the presumption itself has persisted with few challenges.

A confluence of trends and events has laid the groundwork for a productive reexamination of these assumptions. First, the maturation of structural and sequence databases has prompted increasingly critical looks at our understanding of LCRs [26] and IDRs [27]. In parallel, specific examples have accumulated of well-defined structure in sequences which would, by existing heuristics, be overwhelmingly predicted to be disordered: alpha-helices in myosin [28] and caldesmon [29], and a coiled-coil region in the mRNA export protein GLE1 [30]. Indeed, there is a long-established connection between charge patterning and helix formation [10–13,31,32]. Finally, new methods now permit more reliable and farther-reaching assessment of the disorder presumption for highly charged regions, notably high-quality structure prediction [33,34].

So motivated, we return to the root issues: to what extent are naturally occurring highly charged protein regions structured versus disordered? What is the empirical relationship between the fraction and patterning of charged residues and the biophysical properties of a region? Are highly charged regions conserved over evolutionary time, as we would expect for biologically important properties? And how easily can one distinguish between structured and disordered regions on the basis of simple sequence heuristics?

To answer these questions, we systematically identify highly charged regions proteome-wide in related eukaryotes (budding yeasts) and characterize their sequence properties, predicted structure, and evolution. In contrast to previous studies [3,9,10,35], our work examines the entire proteome, permitting us to quantify the frequency and, with proteome-scale homology, evolutionary conservation of charged regions at the genomic scale. We find that naturally occurring polyampholytes are highly prevalent and, despite being low-complexity with often-poor sequence conservation, these regions are often predicted to form or contain alpha-helices. We confirm these predictions through comparison with experimentally derived structures. These results demonstrate that certain LCRs, even those enriched for charged amino acids thought to be important for intermolecular interactions driving phase separation, may adopt a well-defined structure in the right sequence or physicochemical context. More broadly, we show that it is important to consider structural properties explicitly when evaluating the other properties of an LCR.

Results

Highly charged regions of low sequence complexity are prevalent in the yeast proteome

We first established criteria for regions to be "highly charged" and used them to identify regions in the *S. cerevisiae* proteome, our departure point owing to its extensive experimental and evolutionary characterization. Examining amino acid usage (Fig 1A), we found that the charged residues (glutamate, aspartate, lysine, and arginine) together constitute 23% of the total amino acids. Unlike all other categories of amino acids, the frequency of each charged
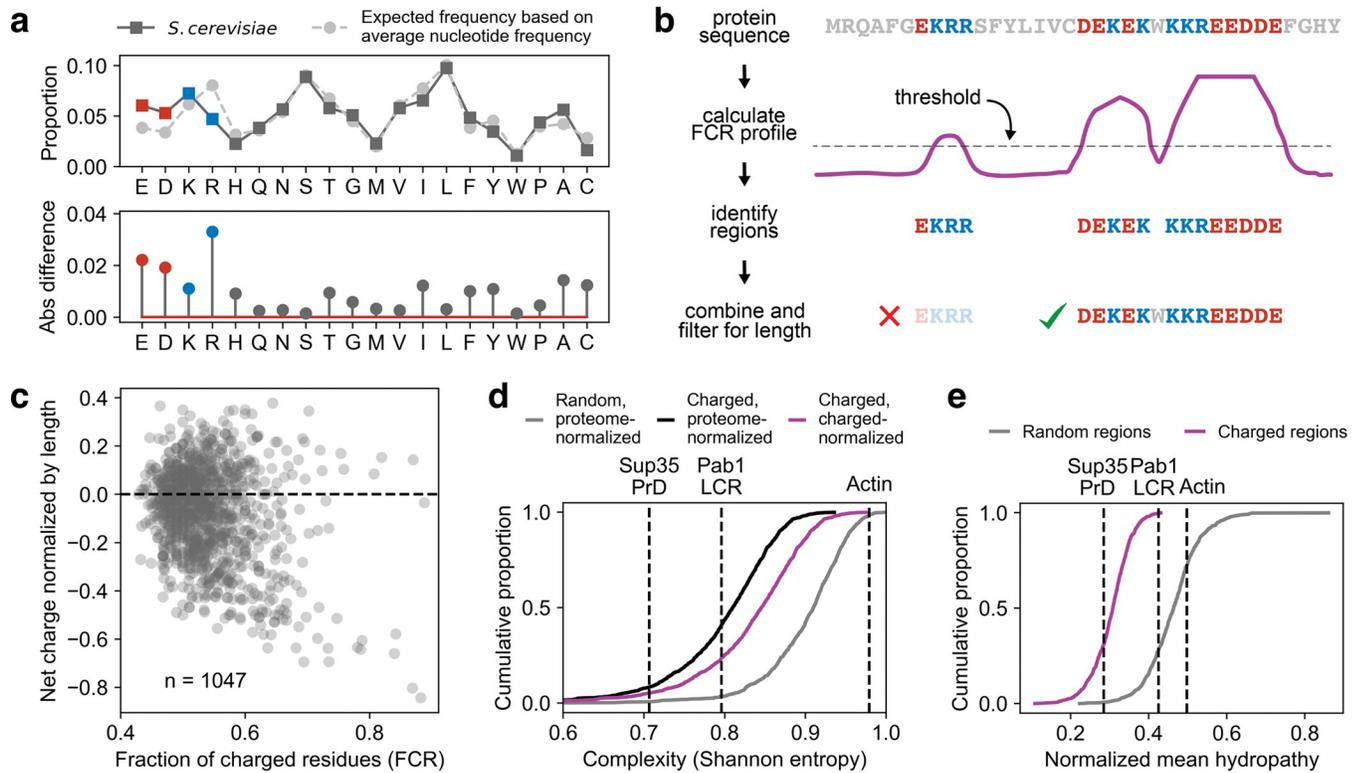
Fig 1. Regions high in charge are prevalent in the yeast proteome and are low complexity. a Top: the average frequency of amino acids in the yeast proteome. Light gray trace is the expected frequency based on the average nucleotide frequency and the codons which code for each amino acid. Bottom: The difference between the dark and light traces in the top panel. **b** Cartoon of the algorithm which detects highly charged regions. **c** Summary of the per-region fraction of charged residues (FCR) and normalized net charge for identified regions. **d** Complexity of the charged region calculated according to the compositional entropy defined in [14]. Gray and black distributions are normalized to the yeast proteome entropy, the purple distribution is normalized to the entropy of a charged-enriched hypothetical proteome. Reference sequences are known low complexity domains from Sup35 (a typical prion-like low-complexity protein) and Pab1 (atypical hydrophobic-rich intrinsically disordered region), and a folded sequence (actin). **e** Normalized hydropathy of the charged regions (purple) and length-matched randomly drawn regions (gray). Reference sequences are the same as those in **d**.

amino acid (as determined from the frequency of observed codons) deviates strongly from expectation based on the underlying nucleotide frequency (Fig 1A, light gray points), evidence for evolutionary selection.

To isolate highly charged regions, we took a sliding-window approach, moving a 12-amino-acid window across all protein sequences and selecting regions with a fraction of charged residues (FCR) above 0.4, with some tolerance for transient deviations (see Fig 1B and Methods). After trimming uncharged ends off these segments, the resulting highly charged regions have a median length of 50 and a FCR $\geq$ 0.43 (Figs 1C and S1C), more stringent than, for example, a published definition of a strong polyampholyte (FCR > 0.30) [36]. We identified 1,047 regions in 800 proteins; about 14% of protein-coding genes encode at least one highly charged region. The FCR in these regions is just over two standard deviations above that for randomly chosen regions in the proteome (S1A Fig); the regions also have substantially higher charge density than the proteins which contain them (S1B Fig).

We examined the distribution of both FCR and normalized net charge across all regions; it is more common for a region to have a net charge close to zero, although there is a significant number of net negatively charged regions (Fig 1C). Examples of neutral regions and those that carry a net charge can be found in Table 1.
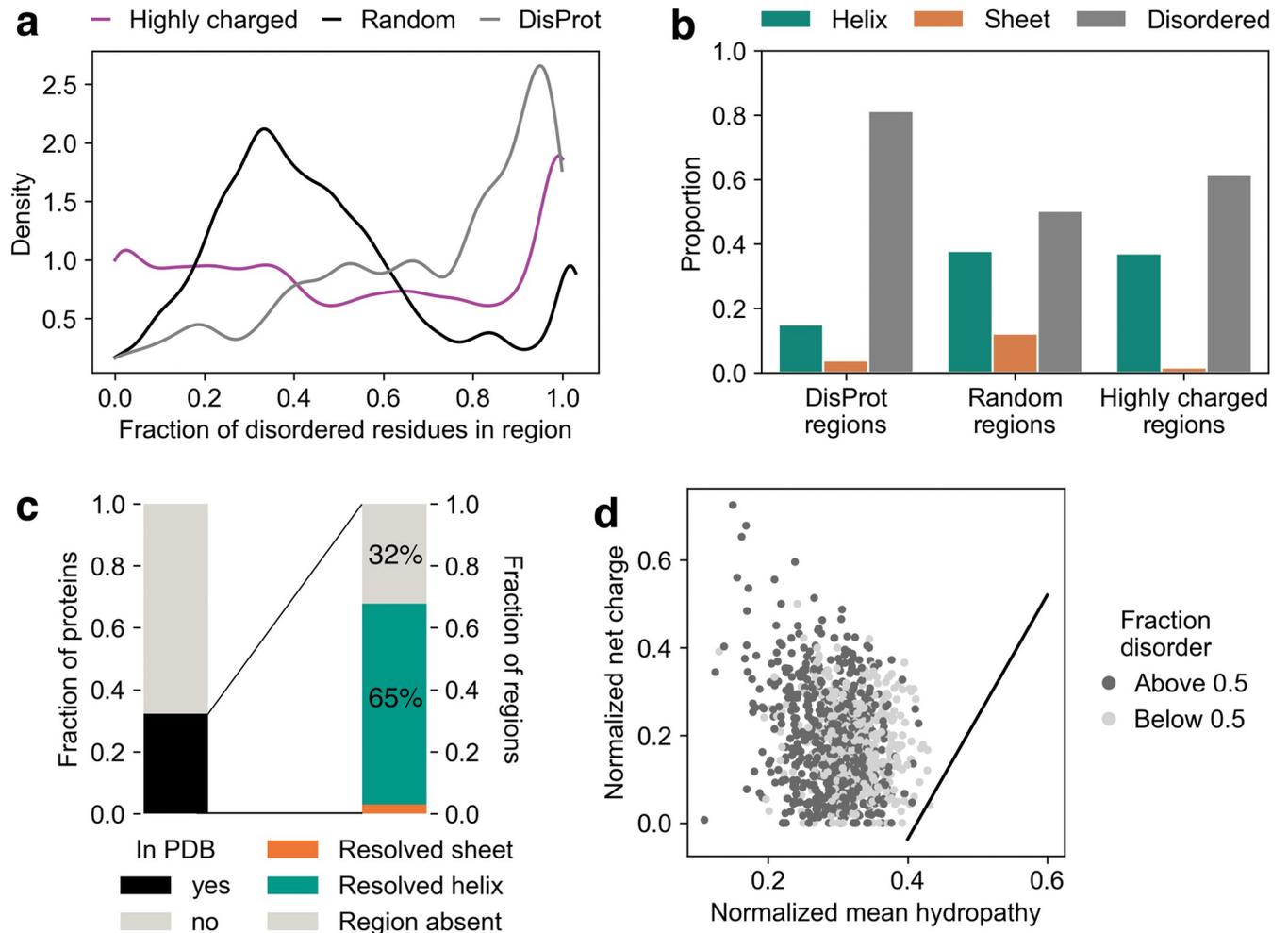
Table 1. Example highly charged regions from the yeast proteome.

