GENERATIVE MODELING OF INTRINSICALLY DISORDERED PROTEIN REGIONS BY REINFORCING SPARSE AUTOENCODER FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Intrinsically disordered protein regions (IDRs) 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 localization, yet their functional design remains challenging for existing generative methods. Structure-based design methods do not readily apply to IDRs, and existing protein language models are trained on full-length sequences, learning priors which are biased towards folded domains. Here, we present **IDiom**, an autoregressive protein language model trained on IDiom-DB, a dataset of 54 million predicted IDRs curated from the AlphaFold Database, enabling generation of diverse and biologically realistic IDRs. For explicit control over sequence features, we also introduce reinforcement learning with sparse autoencoder (SAE) features (**RL-SAE**), which rewards the model for generating sequences that activate a specified set of SAE features. Together, IDiom and RL-SAE enable the interpretable and composable design of IDRs that carry the sequence motifs and predicted functionality of natural IDRs. More broadly, RL-SAE could extend to other protein design settings where interpretable features provide useful design targets.

1 INTRODUCTION

Intrinsically disordered protein regions (IDRs) are ubiquitous, constituting roughly 30% of residues in the human proteome, with over 60% of human proteins containing at least one IDR (Tesei et al., 2024; Alderson et al., 2023). It is now clear that IDRs play essential roles in almost every aspect of cellular function. Because 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 (Wei et al., 2017), and they play crucial roles in diverse biological processes such as signal transduction, transcriptional regulation, chromatin organization, and subcellular localization (Holehouse & Kragelund, 2024).

Rational design of IDRs would unlock new opportunities in bioengineering, including tunable condensate formation, precise regulation of cell signaling, and targeted protein localization (Tesei et al., 2026). Tremendous progress has been seen recently in machine learning-based design of folded proteins, driven by accurate structure prediction (Jumper et al., 2021; Abramson et al., 2024) and structure-based generative models (Ahern et al., 2026; Geffner et al., 2026). However, direct application of these approaches to IDRs remains challenging due to their structurally disordered nature.

Protein language models (PLMs) have recently been explored for applications in protein design. These models learn rich evolutionary statistics over amino acid sequences by training on large corpora of protein sequences from databases such as UniProt (The UniProt Consortium et al., 2025), enabling them to generate novel proteins and functional variants (Hayes et al., 2025; Bhatnagar et al., 2025). However, these models are primarily trained on full-length proteins and possess a generative prior that is biased towards folded domains (Ferruz et al., 2022). Training on an IDR-specific corpus would instead focus the generative prior on the sequence statistics of disordered regions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Control over the incorporation of interpretable sequence features into generated proteins also remains a key challenge. PLMs can be fine-tuned for functional design using online reinforcement learning (RL) against computational oracles (Stocco et al., 2025; Angermueller et al., 2020) and offline alignment with datasets of sequences and their measured or calculated properties (Chennakesavalu et al., 2025; Ibarraran et al., 2026). While these methods can effectively optimize functional properties, their objectives do not explicitly identify or control the sequence-level features underlying those properties. Explicit control over interpretable features would enable specific sequence grammars to be installed and features associated with distinct functions to be composed.

Key contributions: Here, we address these gaps in our current ability to design IDRs with three primary contributions:

- **IDiom**, a protein language model pretrained on IDiom-DB, a corpus of 54M predicted IDRs curated from the AlphaFold Database. IDiom enables the generation of diverse sequences that recapitulate natural IDR composition, patterning, motifs, and predicted disorder.
- Reinforcement learning with sparse autoencoder (SAE) features (**RL-SAE**), a sequence design method that rewards the simultaneous activation of specified sets of interpretable SAE features.
- Computational experiments demonstrating that RL-SAE improves biological design performance over activation steering and supervised fine-tuning while providing explicit control over incorporated sequence features.

2 Related work

Design of intrinsically disordered regions Current sequence-based IDR design techniques primarily employ residue sampling strategies to construct disordered sequences and variants that satisfy imposed constraints such as compositional rules or sequence-level metrics (Pesce et al., 2024; Hunter et al., 2026). However, these techniques impose disorder through hand-specified constraints and do not necessarily capture the evolutionary statistics of natural IDRs. In contrast, IDiom directly learns these sequence statistics, and with RL-SAE, this enables biological design using function-associated features rather than manually specified sequence rules.

Interpreting and steering protein language models Sparse autoencoders trained on PLM activations have been shown to find approximately monosemantic features corresponding to interpretable biological concepts (Candido et al., 2026; Simon & Zou, 2025), and activation steering along such feature directions has been applied to sequence design (Corominas et al., 2025; Villegas Garcia & Ansuini, 2025). However, combining multiple steering directions does not ensure that generated sequences activate all targeted features. In contrast, RL-SAE activates a substantially greater fraction of target features than activation steering, and this enables the design of IDRs that combine interpretable features associated with distinct biological functions, such as localization and transcriptional regulation.

Reinforcement learning with sparse autoencoders Several recent works in the large language model literature have combined SAEs with reinforcement learning. For example, Cho et al. (2026) train a controller to select and amplify SAE feature directions during inference, and Zhang et al. (2025) construct a reward model using SAE feature activations fit to human preference pairs for improved reward interpretability. Unlike these methods, RL-SAE directly rewards the activation of prespecified sets of target SAE features, allowing for explicit control over the incorporation of these features in generated sequences.

3 Generative modeling of intrinsically disordered regions

A corpus of intrinsically disordered region sequences To curate the IDiom-DB dataset of IDR sequences for training, we use AlphaFold2 (AF2) predicted structures from the AlphaFold Database (AFDB) (Varadi et al., 2024). 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

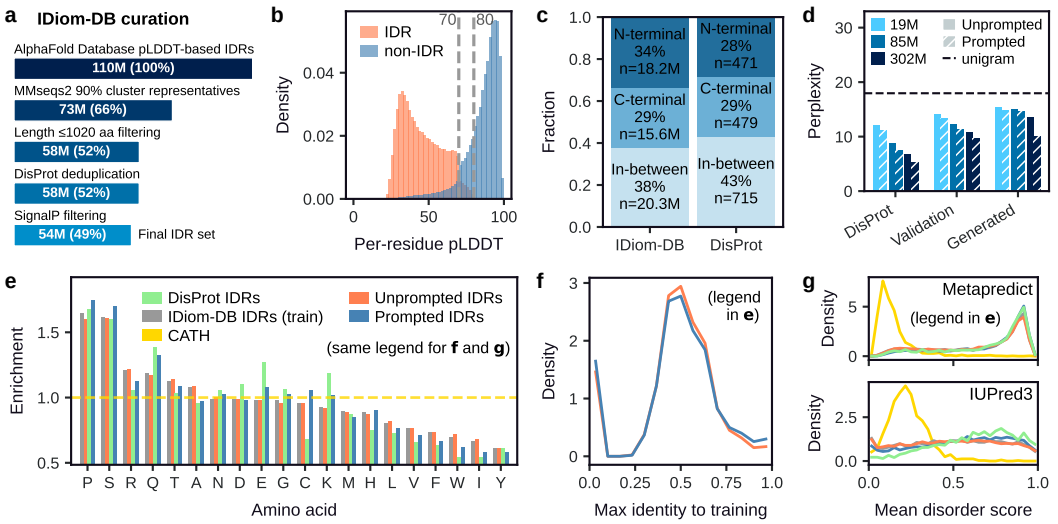

Figure 1: Training on 54M curated IDRs enables the generation of diverse and biologically plausible IDRs. **(a)** IDiom-DB curation workflow. **(b)** AlphaFold2 pLDDT distribution of curated IDRs and non-IDRs. **(c)** Fraction of curated training IDRs and experimental DisProt IDRs which are located at the N-terminus, C-terminus, and internally. **(d)** Model perplexity on DisProt, validation, and generated IDRs. Solid and hatched bars indicate unprompted and prompted IDRs, respectively. **(e)** Amino acid compositional enrichment for training, generated, and DisProt IDRs, relative to the folded CATH sequences' compositions. This panel's legend also applies to panels (f) and (g). **(f)** Sequence identity of generated IDRs relative to training set IDRs, calculated with MMseqs2. **(g)** Predicted disorder scores of training, generated, and DisProt IDRs, alongside CATH sequences (higher is more disordered).

measurements of disorder (Wilson et al., 2022; Zhao et al., 2023). We first cluster AFDB sequences at 90% sequence identity using MMseqs2 (Steinegger & Söding, 2017) and then apply a windowed threshold at pLDDT = 70 to identify IDRs (Tesei et al., 2024) (Figure 1b), with proteins containing multiple IDRs contributing multiple records. We discard IDRs shorter than 30 residues, IDRs residing in proteins longer than 1,020 residues, proteins that are entirely low-pLDDT, and predicted N-terminal signal peptides. This yields 54M IDRs and their positions within their full-length host proteins (Figure 1a). Additional details are in Appendix A.

Architecture and training We pretrain IDiom, a decoder-only transformer, at three sizes: 19M, 85M, and 302M parameters (final model perplexities in Figure 1d). We use a context length of 1,024 tokens, where each amino acid residue is assigned an individual token. Training uses a next-token objective on a fill-in-the-middle (FIM) transformed (Bavarian et al., 2022) version of IDiom-DB, which places N- and C-terminal contexts before the IDR to enable generation conditioned on fixed flanking sequences. During training, we provide 50% of sequences with their flanking contexts and 50% without, enabling both context-prompted and unprompted IDR generation. All analysis in this paper focuses on the 302M parameter model (IDiom-300M). Additional details are in Appendix B.

IDiom generates diverse disordered regions To characterize the pretrained model, we generate 100,000 unprompted IDRs and 166,300 context-prompted IDRs, using the surrounding flanking contexts of 1,663 experimentally validated IDRs from DisProt as prompts (100 IDRs per prompt) (Nugnes et al., 2026). The distribution of maximum sequence identities between generated and training IDRs peaks broadly around 50%, indicating that most generated sequences are substantially dissimilar to any sequence seen during training (Figure 1f).

To confirm that the generated sequences are indeed disordered, we compare them to the 1,663 experimentally validated IDRs from DisProt as well as 1,000 folded CATH S60 domains (Waman et al., 2025). We find that across AF2-predicted pLDDTs (Mirdita et al., 2022) and the sequence-based disorder predictors Metapredict V3 (Lotthammer et al., 2024) and IUPred3 (Erdős et al., 2021), generated sequences closely track the DisProt set, attaining substantially higher disorder scores than CATH domains (Figure 1g).

We additionally compare IDiom-generated IDRs to DisProt and CATH sequences along compositional and patterning metrics. We find that IDiom IDRs are substantially enriched in the disorder-promoting proline and serine residues while being depleted in order-promoting aliphatics and aromatics (Theillet et al., 2013) (Figure 1e). IDiom IDRs also recapitulate known patterning properties that govern IDR behavior, such as charge segregation (Das & Pappu, 2013), reduced hydrophobic clustering (Zheng et al., 2020), and low sequence complexity (Wootton & Federhen, 1993). Close agreement with DisProt is observed across all of these metrics. Additional details are in Appendix C.

4 Reinforcement learning with sparse autoencoder features

Motivation The design of functional IDRs requires control over combinations of sequence-level features. SAEs provide interpretable features that can be used as design targets (Adams et al., 2025; Simon & Zou, 2025), but reliably inducing the activation of many features simultaneously within generated sequences remains challenging (Arad et al., 2025; Xie, 2025). Here, we introduce reinforcement learning with sparse autoencoder features (RL-SAE), a post-training technique that rewards the model for generating sequences that activate a specified set of SAE features. Our experiments demonstrate that RL-SAE enables IDiom to generate sequences that activate tens of function-associated SAE features simultaneously and outperforms activation steering and supervised fine-tuning on biological design tasks.

