Generative design of intrinsically disordered protein regions with IDiom

Jason X. Liu [1] **Sebastian Ibarraran** [†2] **Frank Hu** [†2] **Abigail Park** [†2] **Alexander R. Dunn** [1] **Grant M. Rotskoff** [2 3]

Abstract

Intrinsically disordered protein regions are ubiquitous across all kingdoms of life. These structurally heterogeneous regions play central roles in cellular processes such as transcriptional regulation, cellular signaling, and subcellular organization, yet they have remained largely inaccessible to rational design. Structure-based generative methods are not applicable to proteins that lack a stable fold, and existing sequence-based approaches for disordered regions rely on sampling methods that do not capture the evolutionary statistics of natural disordered regions. Here, we introduce IDiom, an autoregressive protein language model trained on 37 million intrinsically disordered region sequences curated from the AlphaFold Database. Trained using a fill-in-the-middle data augmentation, IDiom generates disordered region sequences conditioned on their surrounding structured context, as well as fully disordered proteins without any context. The model generates diverse sequences that recapitulate biologically relevant sequence features of natural disordered regions, and we demonstrate that post-training via reinforcement learning with a subcellular localization reward model produces sequences with features which are consistent with known sequence determinants of compartment-specific localization. These results establish IDiom as a general platform for the generative design of intrinsically disordered proteins and regions.

[†]These authors contributed equally. [1]Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA [2]Department of Chemistry, Stanford University, Stanford, CA 94305, USA [3]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA. Correspondence to: Grant M. Rotskoff <rotskoff@stanford.edu>, Alexander R. Dunn <ardunn@stanford.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Intrinsically disordered protein regions are ubiquitous across all kingdoms of life. The functional relevance of intrinsically disordered regions (IDRs) has become increasingly clear recently, in spite of the classical dogma of protein biology that structure implies function (Holehouse & Kragelund, 2024). While IDRs do not adopt well-defined folds, these sequences can act as flexible linkers (González-Foutel et al., 2022), multivalent signaling hubs (Wright & Dyson, 2015), and drivers of biomolecular condensate formation (Martin et al., 2020; Elbaum-Garfinkle et al., 2015; Wei et al., 2017), playing crucial roles in biological processes such as transcriptional regulation, chromatin organization, and subcellular organization (Holehouse & Kragelund, 2024; Banani et al., 2017).

Rational design of IDRs would unlock new functional controls in bioengineering, including tunable condensate phase behavior, precise regulation of cell signaling, and targeted protein localization (Tesei et al., 2026). While much of the recent progress in protein design has been driven by accurate structure prediction (Jumper et al., 2021; Abramson et al., 2024) and diffusion-based generative models (Watson et al., 2023; Chu et al., 2024; Pacesa et al., 2025), direct application of these approaches to IDRs is untenable due to their lack of a stable folded structure.

Sequence-based generative models have also been extensively explored recently for applications in protein design. Transformer-based (Vaswani et al., 2017) protein language models (PLMs) have been trained on corpora of full length protein sequences from databases such as the UniProt Reference Clusters and the Big Fantastic Database. These models learn rich evolutionary statistics over amino acid sequences and have enabled the design of novel proteins and functional variants (Ferruz et al., 2022; Lin et al., 2023; Hayes et al., 2025; Madani et al., 2023; Nijkamp et al., 2023; Bhatnagar et al., 2025). Nevertheless, the design of intrinsically disordered regions with existing PLMs is not straightforward: since structured domains outnumber intrinsically disordered regions in the sequence databases on which current PLMs are trained, the generative prior is largely biased towards folded domains (Ferruz et al., 2022). Alternative sequence-based approaches to IDR design have attempted

to use sampling-based methods to construct IDRs from compositional rules or simple statistical models (Pesce et al., 2024; Krueger et al., 2024; Kilgore et al., 2025; Hunter et al., 2026). However, these approaches cannot be conditioned on any surrounding sequence context of generated IDRs, and they do not capture the evolutionary statistics which emerge from training on large corpora of natural protein sequences.

Here, we address this gap by training a 122M parameter autoregressive, decoder-only protein language model called `IDiom` using a dataset of 37 million intrinsically disordered regions curated from the AlphaFold Database (Figure 1a) (Varadi et al., 2022; 2024). We apply a fill-in-the-middle transformation to the training data (Bavarian et al., 2022; Li et al., 2023) to enable the model to generate IDR spans conditioned on their surrounding context, a capability essential for the design of disordered regions (Tesei et al., 2026). Training on this dataset of disordered sequences allows `IDiom` to learn a generative prior over the sequence statistics of natural disordered regions, and we demonstrate that the model generates diverse sequences that recapitulate the compositions, patterning, and motifs of natural intrinsically disordered regions. We further show that `IDiom` learns in-context: given the flanking sequence context of a specific protein, the model generates disordered spans whose sequence features are more appropriate for that context than unprompted generations. Finally, we demonstrate that `IDiom` can be post-trained using reinforcement learning, and we apply this to design disordered sequences with targeted subcellular localization (Kilgore et al., 2025). Together, these results establish `IDiom` as a general platform for the generative design of intrinsically disordered proteins and regions.

2. Results

2.1. Protein language modeling for intrinsically disordered proteins and regions

To curate a dataset of intrinsically disordered region (IDR) sequences for model training, we first use AlphaFold2 (AF2) predicted structures from the AlphaFold Database (AFDB) (Varadi et al., 2024) to identify IDRs of proteins. We use low AF2 predicted local distance difference test (pLDDT) values as a predictor of disorder, as this has been demonstrated to correlate strongly with experimental measurements of disorder (Ruff & Pappu, 2021; Wilson et al., 2022; Zhao et al., 2023). To curate these IDRs from the database, we first cluster AFDB sequences at 90% sequence identity before applying a windowed pLDDT-based threshold to identify IDRs (Tesei et al., 2024), with proteins containing multiple IDRs contributing multiple records to the dataset. Finally, we discard IDRs shorter than 30 residues, IDRs which reside in proteins whose full length is greater than 512 residues, and proteins whose entire length is low-pLDDT. This process yields a dataset of 37M IDRs and their positions within their associated full length proteins (see Methods for more details). Figure 1c (upper) depicts an example low-pLDDT IDR, in orange, extracted by this process for the human protein NPM1 (UniProt: P06748).

We use two baselines to validate our data curation strategy. First, we use 1,017 experimentally validated IDRs from the DisProt database as a ground-truth dataset of IDRs (Nugnes et al., 2026). Second, we use 1,000 randomly chosen sequences from the Class, Architecture, Topology, and Homologous superfamily (CATH) database clustered at 60% identity (S60) as a ground-truth dataset of protein domains with well-defined folded structures (Waman et al., 2025). Using secondary structure content calculations, we show that this method extracts IDRs with substantially lower secondary structural content compared to CATH sequences (Figure BS1). We note that our analysis shows that AF2 assigns high secondary structural content (with low confidence) to a fraction of the extracted IDRs, although this effect is also reflected in the set of DisProt IDRs. We additionally run orthogonal disorder predictions on a subset of the dataset and further validate that the curated sequences score similarly to DisProt sequences on these metrics (Figures BS2, BS3).

Intrinsically disordered regions are naturally located at various locations within a protein sequence (Figure 1b). To enable `IDiom` to infill sequences of intrinsically disordered regions at arbitrary locations within a protein, we employ a fill-in-the-middle data transformation (Bavarian et al., 2022; Li et al., 2023). In this approach, we prepend the special token `<N>` to residues in the N-terminal flanking context before the IDR, the token `<C>` to residues in the C-terminal context after the IDR, and the token `<I>` to residues of the IDR span itself. Then, we transform the sequence by relocating `<I>` and the IDR span to the end of the entire sequence, thus allowing the standard causal language modeling objective to generate IDR spans conditioned on any preceding N-terminal and succeeding C-terminal flanking context (see Figure 1a). In addition, to enable the generation of intrinsically disordered proteins (IDPs) with no surrounding context, we augment the dataset by creating records in which we take each curated IDR and delete its N- and C-terminal flanking context, enabling unconditioned generation. The final dataset comprises 74M sequences in total (37M IDRs and 37M IDPs), which we use to pretrain `IDiom`. Additional details of the data augmentation, model architecture, and training procedure are provided in the Methods section.

2.2. `IDiom` generates diverse disordered regions and proteins

To characterize the pre-trained model, we first generate two sets of sequences: 100,000 unprompted IDPs (referred to

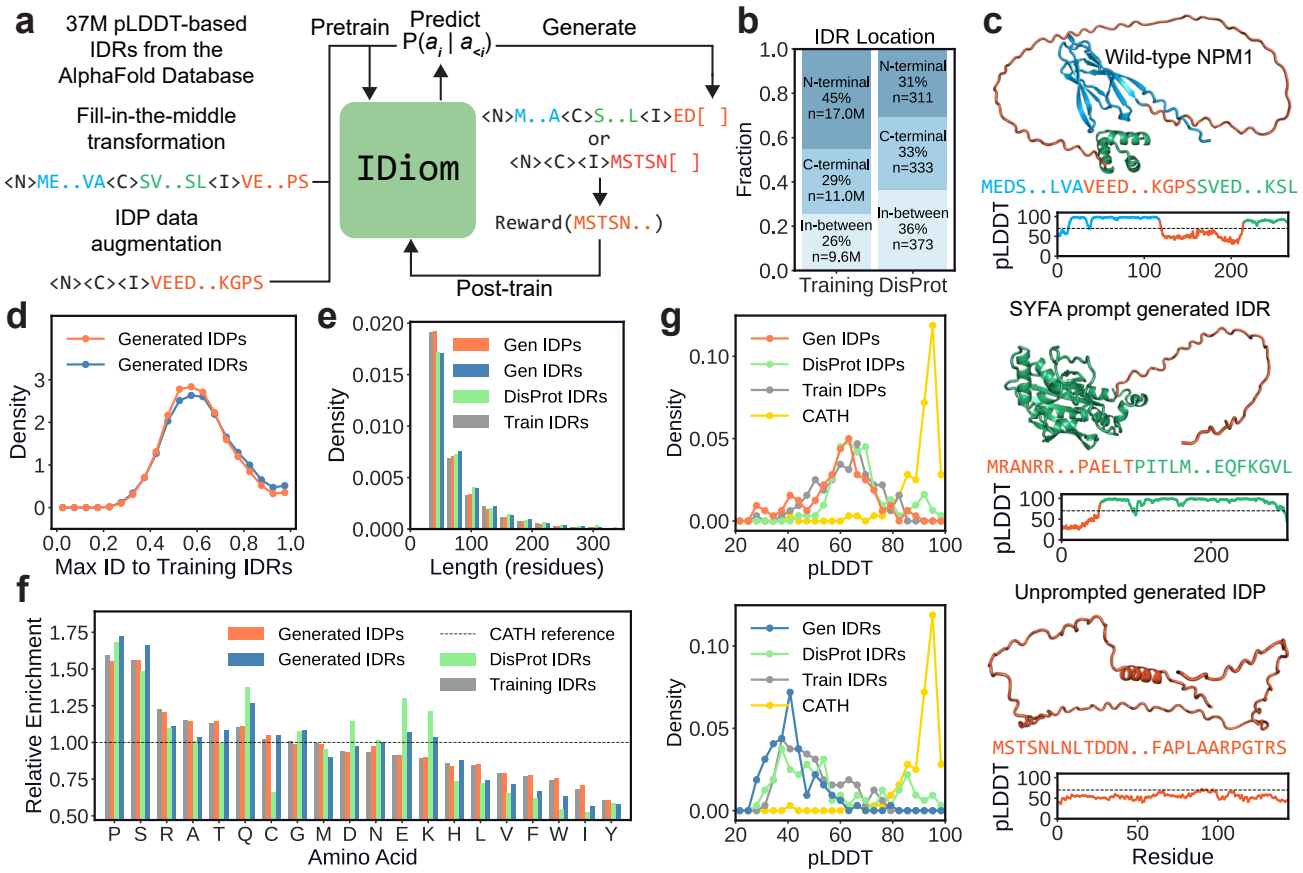

Figure 1. **Data curation, training, and generative modeling of intrinsically disordered regions (IDRs) and proteins (IDPs) using `IDiom`. (a)** Schematic depicting data preparation, pre-training, sequence generation, and post-training of `IDiom`. **(b)** Distribution of the sequence locations of 37M IDRs curated from the AlphaFold Database (AFDB), as well as the locations of experimentally validated DisProt IDRs. **(c)** AlphaFold2 (AF2)-predicted structures, amino acid sequences, and plots of predicted local distance difference test (pLDDT) values of an example IDR curated from AFDB (NPM1, upper), an IDR generated by `IDiom` using SYFA as the prompt (middle), and an unprompted generated IDP (lower). Blue and green regions correspond to the N-terminal and C-terminal flanking context around the IDR, which is orange. **(d)** Distribution of maximum sequence identities of unprompted generated IDPs and DisProt-prompt generated IDRs relative to the 37M IDRs of the training set, calculated using MMseqs2. **(e)** Distribution of sequence lengths of unprompted generated IDPs, DisProt-prompt generated IDRs, natural DisProt IDRs, and training set IDRs. **(f)** Relative enrichment of amino acid compositions of unprompted generated IDPs, DisProt-prompt generated IDRs, natural DisProt IDRs, and training set IDRs. The horizontal dashed line is the reference composition of folded CATH domain sequences. **(g)** (Upper): AF2 prediction pLDDTs of generated unprompted IDPs, natural DisProt IDRs removed from their surrounding context (DisProt IDPs), training set IDRs removed from their surrounding context (train IDPs), and CATH domains. (Lower): AF2 prediction pLDDTs of DisProt-prompt generated IDRs, natural DisProt IDRs within their context, training set IDRs within their context, and CATH domains.

as generated IDPs), and a set of context-prompted IDRs in which 100 IDRs are generated for each of 1,017 experimentally validated IDRs from the DisProt set, using the IDR flanking contexts as prompts (referred to as generated IDRs). Representative examples of a DisProt context-prompted IDR and an unprompted IDP are shown in Figure 1c (middle) and 1c (lower), respectively.