Category	Gene	Kappa	Sequence
Net negative	SIS2	0.756	KNNEEEDDDEDEEEDDDEEEDTEDKNENNNDDDDDDDDDDDDDDDDDDDDDDDDDDDEDEDEAE
	RNA1	0.393	DLLQSKFDDLEVDDFEEVDSEDEEGEDEEDEDEDEKLEEIETERLEKELLEVQVDDLAERLAETEIK
	COP1	0.064	EEPLGEENFNDEDIGEDEGAWDLGDEDLDVGEELPEEVEQGE
~Neutral	RIO1	0.433	EEFSDDEEDGSSGSEEDDEEEGEYYDDDEPKVLKGKKHEDKDLKKLRKQEAKDAKREKRKTKVKKHIKKKLVKKTKSKK
	SEC3	0.106	EVNKRYELEQQQQQEEAELRRLEEQKRLQLQKENEMKRLEEERRIKQEERKRQMELEHQRQLEEEERKRQMEL
			EAKKQMELKRQRQFEEEQRLKKERELLEIQRKQREQETAERLKKEEQEALAKKEEEEKSKRNKVDNE
	MNN4	0.061	KLLEERKRREKKKKEEEEKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKQEEEEKKKKEEEEKKKQ
			EEGEKMKNEDEENKKNEDEEKKKNEEEEKKKQEEKNKKNEDEEKKKQEEEEKKKNEEEEKKKQEE
Net positive	TMA23	0.317	KGKKRRRRDEDDNKVKRKKLKKDKKTSNDSESKKKKKKKSKKESKKGKKSKHSSDEGDKSKHKKSKKSKKHKKEESSARRDRKE
	ZDS2	0.185	KKNSLEKRLAKLFKRKQHNGTCKSDVKVIKKSVKKELKKK
	FAF1	0.071	KVKKGQFRKIESTYKKDIERRIGGSIKARDKEKATKRERGLK

We expected that the highly charged regions would have lower complexity and hydrophobicity than random regions, because they must be enriched for a small subset of (charged) amino acids, and we confirmed that this is the case (Fig 1D and 1E gray and black traces). Because the highly charged regions could be low complexity merely due to the selection bias imposed by looking for enrichment for a few amino acids, we calculated the complexity normalized by the entropy of a synthetic, mostly charged proteome (50% charged amino acids and all other amino acids equally represented, see Methods for details, Fig 1D purple trace). Even with this correction, the regions are far less complex than randomly drawn regions ($P < 10^{-6}$, Wilcoxon rank sum test), suggesting a further bias in amino acid usage in these regions beyond that explained by their enrichment for charge.

We also examined the distribution of the proteins containing these regions within the cell. We found that they were enriched in the nucleus, and especially in the nucleolus (S1D Fig), consistent with recent findings that across several species nucleolar proteins are enriched for charge-rich low-complexity sequences [37].

In summary, we find that highly charged regions which exceed even stringent definitions of polyampholytes are common in the yeast proteome, are on average less hydrophobic and less complex than average sequences, and are enriched in specific nuclear compartments.

Secondary structure is pervasive in highly charged regions

Given the historical and intuitive associations between low hydrophobicity/high charge density and disorder, we predicted that the vast majority of the regions we identified would not adopt a well-defined structure. We thus set out to determine what proportion of the regions we identified were IDRs using experimentally derived structures and recently available proteome-wide structure prediction (AlphaFold) [33]. Although the biophysical properties of disordered regions cannot be accurately determined using AlphaFold structures [38], disorder can be inferred in two ways. The first is to impute disorder to residues with a low AlphaFold confidence score [38]; the second is to ascribe disorder to regions with high-confidence coil (e.g., not helix or sheet) predictions. We employ both methods. To validate of these choices, we analyzed the predicted AlphaFold structure of protein regions from the DisProt database (which contains proteins that have been empirically measured to be disordered through a variety of experimental means including circular dichroism and NMR) and found that the vast majority of confidently predicted residues

Fig 2. Secondary structure is pervasive in highly charged, low complexity regions. a The distribution of the fraction of a region predicted to be disordered with AlphaFold or present in DisProt. See Methods for secondary structure scoring. **b** The number of predicted helical, sheet, and disordered (which includes coil; see Methods) residues for the highly charged regions, randomly-selected regions, and experimentally verified disordered regions from DisProt. **c** (Left) The proportion of the top 200 most structured regions for which the protein that contains them is in the PDB. (Right) For those that are in the PDB, the proportion of regions that is a helix, sheet, or missing. **d** Uversky plot of the highly charged regions. The dividing line is from [2].

https://doi.org/10.1371/journal.pcbi.1011565.g002

in these regions were scored as "coil" by DSSP (S2A Fig). Thus, by using AlphaFold we were able to assess both structure and disorder—with the same method—proteome-wide.

Returning to the highly charged regions, we used the AlphaFold predictions to classify each residue as either disordered (low confidence, or high-confidence and scored as coil) or ordered (high-confidence and scored as helix or sheet; see Methods for a complete description of scoring cutoffs). While a significant number of the highly charged regions were almost completely composed of residues classified as disordered, in many regions (40% of the total) more than half of the residues were predicted to be structured (Fig 2A). We examined the secondary structure classification for all confidently predicted residues (45% of the total) across the entire dataset. The highly charged regions were markedly enriched for alpha-helical secondary structure compared to disordered regions from DisProt and had a similar frequency to length-matched randomly drawn regions (Figs 2B and S2A). There were seven (out of 800) cases where proteins predicted to have high helical content were also found in the DisProt database. We investigated these more deeply

and found that in only three cases was there overlap between the region we identified and the region annotated in DisProt. The full results of this investigation are found in S1 Table.

We examined the subset of proteins which have empirically determined structures as a check on AlphaFold's predictions, to guard against hallucination or other systematic error in highly charged regions. We searched the Protein Data Bank (PDB) for the 200 regions with the highest predicted fraction of structured residues (>92% predicted structured, 19% of the total regions). 27% of the proteins in this subset could be found in the PDB and, of those that were found, 42 (68%) had the highly charged region resolved. In all of these cases but two (3%), the region predicted to be a helix was experimentally determined to be a helix (Fig 2C). That is, where experimental evidence is available for these regions, odds are better than 20:1 that the helical prediction will be confirmed by experimental data.

In roughly a third of cases the region is absent from the PDB, and because disordered structures frequently evade structural resolution, it is possible that these regions are disordered under some conditions. In particular, solvent conditions (e.g., pH) and sequence context could modulate the net charge on each amino acid, altering the propensity for structure. Indeed, recent simulations suggest just such a mechanism for a highly-charged LCR in Hero11 [39], and experimental studies have demonstrated that a transient helix forms in the acidic activation domain of Gcn4 upon interaction with a binding partner [40]. Regions may also be modified post-translationally, potentially altering net charge, charge patterning, and localization. Given that these local effects are challenging to predict [41], and in principle could both promote or inhibit structure formation, we consider our estimate a reasonable lower bound on the propensity for helix formation in these regions. Lending further credence to our estimate, recent work comparing AlphaFold predictions to NMR structures has demonstrated that AlphaFold can reasonably and with high confidence predict conditionally folded structures [42]. It is also possible that the helical region is stable within a disordered element, like a pipe on a jump rope [43], or that only a portion of the protein lacking the highly charged region was expressed and characterized. From our analysis, we conclude that there is no evidence to suggest structural predictions are inaccurate for these regions, and we confirm the presence of many highly charged helices in experimental data.