A sparse dictionary over IDiom activations To perform RL-SAE on IDiom, we train IDiomSAE, a Top-k SAE with $k = 32$ and 16,384 latents in $\boldsymbol{z}$ (Gao et al., 2024). IDiomSAE is trained on IDiom's residual stream activations at layer-18 of 24 and only uses activations from IDR residue positions, thus focusing its learned dictionary on sequence features within disordered regions. Given a residual stream activation $\boldsymbol{x}$, IDiomSAE calculates:

$$\boldsymbol{z} = \mathrm{TopK}(\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{enc}}(\boldsymbol{x} - \boldsymbol{b}_{\mathrm{dec}}) + \boldsymbol{b}_{\mathrm{enc}})), \qquad \hat{\boldsymbol{x}} = \boldsymbol{W}_{\mathrm{dec}}\boldsymbol{z} + \boldsymbol{b}_{\mathrm{dec}}, \tag{1}$$

where the encoder and decoder weights and biases $\boldsymbol{W}_{\mathrm{enc}}$, $\boldsymbol{W}_{\mathrm{dec}}$, $\boldsymbol{b}_{\mathrm{enc}}$, and $\boldsymbol{b}_{\mathrm{dec}}$ are trained to minimize the fraction of variance unexplained, $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 / \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2$, where $\bar{\boldsymbol{x}}$ is the mean activation over the training set. Additional details are in Appendix D.

Identification of enriched SAE feature sets We hypothesize that SAE features enriched in sequences with a labeled functionality capture functionally relevant sequence patterns, and we use these features as reward targets in RL-SAE. To identify salient SAE features, we rank each feature j by the log odds ratio (LOR) of its firing prevalence in the positively labeled set $\mathcal{P}$ to its prevalence within a negative set $\mathcal{N}$ (Candido et al., 2026). We count a feature as firing in a sequence y if it activates at any residue position t:

$$\mathrm{fire}_j(y) = \mathbf{1}\left[\max_t z_j(y)_t > 0\right], \qquad \mathrm{LOR}_j = \log_2 \frac{\Pr_{\mathcal{P}}[\mathrm{fire}_j] \,/\, (1 - \Pr_{\mathcal{P}}[\mathrm{fire}_j])}{\Pr_{\mathcal{N}}[\mathrm{fire}_j] \,/\, (1 - \Pr_{\mathcal{N}}[\mathrm{fire}_j])}. \tag{2}$$

We then rank features by LOR and choose the top-N features for RL-SAE. Additional details are in Appendix E.

A sparse feature-matching reward We reward generated completions y for the fraction of target features that fire, with quadratic penalties on relative deviations from the target sequence length L_{target} and entropy H_{target}. Thus, our reward $R(y)$ is:

$$R(y) = R_S(y) - \left[(L(y) - L_{\mathrm{target}})\,/L_{\mathrm{target}}\right]^2 - \left[(H(y) - H_{\mathrm{target}})\,/H_{\mathrm{target}}\right]^2, \tag{3}$$

where

$$R_S(y) = \frac{1}{N} \sum_{j \,\in\, \text{top-}N} \mathrm{fire}_j(y) \tag{4}$$

is the fraction of top-N features that fire anywhere within y, $L(y)$ is the generated sequence length, and $H(y)$ is the Shannon entropy of the sequence's amino acid composition.

Reinforcement learning optimization We post-train IDiom with group-relative policy optimization (Shao et al., 2024) to optimize $R(y)$ with KL regularization relative to the base model:

$$\mathcal{J}(\theta) = \mathbb{E}_{y \sim \pi_\theta}[R(y)] - \beta\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{base}}) \tag{5}$$

where π_θ is the trained policy and π_{base} is the base IDiom model.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Functional sequence sets and oracles To demonstrate RL-SAE for IDR design, we focus on two IDR-driven biological functions: association with biomolecular compartments and transcriptional regulation. We use six human sequence sets with annotated localization to the nucleolus (629 sequences), chromosomes (846), nuclear speckles (586), stress granules (664), P-bodies (329), and the nuclear pore complex (101) (Kilgore et al., 2025), with 100,000 randomly selected, length-matched IDRs from IDiom-DB as the negative set. For transcriptional regulation, we use experimentally validated human activation domain (AD, 288) and repressor domain (RD, 408) sequences from a large-scale gene regulatory activity tiling scan (DelRosso et al., 2023), with 4,920 inert tiles from the same assay as the negative set.

We characterize the generated sequences using ProtGPS (Kilgore et al., 2025) for compartment association, DeepLoc 2.1 for subcellular localization (Ødum et al., 2024), PADDLE for transcriptional activation (Sanborn et al., 2021), and known corepressor-recruiting motifs from DelRosso et al. (2023) for repression. We also assess compartment-associated sequence motifs and we predict disorder using Metapredict V3 and IUPred3. Dataset details are in Appendix F.

RL-SAE training For all 8 sequence sets, we run RL-SAE and activation steering on the top-$N = 30$ enriched features, referred to as top-30 sets. We also evaluate feature sets consisting of the top 30 enriched features that are unique to each compartment, referred to as private-30 sets. All post-training runs begin with the base IDiom model and generate unprompted IDRs. We target a sequence length of 100 residues for compartment-associated IDRs and 80 residues for effector domains to match the experimental tile length. We set the entropy target to 3.65 bits, the mean sequence entropy of IDRs across IDiom-DB. Training details and other top-N values are in Appendix G.

Baseline methods As a direct baseline, we compare RL-SAE to activation steering using the same target feature sets. We sum and normalize the IDiomSAE decoder vectors into the steering direction, which we add to the layer-18 residual stream at every token position, scaled by α times the original activation norm before renormalizing to that norm. We sweep α from 0 to 1.5, discarding cases whose generated sequences have mean entropy below the fifth percentile of IDiom-DB (3.08 bits), and report the remaining α which achieves the highest score (Appendix H).

For the supervised fine-tuning (SFT) baseline, we fine-tune the base IDiom model using the positive IDR sequence sets from which the top-N features were extracted. We additionally evaluate RL using ProtGPS scores as the reward for compartment association (RL-GPS). Additional details for baseline methods are in Appendix H.

5.2 RESULTS

5.2.1 ENRICHED FEATURES RECOVER CANONICAL IDR SEQUENCE GRAMMARS

Feature enrichment and sequence grammars Figure 2 shows feature enrichment plots and sequence logos of the top-four enriched features for four example sequence sets. We find that many of the top enriched features correspond to canonical IDR sequence grammars, such as FG/GLFG repeats in NPC sequences (Frey & Görlich, 2009), transportin-associated RS repeats of speckle sequences (Lai et al., 2001), the acidic exposure model patterning of activation domains (Staller et al., 2022), and SUMOylation motifs in repression domains (DelRosso et al., 2023). Additional enrichment and logo plots are in Appendix R. These examples show that feature enrichment identifies recognizable sequence grammars associated with the annotated functions, thus providing interpretable targets for RL-SAE. We next test whether RL-SAE increases target feature activation and improves predicted biological functionality in generated IDRs.

5.2.2 DESIGNING BIOMOLECULAR CONDENSATE-ASSOCIATED IDRS

Feature installation and compartment association First, we confirm the ability of RL-SAE to install the target feature set. We find that across the 8 sets, sequences generated after RL-SAE activate on average 90% of the targeted top-30 features, compared to 24% for activation steering.

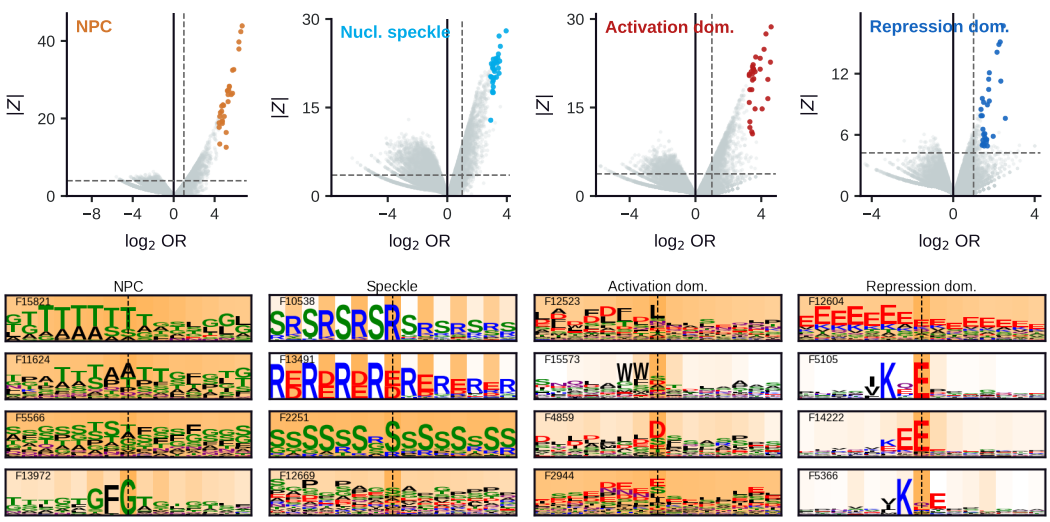

Figure 2: Enriched SAE features recover canonical IDR sequence grammars. Top row: Statistical significance $|Z|$ versus $\log_2$ odds ratio (LOR) for all 16,384 IDiomSAE features in nuclear pore complex, nuclear speckle, transcriptional activation, and repression sequence sets. Dashed lines mark LOR $= 1$ (vertical) and the false discovery rate threshold of 10^{-3} (horizontal). Features in the upper right quadrant pass both thresholds. **Bottom row:** Information content logos for each set's top-four enriched features, showing residue likelihoods within a ± 7-residue window around peak firing positions (vertical dashed line position). Orange shading indicates the feature activation strength at each residue position. Logo construction is described in Appendix R.

SFT activates only 15% of the top-30 on average, and the natural sequences activate 21%. Additional details are in Appendices J and K.

Next, we assess whether the activation of targeted features is accompanied by known functional motifs and changes in predicted biological function. Table 1 shows the predicted ProtGPS localization scores for sequences generated by the different techniques, with predicted disorder for all methods given in Appendix N. We find that for all compartments except P-bodies, RL-SAE using the top-30 enriched features produces sequences with a higher mean predicted localization score than SFT, activation steering, and the natural IDR sequences. Composition-matched scrambles of the RL-SAE sequences show that these predicted scores are largely driven by specific grammars in all compartments except stress granules and P-bodies, whose predicted localization is partly compositional (Appendix M).

The experiments demonstrate that activation steering is limited by incomplete feature installation and decreased sequence entropy at higher steering strengths. While ProtGPS scores increase with α at low α, the effect is limited prior to substantial sequence entropy collapse (Appendix J). We find

					RL-SAE	
Compartment	Base	Nat.	SFT	Steering (α)	top-30	private-30
Nucleolus	$0.14_{.01}$	$\underline{0.48}_{.02}$	$0.30_{.01}$	$0.41_{.01}$ (0.35)	$\mathbf{0.99}_{.00}$	$0.94_{.07}$
Chromosome	$0.26_{.01}$	$\underline{0.52}_{.01}$	$0.37_{.00}$	$0.51_{.02}$ (0.21)	$\mathbf{0.71}_{.03}$	$0.67_{.08}$
Speckle	$0.00_{.00}$	$\underline{0.15}_{.01}$	$0.07_{.00}$	$0.11_{.01}$ (0.20)	$\mathbf{0.36}_{.15}$	$0.30_{.24}$
Stress gran.	$0.05_{.00}$	$0.15_{.01}$	$0.06_{.01}$	$\underline{0.16}_{.01}$ (0.15)	$0.32_{.09}$	$\mathbf{0.51}_{.36}$
P-body	$0.17_{.01}$	$\underline{0.27}_{.02}$	$0.20_{.01}$	$0.19_{.01}$ (0.00)	$0.05_{.02}$	$\mathbf{0.32}_{.10}$
NPC	$0.08_{.01}$	$\underline{0.42}_{.05}$	$0.40_{.00}$	$0.33_{.01}$ (0.20)	$0.84_{.11}$	$\mathbf{0.90}_{.05}$

Table 1: RL-SAE improves predicted association with target biomolecular compartments. ProtGPS-predicted biomolecular compartment association scores (0 to 1). Bold indicates the best score per row, underline indicates the best non-RL-SAE method. The private-30 column targets the top 30 private features of each compartment. Subscripted values are standard deviations over three seeds, except for natural IDRs (Nat.), which report the standard error of the mean (Appendix I).

that SFT raises the score above the base model, but SFT scores remain below those of the natural sequences for all six compartments, potentially due to the limited dataset size.