We find that across multiple metrics, `IDiom` generates sequences that closely resemble natural DisProt IDRs while remaining diverse and distinct from the training data. The distribution of maximum sequence identities to training set IDRs peaks broadly around 60%, indicating that most generated sequences are substantially dissimilar to any sequence seen during training (Figure 1d). Generated IDR and IDP lengths are also consistent with those of both the training set and DisProt IDRs, with most sequences below 100 residues long and a tail extending to approximately 300 residues long (Figure 1e).

To analyze the compositional biases of the IDRs and IDPs, we compute the amino acid enrichment of training, generated, and DisProt sequences relative to a baseline composition of the folded CATH domains (Figure 1f). Consistent

with established compositional biases of disordered regions (Theillet et al., 2013; Ruff et al., 2026), generated sequences are strongly enriched in proline and serine, and depleted in order-promoting aliphatics such as leucine, isoleucine, and valine, as well as the aromatics phenylalanine, tryptophan, and tyrosine. The trends for training and generated sequences closely match natural DisProt IDRs across most amino acids, with the closest agreement observed for IDRs generated with the DisProt flanking contexts as prompts. We note that the discrepancies with DisProt in the generated compositions of cysteine and charged residues (glutamic acid, lysine, and aspartic acid), may be attributed to the small size and selection bias of the DisProt dataset.

We next assess the extent to which the generated sequences are disordered by predicting their structures using ColabFold (Mirdita et al., 2022); the resulting pLDDT distributions are shown in Figure 1g. We run the structure predictions in two settings. In the first, we compare the unprompted generated IDPs to DisProt and training IDRs which are provided to ColabFold as standalone sequences, without their flanking context (i.e. as IDPs). In this setting, the pLDDT for a given sequence is averaged across the entire IDP, generated or extracted (Figure 1g (upper)). In the second, we compare the DisProt-prompt generated IDRs to the full protein DisProt and training sequences. In this setting, we provide the full generated, training, or DisProt sequence to ColabFold, and the pLDDT is only averaged across the IDR span (Figure 1g (lower)). Across both settings, generated sequences exhibit pLDDT distributions that closely mirror those of DisProt as well as the training set, confirming that `IDiom` is able to generate sequences which are disordered to the same extent as natural IDRs. We note that the pLDDTs of IDPs in isolation are higher than for IDRs within their flanking context, which may be due to dataset biases in the AF2 training data. We additionally calculate secondary structure metrics and conduct orthogonal disorder predictions on these generated sequences, and we find that the metrics compare similarly to DisProt IDRs (Figures BS1–BS3).

2.3. Generated sequences capture the residue patterning of natural disordered regions

IDRs exhibit sequence patterning features that differ substantially from those of folded domains, including characteristic charge distributions, hydrophobic residue patterning, and low-complexity compositions (Langstein-Skora et al., 2026; Ruff et al., 2026; Das et al., 2015). Figure 2a shows representative generated sequences which illustrate canonical IDR features such as Q/N-rich low-complexity regions (van der Lee et al., 2014), polyampholyte charge block patterning (Mitrea et al., 2018), prion-like aromatic/glycine patterning (Martin et al., 2020), and proline enrichment (Theillet et al., 2013). To quantify how well `IDiom` recapitulates these properties, we computed the distributions of several sequence-level metrics for the generated, training, and DisProt disordered sequences, as well as the folded CATH domain sequences (Figure 2b-e).

Electrostatic interactions strongly influence the conformational behavior of IDRs, and both the overall charge content and its linear patterning are closely linked to physical properties and biological function (Mantonico et al., 2024; Lin et al., 2017; Mitrea et al., 2018; Das et al., 2015). The fraction of charged residues (FCR, Figure 2b) distinguishes strongly charged polyampholytes from weakly charged sequences. We find that the generated and natural IDRs and IDPs span a wider range of FCR values than folded CATH domains, reflecting the heterogeneity of natural IDRs, which range from highly charged sequences in which electrostatic repulsion drives disorder, to weakly charged low-complexity sequences (Das & Pappu, 2013).

To characterize charge patterning, we calculate the linear charge patterning parameter κ, which quantifies the deviation of a given sequence from a maximally charge segregated permutation of the same sequence (Das & Pappu, 2013). $\kappa \approx 0$ indicates well-mixed opposite charges and $\kappa \approx 1$ indicates segregation into blocks of the same charge, and we plot the distributions of these values in Figure 2c. Consistent with prior work linking charge segregation in IDRs to intermolecular interactions and phase behavior (Mitrea et al., 2018; Ginell et al., 2025), natural IDRs from the training and DisProt sets exhibit a tail toward high κ values relative to CATH domains, a feature that `IDiom` reproduces with its generated sequences.

We next examined hydrophobic patterning using the sequence hydropathy decoration (SHD) metric, which quantifies the spatial clustering of hydrophobic residues along the chain (Zheng et al., 2020). Generated sequences exhibit substantially lower SHD values than folded CATH domains (Figure 2d), consistent with the reduced hydrophobic clustering in disordered regions that prevents hydrophobic collapse (Zheng et al., 2020). This trend closely matches natural DisProt and training set IDRs, and it contrasts with folded CATH domains, which exhibit higher SHD values, reflecting the locally concentrated hydrophobic residues required to stabilize buried protein cores (Dyson et al., 2006).

Finally, we assessed sequence complexity using the SEG algorithm, which computes the average compositional entropy over a sliding window (Wootton & Federhen, 1993). IDRs frequently contain low-complexity segments (Romero et al., 2001), and natural DisProt and training set IDRs show lower complexity than folded CATH domains. Generated IDRs and IDPs closely reproduce this shift, with the complexities of generated sequences matching the DisProt distribution well (Figure 2e). All together, these results demonstrate that `IDiom` has learned the sequence grammar of disordered regions across multiple metrics, and that generated sequences

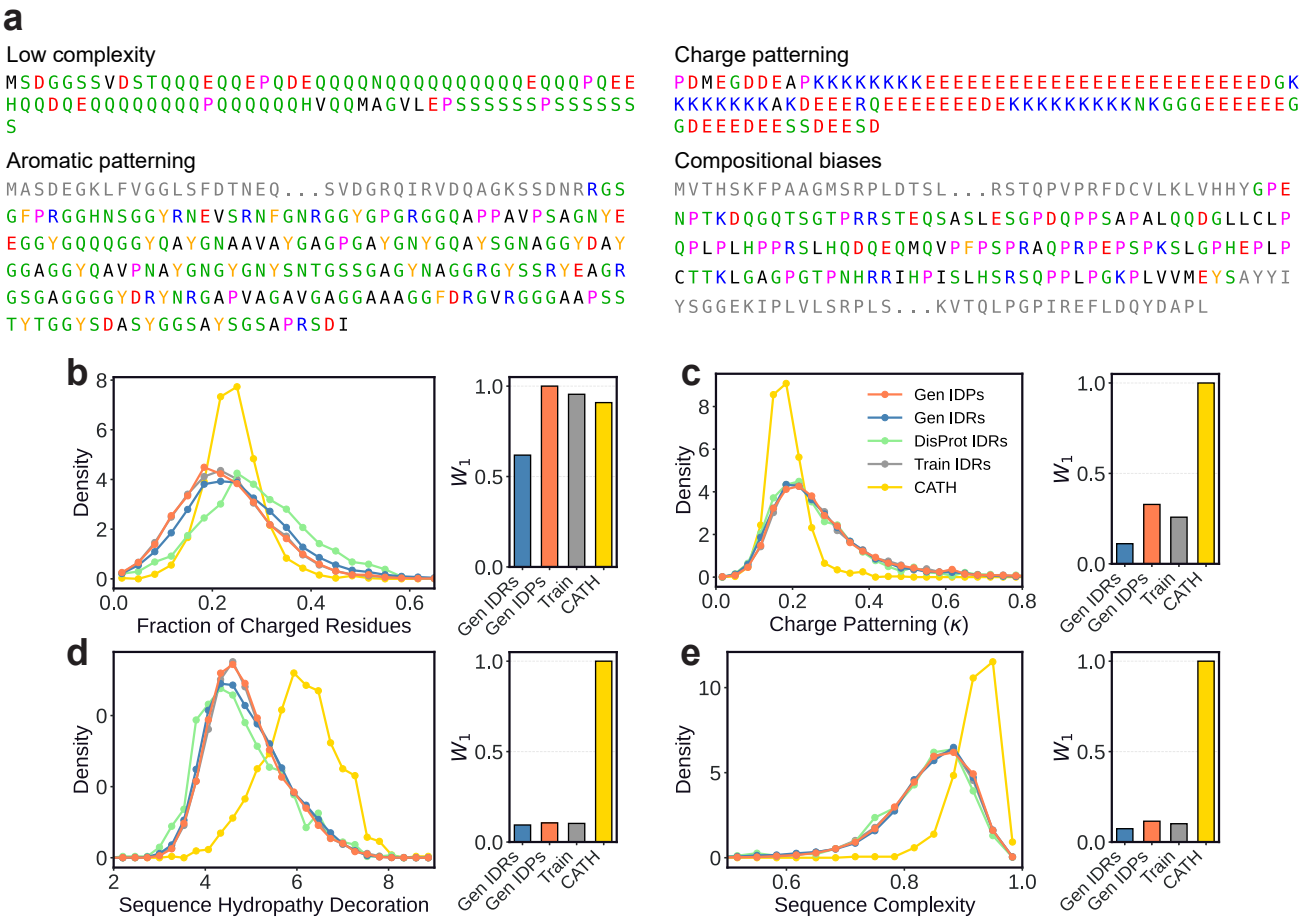

Figure 2. **IDiom generates intrinsically disordered regions and proteins which capture the sequence patterning features of natural sequences. (a)** Example IDPs (upper row) and IDRs (lower row) generated using IDiom which exhibit canonical sequence features of intrinsically disordered regions, including low complexity regions, charge blockiness, aromatic patterning, and compositional biases. Greyed-out residues correspond to the flanking context of the IDR. **(b)–(e)** Distributions of various sequence metrics for generated IDPs and IDRs, training set IDRs, DisProt IDRs, and folded CATH domains. Right subplots: Normalized Wasserstein-1 (W_1) distance between the distributions of DisProt IDRs and all other distributions. **(b)** Fraction of charged residues (FCR). **(c)** Charge patterning κ parameter. **(d)** Sequence hydropathy decoration (SHD). **(e)** Sequence complexity quantified by the SEG algorithm.

closely recapitulate the sequence patterning features of natural IDRs.

2.4. Conditioned generation recapitulates context-specific IDR sequence features

The previous analysis demonstrates that IDiom can generate sequences with patterning features and biophysical properties which are similar to natural IDRs. To quantify the agreement between generated and natural IDRs, we compute the normalized Wasserstein-1 distance (W_1) (Panaretos & Zemel, 2019) between the distributions of each sequence metric and the DisProt reference distribution (right subplots of Figure 2b-e). Across most metrics, W_1 distances for generated sequences are low relative to folded CATH domains, confirming that IDiom produces sequences that are statistically much closer to natural IDRs than to folded proteins. We find that in Figure 2b, the shift in mean values for generated and training sequences relative to DisProt leads to relatively larger W_1 values compared to the other metrics we consider, but we note that the shift relative to the training data likely results from the relatively small set of IDRs in DisProt. We also note that the shape of the broad distribution of FCR values for generated and training sequences matches that of DisProt more closely than the narrower CATH distribution. Furthermore, W_1 distances for DisProt context-prompted IDRs are consistently lower than for unprompted IDPs across all metrics, demonstrating that conditioning on flanking sequence context shifts IDiom's generations toward the sequence features of the natural IDRs that reside inside those contexts. A detailed case study of context-prompted generation using the NPM1 protein is provided in the Supplementary Information.