The presence of substantial helical structure in these highly charged, low-hydrophobicity regions raised questions as to how these regions would be scored by the metrics which initially established connections between disorder and these sequence features. As a set, these regions contradict the argument that large amounts of uncompensated charge predicts disorder, captured in the popular charge/hydropathy or Uversky plot [2]. We created a Uversky plot of the highly charged regions, and all but three fell above the dividing line into the "natively unfolded" region (Fig 2D). Thus classic methods for determining whether a region is ordered or disordered have virtually no predictive power for regions of this composition—a surprising result given that these regions would appear perfectly suited for such a heuristic.

More recently developed metrics have been used to assign biophysical properties to highly charged regions. In particular, connections have been made between the patterning of charges (κ) and the predicted radius of gyration (R_g, a measure of compaction) [3,9]. R_g is used to characterize ensembles of disordered conformations; the implicit argument appears to be that because most IDRs are polyampholytes, analysis of polyampholyte conformations can be carried out productively without considering structured conformations. Yet in specific cases, the very polyampholyte sequences being assigned to various disordered conformational states by computational analysis due to their charge patterning [3] are known to be helical—such as the (EEEKKK)n and (EEEEKKKK)n polymers [32,41,44]. Remarkably, this is more than a mere conceptual curiosity: our set of highly charged sequences in budding yeast contains the sequence KKKKEEEEKKK KEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKQEEEEKKKKEEEEKKKQ in the

protein Mnn4, a region which is, with modifications, conserved in other fungal species (S2C Fig). Such sequences and their relatively well-studied biophysical behavior offer additional evidence of the importance of considering helical structure in biologically relevant highly charged regions. To emphasize this point, we calculated the FCR and κ values for all the regions in our dataset, and compared how the helical and disordered regions fell in that space. Although the helical regions on average had lower κ values than the disordered regions, the disordered regions spanned the entire κ space; thus it is important to establish the structural or predicted structural state of a sequence before interpreting how the κ value relates to the radius of gyration (S2B Fig).

Many of the sequences we identify show the hallmarks of so-called single alpha helices (SAH), helices of length 25–200 residues frequently, though not exclusively, formed by $(E_4K_4)_n$ repeats [9,34,35]. Below, we establish the evolutionary conservation of these regions, suggesting their structure, as well as high charge density, are likely to confer a fitness benefit.

An ancient translation initiation factor contains a conserved highly charged helix with sequence properties similar to an IDR

To more deeply investigate the sequence and structural properties of highly charged regions, we focused on a specific example where a region predicted to be a helix by AlphaFold had a solved empirical structure for comparison. We chose the broadly conserved eukaryotic translation initiation factor eIF3A (Rpg1 in *S. cerevisiae*) in which we identified several highly charged regions. One such region was predicted to be almost entirely helical, which is confirmed in the cryo-electron microscopy (cryoEM) structure (Fig 3B) [45].

When we created a sequence alignment of all the homologous proteins for which a structure had been predicted from AlphaFold, we found significant variation in both the length and the sequence of the region (Fig 3A). Such variability is typical in disordered low-complexity regions and seen as the accumulation of many insertions and deletions in a multiple sequence alignment, but we were surprised to see this variation because the yeast version of this sequence was structured. To determine whether the homologous sequences were likely to be structured as well, we used DSSP (implemented in the MDTraj package for Python [46]) to classify the secondary structure predicted by AlphaFold and projected the predictions into alignment space (Fig 3C). Despite the lack of conservation at the sequence level, the helical nature of the region was conserved across all the homologs. However, the length of the helix varied significantly: the coefficient of variation of the highly charged helix was 0.17, compared to 0.11 for a reference helix located elsewhere in the same protein (Fig 3D).

This analysis demonstrates that a region with all the features typically associated with IDRs (high length variation as indicated by gaps in the alignment, poor sequence conservation, low complexity, low hydrophobicity) can be associated with a charged region that is in fact structured and retains this structure across evolutionary time.

Helical highly charged regions can be predicted from amino acid composition

Given the prevalence of structure in the highly charged regions that we detected, and the failure of existing composition-based heuristics to discriminate between the two categories (disordered and helical), one might guess that more sophisticated and general methods might be needed. An alternative possibility is that alternative simple heuristics exist, a possibility which would be demonstrated by producing such a heuristic. We therefore asked whether we could build a simple predictor of helical structure in highly charged regions from composition alone. We stress that our objective here is to demonstrate a principle, and to gain some insight into
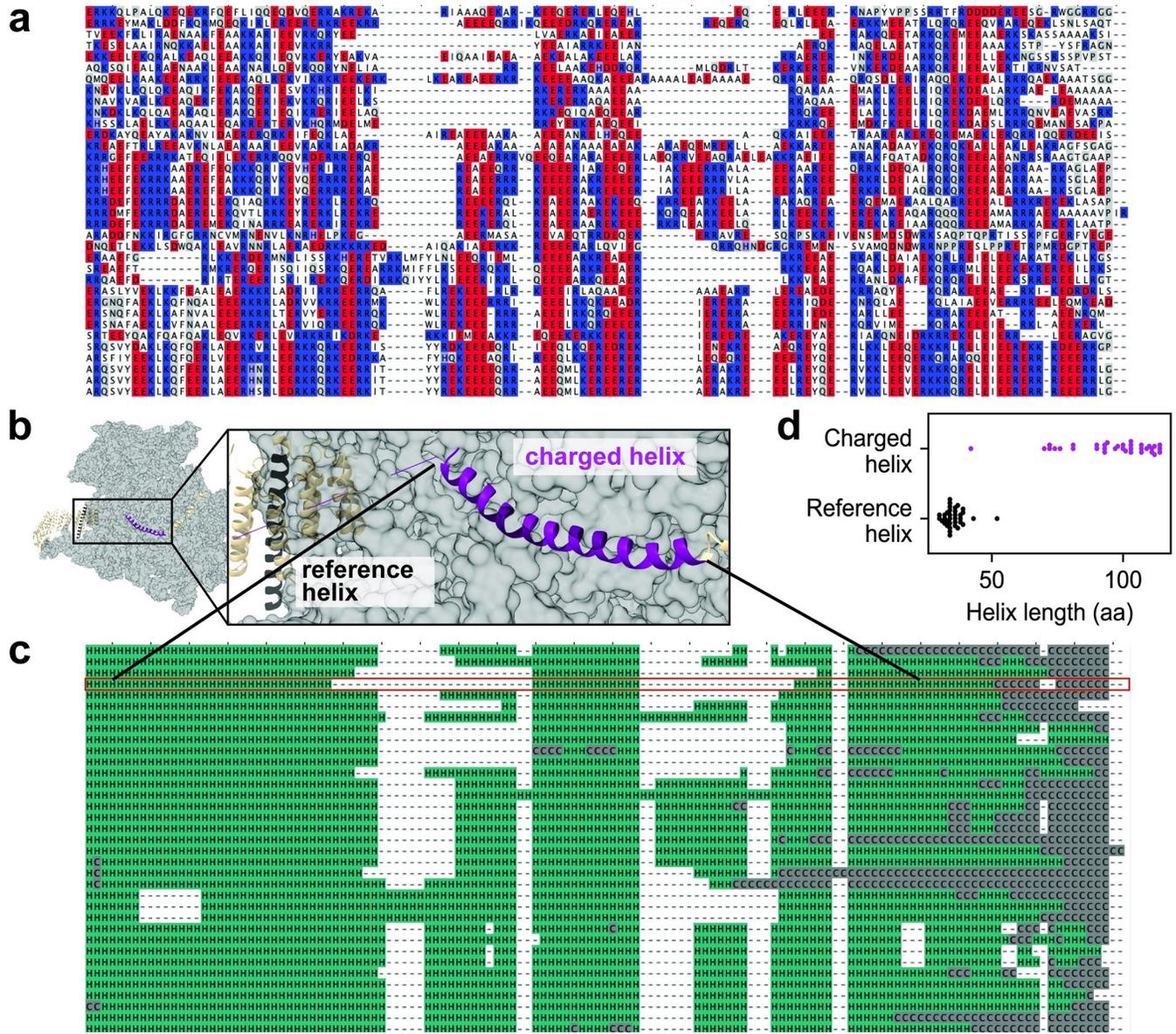
Fig 3. eIF3A, an essential eukaryotic translation initiation factor, contains a conserved, highly charged helix that varies in length but not in secondary structure. a Alignment of the eIF3A highly charged region (orthologs from all distantly-related species with predicted proteomes in AlphaFold) with negatively charged residues colored red, positively charged residues colored blue, and gaps and all other amino acids in white. Although the highly charged nature of the region is conserved, the sequence itself is variable. **b** Representative image of the cryoEM structure of yeast eIF3A [42] with the highly charged region (resolved as a helix) shown in purple, and a reference helix shown in black. **c** Alignment of eIF3A (same as in a) colored by the secondary structure predicted by AlphaFold; *S. cerevisiae* sequence is highlighted in red. Note that the highly charged region is predicted to be helical in every species represented. **d** Despite strong secondary structure conservation, the length of the highly charged helix varies significantly more than a reference helix from the same protein.

the factors that determine disorder versus structure, rather than to compete with the many examples of sophisticated software designed to broadly predict disorder [33,34,47,48].