Since the IDR sequences used for compartment association feature enrichment are extracted from the same full-length sequences used to train ProtGPS, there is circularity in using ProtGPS as a downstream evaluator. Therefore, we additionally score the generated sequences with DeepLoc 2.1, an orthogonal localization predictor (Ødum et al., 2024). Since DeepLoc only predicts coarse localization, we assess sequences on their predicted nuclear versus cytoplasmic localization. We find that RL-SAE targeting the top-30 or private-30 achieves the best or second-best score for the biologically expected label in all six compartments (Appendix O). For P-body associating sequences, targeting the private-30 rather than the top-30 features increases the cytoplasmic DeepLoc prediction, consistent with the increased ProtGPS-predicted P-body score in Table 1.

In addition to ProtGPS and DeepLoc, we evaluate the generated sequences for the presence of several critical biological motifs: classical nuclear localization signals (cNLSs) for the nucleolus and chromosome (Lange et al., 2007), RNA-interacting RG/RGG boxes for P-bodies and stress granules (Decker & Parker, 2012), and the aforementioned transportin RS repeats and GLFG nucleoporin repeats (regular expressions in Appendix Q). Appendix P shows that RL-SAE installs the expected motif in nearly all generated sequences for five of the six motifs, substantially outperforming other methods. The exception is the P-body RG/RGG motif, which is only weakly present within the top enriched features (Appendix R).

Private feature selection reduces predicted off-target association Our feature enrichment analysis lets us explicitly identify features shared between different sequence sets. Figure 3a shows the number of shared top-30 features among the 6 biomolecular compartments. The nucleolus and chromosome share many enriched features, consistent with their nuclear localization and basicity. P-bodies share charged features with the nucleolus and aromatic features with stress granules, consistent with a hydrophobic sticker grammar in granule IDRs, and the NPC shares the fewest features, consistent with its unique FG-based grammar. Figure 3b shows that this feature cross-talk is reflected in the off-target ProtGPS scores. In particular, chromosome and P-body targeting sequences both score highly for the nucleolus, and stress granule targeting sequences score highly for P-bodies.

To test whether these shared features contribute to ProtGPS cross-talk, we run RL-SAE targeting the private-30 feature sets. We find that this substantially reduces off-target nucleolus scores for chromosome- and P-body-targeting sequences, and off-target P-body scores for stress-granule-targeting sequences (Figure 3c). This reduced cross-talk occurs while maintaining strong on-target scores: private-30 sequences achieve higher mean on-target scores than SFT, activation steering, and natural IDRs across all six compartments (Table 1). These experiments support the hypothesis that shared SAE features contribute to predicted cross-talk, and that feature selection can improve predicted compartment specificity. However, we also note that off-target scores are not uniformly reduced, and some off-target pairs increase modestly.

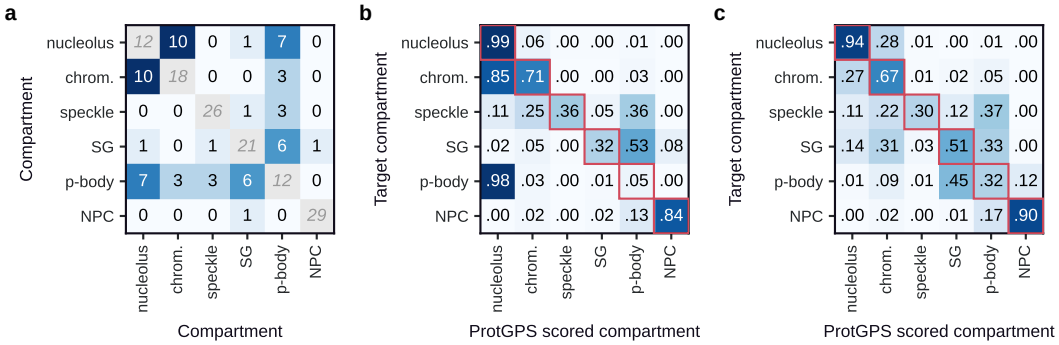

Figure 3: Targeting private SAE features reduces predicted cross-talk between compartments. **(a)** Number of private (greyed diagonal) and shared (off-diagonal) features among the top-30 between the 6 considered compartments. **(b)** Heatmap of off-target ProtGPS scores for sequences generated after RL-SAE on the top-30 features of each compartment. **(c)** Heatmap of off-target ProtGPS scores of sequences generated after RL-SAE on the top-30 private features of each compartment.

5.2.3 DESIGNING AND COMPOSING TRANSCRIPTIONAL EFFECTOR DOMAINS

Designing activation and repression domains Our experiments on transcriptional activation and repression domains show similarly strong performance with RL-SAE. Using the PADDLE predictor to estimate activation strength, we show in Table 2 that RL-SAE generates sequences with substantially higher PADDLE scores than those from SFT and steering. SFT, steering, and RL-SAE generate sequences with mean PADDLE scores approximately 89%, 128%, and 209% of those of natural ADs, respectively. Inspection of generated sequences (Appendix S) reveals that both steering and RL-SAE generate canonical ADs with hydrophobic stickers interspersed within acidic stretches (DelRosso et al., 2023; Staller et al., 2022), justifying the strong predicted activation strength.

For repression, which is frequently driven by SUMO (Small Ubiquitin MOdifier) modification and subsequent corepressor recruitment (DelRosso et al., 2023), we score sequences for the presence of ψKxE SUMOylation consensus motifs and their negative charge-dependent (NDSM) variants (Yang et al., 2006) (regexes in Appendix Q). We find that RL-SAE generates over 5 motifs per sequence, compared to 0.60 per sequence for natural sequences and 0.56 per sequence for steered generations.

SUMO modification is highly dynamic (Vertegaal, 2022), and interactions between SUMOylated proteins and their binding partners are weak and heavily dependent on avidity across multiple modified sites (Psakhye & Jentsch, 2012). Thus, a greater density of SUMOylation motifs is expected to substantially increase corepressor recruitment, both by stabilizing the SUMO modification itself, and by enabling multivalent interactions with corepressors (Banani et al., 2016), as is observed in some of the strongest experimentally measured human repressor sequences (DelRosso et al., 2023).

	Readout	Base	Nat.	SFT	Steering (α)	RL-SAE
Activation	PADDLE	$1.03_{.03}$	$5.34_{.18}$	$4.75_{.08}$	$\underline{6.81}_{.01}$ (0.43)	$\mathbf{11.16}_{.11}$
Repression	ψKxE	$0.07_{.02}$	$\underline{0.60}_{.05}$	$0.37_{.04}$	$0.56_{.02}$ (0.35)	$\mathbf{5.63}_{.60}$
	NDSM	$0.01_{.01}$	$0.10_{.02}$	$0.06_{.00}$	$\underline{0.21}_{.02}$ (0.35)	$\mathbf{1.80}_{.44}$

Table 2: RL-SAE generates sequence features associated with transcriptional regulation. Values are PADDLE scores and densities of SUMOylation motifs per sequence. Bold indicates the best score per row, underline the second best. Motif regexes are in Appendix Q. Subscripted values are standard deviations over three seeds, except for natural IDRs (Nat.), which report the standard error of the mean (Appendix I).

Composing chromosome association and effector features RL-SAE additionally allows us to compose arbitrary feature sets. To demonstrate this, we target the private-30 chromosome features together with the top-30 activation or repression domain features (chromosome + AD/RD) to mimic a synthetic transcriptional regulator that must both localize to chromatin and possess gene regulatory function. The experiments demonstrate that RL-SAE effectively installs on average 53 of the 60 target features in generated sequences (Figure 4a). Additionally, chromosome + AD sequences retain PADDLE scores close to those of AD-only designs while also scoring highly for chromosome association and nuclear localization. Chromosome + RD sequences similarly achieve strong predicted chromosome association and nuclear localization, with multiple SUMOylation motifs per sequence (Table 3).

	Activation	**Repression**		**Chromosome / nuclear**		
RL-SAE target	PADDLE	ψKxE	NDSM	ProtGPS	DeepLoc	cNLS
chromosome + AD	$11.00_{.10}$	$0.00_{.00}$	$0.00_{.00}$	$0.54_{.44}$	$0.76_{.14}$	$0.61_{.53}$
chromosome + RD	$-0.09_{.05}$	$2.89_{.10}$	$1.73_{.77}$	$0.73_{.14}$	$0.92_{.02}$	$0.99_{.02}$

Table 3: RL-SAE combines localization and transcriptional effector features. Evaluation of the two composed designs (chromosome + AD/RD), each targeting nuclear localization together with one gene-regulatory function. Values are PADDLE scores, SUMOylation motif density, predicted localization, and nuclear localization signal presence (cNLS). Grey cells mark the values which are biologically expected to score highly. Subscripted values are standard deviations over three seeds (Appendix I).

Sequence organization reflects feature compatibility We find that how these features are installed depends on the compatibility of the two feature sets with one another. For the chromosome

+ AD set, we find that there is minimal Jaccard overlap between residues that fire for chromosome features and those that fire for activation domain features (Figure 4b, red). Since chromosome associating features are basic and transcriptional activation features are acidic, the model learns to segregate them into two distinct charge blocks (example sequence in Figure 4d, upper), resulting in substantially greater charge blockiness (κ) for chromosome + AD sequences versus their scrambles (Figure 4c, red). On the other hand, substantially more residue overlap is observed between chromosome and repressor domain features: the features which are associated with transcriptional repression are partly basic and the model finds a way to partially overlap those motifs with basic chromosome features in the generated sequences, resulting in greater Jaccard overlap and lower κ (Figure 4b,c, blue). Together, these results suggest that RL-SAE can compose distinct grammars into generated sequences, with the model learning how to organize features to accommodate their composition and patterning requirements without explicit constraints.