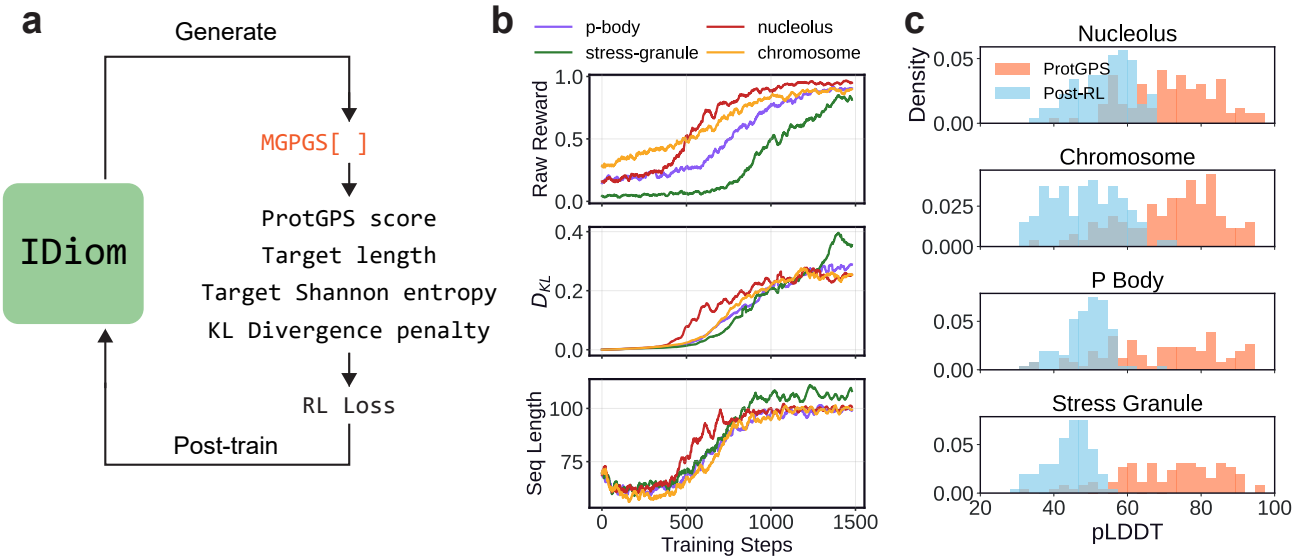

Figure 3. **Post-training `IDiom` with reinforcement learning enables the generation of sequences aligned with an external reward model. (a)** Schematic depicting the reinforcement learning process for post-training `IDiom`. ProtGPS is a reward model which returns a score between 0 and 1 indicating the probability of a given sequence localizing to a chosen subcellular compartment. Quadratic penalties are applied to sequences which deviate from the target ProtGPS score, target length, and target Shannon entropy, and a penalty is applied for an increased Kullback-Leibler (KL) divergence of the post-trained model from the pre-trained base model. These metrics are combined in a reinforcement learning loss which is used to update the `IDiom` policy. **(b)** Training curves depicting the ProtGPS score for a given compartment, magnitude of the KL-divergence, D_{KL}, and the average generated sequence length, as a function of post-training optimizer steps. **(c)** Distribution of AlphaFold2 pLDDTs of sequences generated from `IDiom` checkpoints after 1500 post-training optimizer steps, as well as pLDDTs of the original sequences used to train the ProtGPS reward model.

2.5. Sequence optimization using reinforcement learning

Our results demonstrate that `IDiom` has learned a strong generative prior over IDR sequence space, and that the model produces diverse and biologically realistic sequences. This naturally positions the pre-trained model as a starting point for sequence design: post-training via reinforcement learning allows us to steer generation towards sequences that score well on specified objectives while remaining IDR-like. This approach allows the model to optimize arbitrary reward functions such as computational predictors and reward models trained on experimental data, while incorporating explicit regularization to control sequence diversity, length, and deviation from the base model (Ouyang et al., 2022; Ferruz & Höcker, 2022; Xiong et al., 2025).

As a design target, we focus on subcellular localization, as the ability to engineer protein localization could enable both targeted delivery of therapeutics and modulation of synthetic condensates (Ng et al., 2024; Dai et al., 2023). As the reward model, we use ProtGPS, a neural network trained with ESM2 embeddings to predict the probability of a given protein sequence localizing to each of twelve specific subcellular compartments (Kilgore et al., 2025; Lin et al., 2023). Here, we post-train `IDiom` to optimize localization to four compartments: the nucleolus, chromosomes/chromatin, P-bodies, and stress granules. These compartments were chosen because they are known to be enriched in proteins with IDR-specific sequence features, including charge segregation, RNA-interacting motifs, nuclear localization signals, and post-translational modification sites, thus providing a clear test of whether RL post-training can induce the generation of biologically relevant sequence features.

We optimize `IDiom` using the reinforcement learning algorithm Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Liu et al., 2025), and we generate unprompted IDPs in all post-training runs (overview in Figure 3a). To prevent reward hacking and to maintain sequence diversity, we incorporate three forms of regularization: a Kullback-Leibler (KL) divergence penalty D_{KL} to prevent excess divergence of the post-trained model from the pre-trained base model, a quadratic penalty around a target sequence Shannon entropy of $H = 2.7$ nats to prevent diversity collapse, and a quadratic penalty around a target sequence length of 100 residues. Figure 3b shows training curves for the ProtGPS reward, D_{KL}, and mean sequence length versus post-training optimizer steps. The ProtGPS reward increases steadily across all four compartments, the mean sequence lengths converge to the target value, and D_{KL} remains below 0.4 throughout training, confirming that post-training successfully optimizes the reward without diverging substantially from the pre-trained base distribution.

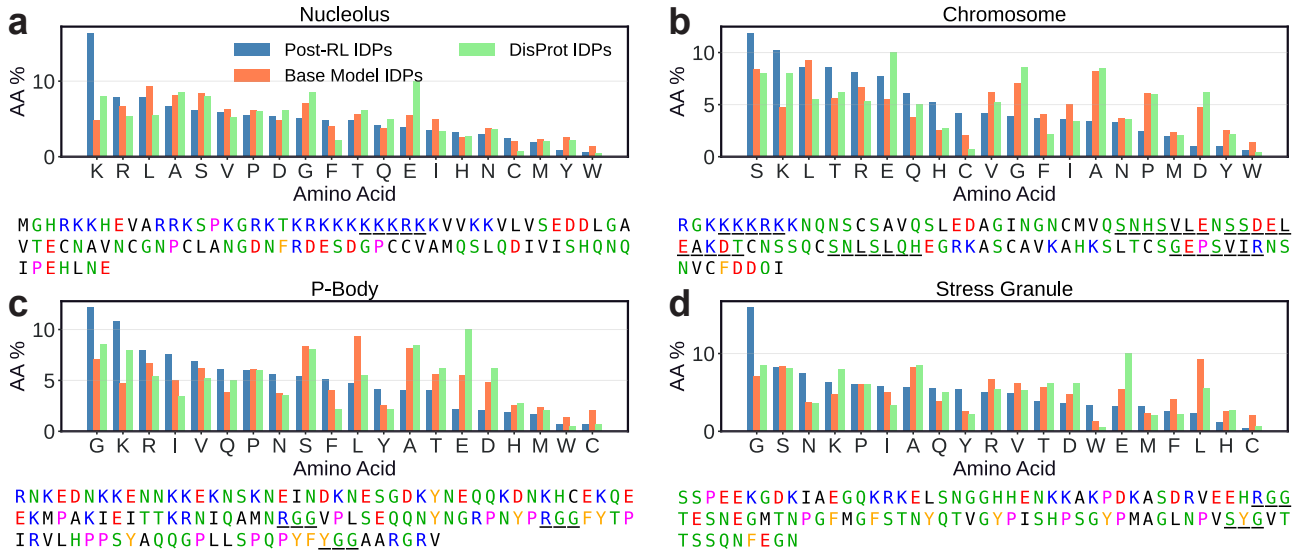

Figure 4. **Post-trained `IDiom` models generate sequences which reproduce the amino acid compositional biases known to drive subcellular localization.** Amino acid composition of sequences generated after post-training `IDiom` to optimize the ProtGPS localization score for the **(a)** nucleolus, **(b)** chromosome, **(c)** P-bodies, and **(d)** stress granules. Underneath each plot is one representative generated sequence for each compartment, with sequence features such as nuclear localization signals, post-translational modification sites, and RNA-binding motifs underlined.

Figure 3c shows the distributions of AlphaFold2 pLDDT values for sequences generated from each post-trained checkpoint, alongside the pLDDT distribution of the full length proteins used to train the ProtGPS predictor. The ProtGPS training sequences have a wide distribution of pLDDTs, with a substantial fraction of high-pLDDT residues, reflecting the fact that ProtGPS was trained on full length proteins containing both folded domains and disordered regions. However, sequences generated by the post-trained `IDiom` models maintain low pLDDT values, comparable to those of natural IDRs, demonstrating that the KL regularization successfully prevents the model from drifting towards the sequence features of folded proteins in the ProtGPS training set. This further confirms that post-training steers generation towards the desired compartment localization while preserving the disordered nature of generated sequences.

2.6. Post-trained models generate sequences with compartment-specific features

After post-training with the ProtGPS reward, we generate 10,000 sequences from each localization-optimized checkpoint and analyze their amino acid compositions. Sequences targeting specific subcellular compartments are expected to have compositional biases that reflect their local biochemical environments, and we find that post-trained generations exhibit biologically interpretable compositional shifts relative to both base model generations and DisProt IDRs (Figure 4).

Nucleolar-targeting sequences are enriched in lysine and arginine, consistent with the prevalence of positively charged nuclear localization signals and the charge-rich low-complexity regions found in nucleolar proteins (Musinova et al., 2015; Martin et al., 2015). Sequences targeting the chromosomes are enriched in serine and threonine, consistent with the high density of phosphorylation sites characteristic of chromatin-associating proteins, which are heavily regulated by post-translational modification (Taverna et al., 2007; Bannister & Kouzarides, 2011). The generated P-body targeting sequences are glycine-rich and basic (lysine- and arginine-rich), consistent with the RNA-binding motifs and arginine-glycine rich regions present in proteins that associate with RNA granules (Chowdhury & Jin, 2023; Thandapani et al., 2013). Finally, stress granule-targeting sequences are similarly glycine-rich, consistent with the low-complexity, RNA-interacting motifs expected in stress granule associating proteins (Millar et al., 2023). Representative sequences generated from each checkpoint are also shown in Figure 4, with specific sequence features such as nuclear localization signals, post-translational modification sites, and RNA-binding motifs underlined, illustrating that the global compositional shifts are accompanied by the generation of specific local sequence features.

To further characterize the sequence features learned during post-training, we analyze the compartment-specific generations for sequence patterning and sequence-specific motifs. Charge segregation, quantified by κ, varies across compartments (Figure 5a), with nucleolus-targeting sequences showing elevated κ relative to all other sequences; this ob-

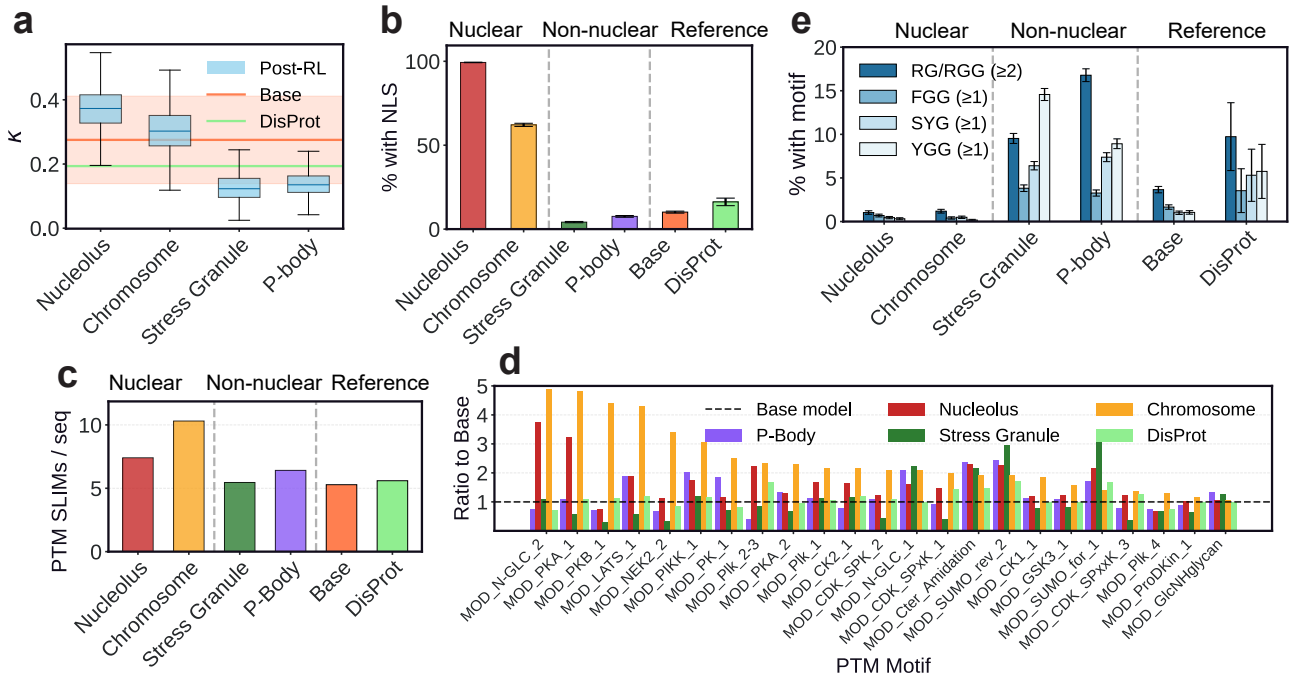

Figure 5. **Proteins generated from post-trained `IDiom` models recapitulate specific sequences features which are characteristic of each target cellular compartment. (a)** Box plot of the distribution of κ charge patterning values for sequences generated from `IDiom` checkpoints optimized for localization to the four labeled ProtGPS compartments. The horizontal orange line and band indicates the mean and standard deviation of κ for sequences generated from the base pre-trained model. The green line indicates the mean κ value of natural DisProt IDRs. **(b)** Plot of the percentage of generated sequences which contain at least one nuclear localization signal from the Eukaryotic Linear Motif (ELM) Resource. **(c)** Plot of the average number of unique post-translational modification (PTM) motifs from the ELM Resource per sequence for sequences generated from various model checkpoints or the DisProt IDRs. **(d)** Plot of the ratio of the number of counts of all PTM motifs from the ELM Resource (MOD) appearing in a generated sequence or DisProt IDRs, normalized to the counts within sequences generated by the base model. **(e)** Plot of the percentage of generated sequences which contain the RNA-interaction sequence motifs indicated in the legend. (≥ 1) indicates the presence of at least one motif within a given sequence. (≥ 2) indicates the presence of at least two motifs within 30 residues of one another within a given sequence.

servation is consistent with the prevalence of charge-block architectures in nucleolar-associating IDRs (Miyagi et al., 2022). In contrast, sequences targeting stress granules and P-bodies show reduced κ relative to baselines, consistent with the weakly charged nature of stress granule proteins (Molliex et al., 2015) and the aromaticity-driven nature of P-body condensate formation (Martin et al., 2020). In addition, we find that sequences targeting the nucleolus and chromosomes are enriched in nuclear localization signals (NLSs), as is expected for these nuclear compartments (Ba¨uerle et al., 2002). We scan generated sequences for a curated set of NLS patterns from the Eukaryotic Linear Motif (ELM) Resource (Kumar et al., 2024) (patterns in the Supplementary Information), and we find that a substantially higher fraction of nucleolus- and chromosome-targeting sequences contain at least one NLS compared to sequences targeting cytoplasmic compartments (Figure 5b).