Using the proteome-wide predictions of structure from AlphaFold, we created a dataset of regions which were predicted to be either completely disordered or completely helical (13,437 sequences from 63.6% of the proteome). On a Uversky plot, the helices and IDRs drawn from the yeast proteome, like the highly charged regions, overlap significantly (Fig 4A).
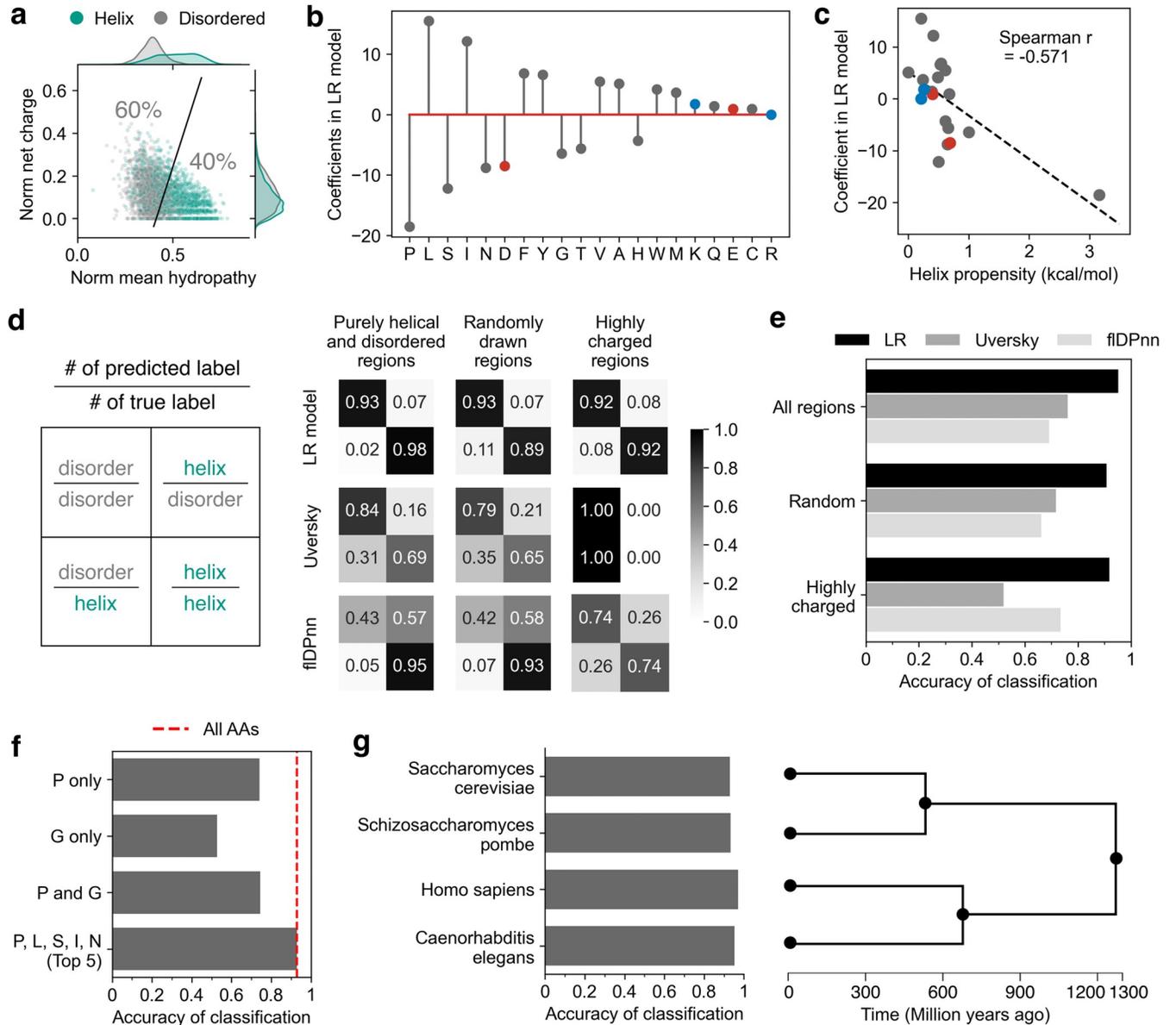
Fig 4. Helical regions can be predicted from composition. a Uversky plot of all regions used to train the LR model. The marginals of the distribution are shown on the plot border. **b** The coefficients of the logistic regression (LR) model which predicts whether a region is helical or disordered on the basis of amino acid composition. The model was trained on purely helical and disordered regions (predicted by AlphaFold) selected from the *S. cerevisiae* proteome. Amino acids with a positive coefficient are correlated with helices, those with a negative value are correlated with disordered regions. **c** The helix propensity from [46] plotted against the LR model coefficients. **d** Accuracy of both the LR model (top), the Uversky dividing line (middle, from [2]), and flDPnn [47] (bottom) on purely helical and disordered regions (held out data from the training set, n = 3360; left), randomly-drawn regions, which are predicted by AlphaFold to be majority (but not completely) helical or disordered (n = 3405; center), and the highly charged regions (n = 681; right). **e** Summarized accuracy for all categories in **d**. **f** Summarized accuracy of the LR model with only a subset of coefficients, **g** (Left) Accuracy of the LR model prediction of regions from other organisms. (Right) Timetree showing the evolutionary divergence of the organisms.

Using this dataset, we built a logistic regression model which classified regions as helical or disordered on the basis of their amino acid composition (see Methods for model details). The coefficients of this logistic regression (LR) model are shown in Fig 4B; as expected, residues known to affect helical character such as proline and glycine have a large regression coefficient, indicating that their presence is highly predictive. More generally, the coefficients from the model are inversely correlated with the individual amino acid helix propensity [49] (Fig 4C).

We assessed the accuracy of the LR model in several ways. First, we calculated the rate of true and false positives and negatives (Fig 4D) for several classes of sequences. The LR model performed extremely well, correctly identifying both helical and disordered regions in the testing data (25% held out from the original dataset) with an accuracy of 92.5% (Fig 4E). We also assessed its performance on a new set of randomly selected regions which were predicted by AlphaFold to contain both helical and disordered character; the LR model predicted the dominant structural feature from composition alone 86.9% of the time (Fig 4E). Finally, we predicted the highly charged regions, where the LR model performed with an overall accuracy of 90.8% (Fig 4E). Most of this accuracy can be captured using only the top five coefficients of the model (Fig 4F). We also scored a dataset of PDB structures with secondary structure annotation (see Methods); the model performed with an accuracy of ~90% on these experimentally determined structures (S3C Fig).

To see whether there were systematic differences in the relationship between amino acid composition and secondary structure when using real versus predicted structures, we also created a second LR model trained on purely helical or disordered sequences from the PDB-derived dataset and compared it to the original LR model (S3A and S3B Fig). The coefficients of the two models are highly correlated, with the interesting exceptions of the two helix-breakers P (which has the same order of importance in both models but a much larger magnitude coefficient in the PDB LR model) and G (which has a much higher relative importance in the PDB LR model). The PDB LR model performed with an accuracy exceeding 90%, a result fully independent of AlphaFold predictions and as such an important robustness check.

To put our results in context and understand the breakdown of existing heuristics, we compared the accuracy of the LR model to the accuracy of the simple charge/hydropathy (Uversky) model, as well as a state-of-the-art deep-learning based disorder-predictor called flDPnn [47]. As expected, the Uversky model performed better than chance but worse than the LR model on the same sets of randomly drawn regions, but was completely non-predictive for highly charged regions (Fig 4E). This reinforces the idea that normalized net charge is not predictive of disorder (see marginal distributions in the right hand side of Fig 4C), at least for regions with this length distribution. In sum, virtually all the predictive power in the Uversky charge/hydropathy heuristic comes from hydropathy. Although the flDPnn predictions were significantly better than the Uversky prediction for the highly charged regions, they did not exceed the performance of the LR model.

Finally, we were curious whether our model was specific to amino acid usage in yeast, or if it could be extended to other proteomes. Using three other proteomes for which structures have been predicted, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, and *Homo sapiens*, we performed the same procedure of random region selection, labeling using the AlphaFold predictions and confidences, and classification with the LR model trained on AlphaFold predictions of yeast regions. The prediction accuracy was nearly identical to or slightly higher than the *S. cerevisiae* accuracy (Fig 4G). This simple model based only on the composition of a region is sufficient to predict helical or disordered character in proteomes that diverged over a billion years ago.

Highly charged regions are evolutionarily conserved

The unique evolutionary signatures suggested by our analysis of eIF3A, coupled with the consistency in predictive power of the LR model across vast evolutionary distances, led us to broader questions about the conservation of the regions we had identified. To what extent do highly charged regions retain their sequence properties and structure as organisms evolve?