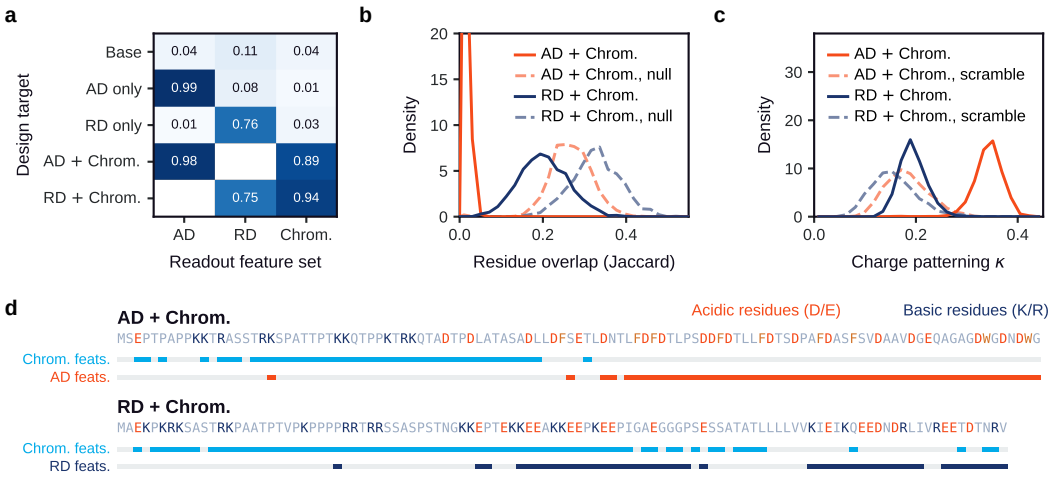

Figure 4: Feature compatibility determines the organization of sequences with composed features. **(a)** Fraction of features activated in each target feature set by sequences generated under each RL-SAE condition. The chromosome feature set is private-30. **(b)** Distribution of the per-sequence Jaccard overlap between residues which fire on chromosome association features versus on activation domain or repression domain features. Curves show each design and a position-shuffled null. **(c)** Distribution of charge patterning κ parameter for chromosome + AD/RD, and scrambled variants. **(d)** Example generated chromosome + AD/RD sequences showing feature firing positions for chromosome association (light blue) and the activation domain (AD, upper, red) or repression domain (RD, lower, dark blue).

6 Conclusions

We introduce **IDiom**, a protein language model for the generative design of IDRs, and **RL-SAE**, a post-training method that enables a generative model to activate specified sets of SAE features within generated sequences. Our experiments demonstrate that IDiom generates diverse sequences which recapitulate the properties of natural IDRs, while RL-SAE effectively installs tens of target features and produces sequences with strong predicted biological function. Together, IDiom and RL-SAE provide a framework for identifying and composing interpretable sequence features for IDR design, and future experimental validation could establish their utility for engineering biological function. More broadly, RL-SAE could enable the composable design of proteins by jointly targeting interpretable features, such as those associated with catalysis, binding, or stability.

Limitations and future work One limitation of RL-SAE is that it requires training, and scanning across many combinations of feature sets could be computationally intensive. Also, our top-N feature selection is simplistic and may omit functionally relevant features or include redundant ones. Finally, the evaluations in this paper are computational and experimental validation is needed to confirm whether the predicted properties translate into biological function.

Reproducibility statement

Code is available in the attached supplementary files. Dataset curation is described in Appendices A and F. Model and training details are described in Appendices B, D, and G. Feature selection is described in Appendix E. Baseline methods are described in Appendices G, H, and L. Evaluation procedures are provided in Appendices C, N–Q, and I. Trained checkpoints and IDiom-DB are not included in the supplementary files for review, but will be included in the public release.

Ethics statement

This work develops a computational method which enables the generative modeling of intrinsically disordered protein regions. The evaluations in this work are computational, and generated sequences must undergo appropriate experimental validation and biosafety assessment before future use. In addition, such protein sequence generation methods may be repurposed for harmful use, and this potential should be considered when distributing and applying these methods. We confirm that we comply with the licenses and attribution requirements of all datasets used in this work.

AI use statement

In this work, we used generative AI tools to assist in finding relevant literature sources, to aid in editing manuscript text, and to help implement research code. We extensively revised all AI-edited text and manually verified all AI-implemented code. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 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.

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025. doi: 10.1101/2025.02.06.636901.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M. Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, Magnus S. Bauer, Regina Barzilay, Tommi S. Jaakkola, Rohith Krishna, and David Baker. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods*, 23(1):96–105, January 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02975-x. URL https://www.nature.com/articles/s41592-025-02975-x.

T. Reid Alderson, Iva Pritišanac, Đesika Kolarić, Alan M. Moses, and Julie D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proceedings of the National Academy of Sciences*, 120(44):e2304302120, October 2023. doi: 10.1073/pnas.2304302120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2304302120.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International*

Conference on Learning Representations (ICLR), 2020. URL https://openreview.net/forum?id=HklxbgBKvr.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of EMNLP*, 2025. URL https://arxiv.org/abs/2505.20063.

Salman F. Banani, Allyson M. Rice, William B. Peeples, Yuan Lin, Saumya Jain, Roy Parker, and Michael K. Rosen. Compositional Control of Phase-Separated Cellular Bodies. *Cell*, 166 (3):651–663, July 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.06.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416307395.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient Training of Language Models to Fill in the Middle, July 2022. URL http://arxiv.org/abs/2207.14255. arXiv:2207.14255 [cs].

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C Curran, Alexander M Hoffnagle, Kyle S Ching, Michael Martyn, Stephen Nayfach, Jeffrey A Ruffolo, and Ali Madani. Scaling Unlocks Broader Generation and Deeper Functional Understanding of Proteins. *NeurIPS*, 2025.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language Modeling Materializes a World Model of Protein Biology, June 2026. URL https://www.biorxiv.org/content/10.64898/2026.06.03.729735v1. ISSN: 2692-8205 Pages: 2026.06.03.729735 Section: New Results.

Shriram Chennakesavalu, Frank Hu, Sebastian Ibarraran, and Grant M. Rotskoff. Aligning Transformers with Continuous Feedback via Energy Rank Alignment. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL http://arxiv.org/abs/2405.12961.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. Control reinforcement learning: Token-level mechanistic analysis via learned SAE feature steering. 2026. URL https://arxiv.org/abs/2602.10437.

Gerard Corominas, Filippo Stocco, and Noelia Ferruz. Sparse autoencoders in protein engineering campaigns: Steering and model diffing. In *2nd Generative AI and Biology Workshop, International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=rnJ6Nn1Wf5.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. URL http://arxiv.org/abs/2205.14135. arXiv:2205.14135 [cs].

Rahul K. Das and Rohit V. Pappu. 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.

Rahul K Das, Kiersten M Ruff, and Rohit V Pappu. 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.

Carolyn J. Decker and Roy Parker. 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/.

Nicole DelRosso, Josh Tycko, Peter Suzuki, Cecelia Andrews, Aradhana, Adi Mukund, Ivan Liongson, Connor Ludwig, Kaitlyn Spees, Polly Fordyce, Michael C. Bassik, and Lacramioara Bintu. 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.

H. Jane Dyson, Peter E. Wright, and Harold A. Scheraga. 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.

Gábor Erdős, Mátyás Pajkos, and Zsuzsanna Dosztányi. 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.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 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.

Steffen Frey and Dirk Görlich. FG/FxFG as well as GLFG repeats form a selective permeability barrier with self-healing properties. *The EMBO Journal*, 28(17):2554–2567, September 2009. ISSN 1460-2075. doi: 10.1038/emboj.2009.199. URL https://doi.org/10.1038/emboj.2009.199.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching, May 2026. URL http://arxiv.org/abs/2507.09466. arXiv:2507.09466 [cs.LG].

Garrett M. Ginell, Ryan J. Emenecker, Jeffrey M. Lotthammer, Alex T. Keeley, Stephen P. Plassmeyer, Nicholas Razo, Emery T. Usher, Jaqueline F. Pelham, and Alex S. Holehouse. 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.

Nicolás S. González-Foutel, Juliana Glavina, Wade M. Borcherds, Matías Safranchik, Susana Barrera-Vilarmau, Amin Sagar, Alejandro Estaña, Amelie Barozet, Nicolás A. Garrone, Gregorio Fernandez-Ballester, Clara Blanes-Mira, Ignacio E. Sánchez, Gonzalo de Prat-Gay, Juan Cortés, Pau Bernadó, Rohit V. Pappu, Alex S. Holehouse, Gary W. Daughdrill, and Lucía B. Chemes. 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.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. 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.

Alex S. Holehouse. 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.

Alex S. Holehouse and Birthe B. Kragelund. 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.

Kara Hunter, Trevor Brandt, Karina Guadalupe, Kavindu Chathuranga Kolamunna, Jeffrey M. Lotthammer, Nora M. Shamoon, Brooke Nicholson, Lea Day, Alec Martinez, Alex S. Holehouse, Shahar Sukenik, and Ryan J. Emenecker. 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.

Sebastian Ibarraran, Shriram Chennakesavalu, Frank Hu, and Grant M. Rotskoff. Efficient, Few-Shot Directed Evolution with Energy Rank Alignment. *Journal of Chemical Information and Modeling*, 66(17):10609–10621, September 2026. doi: 10.1021/acs.jcim.6c01333. URL https://doi.org/10.1021/acs.jcim.6c01333.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 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.

Henry R. Kilgore, Itamar Chinn, Peter G. Mikhael, Ilan Mitnikov, Catherine Van Dongen, Guy Zylberberg, Lena Afeyan, Salman F. Banani, Susana Wilson-Hawken, Tong Ihn Lee, Regina Barzilay, and Richard A. Young. 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.

Ming-Chih Lai, Ru-Inn Lin, and Woan-Yuh Tarn. Transportin-SR2 mediates nuclear import of phosphorylated SR proteins. *Proceedings of the National Academy of Sciences*, 98(18):10154–10159, August 2001. doi: 10.1073/pnas.181354098. URL https://www.pnas.org/doi/10.1073/pnas.181354098.

Allison Lange, Ryan E. Mills, Christopher J. Lange, Murray Stewart, Scott E. Devine, and Anita H. Corbett. Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin alpha *. *Journal of Biological Chemistry*, 282(8):5101–5105, February 2007. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.R600026200. URL https://www.jbc.org/article/S0021-9258(20)68801-9/abstract.

Iris Langstein-Skora, Andrea Schmid, Frauke Huth, Drin Shabani, Lorenz Spechtenhauser, Mariia Likhodeeva, Franziska Kunert, Felix J. Metzner, Ryan J. Emenecker, Mary O. Richardson, Wasim Aftab, Maximilian J. Götz, Sarah K. Payer, Niccoló Pietrantoni, Valentina Valka, Sakthi K. Ravichandran, Till Bartke, Karl-Peter Hopfner, Ulrich Gerland, Philipp Korber, and Alex S. Holehouse. 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.

Yi-Hsuan Lin, Jacob P Brady, Julie D Forman-Kay, and Hue Sun Chan. 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.

Jeffrey M. Lotthammer, Jorge Hernández-García, Daniel Griffith, Dolf Weijers, Alex S. Holehouse, and Ryan J. Emenecker. 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.

Malisa Vittoria Mantonico, Federica De Leo, Giacomo Quilici, Liam Sean Colley, Francesco De Marchis, Massimo Crippa, Rosanna Mezzapelle, Tim Schulte, Chiara Zucchelli, Chiara Pastorello, Camilla Carmeno, Francesca Caprioglio, Stefano Ricagno, Gabriele Giachin, Michela Ghitti, Marco Emilio Bianchi, and Giovanna Musco. 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.

Erik W. Martin, Alex S. Holehouse, Ivan Peran, Mina Farag, J. Jeremias Incicco, Anne Bremer, Christy R. Grace, Andrea Soranno, Rohit V. Pappu, and Tanja Mittag. 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.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 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.

Diana M. Mitrea, Jaclyn A. Cika, Christopher B. Stanley, Amanda Nourse, Paulo L. Onuchic, Priya R. Banerjee, Aaron H. Phillips, Cheon-Gil Park, Ashok A. Deniz, and Richard W. Kriwacki. 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.

Maria Victoria Nugnes, Kamel Eddine Adel Bouhraoua, Mehdi Zoubiri, Rita Pancsa, Erzsébet Fichó, DisProt Consortium, Peter Tompa, Damiano Piovesan, Silvio C E Tosatto, and Maria Cristina Aspromonte. 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.