Chromosome-associating proteins are heavily regulated through post-translational modifications (PTMs), and we probe whether chromosome-targeting sequences are correspondingly enriched in PTM motifs (Pejaver et al., 2014). We scan generated sequences for all 40 PTM motif classes from the ELM Resource (Supplementary Information). Chromosome-targeting sequences show a pronounced increase in MOD-motif density, with over 10 putative unique PTM sites per sequence on average, compared to approximately 5-7 sites per sequence for other compartments (Figure 5c). In Figure 5d we quantify enrichment as the ratio of motif counts in post-trained generations relative to the base model, finding that the enriched motifs span multiple kinase families, including AGC-class sites, PIK/PIKK-associated phosphorylation sites, acidophilic CK2 motifs, and proline-directed CDK-class motifs. This broad enrichment across kinase families is consistent with PTM-driven regulatory control being a crucial role of chromatin-interacting proteins (Taverna et al., 2007; Bannister & Kouzarides, 2011), and it demonstrates that post-training with a localization reward is sufficient to induce the generation of these specific sequence features.

P-bodies and stress granules are RNA-rich condensates

with central roles in mRNA regulation and decay (Decker & Parker, 2012), and we find that post-trained sequences targeting these compartments are enriched in short RNA-interaction motifs including RG/RGG tracts, F/YGG motifs, and SYG motifs (Figure 5e). RG/RGG-rich regions are widely implicated in RNA binding and recruitment to ribonucleoprotein assemblies (Chowdhury & Jin, 2023; Thandapani et al., 2013), while F/YGG (Van Lindt et al., 2022; Sänger et al., 2021) and SYG (Bressin et al., 2019; Kato et al., 2012) motifs provide aromatic sticker elements that mediate π-π and cation-π interactions which drive phase separation in low-complexity IDRs. The emergence of these motifs through post-training indicates that optimizing for RNA granule localization is also sufficient to induce the generation of RNA-interaction sequence grammars.

Altogether, these results demonstrate that RL post-training using the ProtGPS predictor as a reward model is able to teach `IDiom` the global amino acid compositions, sequence patterning, and specific motifs which are necessary for compartment-specific localization. Each of the changes we identify is biologically interpretable and consistent with the known sequence determinants of localization to or interaction with the corresponding compartment, and these features emerge without any explicit supervision. This indicates that the post-training of `IDiom` to optimize the ProtGPS score alone is sufficient to learn the necessary compartment-specific sequence grammars for subcellular localization of IDRs.

3. Discussion

`IDiom` demonstrates that a protein language model trained exclusively on intrinsically disordered region sequences can faithfully capture the highly contextual evolutionary statistics of natural IDRs. The pre-trained model generates diverse sequences that recapitulate the compositional biases, charge patterning, hydrophobic decoration, and low-complexity features of natural disordered regions. Crucially, the model is able to generate highly plausible disordered sequences while remaining substantially dissimilar to the training examples. Furthermore, `IDiom` learns in-context and is able to generate IDRs conditioned on flanking sequence context that are more similar to the natural IDRs that exist within that flanking context, as demonstrated by the DisProt context-prompted IDRs.

Beyond generation from the base pre-trained model, we demonstrate that `IDiom` can be post-trained using reinforcement learning to optimize arbitrary external reward functions. While here we demonstrate this using ProtGPS as the reward signal for subcellular localization, the post-training approach is general. `IDiom` can be combined with any objective, such as sequence-based predictors of conformational ensembles (Novak et al., 2026), intermolecular interactions (Ginell et al., 2025), phase behavior (Von Bülow et al., 2025), as well as machine-learned predictors of experimentally characterized biological function such as transcriptional activity (Sanborn et al., 2021; DelRosso et al., 2023) or direct binding affinity (DelRosso et al., 2024). Post-training can also be applied directly to experimental data, using approaches such as direct preference optimization (Xiong et al., 2025; Rafailov et al., 2024) or energy rank alignment (Ibarraran et al., 2026; Chennakesavalu et al., 2025). Such approaches would enable iterative optimization of the generative model as new experimental measurements become available. Due to the often modular nature of IDRs, a designed IDR could be inserted or appended to a full length protein to tune localization, phase behavior, or signaling properties without redesigning the folded domains. Additionally, the prevalence of short IDRs in our training data (Figure 1e) also makes `IDiom` well-suited for the design of disordered peptides for therapeutic applications. Finally, `IDiom` offers interesting opportunities for shrinking proteins containing functional IDRs, an important objective in protein delivery (Baron et al., 2026).

`IDiom` can also enable the automated discovery of evolutionarily and biologically important sequence features within IDRs. In the examples with the ProtGPS reward model, we manually identified sequence features to analyze in the post-trained sequences. Future work applying techniques such as feature learning with sparse autoencoders (Cunningham et al., 2023) would enable automatic identification of prominent sequence features learned by `IDiom`. While prior work has studied the features learned by protein language models trained on full length sequences, intrinsically disordered regions are poorly resolved in those learned feature spaces (Simon & Zou, 2025). In contrast, sparse autoencoders trained on `IDiom` representations would be free from folded domain features, allowing for more precise identification of sequence grammars that underlie intrinsically disordered regions. When applied to post-trained models, this approach could further reveal the function-specific features that underlie biological function, such as subcellular localization.

Intrinsically disordered regions play central roles in cellular processes such as gene regulation, subcellular compartmentalization, and signaling, yet they have remained largely inaccessible to rational design. `IDiom` provides a generative framework for disordered sequence design that can be steered towards specific functional objectives through post-training. Combined with high-throughput experimental assays and machine learned reward models trained on experimental data, this platform offers a path toward the systematic design and engineering of intrinsically disordered proteins and regions, opening new avenues in synthetic biology, targeted therapeutics, and the design of programmable condensates.

Acknowledgements

J.X.L. acknowledges support from the National Institute of Health T32 award number T32HL094274. S.I. acknowledges support from the National Science Foundation Graduate Research Fellowship Program and the Shoucheng Zhang Graduate Fellowship. F.H. acknowledges support from the Stanford Center for Molecular Analysis and Design Fellowship. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number 1R35GM159834-01 (G.M.R.) and 5R35GM130332 (A.R.D.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge the use of the Stanford Sherlock compute cluster, and the authors acknowledge insightful discussions with Prof. Mikko Haataja.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w.

Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5):285–298, May 2017. ISSN 1471-0080. doi: 10.1038/nrm.2017.7. URL https://www.nature.com/articles/nrm.2017.7.

Bannister, A. J. and Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, March 2011. ISSN 1748-7838. doi: 10.1038/cr.2011.22. URL https://www.nature.com/articles/cr201122.

Baron, E., Amin, A. N., Weitzman, R., d'Oelsnitz, S., Marks, D. S., and Wilson, A. G. Shrinking proteins with diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=quxeCxJwKm.

Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., and Chen, M. Efficient Training of Language Models to Fill in the Middle, July 2022. URL http://arxiv.org/abs/2207.14255. arXiv:2207.14255 [cs].

Ba¨uerle, M., Doenecke, D., and Albig, W. The Requirement of H1 Histones for a Heterodimeric Nuclear Import Receptor*. *Journal of Biological Chemistry*, 277(36):32480–32489, September 2002. ISSN 0021-9258. doi: 10.1074/jbc.M202765200. URL https://www.sciencedirect.com/science/article/pii/S002192582074384X.

Bhatnagar, A., Jain, S., Beazer, J., Curran, S. C., Hoffnagle, A. M., Ching, K. S., Martyn, M., Nayfach, S., Ruffolo, J. A., and Madani, A. Scaling Unlocks Broader Generation and Deeper Functional Understanding of Proteins. *NeurIPS*, 2025.

Bressin, A., Schulte-Sasse, R., Figini, D., Urdaneta, E. C., Beckmann, B. M., and Marsico, A. TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Research*, 47(9):4406–4417, May 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz203. URL https://doi.org/10.1093/nar/gkz203.

Chennakesavalu, S., Hu, F., Ibarraran, S., and Rotskoff, G. M. Aligning Transformers with Continuous Feedback via Energy Rank Alignment, October 2025. URL http://arxiv.org/abs/2405.12961. arXiv:2405.12961 [cs].

Chowdhury, M. N. and Jin, H. The RGG motif proteins: Interactions, functions, and regulations. *WIREs RNA*, 14(1): e1748, 2023. ISSN 1757-7012. doi: 10.1002/wrna.1748. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1748. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1748.

Chu, A. E., Kim, J., Cheng, L., El Nesr, G., Xu, M., Shuai, R. W., and Huang, P.-S. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, July 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2311500121. URL https://pnas.org/doi/10.1073/pnas.2311500121.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

Dai, Y., You, L., and Chilkoti, A. Engineering synthetic biomolecular condensates. *Nature Reviews Bioengineering*, pp. 1–15, April 2023. ISSN 2731-6092. doi: 10.1038/s44222-023-00052-6. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10107566/.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. URL http://arxiv.org/abs/2205.14135. arXiv:2205.14135 [cs].

Das, R. K. and Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, August 2013. doi: 10.1073/pnas.1304749110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1304749110.

Das, R. K., Ruff, K. M., and Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 32:102–112, June 2015. ISSN 0959-440X. doi: 10.1016/j.sbi.2015.03.008. URL https://www.sciencedirect.com/science/article/pii/S0959440X15000354.

Decker, C. J. and Parker, R. P-Bodies and Stress Granules: Possible Roles in the Control of Translation and mRNA Degradation. *Cold Spring Harbor Perspectives in Biology*, 4(9):a012286, September 2012. ISSN 1943-0264. doi: 10.1101/cshperspect.a012286. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3428773/.

DelRosso, N., Tycko, J., Suzuki, P., Andrews, C., Aradhana, Mukund, A., Liongson, I., Ludwig, C., Spees, K., Fordyce, P., Bassik, M. C., and Bintu, L. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature*, 616(7956):365–372, April 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05906-y. URL https://www.nature.com/articles/s41586-023-05906-y.

DelRosso, N., Suzuki, P. H., Griffith, D., Lotthammer, J. M., Novak, B., Kocalar, S., Sheth, M. U., Holehouse, A. S., Bintu, L., and Fordyce, P. High-throughput affinity measurements of direct interactions between activation domains and co-activators, August 2024. URL https://www.biorxiv.org/content/10.1101/2024.08.19.608698v1. Pages: 2024.08.19.608698 Section: New Results.

Dyson, H. J., Wright, P. E., and Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences*, 103(35):13057–13061, August 2006. doi: 10.1073/pnas.0605504103. URL https://www.pnas.org/doi/10.1073/pnas.0605504103.

Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C. C.-H., Eckmann, C. R., Myong, S., and Brangwynne, C. P. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences*, 112(23):7189–7194, June 2015. doi: 10.1073/pnas.1504822112. URL https://www.pnas.org/doi/10.1073/pnas.1504822112.

Erdős, G., Pajkos, M., and Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, 49(W1):W297–W303, July 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab408. URL https://doi.org/10.1093/nar/gkab408.

Ferrolino, M. C., Mitrea, D. M., Michael, J. R., and Kriwacki, R. W. Compositional adaptability in NPM1-SURF6 scaffolding networks enabled by dynamic switching of phase separation mechanisms. *Nature Communications*, 9(1):5064, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07530-1. URL https://www.nature.com/articles/s41467-018-07530-1.

Ferruz, N. and Höcker, B. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, June 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00499-z. URL https://www.nature.com/articles/s42256-022-00499-z.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32007-7. URL https://www.nature.com/articles/s41467-022-32007-7.