To address this question, we turned to AYbRAH, a curated database of protein homologs and paralogs in 33 fungal species spanning 600 million years of evolution [50]. This dataset
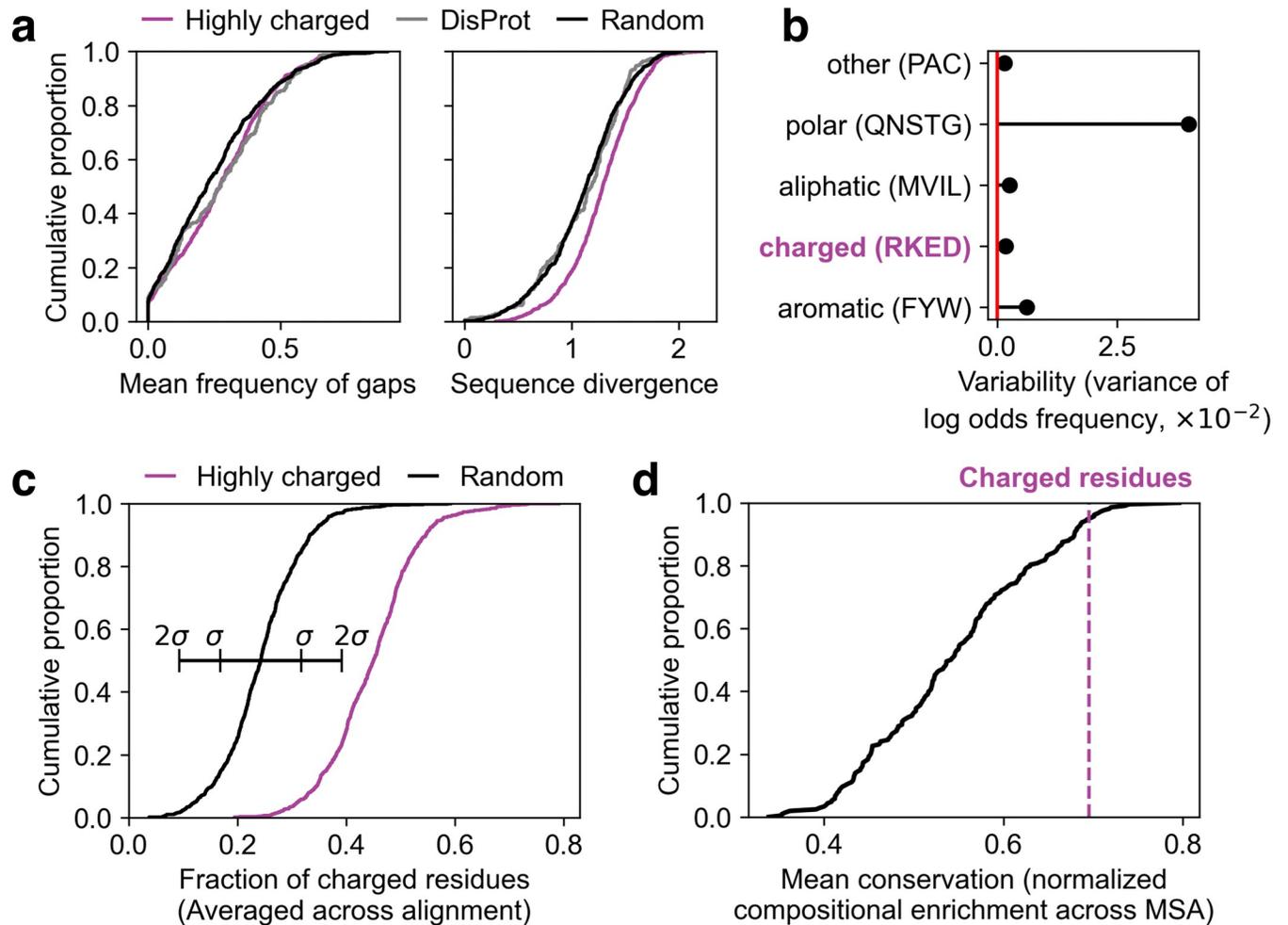
Fig 5. Highly charged regions are evolutionarily conserved. a The distribution of gap frequencies (left) and average position-wise entropy (right) summarizing multiple sequence alignments (MSAs) of the highly charged regions (purple), randomly drawn regions from the rest of the proteins that contain them (black), and IDRs from DisProt (gray). **b** Summary of the variance (in log odds space) of usage of categories of amino acids for all proteomes in the AYbRAH database. **c** FCR values, averaged across AYbRAH alignments, for the highly charged regions identified in the *S. cerevisiae* proteome and length-matched randomly-drawn regions and their associated AYbRAH MSA. **d** The average compositional conservation of regions enriched for all sets of four amino acids with the same total frequency as the charged amino acids, plotted as a CDF. Higher values indicate less drift, and lower values indicate more drift (regression to the proteome average).

combines automated homology detection and manual curation to achieve high-confidence predictions of highly diverged orthologs.

First, we quantified the sequence conservation of the regions in question by examining their alignments and calculating both the frequency of gaps and the sequence divergence (average position-wise entropy). Sequences with high insertion and deletion rates, a known feature of IDRs, will have higher alignment gap frequencies. Those with a high point mutation rate will have high divergence. A low value of both metrics indicates sequence conservation. We calculated these values for all the highly charged regions, and compared them to length-matched, randomly drawn regions from the rest of the same proteins. We found that the charged regions have significantly more gaps ($P<0.001$, Mann-Whitney U test) and are more divergent ($P<0.001$, Mann-Whitney U test) than the proteins in which they are found (Fig 5A). We compared these distributions to the same values calculated for experimentally verified IDRs from DisProt, and found that although the charged regions have similar gap

frequency to IDRs, they are even more divergent at the sequence level. Thus simply viewing an alignment of these LCRs, without using a secondary structure prediction algorithm, one might conclude that they are disordered.

Despite this apparent lack of conservation, we were curious whether any aspects beyond sequence of the highly charged regions were conserved. These regions were identified because of their unique sequence composition, so we tested the degree to which they retained this composition over evolutionary time. To first determine the expected compositional variation, we measured the variation in the total proportion of each amino acid across the species represented in the AYbRAH database (S4A Fig). We found that as a group, the charged amino acids had very little variation in proportion of usage (Fig 5B). Consistent with selection, a high fraction of charged residues was preserved across species and substantially differed from randomly drawn regions in the same species (Fig 5C).

The distribution in Fig 5C contains some regions that on average fall below the threshold of 0.4 FCR that we established for the original search in the yeast proteome; this is not surprising given that all sequences are subject to drift, which pulls them towards the proteome average for any given trait unless selection intervenes. Therefore, we created a method to quantify drift in charged regions relative to other compositionally extreme regions.

We first identified regions in the yeast proteome enriched for all groups of four amino acids with a combined frequency within +/–0.01 of the combined frequency of the charged amino acids (0.233). For each of the 209 datasets, we calculated the mean proportion of the amino acids in question for each identified region across the AYbRAH alignment (note that FCR is a special case of this property where the four amino acids in question are glutamate, aspartate, lysine, and arginine) If the composition (enrichment of the four amino acids in question) is conserved, we should expect that the mean enrichment score across regions and alignments should be close to the mean of the original enriched dataset (close to or higher than the 0.4 threshold). In contrast, if the property is not conserved, we should expect this enrichment score to be close to the proteome average. To compare directly between datasets, we scaled this enrichment score to a unit scale between an effective 0 (the proteome average), and 1 (the median of the enriched regions detected from the *S. cerevisiae* proteome). Sets of four with a conservation score close to 1 are highly conserved for that property, while those that are close to 0 have experienced high levels of drift, indicating that they are not conserved. We find that the set of four charged amino acids falls within the top 5% of these scores (Fig 5D). From this analysis we conclude that in the highly charged regions, the charge density is extremely well-conserved.

Discussion

To understand the biology of proteins and their subdomains, heuristics are almost inevitably used: comparison to other proteins to infer similarity by homology, motif identification to predict binding partners, and so on. In the case of highly charged protein regions, several heuristics appear to converge on the conclusion that such regions are overwhelmingly likely to be disordered. By virtue of their strongly biased sequence composition, they tend to fall into the class of low-complexity sequences associated with lack of stable structure [51]; they tolerate insertion/deletion events at higher rates than typical well-folded sequences; their high charge favors interactions with solvent, and low hydrophobicity suggests the absence of a solvent-protected hydrophobic core. Consistent with this, many analyses of such regions proceed as though structure can be mostly or completely ignored.

Here, we have shown that naturally occurring highly charged regions are predicted to adopt helical structure to a degree which cannot be neglected—~40% in a proteome-wide analysis—
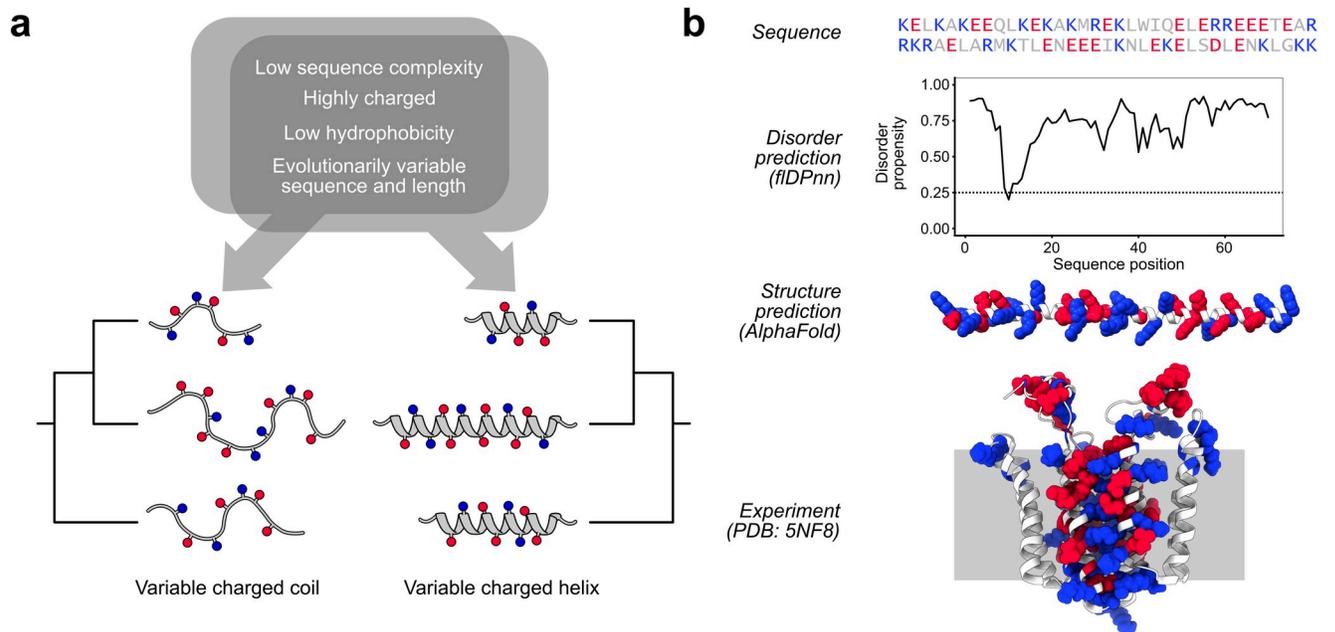
Fig 6. Rethinking assumptions of disorder in highly charged regions. a The same set of predictors that have been associated with disordered regions turn out to also be compatible with a fully structured helical region. **b** Present leading-edge methods for disorder and structure prediction disagree completely and also fail to capture experimental reality. The highly charged region of Rcf1 is alternatively predicted to be near-completely disordered or completely helical; experimentally and biologically, it forms most of the membrane-spanning helices and exposed loops in this dimeric mitochondrial inner membrane protein.

and that these predictions are supported by existing experimental data for both structured and disordered sequences. Moreover, we show that all these heuristic signals of disorder are in fact compatible with fully structured polypeptides: extended charged helices which have no hydrophobic core, grow and shrink in length over evolutionary time, interact with solvent on all sides, and form from sequences of two or even one type of amino acid (Fig 6A). Together, our results indicate that understanding the biology of highly charged sequences requires integrating insights from both structural and disorder-based approaches.