Francesco Pesce, Anne Bremer, Giulio Tesei, Jesse B. Hopkins, Christy R. Grace, Tanja Mittag, and Kresten Lindorff-Larsen. 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.

Zhixuan Piao, Yuan Liu, Xiaoyong Pan, and Hong-Bin Shen. A Co-attention Mechanism of Jointly Capturing Sequence and Structure Features for Signal Peptide Prediction. *Computational and Structural Biotechnology Journal*, 35(1):0120, July 2026. doi: 10.34133/csbj.0120. URL https://spj.science.org/doi/10.34133/csbj.0120.

Ivan Psakhye and Stefan Jentsch. Protein Group Modification and Synergy in the SUMO Pathway as Exemplified in DNA Repair. *Cell*, 151(4):807–820, November 2012. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2012.10.021. URL https://www.cell.com/cell/abstract/S0092-8674(12)01241-X.

Pedro Romero, Zoran Obradovic, Xiaohong Li, Ethan C. Garner, Celeste J. Brown, and A. Keith Dunker. 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.

Kiersten M. Ruff, Matthew R. King, Alexander W. Ying, Vicky Liu, Avnika Pant, Whitney E. Lieberman, Min Kyung Shinn, Xiaolei Su, Cigall Kadoch, and Rohit V. Pappu. 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.

Adrian L Sanborn, Benjamin T Yeh, Jordan T Feigerle, Cynthia V Hao, Raphael JL Townshend, Erez Lieberman Aiden, Ron O Dror, and Roger D Kornberg. 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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. 2024. URL https://arxiv.org/abs/2402.03300.

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

Elana Simon and James Zou. 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.

Max V. Staller, Eddie Ramirez, Sanjana R. Kotha, Alex S. Holehouse, Rohit V. Pappu, and Barak A. Cohen. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Systems*, 13(4):334–345.e5, April 2022. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2022.01.002. URL https://www.cell.com/cell-systems/abstract/S2405-4712(22)00002-3.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988.

Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding Generative Protein Language Models with Reinforcement Learning, November 2025. URL http://arxiv.org/abs/2412.12979. arXiv:2412.12979 [q-bio.BM].

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, November 2023. URL http://arxiv.org/abs/2104.09864. arXiv:2104.09864 [cs].

Giulio Tesei, Anna Ida Trolle, Nicolas Jonsson, Johannes Betz, Frederik E. Knudsen, Francesco Pesce, Kristoffer E. Johansson, and Kresten Lindorff-Larsen. 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.

Giulio Tesei, Francesco Pesce, and Kresten Lindorff-Larsen. 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.

Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40(7):1023–1025, July 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01156-3. URL https://www.nature.com/articles/s41587-021-01156-3.

The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose Da Costa Gonzales, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu, Rossana Zaru, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard De Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Anastasia Sveshnikova, Cathy H Wu, Cecilia N Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Yuqi Wang, and Jian Zhang. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkae1010. URL https://academic.oup.com/nar/article/53/D1/D609/7902999.

Francois-Xavier Theillet, Lajos Kalmar, Peter Tompa, Kyou-Hoon Han, Philipp Selenko, A. Keith Dunker, Gary W. Daughdrill, and Vladimir N. Uversky. 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.

Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. 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.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. 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.

Alfred C. O. Vertegaal. Signalling mechanisms and cellular functions of SUMO. *Nature Reviews Molecular Cell Biology*, 23(11):715–731, November 2022. ISSN 1471-0080. doi: 10.1038/s41580-022-00500-y. URL https://www.nature.com/articles/s41580-022-00500-y.

Edith Natalia Villegas Garcia and Alessio Ansuini. Interpreting and steering protein language models through sparse autoencoders. 2025. URL https://arxiv.org/abs/2502.09135.

Vaishali P Waman, Nicola Bordin, Andy Lau, Shaun Kandathil, Jude Wells, David Miller, Sameer Velankar, David T Jones, Ian Sillitoe, and Christine Orengo. 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.

Ming-Tzo Wei, Shana Elbaum-Garfinkle, Alex S. Holehouse, Carlos Chih-Hsiung Chen, Marina Feric, Craig B. Arnold, Rodney D. Priestley, Rohit V. Pappu, and Clifford P. Brangwynne.

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.

Carter J. Wilson, Wing-Yiu Choy, and Mikko Karttunen. 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.

John C. Wootton and Scott Federhen. 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.

Peter E. Wright and H. Jane Dyson. 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.

Jiaqing Xie. A Comparative Analysis of Sparse Autoencoder and Activation Difference in Language Model Steering, September 2025. URL http://arxiv.org/abs/2510.01246. arXiv:2510.01246 [cs.CL].

Shen-Hsi Yang, Alex Galanis, James Witty, and Andrew D. Sharrocks. An extended consensus motif enhances the specificity of substrate modification by SUMO. *The EMBO Journal*, 25(21): 5083–5093, November 2006. ISSN 1460-2075. doi: 10.1038/sj.emboj.7601383. URL https://doi.org/10.1038/sj.emboj.7601383.

Shuyi Zhang, Wei Shi, Sihang Li, Jiayi Liao, Hengxing Cai, and Xiang Wang. Interpretable reward model via sparse autoencoder. 2025. URL https://arxiv.org/abs/2508.08746. AAAI 2026.

Bi Zhao, Sina Ghadermarzi, and Lukasz Kurgan. 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.

Wenwei Zheng, Gregory Dignon, Matthew Brown, Young C. Kim, and Jeetain Mittal. 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.

Marius Thrane Ødum, Felix Teufel, Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Research*, 52(W1): W215–W220, July 2024. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkae237. URL https://academic.oup.com/nar/article/52/W1/W215/7642068.

APPENDIX

A IDIOM-DB CONSTRUCTION

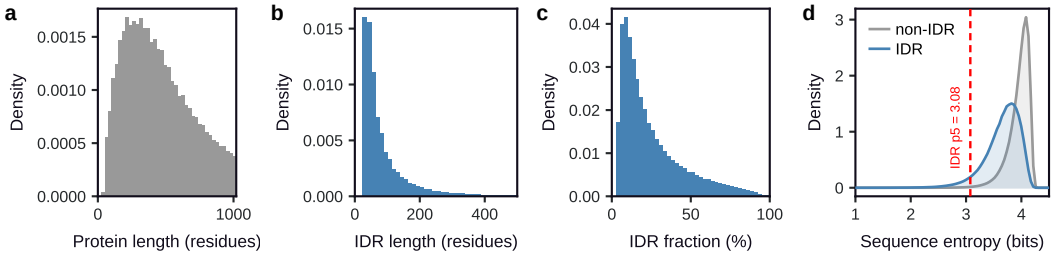

Figure 5: Characterization of IDiom-DB. (a) Length distribution of full-length protein sequences within which the IDRs are located, of the final curated dataset. **(b)** Length distribution of IDRs within the final curated dataset. **(c)** Distribution of the fraction of residues of the full-length sequence which are IDR residues. **(d)** Distribution of sequence entropy for IDR versus non-IDR regions in the final curated dataset. The vertical red dashed line indicates the sequence entropy threshold at the lower 5th percentile.

Data curation We curate IDiom-DB 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 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 $\text{pLDDT} > 80$ as folded, $\text{pLDDT} < 70$ as disordered, and $70 \leq \text{pLDDT} \leq 80$ 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 1,020 residues, IDRs shorter than 30 residues long, and sequences whose entire length is low-pLDDT are discarded. Finally, we deduplicate DisProt sequences from this dataset and filter out signal peptides (both described below). This curation process yields 54M 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 is 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 determine the sequence identity of generated sequences with respect to the training set. Specifically, we compare prompted and unprompted IDR sequences against the 54M IDRs in IDiom-DB (IDR vs IDR only). MMseqs2 is run with the following command to search the generated sequences against the IDiom-DB 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, whose distribution is plotted in Figure 1f.

DisProt deduplication We use the experimentally validated DisProt IDRs as a ground truth IDR reference for both dataset curation as well as for evaluating our generated sequences. To prevent leakage of these sequences into the training data, we remove IDRs from the curated training dataset which have high sequence similarity to a DisProt IDR. DisProt IDRs are first extracted using experimentally annotated disordered regions (consensus structural state "D") (Nugnes et al., 2026). IDRs which are at least 30 residues long, which reside in proteins whose full length is $\leq$ 1,020 residues, and which are not 100% disordered are retained, yielding 1,663 benchmark IDRs. We next use MMseqs2 to search these DisProt IDRs against the IDR spans of the 54M curated training IDRs, and we discard any training record whose IDR aligns with a DisProt IDR with at least 50% sequence identity and 80% bidirectional coverage (`-cov-mode 0`). MMseqs2 is run using the command: `mmseqs`

`search --min-seq-id 0.5 -c 0.8 --cov-mode 0 -s 7 --max-seqs 300`, and this procedure removes 55K records.

Signal peptide filtering Signal peptides are short N-terminal sequences which are often cleaved after protein expression and subsequently degraded. Residues corresponding to N-terminal signal peptides often have low AF2 pLDDT scores and are disordered in solution. However, they are cleaved and thus absent from the fully mature proteins, so we do not consider them as suitable IDRs for our dataset (Piao et al., 2026). We therefore use the SignalP 6.0 predictor to identify potential signal peptide cleavage sites for filtering (Teufel et al., 2022). We consider N-terminal subsequences to be signal peptides if they precede a SignalP cleavage residue that has a predicted probability ≥ 0.9. If an identified IDR contains such a signal peptide, we trim those residues and keep the remainder of the IDR. If the trimmed IDR length falls below 30 residues, then we discard the entry entirely. This process removes 3.6M records, yielding the final set of 54M IDRs.

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 1,024 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 IDiom-DB for model pretraining, 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 unprompted IDRs 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 108M sequences for training (54M prompted IDRs and 54M unprompted IDRs), which are split 99% training, 0.5% validation, and 0.5% test.

B MODEL ARCHITECTURE, PRETRAINING, AND INFERENCE

Model architecture IDiom is a decoder-only transformer which we train in three model sizes: 19M (6 layers, hidden dimension 512), 85M (12 layers, hidden dimension 768), and 302M parameters (24 layers, hidden dimension 1,024). The model employs pre-normalization with RMSNorm, applies query-key normalization within each attention block, and uses the SwiGLU non-linearity (Shazeer, 2020) in all feedforward networks. The feedforward expansion ratio is 8/3, and positional information is encoded using Rotary Position Embeddings (RoPE) (Su et al., 2023). The model processes sequences with a maximum length of 1,024 tokens, and shorter sequences are padded to this length. Multi-head attention is calculated using Flash Attention (Dao et al., 2022).

Pretraining We pretrain IDiom on the aforementioned 108M sequences. The data are randomly split into 99% training, 0.5% validation, and 0.5% test sets. All sequences are padded to a fixed length of 1,024 tokens. The model is trained autoregressively using next-token prediction.

Training is performed using the AdamW optimizer and no weight decay. Learning rates of 4.0×10^{-4}, 3.5×10^{-4}, and 3.0×10^{-4} are used for the 19M, 85M, and 302M parameter models, respectively. The learning rate schedule uses a linear warmup over the first 3,000 steps, followed by cosine annealing decay to a minimum of 10% the initial learning rate over 250,000 total training steps. The global batch size is 1,024 (Distributed Data Parallel training). We use a cross-entropy loss while ignoring contributions from pad or mask tokens. Training is performed in mixed precision (fp32/bfloat16) and with gradient clipping at 1.0. Model validation is performed every 25,000 training steps on the validation split, and training is run for 250,000 optimizer steps.

Model inference We generate IDR sequences from IDiom using autoregressive decoding. At each position t in the sequence, we calculate 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, 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 1,024 tokens.