Ginell, G. M., Emenecker, R. J., Lotthammer, J. M., Keeley, A. T., Plassmeyer, S. P., Razo, N., Usher, E. T., Pelham, J. F., and Holehouse, A. S. Sequence-based prediction of intermolecular interactions driven by disordered regions. *Science*, 388(6749):eadq8381, May 2025. doi: 10.1126/science.adq8381. URL https://www.science.org/doi/10.1126/science.adq8381.

González-Foutel, N. S., Glavina, J., Borcherds, W. M., Safranchik, M., Barrera-Vilarmau, S., Sagar, A., Estaña, A., Barozet, A., Garrone, N. A., Fernandez-Ballester, G., Blanes-Mira, C., Sánchez, I. E., de Prat-Gay, G., Cortés, J., Bernadó, P., Pappu, R. V., Holehouse, A. S., Daughdrill, G. W., and Chemes, L. B. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nature Structural & Molecular Biology*, 29(8):781–790, August 2022. ISSN 1545-9985. doi: 10.1038/s41594-022-00811-w. URL https://www.nature.com/articles/s41594-022-00811-w.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, February 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/10.1126/science.ads0018.

Holehouse, A. S. sparrow: a tool for integrative analysis and prediction from protein sequence data. *Zenodo*, July 2022. doi: 10.5281/zenodo.6891920. URL https://ui.adsabs.harvard.edu/abs/2022zndo...6891920H. ADS Bibcode: 2022zndo...6891920H.

Holehouse, A. S. and Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 25(3):187–211, March 2024. ISSN 1471-0080. doi: 10.1038/s41580-023-00673-0. URL https://www.nature.com/articles/s41580-023-00673-0.

Hunter, K., Brandt, T., Guadalupe, K., Kolamunna, K. C., Lotthammer, J. M., Shamoon, N. M., Nicholson, B., Day, L., Martinez, A., Holehouse, A. S., Sukenik, S., and Emenecker, R. J. Rational design of disordered proteins for systematic sequence-to-function investigation, February 2026. URL https://www.biorxiv.org/content/10.1101/2023.10.29.564547v3. ISSN: 2692-8205 Pages: 2023.10.29.564547 Section: New Results.

Ibarraran, S., Chennakesavalu, S., Hu, F., and Rotskoff, G. M. Efficient, Few-shot Directed Evolution with Energy Rank Alignment, February 2026. URL https://www.biorxiv.org/content/10.64898/2026.02.03.703561v1. ISSN: 2692-8205 Pages: 2026.02.03.703561 Section: New Results.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Kato, M., Han, T., Xie, S., Shi, K., Du, X., Wu, L., Mirzaei, H., Goldsmith, E., Longgood, J., Pei, J., Grishin, N., Frantz, D., Schneider, J., Chen, S., Li, L., Sawaya, M., Eisenberg, D., Tycko, R., and McKnight, S. Cell-free Formation of RNA Granules: Low Complexity Sequence

Domains Form Dynamic Fibers within Hydrogels. *Cell*, 149(4):753–767, May 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.04.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867412005144.

Kilgore, H. R., Chinn, I., Mikhael, P. G., Mitnikov, I., Van Dongen, C., Zylberberg, G., Afeyan, L., Banani, S. F., Wilson-Hawken, S., Lee, T. I., Barzilay, R., and Young, R. A. Protein codes promote selective subcellular compartmentalization. *Science*, 387(6738):1095–1101, March 2025. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adq2634. URL https://www.science.org/doi/10.1126/science.adq2634.

Krueger, R., Brenner, M. P., and Shrinivas, K. Generalized design of sequence-ensemble-function relationships for intrinsically disordered proteins, October 2024. URL https://www.biorxiv.org/content/10.1101/2024.10.10.617695v1. Pages: 2024.10.10.617695 Section: New Results.

Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J., Kalman, Z., Pascarelli, S., Palopoli, N., Dobson, L., Suarez, C., Van Roey, K., Krystkowiak, I., Griffin, J., Nagpal, A., Bhardwaj, R., Diella, F., Mészáros, B., Dean, K., Davey, N., Pancsa, R., Chemes, L., and Gibson, T. ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research*, 52(D1):D442–D455, January 2024. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkad1058. URL https://academic.oup.com/nar/article/52/D1/D442/7420098.

Langstein-Skora, I., Schmid, A., Huth, F., Shabani, D., Spechtenhauser, L., Likhodeeva, M., Kunert, F., Metzner, F. J., Emenecker, R. J., Richardson, M. O., Aftab, W., Götz, M. J., Payer, S. K., Pietrantoni, N., Valka, V., Ravichandran, S. K., Bartke, T., Hopfner, K.-P., Gerland, U., Korber, P., and Holehouse, A. S. Sequence and chemical specificity define the functional landscape of intrinsically disordered regions. *Nature Cell Biology*, 28(2):323–337, February 2026. ISSN 1476-4679. doi: 10.1038/s41556-025-01867-8. URL https://www.nature.com/articles/s41556-025-01867-8.

Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Davaadorj, M., Lamy-Poirier, J., Monteiro, J., Shliazhko, O., Gontier, N., Meade, N., Zebaze, A., Yee, M.-H., Umapathi, L. K., Zhu, J., Lipkin, B., Oblokulov, M., Wang, Z., Murthy, R., Stillerman, J., Patel, S. S., Abulkhanov, D., Zocca, M., Dey, M., Zhang, Z., Fahmy, N., Bhattacharyya, U., Yu, W., Singh, S., Luccioni, S., Villegas, P., Kunakov, M., Zhdanov, F., Romero, M., Lee, T., Timor, N., Ding, J., Schlesinger, C., Schoelkopf, H., Ebert, J., Dao, T., Mishra, M., Gu, A., Robinson, J., Anderson, C. J., Dolan-Gavitt, B., Contractor, D., Reddy, S., Fried, D., Bahdanau, D., Jernite, Y., Ferrandis, C. M., Hughes, S., Wolf, T., Guha, A., Werra, L. v., and Vries, H. d. StarCoder: may the source be with you!, December 2023. URL http://arxiv.org/abs/2305.06161. arXiv:2305.06161 [cs].

Lin, Y.-H., Brady, J. P., Forman-Kay, J. D., and Chan, H. S. Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New Journal of Physics*, 19(11):115003, November 2017. ISSN 1367-2630. doi: 10.1088/1367-2630/aa9369. URL https://doi.org/10.1088/1367-2630/aa9369.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding R1-Zero-Like Training: A Critical Perspective, March 2025. URL http://arxiv.org/abs/2503.20783. arXiv:2503.20783 [cs] version: 1.

Lotthammer, J. M., Hernández-García, J., Griffith, D., Weijers, D., Holehouse, A. S., and Emenecker, R. J. Metapredict enables accurate disorder prediction across the Tree of Life, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.05.622168v1. Pages: 2024.11.05.622168 Section: New Results.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, August 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2.

Mantonico, M. V., De Leo, F., Quilici, G., Colley, L. S., De Marchis, F., Crippa, M., Mezzapelle, R., Schulte, T., Zucchelli, C., Pastorello, C., Carmeno, C., Caprioglio, F., Ricagno, S., Giachin, G., Ghitti, M., Bianchi, M. E., and Musco, G. The acidic intrinsically disordered region

of the inflammatory mediator HMGB1 mediates fuzzy interactions with CXCL12. *Nature Communications*, 15 (1):1201, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45505-7. URL https://www.nature.com/articles/s41467-024-45505-7.

Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, 367(6478):694–699, February 2020. doi: 10.1126/science.aaw8653. URL https://www.science.org/doi/10.1126/science.aaw8653.

Martin, R. M., Ter-Avetisyan, G., Herce, H. D., Ludwig, A. K., Lättig-Tünnemann, G., and Cardoso, M. C. Principles of protein targeting to the nucleolus. *Nucleus*, 6(4):314–325, August 2015. ISSN 1949-1034. doi: 10.1080/19491034.2015.1079680. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4615656/.

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J., and Pande, V. S. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015. doi: 10.1016/j.bpj.2015.08.015.

Millar, S. R., Huang, J. Q., Schreiber, K. J., Tsai, Y.-C., Won, J., Zhang, J., Moses, A. M., and Youn, J.-Y. A New Phase of Networking: The Molecular Composition and Regulatory Dynamics of Mammalian Stress Granules. *Chemical Reviews*, 123(14):9036–9064, July 2023. ISSN 0009-2665. doi: 10.1021/acs.chemrev.2c00608. URL https://doi.org/10.1021/acs.chemrev.2c00608.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://www.nature.com/articles/s41592-022-01488-1.

Mitrea, D. M., Cika, J. A., Stanley, C. B., Nourse, A., Onuchic, P. L., Banerjee, P. R., Phillips, A. H., Park, C.-G., Deniz, A. A., and Kriwacki, R. W. Self-interaction of NPM1 modulates multiple mechanisms of liquid–liquid phase separation. *Nature Communications*, 9(1):842, February 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03255-3. URL https://www.nature.com/articles/s41467-018-03255-3.

Miyagi, T., Yamazaki, R., Ueda, K., Narumi, S., Hayamizu, Y., Uji-i, H., Kuroda, M., and Kanekura, K. The Patterning and Proportion of Charged Residues in the Arginine-Rich Mixed-Charge Domain Determine the Membrane-Less Organelle Targeted by the Protein. *International Journal of Molecular Sciences*, 23(14):7658, January 2022. ISSN 1422-0067. doi: 10.3390/ijms23147658. URL https://www.mdpi.com/1422-0067/23/14/7658.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell*, 163(1):123–133, September 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.09.015. URL https://www.cell.com/cell/abstract/S0092-8674(15)01176-9.

Musinova, Y. R., Kananykhina, E. Y., Potashnikova, D. M., Lisitsyna, O. M., and Sheval, E. V. A charge-dependent mechanism is responsible for the dynamic accumulation of proteins inside nucleoli. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1853(1):101–110, January 2015. ISSN 0167-4889. doi: 10.1016/j.bbamcr.2014.10.007. URL https://www.sciencedirect.com/science/article/pii/S0167488914003668.

Ng, C. S. C., Liu, A., Cui, B., and Banik, S. M. Targeted protein relocalization via protein transport coupling. *Nature*, 633(8031):941–951, September 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07950-8. URL https://www.nature.com/articles/s41586-024-07950-8.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, November 2023. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2023.10.002. URL https://www.cell.com/cell-systems/abstract/S2405-4712(23)00272-7.

Novak, B., Lotthammer, J. M., Emenecker, R. J., and Holehouse, A. S. Accurate predictions of disordered protein ensembles with STARLING. *Nature*, pp. 1–11, February 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10141-2. URL https://www.nature.com/articles/s41586-026-10141-2.

Nugnes, M. V., Bouhraoua, K. E. A., Zoubiri, M., Pancsa, R., Fichó, E., DisProt Consortium, Tompa, P., Piovesan, D., Tosatto, S. E., and Aspromonte, M. C. DisProt in 2026: enhancing intrinsically disordered proteins accessibility, deposition, and annotation. *Nucleic Acids Research*, 54 (D1):D383–D392, January 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1175. URL https://doi.org/10.1093/nar/gkaf1175.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., Cho, Y., Ghamary, K. H., Vinué, L., Yachnin, B. J., Wollacott, A. M., Buckley, S., Westphal, A. H., Lindhoud, S., Georgeon, S., Goverde, C. A., Hatzopoulos, G. N., Gönczy, P., Muller, Y. D., Schwank, G., Swarts, D. C., Vecchio, A. J., Schneider, B. L., Ovchinnikov, S., and Correia, B. E. One-shot design of functional protein binders with BindCraft. *Nature*, 646(8084):483–492, October 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09429-6. URL https://www.nature.com/articles/s41586-025-09429-6.

Panaretos, V. M. and Zemel, Y. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):405–431, March 2019. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-030718-104938. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-030718-104938.

Pejaver, V., Hsu, W.-L., Xin, F., Dunker, A. K., Uversky, V. N., and Radivojac, P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Science*, 23(8):1077–1093, 2014. ISSN 1469-896X. doi: 10.1002/pro.2494. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.2494. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.2494.

Pesce, F., Bremer, A., Tesei, G., Hopkins, J. B., Grace, C. R., Mittag, T., and Lindorff-Larsen, K. Design of intrinsically disordered protein variants with diverse structural properties. *Science Advances*, 10(35):eadm9926, August 2024. doi: 10.1126/sciadv.adm9926. URL https://www.science.org/doi/10.1126/sciadv.adm9926.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics*, 42(1):38–48, 2001. ISSN 1097-0134. doi: 10.1002/1097-0134(20010101)42:1⟨38::AID-PROT50⟩3.0.CO;2-3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0134%2820010101%2942%3A1%3C38%3A%3AAID-PROT50%3E3.0.CO%3B2-3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0134%2820010101%2942%3A1%3C38%3A%3AAID-PROT50%3E3.0.CO%3B2-3.

Ruff, K. M. and Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *Journal of Molecular Biology*, 433(20):167208, October 2021. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167208. URL https://www.sciencedirect.com/science/article/pii/S0022283621004411.

Ruff, K. M., King, M. R., Ying, A. W., Liu, V., Pant, A., Lieberman, W. E., Shinn, M. K., Su, X., Kadoch, C., and Pappu, R. V. Molecular grammars of predicted intrinsically disordered regions that span the human proteome. *Cell*, 189(1):323–342.e17, January 2026. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2025.10.019. URL https://www.cell.com/cell/abstract/S0092-8674(25)01191-2.