A consequence of these results is that they upend multiple well-established heuristics for determining how to think about, and study, a sequence's biological behaviors. The shortcut that charge and hydrophobicity can serve as accurate dimensions for separating structured from disordered sequences, powerfully demonstrated using limited data available at the time [2], does not work for highly charged, low-hydrophobicity sequences. Although many sophisticated methods for detecting disorder or single helices have been developed [33,34,47,48,52], the assumption that, because many disordered sequences are polyampholytes, other polyampholyte sequences can reasonably be treated as if disordered persists in modern work [3]. We emphasize that our results say nothing about the utility of further results built on the assumption of disorder, conditioned on its accuracy. And we further stress that considerable work may properly focus only on disordered sequences with no claims regarding the assessment of disorder. Nevertheless, as for the example of $(E_4K_4)_n$ polymers, it is straightforward to find examples in which sequences known to have well-defined structure are treated as if they did not, evidence for the undue influence of improper heuristics.

Given these results, it might seem inevitable that sophisticated structure-prediction methods would be required to more accurately discern whether particular highly charged sequences adopt a helical conformation. However, we introduce a simple amino acid composition-based

classifier—logistic regression with as few as five inputs—which can predict structure (or its absence) with accuracy above 90%. This model is trained on biologically occurring sequences, a tiny and profoundly biased subset of protein sequence space, such that we do not expect its performance to carry over to arbitrary sequences. Still, as a heuristic method implemented with a handful of numbers, it balances simplicity and accuracy (particularly over the charge/hydropathy heuristic) in a way which is practically useful in diagnosing structure for charged sequences.

The notion that intrinsically disordered proteins or regions sometimes adopt structure is well-understood [35,53], particularly in the case of folding upon binding [54]. Because of this distinction between conformations in isolation versus when bound to a partner, structures in the PDB may tell only a portion of the story. Similarly, AlphaFold specifically predicts structures most likely to appear in the PDB [33,42], rather than, for example, conformations which are occupied most of the time in the biological context. To the extent that our results depend on these resources, they similarly remain inapplicable to questions about the broader conformational ensembles that highly charged sequences may sample. But how stable or frequently adopted might these structures be?

From the perspective of evolutionary conservation, even a conformation which is occupied for a tiny fraction of the time may impose dominant constraints on a sequence, if this conformation contributes to organism fitness. To the extent that we wish to understand the relationship between sequence and biological function, this potential for rare conformations to dictate function may permit most conformations in the ensemble to be neglected—much as recognition of folding upon binding for a disordered region may properly focus attention on the bound state, even if it is fleeting. In the case of highly charged regions which must adopt helical conformations to carry out their functions, certain near-absolute constraints must be satisfied; no matter how unstable the helix, a proline kink in the backbone cannot be straightened, and so depletion of proline from these regions provides an additional signal. On the other hand, presence of a proline powerfully indicates that a straight helix cannot form and is therefore unlikely to be the functional conformation, no matter what other sequence signals exist and no matter how fleeting the helix state is proposed to be.

Even the best methods for predicting structure and for predicting disorder can disagree and fail to capture experimental reality. Consider the highly charged region of the yeast protein Rcf1 (Fig 6B). A top-ranked modern disorder predictor, flDPnn [47,52], predicts this region to be entirely disordered. AlphaFold predicts it to be entirely helical with high confidence. Neither captures reality: experimentally, this region forms most of a dimeric five-pass transmembrane protein in which charged residues, exposed on stable helices, form dimer-stabilizing salt bridges through the mitochondrial inner membrane [55] (Fig 6B). To the extent that cases like this closing example persist [56], the challenges we identify here remain open. Moverover, discrepancies between predicted proteome-wide frequencies of disorder [57] and inference of the same measure from experiment [58] invite deeper investigation of specific cases, as we have done here.

Broadly, while our results uncover previously overlooked structure in highly charged regions, the dual challenges of determining the biologically active configurations of these sequences, and of determining the statistical features of the conformational ensembles they occupy, remain open. Rather than looking at such sequences through the lens of disorder, it appears that both lenses—structure and disorder—will be needed to give the proper depth of focus.

Methods

Extraction of highly charged regions from the yeast proteome

The S288C reference genome was obtained from the Saccharomyces Genome Database (SGD). For each gene in the reference genome, we first computed fractional charge as a

moving average across its sequence using a window size of 12 residues and a triangular weight, where the highest weight was assigned to the middle region of each window. We then searched for highly charged regions in each sequence based on a fractional charge threshold of 0.4 and tolerance of 10 residues. Tolerance refers to the maximum number of residues that we allow to have moving average values below the fractional charge threshold before terminating the region. This tolerance allows for transient deviations from high charge and prevents fragmenting highly charged regions with small insertions of uncharged amino acids. We extracted regions that were longer than a given minimum region length of 30 residues, then trimmed any remaining uncharged residues from the N and C terminal ends of the sequences (these result from the triangular weighting scheme and the tolerance).

Calculating sequence complexity

Sequence complexity was calculated according to [6] using the following equation:

$$K_2 = -\sum_{i=1}^{N} \frac{n_i}{L}\left(log \frac{n_i}{L}\right)$$

where K_2 is the unnormalized complexity, N is the number of possible residues (in this case the 20 natural amino acids), and n_i is the number of each residue in the sequence, which has length L. This value is normalized to the "entropy of the language" (e.g., the yeast proteome), such that a sequence with compositional properties exactly equal to the average frequencies will have a complexity of 1.

The entropy of the language is calculated using the equation

$$H_{language} = -\sum_{i=1}^{N} p_i log \, p_i$$

where p_i is the frequency of letter i in the reference. We used two different languages as references; the first is the yeast proteome (so each p_i represents the average frequency of that amino acid in the proteome). We also used a modified reference enriched for charged residues; each of the amino acids lysine, arginine, aspartate, and glutamate had a frequency of 0.125; the remaining frequency (0.5) was distributed evenly among the other amino acids.

Generation of null distribution for amino acid usage

We counted the number of occurrences of adenine (A), thymine (T), cytosine (C), and guanine (G) in the DNA sequences of all open reading frames in *S. cerevisiae*. The expected frequency of each codon was computed as the product of the frequencies of all nucleotides that appear in that codon and the expected frequency of each amino acid was computed as the sum of the expected frequencies of all codons specifying that amino acid.

Extraction of AlphaFold data

We used proteome-wide structure predictions from AlphaFold to analyze the structure of the regions we identified with high proportions of charged residues. We downloaded the structures for all *S. cerevisiae* proteins from the AlphaFold website (https://alphafold.ebi.ac.uk/download#proteomes-section) [33]. We read the PDB files into Python and used DSSP implemented in MDTraj [46] to score secondary structure. We used custom Python functions to extract the confidence scores from the predicted structure file for each protein.

Construction of a logistic regression (LR) model to predict secondary structure

To classify the secondary structure of a region on the basis of composition, we first constructed a training dataset built from AlphaFold structures. We classified all residues in all structures as either helical (classified as helical by DSSP and with a pLDDT score above 70), disordered (classified as coil by DSSP and with a pLDDT score above 70 or any residue with a pLDDT score less than 50) [38], sheet (classified as sheet by DSSP and with a pLDDT score above 70), or other. The pLDDT cutoff of 70 marks was chosen because this value was used by the creators of AlphaFold to distinguish between "Confident" and "Low" model confidence. We did not explicitly test the dependence of our results on the value of this cutoff since it was defined and extensively validated by the creators of the model.

To generate training and testing data for the LR model, we exhaustively searched for purely helical and disordered regions by identifying regions that were greater than 25 amino acids long and only contained either helical or disordered residues. We extracted 6882 helical regions and 6366 disordered regions from the *S. cerevisiae* proteome. We randomly selected 75% of these regions as training data and 25% of the regions as testing data and built a LR model using amino acid composition as the predictor, that is, each individual amino acid is assigned a weight in the regression. The LR model is built using the scikit-learn package in Python.

We also constructed a LR model based on empirical structures from the PDB; secondary structure and disorder were annotated on a per-residue basis from the experimental 3D structure by the PDB and were obtained by request [59] (see their Methods for details). The final dataset contained 64,804 regions greater than 25 amino acids long and consisting solely of either helix or disorder; the regions were approximately equally split between helical and disordered. Otherwise, the procedure was identical to that used for the AlphaFold predictions.

Classifying regions and computing model accuracy

We used the per-residue secondary structure classifications described above to score the highly charged regions: regions with more than 60% helical or disordered residues were labeled as their dominant type, and all others were labeled as "intermediate." These experimental classifications were taken to be ground truth in the accuracy scores for each method (Uversky, LR, and flDPnn). Out of all the regions, 34% were labeled as disordered and 31% were labeled as helical. We then used the LR model to classify these regions on the basis of their amino acid composition. The ground-truth labels were used to compute the false negative and positive rates, and the overall model accuracy.