C CHARACTERIZATION OF GENERATED SEQUENCES

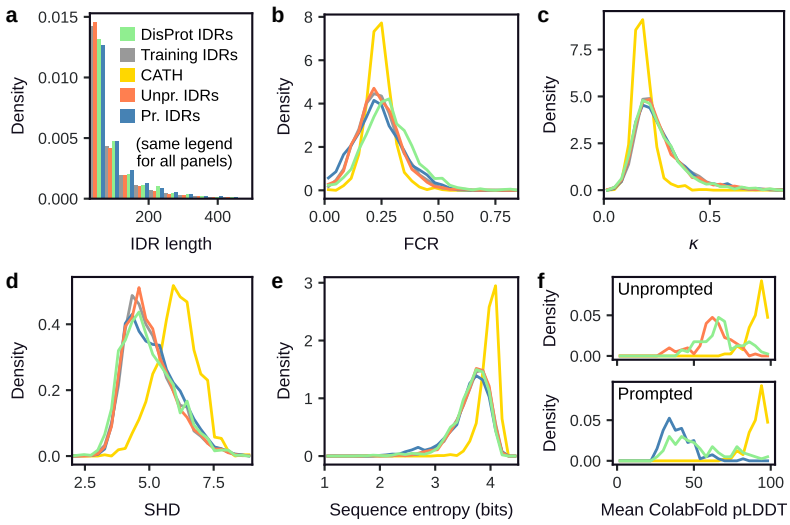

Figure 6: Characterization of training IDRs, IDiom-generated IDRs, DisProt IDRs, and CATH sequences. **(a)** Distribution of training, generated, and DisProt IDR lengths. The legend in this panel applies to all other panels in this figure. **(b)** Distribution of the fraction of charged residues (FCR) for training, generated, and DisProt IDRs, and CATH sequences. **(c)** Distribution of charge patterning κ parameter. **(d)** Distribution of sequence hydropathy decoration (SHD). **(e)** Distribution of sequence entropy. **(f)** Distribution of AlphaFold2 scores for CATH sequences, IDiom-generated IDRs, unprompted (upper) and prompted (lower), as well as DisProt IDRs without surrounding context (upper) and with context (lower).

Generated sequences To characterize the pretrained 302M parameter IDiom model, we generate 100,000 unprompted IDRs and 166,300 IDRs conditioned on the N- and C-terminal flanking contexts of 1,663 DisProt IDRs (100 IDRs per protein).

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

Sequence patterning 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). IDiom-generated sequences 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 assess how well IDiom recapitulates these properties, we calculate the distributions of four sequence-level descriptors (discussed below) for the generated, training, and DisProt disordered sequences, as well as for the folded CATH domain sequences (Figure 6b-e), together with the generated IDR length distribution (Figure 6a) and AF2 pLDDTs (Figure 6f).

Charge content 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 6b) distinguishes strongly charged polyampholytes from weakly charged sequences. We find that the generated and natural IDRs 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).

Charge patterning 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). Values of $\kappa \approx 0$ indicate well-mixed opposite charges and $\kappa \approx 1$ indicates segregation into blocks of the same charge, and we plot these distributions in Figure 6c. 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 in its generated sequences.

Hydrophobic patterning We next examine 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 6d), 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, whose higher SHD values reflect the locally concentrated hydrophobic residues required to stabilize buried protein cores (Dyson et al., 2006).

Sequence entropy Finally, we calculate each sequence's entropy as the Shannon entropy (bits) of its amino acid composition. IDRs frequently contain low-complexity segments (Romero et al., 2001), and natural DisProt and training set IDRs show lower entropy than folded CATH domains. Prompted and unprompted IDRs closely reproduce this shift, with the entropy of generated sequences matching the DisProt distribution well (Figure 6e). Taken together, these results show that IDiom has learned the sequence grammar of disordered regions across multiple descriptors, and that generated sequences recapitulate the sequence patterning features of natural IDRs.

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 the fraction of charged residues, the linear charge patterning parameter κ, and sequence hydropathy decoration.

D SPARSE AUTOENCODER TRAINING AND CHARACTERIZATION

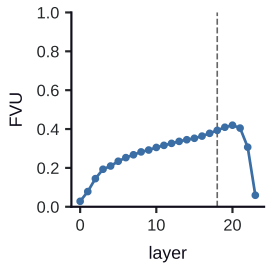

Figure 7: Sparse autoencoder layer sweep. Fraction of variance unexplained (FVU) per layer, evaluated on 10K held-out validation records ($k = 32$, expansion $\times 16$). The layer-18 is marked by the vertical dashed line.

Sparse autoencoder training We train TopK sparse autoencoders (Cunningham et al., 2023; Gao et al., 2024), IDiomSAEs, on the residual stream of all layers of the base IDiom model:

$$\boldsymbol{z} = \mathrm{TopK}\big(\,\mathrm{ReLU}(\boldsymbol{W}_{\text{enc}}(\boldsymbol{x} - \boldsymbol{b}_{\text{dec}}) + \boldsymbol{b}_{\text{enc}})\big), \qquad \hat{\boldsymbol{x}} = \boldsymbol{W}_{\text{dec}}\,\boldsymbol{z} + \boldsymbol{b}_{\text{dec}}$$

with $k = 32$ active latents from 16,384 total ($16\times$ expansion factor), and where $\boldsymbol{b}_{\text{dec}}$ is initialized from the mean of the training activations. The model is trained to minimize the fraction of variance unexplained: $\mathcal{L}_{\text{FVU}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 / \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2$, with an auxiliary AuxK loss (with weight 0.03125) to revive dead latents (Gao et al., 2024). Training activations are taken only from IDR residue positions, and only unprompted IDRs are used for training. IDiomSAEs are trained using Adam with a learning rate of 2×10^{-4} and batch size of 4,096, for 50,000 optimizer steps. We use a 1,000 step warmup. For all analysis in this paper, we use the layer-18 IDiomSAE.

E FEATURE ENRICHMENT PROCEDURE

Significance testing To determine the SAE features which are most enriched in a set of positively labeled sequences $\mathcal{P}$ relative to a set of negatively labeled sequences $\mathcal{N}$, we follow the method of Candido et al. (2026). For each feature j of the 16,384 SAE features, we set a_j and b_j to be the number of sequences in $\mathcal{P}$ and $\mathcal{N}$, respectively, in which j fires, $K_j = a_j + b_j$ to be the total firing count pooled across both sets, and $M = |\mathcal{P}| + |\mathcal{N}|$ to be the total pooled sequence count. Under the null hypothesis that firing is independent of $\mathcal{P}$ or $\mathcal{N}$ membership, drawing $|\mathcal{P}|$ sequences without replacement from the pooled M (of which K_j fire) gives a firing count that is hypergeometrically distributed with mean μ_j and variance σ_j^2. We standardize the observed count a_j against this null:

$$\mu_j = \frac{|\mathcal{P}|\,K_j}{M}, \qquad \sigma_j^2 = \frac{K_j\,(M - K_j)\,|\mathcal{P}|\,(M - |\mathcal{P}|)}{M^2\,(M-1)}, \qquad Z_j = \frac{a_j - \mu_j}{\sqrt{\sigma_j^2}}. \tag{6}$$

We then calculate a two-sided p-value from Z_j and control the false discovery rate with the Benjamini-Hochberg procedure. The log odds ratio (LOR, Equation 2) is calculated with a 0.5 pseudocount added to all four contingency table cells.

Feature filtering and ranking We discard features whose peak activating position lies within 2 residues of either edge of an IDR in $\geq 50\%$ of feature firings, since these are likely artifacts from the process of extracting IDRs from their full sequence context. Then, we filter out features with a false discovery rate of $> 10^{-3}$, similarly remove features with LOR ≤ 1, and only consider features j which fire in at least 5% of sequences in the positive set $\mathcal{P}$. We rank the remaining features by LOR and take the top-N as the feature set for analysis. All results in the main text use $N = 30$.

F COMPARTMENT AND EFFECTOR DOMAIN SEQUENCES

Biomolecular compartments To curate a positive set $\mathcal{P}$ of IDR sequences with annotated localization to various biomolecular compartments, we start with the dataset curated by Kilgore et al. (2025). That dataset consists of a curated set of human full-length protein sequences with UniProt or CD-CODE annotated association with 12 biomolecular condensates. We further process the dataset by first selecting six compartments for analysis: nucleolus, chromosome, stress granule, P-body, nuclear speckle, and the nuclear pore complex (NPC), and then removing sequences with annotated association with more than one compartment. Next, we run the Metapredict V3 disorder predictor on these sequences to isolate the IDRs and only keep IDRs with length 30–1,020 residues. We then use these sequences as the $\mathcal{P}$ set. The $\mathcal{N}$ set is 100,000 length-matched IDRs randomly selected from IDiom-DB.

Effector domains To curate a positive set $\mathcal{P}$ of human transcriptional effector domain sequences, we start with a dataset of 80-residue sequences from a high-throughput tiling screen (DelRosso et al., 2023). For both the activation domain and repression domain datasets, we apply the same processing. We first merge overlapping hit sequences to avoid duplication. Next, we run the Metapredict V3 disorder predictor on the full-length parent proteins of the sequences and keep only sequences where at least 50% of residues are predicted disordered. This yields 288 AD sequences and 408 RD sequences. For the negative set, we use sequences which are non-hits in all three experimental screens (ON+OFF counts > 50 and average score below the control mean $+$ 2 SD), without merging overlaps, and with the same Metapredict V3 IDR filtering. This yields 4,920 inert tiles as the $\mathcal{N}$ set.

G RL-SAE DETAILS

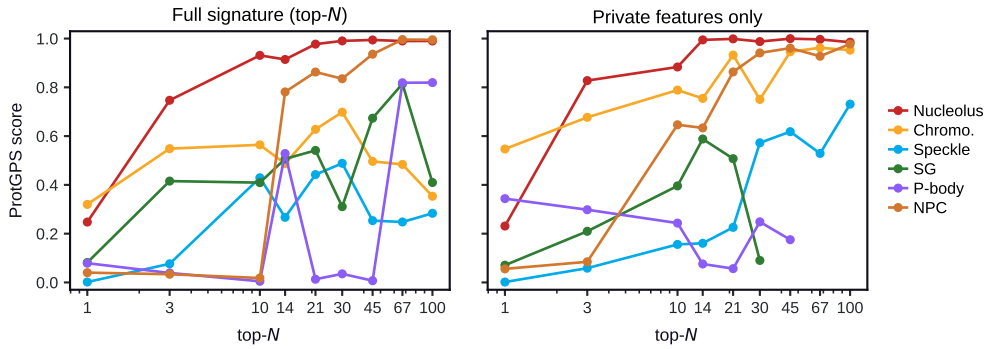

Figure 8: ProtGPS dose-response against top-N. Left: Average ProtGPS score for sequences generated after RL-SAE on the top-N features per compartment. Right: Average ProtGPS score for sequences generated after RL-SAE on the top-N features which are private to the given compartment. Stress granule and P-body only have 36 and 55 private features, respectively.