Sanborn, A. L., Yeh, B. T., Feigerle, J. T., Hao, C. V., Townshend, R. J., Lieberman Aiden, E., Dror, R. O., and Kornberg, R. D. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife*, 10:e68068, April 2021. ISSN 2050-084X. doi: 10.7554/eLife.68068. URL https://doi.org/10.7554/eLife.68068.

Schrödinger, LLC. *The PyMOL Molecular Graphics System*, 2025. PyMOL (Version 3.1).

Shazeer, N. GLU Variants Improve Transformer, February 2020. URL http://arxiv.org/abs/2002.05202. arXiv:2002.05202 [cs].

Simon, E. and Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, October 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02836-7. URL https://www.nature.com/articles/s41592-025-02836-7.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. RoFormer: Enhanced Transformer with Rotary Position Embedding, November 2023. URL http://arxiv.org/abs/2104.09864. arXiv:2104.09864 [cs].

Sänger, L., Bender, J., Rostowski, K., Golbik, R., Lilie, H., Schmidt, C., Behrens, S.-E., and Friedrich, S. Alternatively spliced isoforms of AUF1 regulate a miRNA–mRNA interaction differentially through their YGG motif. *RNA Biology*, 18(6):843–853, June 2021. ISSN 1547-6286. doi: 10.1080/15476286.202

0.1822637. URL https://doi.org/10.1080/15476286.2020.1822637. _eprint: https://doi.org/10.1080/15476286.2020.1822637.

Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D., and Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature structural & molecular biology*, 14(11): 1025–1040, November 2007. ISSN 1545-9993. doi: 10.1038/nsmb1338. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4691843/.

Tesei, G., Trolle, A. I., Jonsson, N., Betz, J., Knudsen, F. E., Pesce, F., Johansson, K. E., and Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-07004-5. URL https://www.nature.com/articles/s41586-023-07004-5.

Tesei, G., Pesce, F., and Lindorff-Larsen, K. Computational design of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 96:103210, February 2026. ISSN 0959-440X. doi: 10.1016/j.sbi.2025.103210. URL https://www.sciencedirect.com/science/article/pii/S0959440X25002283.

Thandapani, P., O'Connor, T. R., Bailey, T. L., and Richard, S. Defining the RGG/RG Motif. *Molecular Cell*, 50(5): 613–623, June 2013. ISSN 1097-2765. doi: 10.1016/j.molcel.2013.05.021. URL https://www.sciencedirect.com/science/article/pii/S1097276513004085.

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., and Uversky, V. N. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins*, 1 (1):e24360, 2013. ISSN 2169-0693. doi: 10.4161/idp.24360.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13):6589–6631, July 2014. ISSN 0009-2665. doi: 10.1021/cr400525m. URL https://doi.org/10.1021/cr400525m.

Van Lindt, J., Lazar, T., Pakravan, D., Demulder, M., Meszaros, A., Van Den Bosch, L., Maes, D., and Tompa, P. F/YGG-motif is an intrinsically disordered nucleic-acid binding motif. *RNA Biology*, 19(1):622–635, December 2022. ISSN 1547-6286. doi: 10.1080/15476286.2022.2066336. URL https://doi.org/10.1080/15476286.2022.2066336. _eprint: https://doi.org/10.1080/15476286.2022.2066336.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., and Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., and Velankar, S. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Von Bülow, S., Tesei, G., Zaidi, F. K., Mittag, T., and Lindorff-Larsen, K. Prediction of phase-separation propensities of disordered proteins from sequence. *Proceedings of the National Academy of Sciences*, 122(13): e2417920122, April 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2417920122. URL https://pnas.org/doi/10.1073/pnas.2417920122.

Waman, V., Bordin, N., Lau, A., Kandathil, S., Wells, J., Miller, D., Velankar, S., Jones, D., Sillitoe, I., and Orengo, C. CATH v4.4: major expansion of CATH by experimental and predicted structural data. *Nucleic Acids Research*, 53(D1):D348–D355, January 2025. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkae1087. URL https://academic.oup.com/nar/article/53/D1/D348/7905304.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte,

R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8.

Wei, M.-T., Elbaum-Garfinkle, S., Holehouse, A. S., Chen, C. C.-H., Feric, M., Arnold, C. B., Priestley, R. D., Pappu, R. V., and Brangwynne, C. P. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nature Chemistry*, 9(11): 1118–1125, November 2017. ISSN 1755-4349. doi: 10.1038/nchem.2803. URL https://www.nature.com/articles/nchem.2803.

Wilson, C. J., Choy, W.-Y., and Karttunen, M. AlphaFold2: A Role for Disordered Protein/Region Prediction? *International Journal of Molecular Sciences*, 23(9):4591, January 2022. ISSN 1422-0067. doi: 10.3390/ijms23094591. URL https://www.mdpi.com/1422-0067/23/9/4591.

Wootton, J. C. and Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17(2):149–163, June 1993. ISSN 00978485. doi: 10.1016/0097-8485(93)85006-X. URL https://linkinghub.elsevier.com/retrieve/pii/009784859385006X.

Wright, P. E. and Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, January 2015. ISSN 1471-0080. doi: 10.1038/nrm3920. URL https://www.nature.com/articles/nrm3920.

Xiong, J., Nisonoff, H., Lukarska, M., Gaur, I., Oltrogge, L. M., Savage, D. F., and Listgarten, J. Guide your favorite protein sequence generative model, July 2025. URL http://arxiv.org/abs/2505.04823. arXiv:2505.04823 [cs].

Zhao, B., Ghadermarzi, S., and Kurgan, L. Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. *Computational and Structural Biotechnology Journal*, 21:3248–3258, January 2023. ISSN 2001-0370. doi: 10.1016/j.csbj.2023.06.001. URL https://www.sciencedirect.com/science/article/pii/S2001037023002143.

Zheng, W., Dignon, G., Brown, M., Kim, Y. C., and Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *The Journal of Physical Chemistry Letters*, 11 (9):3408–3415, May 2020. doi: 10.1021/acs.jpclett.0c00288. URL https://doi.org/10.1021/acs.jpclett.0c00288.

A. Methods

Data Curation We curate our dataset of intrinsically disordered regions from the AlphaFold Database (AFDB), version 4. First, we use MMseqs2 to cluster the 214M AFDB sequences at 90% identity and 80% coverage (other MMseqs2 parameters below). Next, we follow the method of (Tesei et al., 2024) to determine the sequences and locations of pLDDT-based IDRs within the MMseqs2 cluster representative proteins. In this method, we first apply a 15 residue-wide averaging filter to the pLDDT values. Next, we mark residues with pLDDT > 80 as folded, pLDDT < 70 as disordered, and $70 <$ pLDDT < 80 as gap regions. Folded and disordered regions with a length shorter than ten residues are reclassified as gaps. If a gap region is flanked by two disordered regions, or if it is N- or C-terminal and is adjacent to a disordered region, we relabel it as disordered. All other gap regions are relabeled as folded. Each record within the dataset corresponds to a different IDR, and any given protein may yield ≥ 1 IDR. IDRs which are located in proteins whose full length is greater than 512 residues, IDRs shorter than 30 residues long, and sequences whose entire length is low-pLDDT are discarded. This curation process yields 37M IDRs and their associated N- and C-terminal flanking contexts.

Sequence Clustering and Percent Identity Characterization We use MMseqs2 to cluster sequences in the AlphaFold Database at 90% identity and 80% coverage. MMseqs2 was run using the following command: `mmseqs linclust --min-seq-id 0.9 --cov-mode 0 -c 0.8 --cluster-mode 2`.

We also use MMseqs2 to identify the sequence identity of generated sequences with respect to the training set. Specifically, we compare generated IDR and IDP sequences against the 37M AFDB training set IDRs without their surrounding context. MMseqs2 was run with the following command to search the generated sequences against the AFDB training set IDRs: `mmseqs search --max-seqs 1 -e 1e3 --min-seq-id 0.0 -c 0.0`. For each generated sequence, this command returns the sequence identity of the closest match in the training set, which is plotted in Figure 1.

Tokenization We tokenize protein sequences using a simple alphabet in which each amino acid is represented by a single token. We introduce the tokens `<N>`, `<C>`, and `<I>` to denote the N-terminal flanking context of an IDR, the C-terminal flanking context, and the IDR span itself, respectively. We additionally add the standard beginning-of-sequence `<bos>` and end-of-sequence `<eos>` tokens to all sequences, and we pad all sequences to the maximum length of 512 tokens using the `<pad>` token. The total size of the alphabet we use is 27 tokens (20 amino acids, `<N>`, `<C>`, and `<I>`, and `<bos>`, `<eos>`, `<pad>`, and `<mask>`).

Data Augmentations To process the curated AFDB IDRs for model pre-training, we transform the protein sequences into a fill-in-the-middle format. We prepend the token `<N>` to any N-terminal context of the IDR, prepend `<I>` to the IDR itself, and prepend `<C>` to any C-terminal context of the IDR. We then rearrange the sequence in the order `<N><N-terminal context><C><C-terminal context><I><IDR span>`. An example sequence is `<N>MEDS..HLVA<C>SVED..RKSL<I>VEED..KGPS`.

We augment the data with intrinsically disordered proteins by duplicating the set of IDRs and removing their N- and C-terminal flanking contexts. An example sequence from this data augmentation is `<N><C><I>VEED..KGPS`. In all, this produces 74M sequences for training (37M IDRs and 37M IDPs).

Model Architecture `IDiom` is a 12-layer decoder-only Transformer with 14 attention heads and a hidden dimension of $d_{\text{model}} = 896$. The model employs pre-LayerNorm and utilizes the SwiGLU non-linearity (Shazeer, 2020) in all feedforward networks. The feedforward network expansion ratio is 8/3, resulting in a total of 122M trainable parameters. Positional information is encoded using Rotary Position Embeddings (RoPE) (Su et al., 2023). The model processes sequences with a maximum length of 512 tokens, and shorter sequences are padded to this length. Multi-head attention is computed using Flash Attention (Dao et al., 2022).

Pre-training We pre-train `IDiom` on the aforementioned 74M sequences. The data are randomly split into 99% training, 0.5% validation, and 0.5% test sets. All sequences are padded or truncated to a fixed length of 512 tokens. The model is trained autoregressively using next-token prediction.

Training is performed using the AdamW optimizer with a learning rate of 4.0×10^{-4} and no weight decay. The learning rate schedule uses a linear warmup over the first 3,000 steps, followed by cosine annealing decay to a minimum of 4.0×10^{-5} (10% the initial learning rate) over 250,000 total training steps. The global batch size is 1,024 (Distributed Data Parallel training with 128 sequences per GPU across 8 NVIDIA H100 GPUs), with no gradient accumulation. We use a cross-entropy

loss while ignoring contributions from pad or mask tokens. Training is performed in mixed precision (fp32/bfloat16) and no gradient clipping is applied. Model validation is performed every 25,000 training steps on the validation set, and training is run for 250,000 optimizer steps. The training and validation curves are presented in the Supplementary Information.

Sequence Generation We generate IDR sequences from `IDiom` using autoregressive decoding. At each position t in the sequence, we compute the model logits z_t and convert them to a probability distribution via $p(x_t|x_{<t}) = \text{softmax}(z_t/T)$, then sample the next token from this full categorical distribution over the vocabulary. We use a fixed sampling temperature of $T = 1.0$ for all generations.

Sequence generation supports both prompted and unprompted modes. For prompted generation, the N- and C-terminal flanking contexts are provided as the prompt in fill-in-the-middle format: `<N><N-terminal context><C><C-terminal context><I>`, and the model generates the IDR span autoregressively following the `<I>` token. For unprompted generation of fully disordered proteins, the prompt consists of only the three special tokens without any flanking context: `<N><C><I>`, and the model generates the disordered sequence. In both cases, generation terminates upon sampling the `<eos>` token or upon reaching the maximum sequence length of 512 tokens.

Post-training We fine-tuned `IDiom` using Group Relative Policy Optimization (GRPO) with the Decoupled Advantage Policy Optimization (DAPO) modification. During GRPO training, for each batch of sequence prompts, the model generates multiple completions per prompt (group size = 8). Rewards are computed for each generated sequence using a reward function that combines three rewards: 1) A quadratic reward around a target ProtGPS score of 0.9 for the desired target compartment, 2) A quadratic reward around a target sequence length of $L_{\text{target}} = 100$ residues, and 3) A quadratic reward around a target sequence Shannon entropy of $H_{\text{target}} = 2.7$ nats to prevent diversity collapse.

Within each group, advantages are computed as normalized relative rewards: $A_i = \frac{r_i - \bar{r}_g}{\sigma_g}$, where $\bar{r}_g$ and σ_g are the group-wise mean and standard deviation. The GRPO loss combines a clipped policy gradient term (PPO-style clipping with $\epsilon_{\text{clip}} = 0.2$) weighted by advantages, and a KL-divergence penalty $\mathcal{L}_{\text{KL}} = \beta_{\text{KL}} D_{\text{KL}}(p_{\text{ref}}||p_{\text{current}})$ with strength $\beta_{\text{KL}} = 0.02$. The total loss is optimized per-token on generated completions only (any prompt tokens are masked).