To directly compare to the LR dataset, we extracted all purely helical or purely disordered regions as well as random regions from the AlphaFold dataset. We used the same scoring conditions as that used for the highly charged regions (>60% helical or disordered), scored each region using the LR model, and computed model accuracy in the same way as described above.

To compare between the accuracy of the LR model and the Uversky model, we also used the Uversky model to classify the three sets of protein regions as helical or disordered. In particular, we calculated the normalized net charge and normalized mean hydropathy for each of the regions. The regions that fall above and to the left of the dividing line are classified as disordered and the remaining regions are classified as helical. The accuracy of the Uversky model is computed in the same way as for the LR model.

Finally, we used flDPnn to classify each of the highly charged and random regions as disordered or ordered. We averaged the predicted score for disorder for each of the residues in each

given region and classified regions with a mean disorder score of 0.3 (the threshold chosen by flDPnn) or higher as disordered. The remaining non-disordered regions were assumed to be compatible with helices, the only order type in this dataset. The accuracy of flDPnn predictions was then computed in the same way as above.

Evolutionary analysis: sequence properties

We used custom Python scripts to extract multiple sequence alignments (MSAs) for all proteins in the AYbRAH alignment [50]. A region was considered "present" in a protein in the MSA if the region which aligned to the *S. cerevisiae* sequence that we identified as highly charged contained at least 30 amino acids (the same length minimum length as was required for a region detected by the algorithm).

To compute alignment quality, the longest and shortest sequences in each alignment were removed, and any resulting columns containing only gaps (represented by the symbol "-" in the alignment) were removed. We then quantified the frequency of gaps as well as sequence column-wise entropy, which we refer to as "sequence divergence." The mean frequency of gaps was computed as the number of "-" characters divided by the total number of characters across all sequences in an alignment. Sequence divergence was computed by summarizing the frequency of amino acids in each column as a one-dimensional probability distribution and calculating the Shannon entropy of that distribution.

Evolutionary analysis: compositional drift

To test compositional drift for regions enriched in selected amino acids, we identified all unique sets of four amino acids (excluding the charged amino acids) with the same combined frequency as the four charged amino acids. We modified the charged region algorithm to detect regions enriched for these sets of four. We used an enrichment threshold of 0.35 (35% of the region composed of the amino acids in question), minimum length of 30 amino acids, and a tolerance of 15; these parameters were selected to yield datasets that most closely matched the number of hits and length distribution of the original dataset (enrichment for E, D, R, and K). For each of these datasets, we calculated the conservation by averaging the enrichment score (percent composition of the specific amino acids) across the alignment in each region, taking the mean of that distribution, and scaling it between the *S. cerevisiae* proteome average (0) and the average enrichment of the hits detected by the algorithm (1). This allowed us to compare datasets directly: sets of amino acids with values close to 0 experienced enough drift that they approach the proteome average; those with values close to 1 stay far from the average and close to the (rescaled) enrichment threshold and thus are likely conserved.

Analysis of the eIF3A charged helix

We used MUSCLE version 3.8.31 [60] with default parameters to generate an alignment of all the eIF3A homologs for which a structure was available on AlphaFold as of April 2022 (35 species of model organisms including bacteria, yeast, mold, rice, soybean, mouse, and human). This dataset was used in Fig 3.

A cryoEM structure of eIF3A in complex with the ribosome was used for structural analysis. The structure has PDB code 6ZCE (http://doi.org/10.2210/pdb6ZCE/pdb) [45].

Secondary structure was scored with DSSP as described above. The length of the charged helix was calculated by identifying the start of the highly charged region and then counting the amino acids until a run of more than three non-helix characters was encountered. A reference helix from earlier in the sequence was chosen as a comparison.

Analysis of Rcf1

The 70-residue Rcf1 charged region (Fig 6B) was used for disorder predictions with flDPnn [47] at http://biomine.cs.vcu.edu/servers/flDPnn/ with default settings, and for structure prediction with AlphaFold through ColabFold [48].

Statistical tests

All p-values were calculated with the Mann-Whitney U Test (Wilcoxon Rank Sum Test), either two-sided if no hypotheses were formed about the relationship between the two distributions or one-sided otherwise. The evaluation of these tests was done in Python and can be found (along with the data) in the Jupyter notebook for the relevant figure.

Supporting information

S1 Fig. Regions high in charge are prevalent in the yeast proteome and are low complexity. a The distribution of FCR values for randomly-drawn regions of different length in the yeast proteome (black, purple), and randomly generated regions using the average S. cerevisiae amino acid frequencies (gray). Vertical lines show two standard deviations above the mean. The minimum FCR of the highly charged regions after trimming is shown with a red dashed line. b The distribution of FCR values for the regions detected (purple), the proteins that contain them (green and gray), and the per-protein FCR for all yeast proteins (black). c Distribution of region lengths for highly charged regions. d Enrichment for proteins containing highly charged regions in different cell compartments. The proportion of the proteins in the category with at least one highly charged region is shown in black. The expected proportion (based on the number of unique genes detected by the algorithm and the total number of yeast proteins) is shown as a horizontal white line.
(TIFF)

S2 Fig. Secondary structure is pervasive in highly charged, low complexity regions. a The frequency of (per-residue) secondary structure for all the confidently predicted (pLDDT > 70) residues from the AlphaFold structure for highly charged regions, length-matched random regions, and experimentally-verified disordered regions from DisProt. The majority of confidently prediction regions in disordered (DisProt) regions are coil. b The κ value of the region versus the proportion predicted helix for all regions that are more than 60% helix or disordered. c Alignment of MNN4 C-terminal region in fungal species in the AYbRAH database.
(TIFF)

S3 Fig. Helical regions can be predicted from composition. a Coefficients from the LR model trained on helical and coil/disordered regions randomly drawn from the PDB. **b** Comparison between model coefficients in both the PDB and AlphaFold trained models. The importance of P and G varies the most between the two; P in magnitude and G in order of importance. **c** Accuracy of the AlphaFold and PDB trained models and the accuracy of the AlphaFold model when used to classify PDB structures.
(TIFF)

S4 Fig. Highly charged regions are evolutionarily conserved. a Variation in frequency for each individual amino acid. Distribution is the frequency of usage in each proteome in the AYbRAH database. **b** Absolute coverage (number of species in the AYbRAH alignment which contain non-gap characters in the detected highly charged region relative to the maximum number of species possible) and relative coverage (number of species in the AYbRAH alignment which have non-gap characters in the highly-charged region relative to the number of

species with a homolog of the protein) for each protein with a detected highly charged region. **c** Length variation (between AYbRAH homologs) in regions labeled as helix (teal) or disordered (gray) based on their predicted structure from AlphaFold and experimentally-verified yeast disordered regions from DisProt (black).
(TIFF)

S1 Table. Overlap between predicted helical highly charged regions and proteins found in DisProt. The proteins listed were predicted to contain highly helical regions (summarized in Fig 2C) and also found in the Disprot database (a reference for experimentally verified disordered proteins and regions).
(XLSX)

Acknowledgments

The authors thank Alex Holehouse for helpful discussions, and Alexander Cope for providing the structure-annotated PDB data.

Author Contributions

Conceptualization: Catherine G. Triandafillou, Aaron R. Dinner, D. Allan Drummond.

Formal analysis: Catherine G. Triandafillou, Rosalind Wenshan Pan.

Funding acquisition: Catherine G. Triandafillou, Rosalind Wenshan Pan, Aaron R. Dinner, D. Allan Drummond.

Investigation: Catherine G. Triandafillou, Rosalind Wenshan Pan.

Methodology: Catherine G. Triandafillou, D. Allan Drummond.

Supervision: Aaron R. Dinner, D. Allan Drummond.

Visualization: Catherine G. Triandafillou, Rosalind Wenshan Pan, D. Allan Drummond.

Writing – original draft: Catherine G. Triandafillou, Rosalind Wenshan Pan, Aaron R. Dinner, D. Allan Drummond.

Writing – review & editing: Catherine G. Triandafillou, Rosalind Wenshan Pan, Aaron R. Dinner, D. Allan Drummond.

References

1. Babu MM, Kriwacki RW, Pappu RV. Structural biology. Versatility from protein disorder. Science. 2012; 337: 1460–1461. https://doi.org/10.1126/science.1228775 PMID: 22997313

2. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000; 41: 415–427. https://doi.org/10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7 PMID: 11025552

3. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A. 2013; 110: 13392–13397. https://doi.org/10.1073/pnas.1304749110 PMID: 23901099

4. Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett. 2004; 576: 348–352. https://doi.org/10.1016/j.febslet.2004.09.036 PMID: 15498561

5. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001; 42: 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::aid-prot50>3.0.co;2-3 PMID: 11093259

6. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem. 1994; 18: 269–285. https://doi.org/10.1016/0097-8485(94)85023-2 PMID: 7952898

7. Pak CW, Kosno M, Holehouse AS, Padrick SB, Mittal A, Ali R, et al. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. Mol Cell. 2016; 63: 72–85. https://doi.org/10.1016/j.molcel.2016.05.042 PMID: 27392146

8. Huang F, Oldfield CJ, Xue B, Hsu W-L, Meng J, Liu X, et al. Improving protein order-disorder classification using charge-hydropathy plots. BMC Bioinformatics. 2014; 15 Suppl 17: S4. https://doi.org/10.1186/1471-2105-15-S17-S4 PMID: 25559583

9. Sawle L, Ghosh K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. J Chem Phys. 2015; 143: 085101. https://doi.org/10.1063/1.4929391 PMID: 26328871

10. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2010; 107: 8183–8188. https://doi.org/10.1073/pnas.0911107107 PMID: 20404210

11. Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife. 2019; 8. https://doi.org/10.7554/eLife.46883 PMID: 31264965

12. Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu RV. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc Natl Acad Sci U S A. 2016; 113: 5616–5621. https://doi.org/10.1073/pnas.1516277113 PMID: 27140628

13. Wang Y, Zhou H, Sun X, Huang Q, Li S, Liu Z, et al. Charge Segregation in the Intrinsically Disordered Region Governs VRN1 and DNA Liquid-like Phase Separation Robustness. J Mol Biol. 2021; 433: 167269. https://doi.org/10.1016/j.jmb.2021.167269 PMID: 34571015

14. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem. 1993; 17: 149–163.