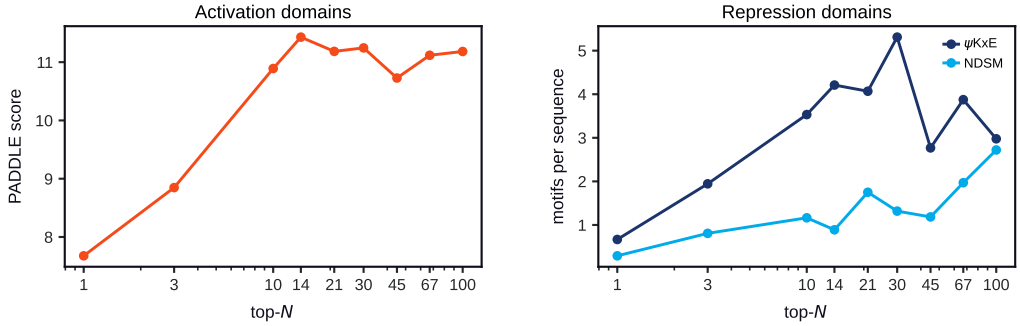

Figure 9: Transcriptional effector domain dose-response against top-N. Left: PADDLE score for activation domains generated after RL-SAE on the top-N activation domain features. Right: ψKxE and NDSM motif density in sequences generated after RL-SAE on the top-N repression domain features.

Post-training We post-train IDiom using group relative policy optimization (GRPO) (Shao et al., 2024). During GRPO training, for each batch of sequence prompts, the model generates multiple IDRs per prompt. Rewards are calculated for each generated sequence using the reward function defined in Equation 3. For RL-SAE, feature activations are computed using the frozen pretrained IDiom model and IDiomSAE. When performing RL on the ProtGPS oracle baseline, $R_S(y)$ is replaced with the ProtGPS score (value between 0 and 1) for the target compartment.

Advantages are calculated as normalized relative rewards per group: $A_i = \frac{r_i - \bar{r}_g}{\sigma_g}$, where $\bar{r}_g$ and σ_g are the group-wise mean and standard deviation. The GRPO loss combines PPO-style clipping ($\epsilon_{\text{clip}} = 0.2$) with a KL penalty ($\beta_{\text{KL}} = 0.02$), applied per-token on completions only. We use a learning rate of 5×10^{-6} with AdamW, group size of 8, and global batch size of 8. Training is performed for 3,000 steps, and all runs generate unprompted IDRs with a temperature of 1.0.

Top-N score dependence In Figures 8 and 9 we plot the ProtGPS, PADDLE, and repression domain motif scores as a function of top-N features on which RL-SAE is applied. Across most metrics, we observe that there is generally an upward trend of score versus top-N, but with substantial variability in ProtGPS scores. For all experiments in the main text, we choose the top-30 features for analysis because it is the largest N on our sweep grid at which all six compartments have sufficient private features (Figure 8, right panel).

H BASELINE METHODS

Activation steering For the same target feature sets as in RL-SAE, we sum and normalize the decoder vectors $\boldsymbol{W}_{\text{dec}}$ into the steering direction $\hat{\boldsymbol{d}}_S$:

$$\boldsymbol{d}_S = \sum_{j \in S} \boldsymbol{W}_{\text{dec}}[:, j], \qquad \hat{\boldsymbol{d}}_S = \frac{\boldsymbol{d}_S}{\|\boldsymbol{d}_S\|_2}, \tag{7}$$

where $\boldsymbol{W}_{\text{dec}}[:, j]$ is the decoder column corresponding to feature j. At every token position t during inference, we modify the residual stream vector $\boldsymbol{x}_t$ at layer-18 of IDiom by adding $\hat{\boldsymbol{d}}_S$ scaled by some fraction α of the original residual stream vector's magnitude before renormalizing:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t + \alpha \|\boldsymbol{x}_t\|_2 \, \hat{\boldsymbol{d}}_S, \qquad \boldsymbol{x}'_t = \|\boldsymbol{x}_t\|_2 \, \frac{\tilde{\boldsymbol{x}}_t}{\|\tilde{\boldsymbol{x}}_t\|_2} \tag{8}$$

and then substitute $\boldsymbol{x}_t$ with $\boldsymbol{x}'_t$ in the forward pass.

We steer using the same top-N feature sets as RL-SAE, sweeping 40 values of α between 0 (base model) and 1.5 which are spaced more densely at low α. Sequences are generated unprompted at a temperature of 1.0. We only keep generated sequences 50–150 residues in length, and we oversample to generate the required number of sequences. We discard any α whose generated sequences have a mean sequence entropy below the lower 5th percentile of IDR sequence entropies in IDiom-DB (3.08 bits), and we report steering results at the remaining α (parenthesized in tables) which achieves the highest score.

Supervised fine-tuning For each target, we fine-tune the base checkpoint on that target's positive set $\mathcal{P}$ (Appendix F), holding out 10% of sequences for validation, and using the unprompted mode with a completion-only loss. We train with AdamW with a learning rate of 1×10^{-5}, weight decay 0.1, gradient clipping at 1.0, with a batch size of 16, for 1000 optimizer steps (number of epochs ranges from roughly 21 to 178 depending on the size of $\mathcal{P}$), with a 100 step warmup and then cosine decay.

Oracle RL baseline For biomolecular compartment association, we use the ProtGPS score (0 to 1) as the baseline reward in lieu of $R_S(y)$ in Equation 3. ProtGPS is trained using the full-length protein sequences of the same per-compartment positive sequence sets. We refer to these runs as RL-GPS, and all hyperparameters are otherwise the same as in RL-SAE.

I REPORTED UNCERTAINTIES

Variability across seeds For the three-seed table comparisons, we report the mean and standard deviation of a score across three seeds for each generative method. For RL-SAE, RL-GPS, and SFT, the three seeds are independent training runs with different random seeds. We score 1000 sequences generated from each final checkpoint, calculate a per-run average, and report the mean and standard deviation of the three per-run averages. For the base model and activation steering (no training), the three seeds are independent generation seeds (of 1000 sequences), and we report the mean and standard deviation across the three generated sets in the same way. The scrambled control is only done for RL-SAE and is calculated in the same way. Scores are calculated over 1000 generated sequences for all metrics except DeepLoc predictions, for which we use a 200 sequence subset due to computational cost. For natural IDR sets we report the mean and its standard error over all sequences in the set.

J TOP-N COVERAGE

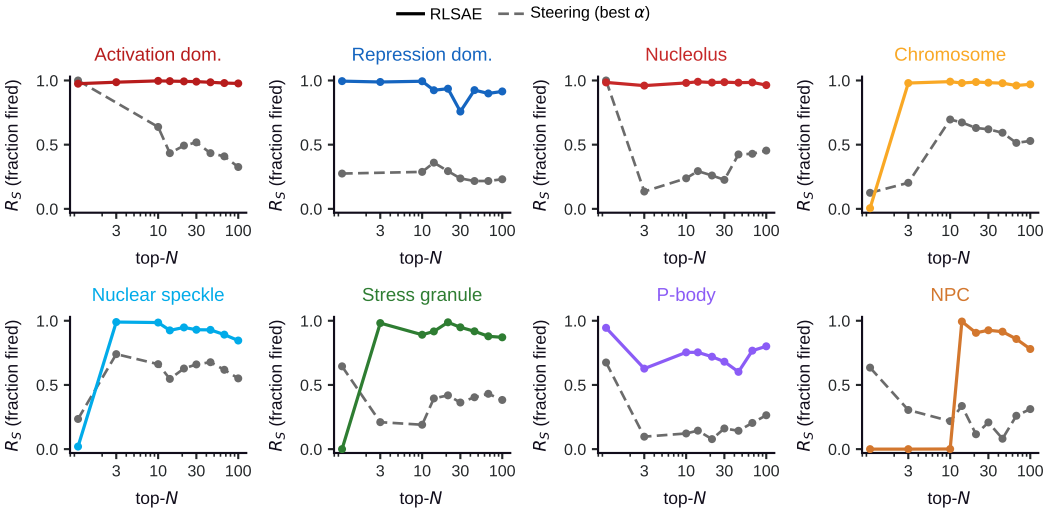

Figure 10: Feature activation fraction versus top-N. For all 8 sequence sets (2 transcriptional effector sequences and 6 compartment association sequences) the average fraction of target features which fire within generated sequences, R_S, as a function of top-N, for RL-SAE (solid colored lines) and activation steering (grey dashed lines). R_S for steering is presented at the best α value across the sweep, regardless of whether that α passes the sequence entropy threshold (i.e. steering R_S values presented here are a generous upper bound).

Target feature coverage In Figure 10, we plot the average fraction of the target features which fire within generated sequences, $R_S(y)$ (Equation 3), as a function of top-N for RL-SAE versus steering. For this comparison, we plot the best $R_S(y)$ achievable by steering, regardless of whether the corresponding α passes the sequence entropy threshold.

We find that when steering, for 5 of 8 sequence sets (activation domains, nucleolus, stress granule, P-body, NPC), $R_S(y)$ starts at a high value at $N = 1$ and drops as N is increased. When steering towards a number of directions beyond $N = 1$, the activations are pushed towards the centroid of the set of directions. We hypothesize that as N grows, this steered centroid direction becomes less representative of any individual feature's direction, hence the lower values of $R_S(y)$ as N is increased.

On the other hand, we observe that RL-SAE is able to effectively attain $R_S(y)$ values near 1.0 across most compartments and most N values. The exceptions we find are several low-N cases, such as $N = 1$ for chromosome, nuclear speckle, and stress granule, and $N = 1, 3, 10$ for NPC. Potentially, this is due to the rarity of some of these few top features: these features may be infrequently sampled by the base model and are thus difficult to optimize for.

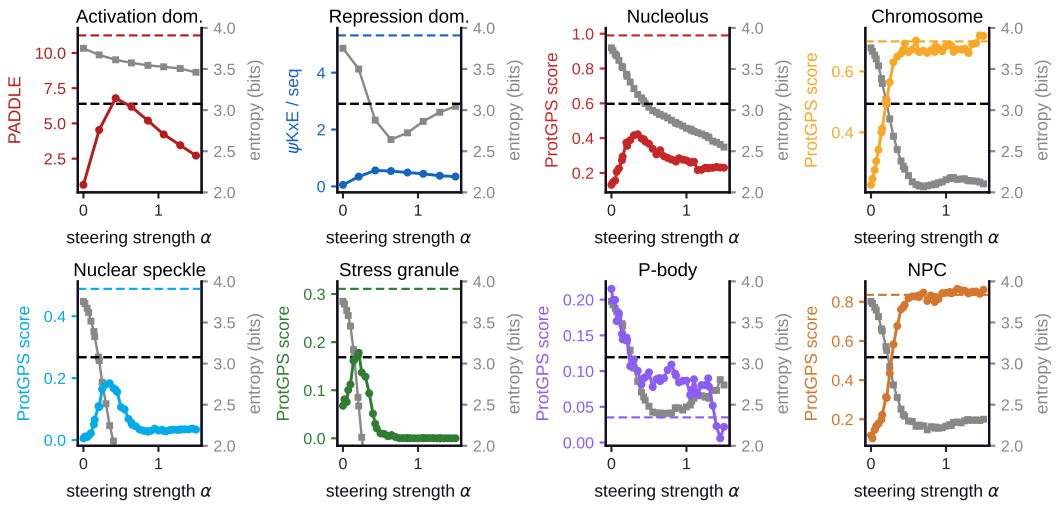

Figure 11: Steering dose-response at $N = 30$. Here, we plot the various scoring metrics (PADDLE, ψKxE, NDSM, and ProtGPS) versus activation steering strength α at top-$N = 30$. The metrics are plotted as colored curves (left axis), and the generated sequence entropies are plotted as grey curves (right axis).

Steering strength curves In Figure 11, we plot various scoring metrics (PADDLE, ψKxE density, NDSM density, ProtGPS score) for sequences generated from the activation steering baseline, as a function of the steering strength α. The most common trend we observe (5 of 8 cases: activation domains, repression domains, nucleolus, nuclear speckle, and stress granule) is an initial increase in score as a function of α, up to a peak value at an α between 0.2 and 0.45, followed by a decrease in the score at larger values of α. This is accompanied by a concomitant reduction in sequence entropy with increased α. Interestingly, we find that for 4 sequence sets (repression, nucleolus, nuclear speckle, and stress granule), the peak and subsequent drop in score occurs at approximately the same point where the sequence entropy falls past the 5% cutoff. We hypothesize that in these cases, the reduction in score occurs simultaneously with the degeneration of sequences beyond the natural range of IDR sequence entropies.