Post-training is performed with a learning rate of 5×10^{-6}, the AdamW optimizer, and a global batch size of 8, for 1,500 optimizer steps. Post-training was performed separately for all 12 target cellular compartments in ProtGPS, although we only analyze the sequences generated from checkpoints trained for localization to the nucleolus, chromosomes, P-bodies, and stress granules.

Structure Prediction We use ColabFold (Mirdita et al., 2022) to perform AlphaFold2 structure predictions on generated sequences and determine the predicted local distance difference test (pLDDT) values.

Sequence Analysis We use the Sparrow package (Holehouse, 2022) to calculate sequence-level metrics such as charge patterning κ, sequence hydropathy decoration, sequence complexity (SEG), and fraction of charged residues. Short linear motif (SLIM) analysis is performed with a regular expression search using the SLIM regular expressions from the Eukaryotic Linear Motif Resource (Kumar et al., 2024). The specific regular expressions are presented in the Supplementary Information.

Visualization Protein structure visualization is performed with PyMOL (Schrödinger, LLC, 2025).

Data and Code Availability The code used for model pre-training, sequence generation, and post-training is available on Github: Code (https://github.com/rotskoff-group/idiom).

The pre-training datasets and pre- and post-trained model checkpoints are available on Hugging Face:

Datasets (https://huggingface.co/datasets/jxliu2/idiom-datasets)

Models (https://huggingface.co/jxliu2/idiom).

Information regarding these artifacts is provided in the Supplementary Information.

B. Supplementary Information

B.1. Datasets

Here we describe the datasets provided on HuggingFace:

Datasets (https://huggingface.co/datasets/jxliu2/idiom-datasets)

Below, we describe the data files under `idr_datasets/training_sequences`:

- `AFDB_IDR_90_reps.fasta` contains the 53M cluster representatives after the initial 214M full length AFDB protein sequences are clustered at 90% identity, 80% coverage.
- `AFDB_IDR_90_alldata.h5` contains 73M IDRs as extracted from the AFDB according to the Tesei logic (Tesei et al., 2024) (see Methods), and after filtering for IDRs belonging to the 53M cluster representatives identified in `AFDB_IDR_90_reps.fasta`. This HDF5 file contains the following keys: `<KeysViewHDF5 ['accession_ids', 'full_avg_plddt', 'full_length', 'full_seq', 'idr_end', 'idr_length', 'idr_plddt', 'idr_start', 'idrs']>`.
- `AFDB_IDR_90_FIM_512.h5` is created from `AFDB_IDR_90_alldata.h5` by filtering out IDRs whose full length sequences are longer than 512 residues. We also find that $\sim$ 1/3 of records in `AFDB_IDR_90_alldata.h5` are fully low-pLDDT sequences, and we filter out those sequences because we find that they are not representative of intrinsically disordered proteins. We only keep sequences with both low- and high-pLDDT regions. We hypothesize that sequences which are fully low-pLDDT are due to AlphaFold2's poor confidence in sequences which are not similar to those seen during training, rather than because they are fully intrinsically disordered proteins. For the remaining 37M IDRs, we apply the fill-in-the-middle (FIM) transformation as well as IDP data augmentation as mentioned in the Methods, and place those records into `AFDB_IDR_90_FIM_512.h5`. We note that we represent the `<N>`, `<C>`, and `<I>` tokens with `1`, `2`, and `3`, respectively, in this HDF5 file as well as in the codebase. This is the final file used for the precompute and pre-training steps.
- `AFDB_IDR_90_FIM_512_full.fasta` contains the 37M full length sequences (in correct order, not FIM-transformed) contained in `AFDB_IDR_90_FIM_512.h5`. The fasta header contains `_IDR_X-Y` where `X` and `Y` are the 1-indexed indices of the start and end (inclusive) of the intrinsically disordered region.
- `AFDB_IDR_90_FIM_512_idrs.fasta` contains only the sequences of the 37M intrinsically disordered regions in `AFDB_IDR_90_FIM_512_full.fasta`, without their surrounding context.

We also provide several datasets of sequences generated by our model under `idr_datasets/generated_sequences`. All generated sequences are provided in FASTA format along with their corresponding autoregressive model log (pickle format).

- Generated IDPs: 100,000 unprompted intrinsically disordered proteins.
- Generated IDRs: 101,700 intrinsically disordered regions generated using 1,017 DisProt flanking contexts prompts (100 generated IDRs per prompt).
- Generated NPM1 IDRs: 100,000 sequences generated using the NPM1 flanking context as the prompt (UniProt: P06748).
- Generated ProtGPS Sequences: 10,000 IDPs generated from post-trained checkpoints. Post-training was done to optimize ProtGPS localization scores for the four target compartments: chromosome, nucleolus, P-body, and stress granule.

B.2. Models

Here we describe the model checkpoints and other files provided on HuggingFace:

Models (https://huggingface.co/jxliu2/idiom)

Below, we describe the directories under `idiom/`:

- `base/` contains the checkpoint of our pre-trained base `IDiom` model, along with its configuration files.
- `post_trained/protgps_reward/` contains the checkpoints of `IDiom` post-trained via reinforcement learning using the ProtGPS reward model, one checkpoint per target compartment. In this paper, we analyzed results for 4 compartments: the nucleolus, stress granules, P-bodies, and chromosomes. However, post-training runs were conducted for all 12 ProtGPS compartments (chromosome, nucleolus, nuclear speckle, nuclear pore complex, P-body, PML body, post-synaptic density, stress granule, Cajal body, RNA granule, cell junction, and transcriptional condensate). We leave analysis of the remaining compartments to future work.
- `protgps/` contains the ProtGPS reward model used during reinforcement learning post-training.
- `data/` contains auxiliary files used during training and inference.

B.3. Secondary Structure Metric Analysis

Here we present analysis of secondary structure metrics for training, generated, DisProt, and CATH sequences. Secondary structure was assigned per residue using the dictionary of secondary structure of proteins (DSSP) algorithm as implemented in MDTraj (McGibbon et al., 2015). Secondary structure content was then defined as the sum of the mean α-helical and mean β-sheet fractions across all residues. Figure BS1 shows histograms of the average secondary structure content of 100 randomly chosen sequences from the various training, generated, DisProt, and CATH sets of proteins.

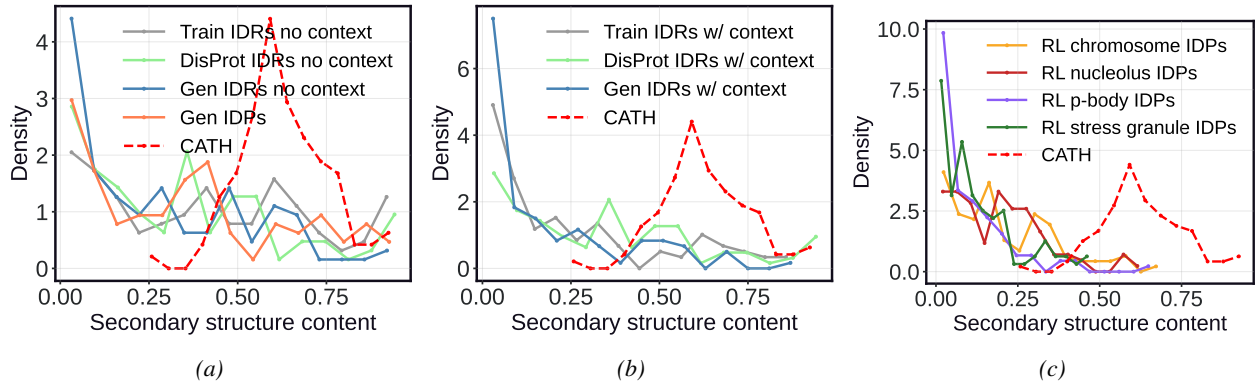

Figure BS1. **Secondary structure content analysis.** Histograms of the average secondary structure content ($\alpha + \beta$) for the AF2-predicted structures of 100 randomly chosen sequences from the following sets of sequences: **(a)** Secondary structure content of training IDRs, generated IDRs, DisProt IDRs, and CATH sequences, with their structures predicted with surrounding context included. **(b)** Secondary structure content of training IDPs, generated IDPs, DisProt IDPs, and CATH sequences, with their structures predicted without their surrounding context. **(c)** Secondary structure content of IDPs generated from post-trained `IDiom` checkpoints and CATH sequences, with their structures predicted with surrounding context included.

B.4. Disorder Predictions

Here, we present orthogonal disorder predictions from Metapredict v3 (Lotthammer et al., 2024) and IUPred3 (Erdős et al., 2021). For both predictors, a higher value represents a higher propensity towards disorder.

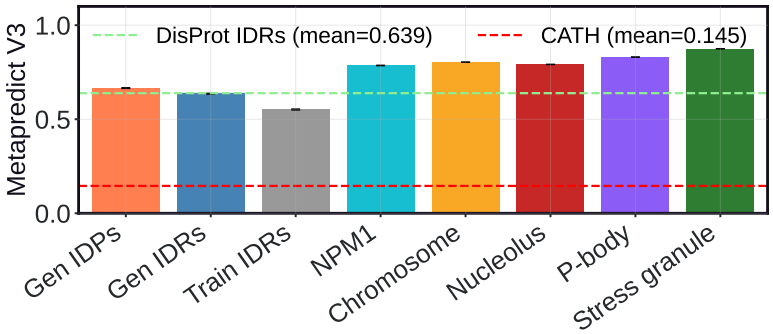

Figure BS2. **Disorder predictions from Metapredict V3**. Higher values correspond to higher propensity towards disorder. The horizontal green and red dashed lines correspond to the predicted Metapredict V3 values for 1,017 DisProt IDRs and 1,000 CATH sequences, respectively. The bars correspond to predicted Metapredict values for 10,000 sequences generated from `IDiom` for each condition, as well as 10,000 training sequences. The sequences generated from `IDiom` include unprompted IDPs, DisProt-prompted IDRs, NPM1 IDRs, and IDPs generated after post-training for localization to the chromosomes, nucleolus, P-bodies, and stress granules.

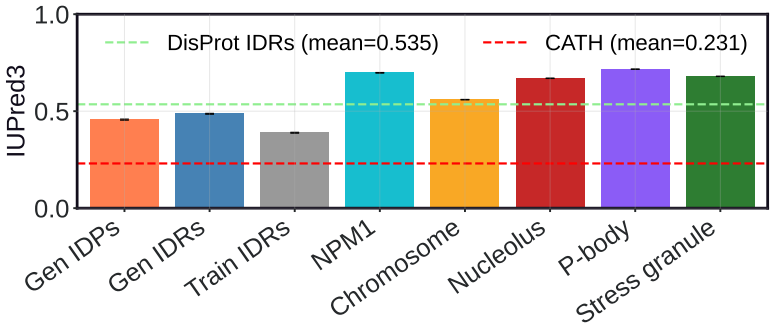

Figure BS3. **Disorder predictions from IUPred3**. Higher values correspond to higher propensity towards disorder. The horizontal green and red dashed lines correspond to the predicted IUPred3 values for 1,017 DisProt IDRs and 1,000 CATH sequences, respectively. The bars correspond to predicted IUPred3 values for 10,000 sequences generated from `IDiom` for each condition, as well as 10,000 training sequences. The sequences generated from `IDiom` include unprompted IDPs, DisProt-prompted IDRs, NPM1 IDRs, and IDPs generated after post-training for localization to the chromosomes, nucleolus, P-bodies, and stress granules.