15. Xiang S, Kato M, Wu LC, Lin Y, Ding M, Zhang Y, et al. The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei. Cell. 2015; 163: 829–839. https://doi.org/10.1016/j.cell.2015.10.040 PMID: 26544936

16. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell. 2012; 149: 753–767. https://doi.org/10.1016/j.cell.2012.04.017 PMID: 22579281

17. Kwon I, Kato M, Xiang S, Wu L, Theodoropoulos P, Mirzaei H, et al. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. Cell. 2013; 155: 1049–1060. https://doi.org/10.1016/j.cell.2013.10.033 PMID: 24267890

18. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017. https://doi.org/10.1038/nrm.2017.7 PMID: 28225081

19. Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol Cell. 2015; 57: 936–947. https://doi.org/10.1016/j.molcel.2015.01.013 PMID: 25747659

20. Elbaum-Garfinkle S, Kim Y, Szczepaniak K, Chen CC-H, Eckmann CR, Myong S, 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. 2015; 112: 7189–7194. https://doi.org/10.1073/pnas.1504822112 PMID: 26015579

21. Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell. 2015; 163: 123–133. https://doi.org/10.1016/j.cell.2015.09.015 PMID: 26406374

22. Szabó AL, Sánta A, Pancsa R, Gáspári Z. Charged sequence motifs increase the propensity towards liquid-liquid phase separation. FEBS Lett. 2022; 596: 1013–1028. https://doi.org/10.1002/1873-3468.14294 PMID: 35072950

23. Chong PA, Vernon RM, Forman-Kay JD. RGG/RG Motif Regions in RNA Binding and Phase Separation. J Mol Biol. 2018; 430: 4650–4665. https://doi.org/10.1016/j.jmb.2018.06.014 PMID: 29913160

24. Wang J, Choi J-M, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell. 2018; 174: 688–699.e16. https://doi.org/10.1016/j.cell.2018.06.006 PMID: 29961577

25. Lin Y-H, Chan HS. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. Biophys J. 2017; 112: 2043–2046. https://doi.org/10.1016/j.bpj.2017.04.021 PMID: 28483149

26. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2020; 21: 458–472. https://doi.org/10.1093/bib/bbz007 PMID: 30698641

27. Moses D, Guadalupe K, Yu F, Flores E, Perez A, McAnelley R, et al. Hidden Structure in Disordered Proteins is Adaptive to Intracellular Changes. 2022. https://doi.org/10.2139/ssrn.4002760

28. Spink BJ, Sivaramakrishnan S, Lipfert J, Doniach S, Spudich JA. Long single alpha-helical tail domains bridge the gap between structure and function of myosin VI. Nat Struct Mol Biol. 2008; 15: 591–597. https://doi.org/10.1038/nsmb.1429 PMID: 18511944

29. Wang CL. Caldesmon and smooth-muscle regulation. Cell Biochem Biophys. 2001; 35: 275–288. https://doi.org/10.1385/cbb:35:3:275 PMID: 11894847

30. Folkmann AW, Collier SE, Zhan X, Aditi, Ohi MD, Wente SR. Gle1 functions during mRNA export in an oligomeric complex that is altered in human disease. Cell. 2013; 155: 582–593.

31. Sivaramakrishnan S, Spink BJ, Sim AYL, Doniach S, Spudich JA. Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif. Proc Natl Acad Sci U S A. 2008; 105: 13356–13361. https://doi.org/10.1073/pnas.0806256105 PMID: 18768817

32. Batchelor M, Paci E. Helical Polyampholyte Sequences Have Unique Thermodynamic Properties. J Phys Chem B. 2018; 122: 11784–11791. https://doi.org/10.1021/acs.jpcb.8b08344 PMID: 30351106

33. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

34. Simm D, Kollmar M. Waggawagga-CLI: A command-line tool for predicting stable single α-helices (SAH-domains), and the SAH-domain distribution across eukaryotes. PLoS One. 2018; 13: e0191924.

35. Fossat MJ, Posey AE, Pappu RV. Uncovering the Contributions of Charge Regulation to the Stability of Single Alpha Helices. Chemphyschem. 2023; e202200746. https://doi.org/10.1002/cphc.202200746 PMID: 36599672

36. Uversky VN. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Frontiers in Physics. 2019;7. https://doi.org/10.3389/fphy.2019.00010

37. Lee B, Jaberi-Lashkari N, Calo E. A unified view of LCRs across species. bioRxiv. 2022. p. 2022.01.06.475253. https://doi.org/10.1101/2022.01.06.475253

38. Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol. 2021; 433: 167208. https://doi.org/10.1016/j.jmb.2021.167208 PMID: 34418423

39. Tan C, Niitsu A, Sugita Y. Repulsive interaction and secondary structure of highly charged proteins in regulating biomolecular condensation. bioRxiv. 2022. p. 2022.11.16.516834. https://doi.org/10.1101/2022.11.16.516834

40. Brzovic PS, Heikaus CC, Kisselev L, Vernon R, Herbig E, Pacheco D, et al. The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. Mol Cell. 2011; 44: 942–953. https://doi.org/10.1016/j.molcel.2011.11.008 PMID: 22195967

41. Fossat MJ, Posey AE, Pappu RV. Quantifying charge state heterogeneity for proteins with multiple ionizable residues. Biophys J. 2021; 120: 5438–5453. https://doi.org/10.1016/j.bpj.2021.11.2886 PMID: 34826385

42. Reid Alderson T, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. bioRxiv. 2023. p. 2022.02.18.481080. https://doi.org/10.1101/2022.02.18.481080

43. Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol. 2004; 338: 1015–1026. https://doi.org/10.1016/j.jmb.2004.03.017 PMID: 15111064

44. Baker EG, Bartlett GJ, Crump MP, Sessions RB, Linden N, Faul CFJ, et al. Local and macroscopic electrostatic interactions in single α-helices. Nat Chem Biol. 2015; 11: 221–228.

45. Kratzat H, Mackens-Kiani T, Ameismeier M, Potocnjak M, Cheng J, Dacheux E, et al. A structural inventory of native ribosomal ABCE1-43S pre-initiation complexes. EMBO J. 2021; 40: e105179. https://doi.org/10.15252/embj.2020105179 PMID: 33289941

46. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J. 2015; 109: 1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015 PMID: 26488642

47. Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021; 12: 4438. https://doi.org/10.1038/s41467-021-24773-7 PMID: 34290238

48. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making Protein folding accessible to all. Nat Methods. 2022. https://doi.org/10.1038/s41592-022-01488-1 PMID: 35637307

49. Pace CN, Scholtz JM. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J. 1998; 75: 422–427. https://doi.org/10.1016/s0006-3495(98)77529-0 PMID: 9649402

50. Correia K, Yu SM, Mahadevan R. AYbRAH: a curated ortholog database for yeasts and fungi spanning 600 million years of evolution. Database. 2019;2019. https://doi.org/10.1093/database/baz022 PMID: 30893420

51. Huntley MA, Golding GB. Simple sequences are rare in the Protein Data Bank. Proteins. 2002; 48: 134–140. https://doi.org/10.1002/prot.10150 PMID: 12012345

52. Necci M, Piovesan D, CAID Predictors, DisProt Curators, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods. 2021; 18: 472–481. https://doi.org/10.1038/s41592-021-01117-3 PMID: 33875885

53. Galea CA, Nourse A, Wang Y, Sivakolundu SG, Heller WT, Kriwacki RW. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. J Mol Biol. 2008; 376: 827–838. https://doi.org/10.1016/j.jmb.2007.12.016 PMID: 18177895

54. Fuxreiter M. Fold or not to fold upon binding—does it really matter? Curr Opin Struct Biol. 2019; 54: 19–25. https://doi.org/10.1016/j.sbi.2018.09.008 PMID: 30340123

55. Zhou S, Pettersson P, Huang J, Sjöholm J, Sjöstrand D, Pomès R, et al. Solution NMR structure of yeast Rcf1, a protein involved in respiratory supercomplex formation. Proc Natl Acad Sci U S A. 2018; 115: 3048–3053. https://doi.org/10.1073/pnas.1712061115 PMID: 29507228

56. O'Neil PT, Vasquez-Montes V, Swint-Kruse L, Baldwin MR, Ladokhin AS. Spectroscopic evidence of tetanus toxin translocation domain bilayer-induced refolding and insertion. Biophys J. 2021; 120: 4763–4776. https://doi.org/10.1016/j.bpj.2021.09.030 PMID: 34555358

57. Lee R van der, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014; 114: 6589–6631. https://doi.org/10.1021/cr400525m PMID: 24773235

58. Leuenberger P, Ganscha S, Kahraman A, Cappelletti V, Boersema PJ, von Mering C, et al. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science. 2017;355. https://doi.org/10.1126/science.aai7825 PMID: 28232526

59. Cope AL, Gilchrist MA. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. BMC Genomics. 2022; 23: 408. https://doi.org/10.1186/s12864-022-08635-0 PMID: 35637464

60. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5: 113. https://doi.org/10.1186/1471-2105-5-113 PMID: 15318951