B.5. ESM3 Comparison

Here we present comparison plots between sequences generated by IDiom and ESM3, using the same 1,017 DisProt flanking domain prompts. ESM3 sequences are generated using iterative decoding. A total of 1,000 sequences are sampled for each prompt. As ESM3 consists of a bidirectional transformer architecture, the length of the generated IDRs is fixed at the length of the ground truth IDR. The number of decoding steps, i.e. forward passes until the sequence is fully unmasked, is set to be the minimum of 20 and the ground truth IDR length for each prompt. Tokens are sampled with a temperature of 1.0, and all other default inference hyperparameters for ESM3 are used. We find that compared to IDiom, ESM3-generated IDRs are extremely low-complexity sequences, with a peak in the SEG complexity distribution around 0.5 (Figure BS4). We show three example ESM3-generated IDRs below, in red:

UniProt P48439:
MNWLFLVSLVFFCGVSTHPALAHFLDLLLLLLLLLLLLLQLILTAILAAIALLLLLLFLL
IVIGILLGLSLGALQLLLLLLLLLLLLSFALQLIFAAILAALLLILLLLLLLLIVIGILLS
LSFGALQLLILLLILLLWLLTLLLAKQLKLALALILAAILAALILLLLLLLLLLLLIVIGI
LFGLSLSALQLLLFLLLLLLLLLLVSFALKLKNPISRIIWATLSTFFIICMISAYMFNQI
RNTQLAGVGPKGEVMYFLPNEFQHQFAIETQVMVLIYGTLAALVVVLVKGIQFLRSHLYP
ETKKAYFIDAILASFCALFIYVFFAALTTVFTIKSPAYPFPLLRLSAPFK

UniProt Q9NS23-4:
MPCHPPPLPPPPPPPSPPPEEEEEEEEIEEEGEEEEPPASPLPPASPPAPEPVEWETPDL
SQAEIEQKIKEYNAQINSNLFMSLNKDGSYTGFIKVQLKLVRPVSVPSSKKPPSLQDARR
GPGRGTSVRRRTSFYLPKDAVKHLHVLSRTRAREVIEALLRKFLVVDDPRKFALFERAER
HGQVYLRKLLDDEQPLRLRLLAGPSDKALSFVLKENDSGEVNWDAFSMPELHNFLRILQR
EEEEHLRQILQKYSYCRQKIQEALHACPLG

UniProt Q14011:
MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDA
KDAMMAMNGKSVDGRQIRVDQAGKSSDNRGGGGGGGGGRGGGGGGGGGGGRGGGGGGRG
GGGSGGGGGGRGGGGGSGGRGGGGGGGGGGGGGGGGGGGRGGGGGGGGRY

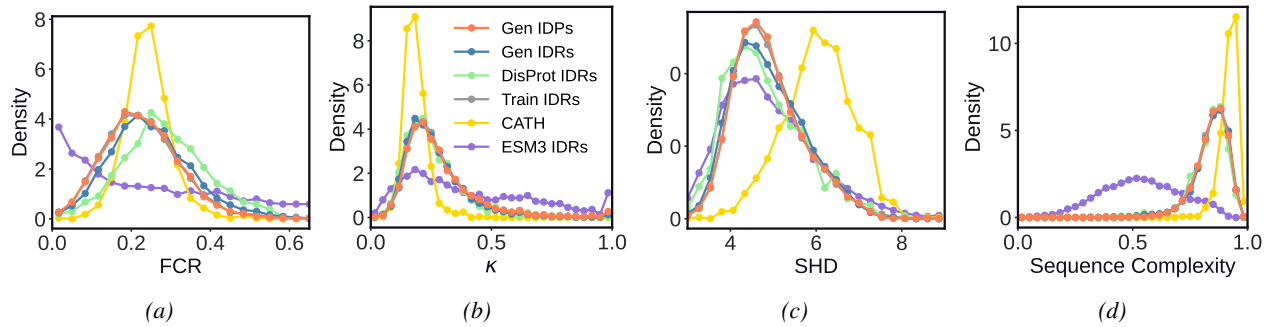

Figure BS4. **Comparison between ESM3-generated IDRs and IDiom-generated IDRs. (a)–(d)** Distributions of various sequence metrics for sequences generated from IDiom versus ESM3. Training set IDRs, natural DisProt IDRs, and folded CATH domains are shown as well. **(a)** Fraction of charged residues (FCR). **(b)** Charge patterning κ parameter. **(c)** Sequence hydropathy decoration (SHD). **(d)** Sequence complexity quantified by the SEG algorithm.

B.6. In-Context Learning Case Study: NPM1

Here, we consider the human protein NPM1 (UniProt: P06748) as a case study to further illustrate the model's ability to learn via in-context conditioning. NPM1 has an IDR (residues 119–242) which drives nucleolar phase separation through charge block patterning that mediates interactions with itself (Mitrea et al., 2018) as well as binding partners such as SURF6 (Ferrolino et al., 2018). Using the flanking regions around this IDR as the prompt (green and blue domains in Figure 1c (upper)), we generated 100,000 IDRs and filtered out any with sequence identity $> 90\%$ to the wild-type (WT) NPM1 IDR (Figure BS5a). Figure BS5b shows the distribution of κ values for NPM1-prompted generations, randomly scrambled versions of those sequences, and the dataset of DisProt IDRs. We find that the κ distribution of NPM1-prompted generations is peaked near the WT NPM1 value, and is substantially shifted toward higher charge segregation than either randomly scrambled sequences or generic DisProt IDRs, indicating that the model has learned to generate sequences with substantial charge block patterning, when conditioned on the NPM1 flanking contexts.

Figure BS5c shows linear plots of net charge per residue (NCPR) for the WT NPM1 IDR and representative generated sequences with high κ values but low sequence identity to the WT IDR. These plots illustrate that `IDiom` generates sequences with alternating blocks of positive and negative charge which closely mirror the architecture of the WT IDR, despite sharing little sequence identity with it. Together, these results demonstrate that `IDiom` is able to use flanking context to reproduce the biologically relevant sequence features of a given IDR, and that the model is able to generate diverse sequences that preserve the key sequence features that underlie biological function.

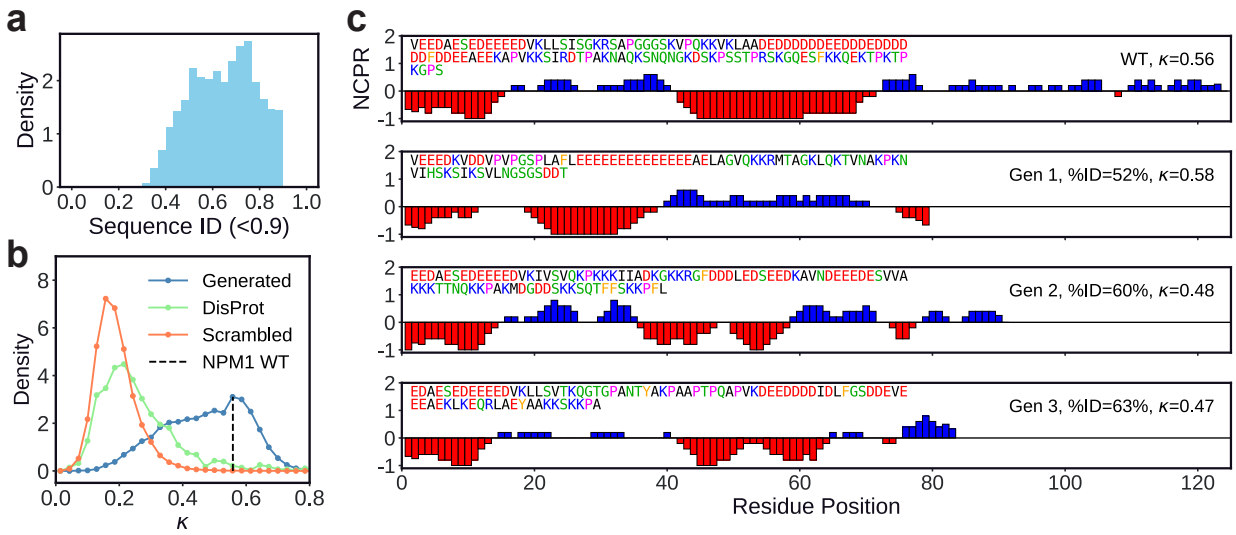

Figure BS5. **Conditioned generation and in-context learning enables generation of disordered regions which capture biologically relevant sequence features. (a)** Distribution of sequence identities between the wild-type (WT) NPM1 IDR and sequences generated with the NPM1 IDR's flanking context as the prompt. Generated sequences with identities greater than 0.9 are filtered out. **(b)** Distribution of κ values for the WT NPM1 IDR (black vertical dashed line), IDRs generated with the NPM1 context as prompt (blue), randomly scrambled versions of the generated IDRs (orange), and DisProt IDRs (green). **(c)** Plots of the net charge per residue along the sequence. The WT NPM1 IDR is in the upper row, and three representative generated IDRs are below. The sequence identities relative to the WT, κ values, and the amino acid sequences themselves are printed within the plots.

B.7. Short Linear Motifs from the Eukaryotic Linear Motif Resource

Here, we list the short linear motifs from the ELM Resource which we scan for, for nuclear localization signals (NLSs) as well as for post-translational modification (PTM) sites (ELM Identifier: MOD).

Nuclear Localization Signals The regular expressions of the 4 NLSs we consider are:

ID	Pattern (regex)
TRG_NLS_Bipartite_1	[KR][KR].{7,15}[DE]((K[RK])|(RK))(([DE][KR])|([KR][DE]))[DE]
TRG_NLS_MonoCore_2	[DE]((K[RK])|(RK))[KRP][KR][DE]
TRG_NLS_MonoExtC_3	[DE]((K[RK])|(RK))(([DE][KR])|([KR][DE]))(([PKR])|([DE][DE]))
TRG_NLS_MonoExtN_4	(([PKR].{0,1}[DE])|([PKR]))((K[RK])|(RK))(([DE][KR])|([KR][DE]))[DE]

Table 1. ELM NLS motifs and their corresponding regex patterns.

Post Translational Modification Motifs The regular expressions of the 40 PTM MOD sites we consider are:

ID	Pattern (regex)
MOD_AAK1BIKe_LxxQxTG_1	[LIVM][D][DEHYWF]Q.(T)G
MOD_ASX_betaOH_EGF	C.([DN]).{4,4}[FY].C.C
MOD_CAAXbox	(C)[DENQ][LIVMF].$
MOD_CDC14_SPxK_1	(S)P.[KR]
MOD_CDK_SPK_2	...([ST])P[RK]
MOD_CDK_SPxK_1	...([ST])P.[KR]
MOD_CDK_SPxxK_3	...([ST])P..[RK]
MOD_CK1_1	S..([ST])...
MOD_CK2_1	...([ST])..E
MOD_CMANNOS	(W)..W
MOD_Cter_Amidation	(.)G[RK][RK]
MOD_DYRK1A_RPxSP_1	R[PSVA].([ST])P
MOD_GlcNHglycan	[ED]{0,3}.(S)[GA].
MOD_GSK3_1	...([ST])...[ST]
MOD_LATS_1	H.[KR]..([ST])[P]
MOD_LOK_YxT_1	[KR][YF][IVEDPGAC](T)[LMIVWFY][RKH]
MOD_NEK2_1	[FLM][PVIED][PVID]([ST])[MLIVF][RKH].
MOD_NEK2_2	[FLMW][P][P]([ST])[PDEGAN][RKH].
MOD_N-GLC_1	.(N)[P][ST]..
MOD_N-GLC_2	(N)[P]C
MOD_NMyristoyl	M{0,1}(G)[EDRKHPFYW]..[STAGCN][P]
MOD_OFUCOSY	C.{3,5}([ST])C
MOD_OGLYCOS	C.(S).PC
MOD_PIKK_1	...([ST])Q..
MOD_PK_1	[RK]..(S)[VI]..
MOD_PKA_1	[RK][RK].([ST])[P]..
MOD_PKA_2	.R.([ST])[P]..
MOD_PKB_1	R.R..([ST])[P]..
MOD_Plk_1	.[DNE][PG][ST](([FYILMVW]..)|([PEDGKN][FWYLIVM]).)
MOD_Plk_2-3	[DE]..([ST])[EDILMVFWY](([DE].)|(.[DE]))
MOD_Plk_4	..[IRFW]([ST])[ILMVFWY][ILMVFWY].
MOD_PRMT_GGRGG_1	GGRGG
MOD_ProDKin_1	...([ST])P..
MOD_SPalmitoyl_2	G(C)M[GS][CL][KP]C
MOD_SPalmitoyl_4	M{0,1}G(C)..S[AKS]
MOD_SUMO_for_1	[VILMAFP](K).E
MOD_SUMO_rev_2	[SDE].{0,5}[DE].(K).{0,1}[AIFLMPSTV]
MOD_TYR_CSK	[TAD][EA].Q(Y)[QE].[GQA][PEDLS]
MOD_TYR_DYR	..[RKTC][IVL]Y[TQHS](Y)[IL]QSR
MOD_WntLipid	[ETA](C)[QERK]..F...RWNC[ST]

Table 2. ELM MOD motifs and their corresponding regex patterns.

B.8. Training Curves

Here, we present additional training curves from pre-training as well as post-training.

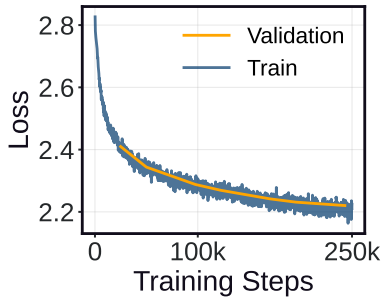

Figure BS6. **Pretraining loss curves.** Training and validation losses vs optimizer steps during pre-training. The final training loss is 2.19. The final validation loss is 2.22.

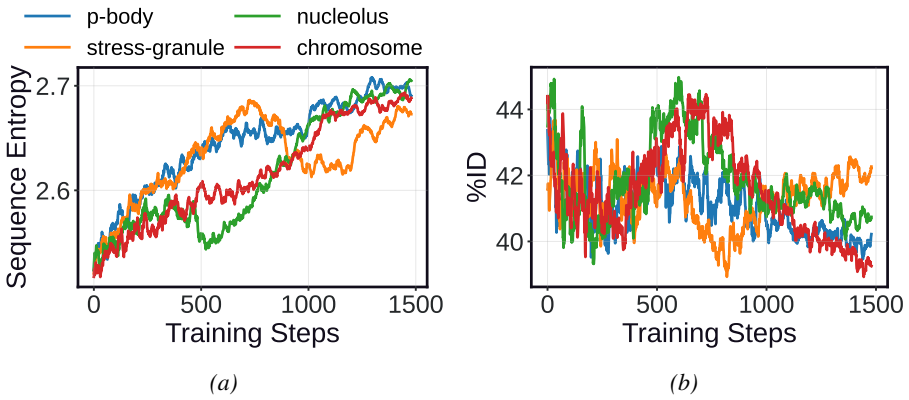

Figure BS7. **Additional post-training curves with the ProtGPS reward model. (a)** Shannon entropy vs. training steps (target $H = 2.7$). **(b)** %ID within a generated batch vs training steps (no target value).