K RANDOM FEATURE SELECTION CONTROL

	Target feature coverage R_S			On-target ProtGPS		
Target	Base	random-30	top-30	Base	random-30	top-30
Nucleolus	$0.03_{.00}$	$0.59_{.04}$	$0.99_{.00}$	$0.14_{.01}$	$0.17_{.28}$	$0.99_{.00}$
Chromosome	$0.04_{.00}$	$0.55_{.06}$	$0.99_{.00}$	$0.26_{.01}$	$0.22_{.26}$	$0.71_{.03}$
Speckle	$0.01_{.00}$	$0.58_{.07}$	$0.93_{.00}$	$0.00_{.00}$	$0.00_{.00}$	$0.36_{.15}$
Stress gran.	$0.01_{.00}$	$0.59_{.05}$	$0.95_{.00}$	$0.05_{.00}$	$0.05_{.06}$	$0.32_{.09}$
P-body	$0.02_{.00}$	$0.61_{.09}$	$0.69_{.04}$	$0.17_{.01}$	$0.27_{.40}$	$0.05_{.02}$
NPC	$0.00_{.00}$	$0.56_{.04}$	$0.94_{.02}$	$0.08_{.01}$	$0.15_{.23}$	$0.84_{.11}$
Activation	$0.05_{.00}$	$0.68_{.07}$	$0.99_{.00}$	–	–	–
Repression	$0.11_{.00}$	$0.62_{.03}$	$0.76_{.00}$	–	–	–

Table 4: RL-SAE on randomly chosen target features (random-30) compared to the base model and to RL-SAE on the enriched top-30. Subscripted values are standard deviations over three seeds (Appendix I).

Random target feature sets As a control, we run RL-SAE on randomly chosen sets of SAE features (random-30). For each of the 8 sequence sets, we sample 30 features uniformly at random from all 16,384 IDiomSAE latents, excluding that set's enriched top-30. We draw three independent random sets per target and then run RL-SAE on them using the same hyperparameters as other runs. We generate and score 1,000 sequences per run. We find that while RL-SAE achieves high target feature coverage for the random features ($R_S \approx 0.55$ to 0.68), the ProtGPS scores remain low, close to the base model's values, as compared to RL-SAE on the top-30 enriched features.

L ORACLE RL BASELINE (RL-GPS)

Compartment	RL-GPS ProtGPS
Nucleolus	$1.00_{.00}$
Chromosome	$0.99_{.00}$
Speckle	$0.98_{.00}$
Stress gran.	$0.99_{.00}$
P-body	$1.00_{.00}$
NPC	$0.99_{.00}$

Table 5: ProtGPS-predicted compartment association scores (0 to 1) for the oracle RL baseline (RL-GPS). Subscripted values are standard deviations over three training seeds (Appendix I).

M COMPOSITION SCRAMBLE CONTROLS

	RL-SAE top-30		RL-SAE private	
Compartment	on-target	scramble	on-target	scramble
Nucleolus	$0.99_{.00}$	$0.50_{.03}$	$0.94_{.07}$	$0.39_{.02}$
Chromosome	$0.71_{.03}$	$0.38_{.04}$	$0.67_{.08}$	$0.44_{.02}$
Speckle	$0.36_{.15}$	$0.11_{.02}$	$0.30_{.24}$	$0.14_{.05}$
Stress gran.	$0.32_{.09}$	$0.68_{.04}$	$0.51_{.36}$	$0.56_{.07}$
P-body	$0.05_{.02}$	$0.33_{.03}$	$0.32_{.10}$	$0.17_{.01}$
NPC	$0.84_{.11}$	$0.50_{.10}$	$0.90_{.05}$	$0.57_{.06}$

Table 6: ProtGPS scores of RL-SAE localization sequences and their composition-matched scrambles. Subscripted values are standard deviations over three training seeds (Appendix I).

Readout	RL-SAE	scramble
PADDLE	$11.16_{.11}$	$8.73_{.10}$
ψKxE	$5.63_{.60}$	$0.49_{.08}$
NDSM	$1.80_{.44}$	$0.12_{.02}$

Table 7: Transcriptional effector readouts of RL-SAE sequences and their composition-matched scrambles. Subscripted values are standard deviations over three training seeds (Appendix I).

N ADDITIONAL DISORDER PREDICTIONS

Target	Nat.	SFT	Steering (α)	RL-SAE	RL-GPS
			Metapredict V3		
Nucleolus	$0.85_{.00}$	$0.81_{.01}$	$0.87_{.00}$ (0.35)	$0.93_{.00}$	$0.95_{.00}$
Chromosome	$0.86_{.00}$	$0.81_{.01}$	$0.88_{.00}$ (0.21)	$0.92_{.01}$	$0.91_{.00}$
Speckle	$0.83_{.01}$	$0.82_{.00}$	$0.89_{.00}$ (0.20)	$0.84_{.07}$	$0.94_{.00}$
Stress gran.	$0.86_{.00}$	$0.80_{.01}$	$0.86_{.01}$ (0.15)	$0.91_{.01}$	$0.93_{.01}$
P-body	$0.82_{.01}$	$0.80_{.00}$	$0.64_{.01}$ (0.00)	$0.92_{.01}$	$0.91_{.01}$
NPC	$0.85_{.01}$	$0.84_{.00}$	$0.81_{.00}$ (0.20)	$0.89_{.01}$	$0.44_{.12}$
Activation	$0.74_{.01}$	$0.75_{.00}$	$0.77_{.00}$ (0.43)	$0.83_{.01}$	–
Repression	$0.73_{.01}$	$0.75_{.00}$	$0.76_{.00}$ (0.35)	$0.90_{.01}$	–
			IUPred3		
Nucleolus	$0.74_{.01}$	$0.68_{.01}$	$0.78_{.00}$ (0.35)	$0.87_{.01}$	$0.91_{.02}$
Chromosome	$0.72_{.01}$	$0.68_{.01}$	$0.81_{.01}$ (0.21)	$0.86_{.01}$	$0.71_{.01}$
Speckle	$0.67_{.01}$	$0.65_{.00}$	$0.79_{.01}$ (0.20)	$0.77_{.03}$	$0.97_{.01}$
Stress gran.	$0.71_{.01}$	$0.67_{.01}$	$0.70_{.01}$ (0.15)	$0.63_{.01}$	$0.87_{.02}$
P-body	$0.70_{.01}$	$0.67_{.01}$	$0.49_{.01}$ (0.00)	$0.89_{.01}$	$0.82_{.02}$
NPC	$0.64_{.02}$	$0.63_{.00}$	$0.54_{.00}$ (0.20)	$0.49_{.04}$	$0.12_{.09}$
Activation	$0.50_{.01}$	$0.52_{.01}$	$0.39_{.00}$ (0.43)	$0.48_{.03}$	–
Repression	$0.57_{.01}$	$0.60_{.01}$	$0.68_{.00}$ (0.35)	$0.78_{.05}$	–

Table 8: Additional disorder predictions for generated and natural sequences, using Metapredict V3 and IUPred3. Values range from 0 to 1, and higher values indicate a prediction of greater disorder. Subscripted values are standard deviations over three seeds, except for natural IDRs (Nat.), which report the standard error of the mean (Appendix I).

Predicted disorder of designed sequences In Table 8 we present Metapredict V3 and IUPred3 disorder predictions on IDR sequences generated from the various methods and sets. Both Metapredict and IUPred consider a sequence to be disordered if the score is > 0.5.

With one exception, we find that across all methods and sequence sets, Metapredict predicts the sequences to be disordered. The exception is RL-GPS generated NPC sequences, with a score of 0.44. As RL-GPS on NPC was able to achieve a high ProtGPS score, we posit that this is due to hacking of the ProtGPS reward.

Sequences are also largely predicted to be disordered by IUPred, with most scores near or above the 0.5 threshold. Exceptions include sequences steered for activation, with a score of 0.39, and RL-GPS NPC sequences, with a score of 0.12. The high hydrophobic residue content in steered activation domain sequences may lead to a low IUPred score, despite their functional relevance to co-activator recruitment. Here, the extremely low NPC score again indicates the possibility of reward hacking.

O ORTHOGONAL LOCALIZATION PREDICTION

					RL-SAE		
Compartment	Base	Nat.	SFT	Steering (α)	top-30	private-30	RL-GPS
	P(Nucleus)						
Nucleolus	$0.43_{.01}$	$0.66_{.01}$	$0.59_{.01}$	$0.47_{.01}$ (0.35)	$\mathbf{0.90}_{.01}$	$\underline{0.87}_{.02}$	$0.70_{.03}$
Chromosome	$0.45_{.01}$	$0.67_{.01}$	$0.58_{.01}$	$0.57_{.01}$ (0.21)	$\mathbf{0.91}_{.00}$	$\underline{0.85}_{.01}$	$0.75_{.02}$
Speckle	$0.44_{.02}$	$0.67_{.02}$	$0.56_{.01}$	$0.63_{.01}$ (0.20)	$0.80_{.04}$	$\underline{0.82}_{.05}$	$\mathbf{0.84}_{.03}$
Stress gran.	$0.43_{.02}$	$0.61_{.01}$	$0.53_{.00}$	$0.54_{.01}$ (0.15)	$0.78_{.02}$	$0.56_{.03}$	$0.61_{.16}$
P-body	$0.44_{.01}$	$0.61_{.01}$	$0.55_{.01}$	$0.45_{.01}$ (0.00)	$0.81_{.05}$	$0.42_{.03}$	$0.49_{.09}$
NPC	$0.44_{.01}$	$\underline{0.59}_{.02}$	$0.58_{.01}$	$0.33_{.01}$ (0.20)	$\mathbf{0.68}_{.04}$	$\mathbf{0.68}_{.06}$	$0.22_{.03}$
	P(Cytoplasm)						
Nucleolus	$0.50_{.01}$	$0.45_{.01}$	$0.49_{.00}$	$0.52_{.01}$ (0.35)	$0.18_{.02}$	$0.30_{.03}$	$0.43_{.05}$
Chromosome	$0.50_{.01}$	$0.48_{.01}$	$0.51_{.00}$	$0.38_{.01}$ (0.21)	$0.20_{.01}$	$0.29_{.02}$	$0.40_{.01}$
Speckle	$0.50_{.00}$	$0.46_{.01}$	$0.50_{.01}$	$0.35_{.01}$ (0.20)	$0.34_{.02}$	$0.33_{.06}$	$0.32_{.02}$
Stress gran.	$0.51_{.00}$	$0.53_{.01}$	$0.51_{.01}$	$0.46_{.02}$ (0.15)	$0.56_{.01}$	$\mathbf{0.67}_{.02}$	$\underline{0.64}_{.08}$
P-body	$0.51_{.01}$	$0.50_{.01}$	$0.52_{.01}$	$0.50_{.00}$ (0.00)	$0.21_{.02}$	$\mathbf{0.65}_{.03}$	$\underline{0.55}_{.02}$
NPC	$0.50_{.01}$	$0.48_{.02}$	$0.48_{.00}$	$0.40_{.01}$ (0.20)	$0.32_{.04}$	$0.30_{.02}$	$0.23_{.05}$

Table 9: Orthogonal DeepLoc localization predictions. Bold indicates the best score and underline the second best within each grey shaded row. The private-30 column targets the top 30 private features of a compartment. RL-GPS is included for comparison here because DeepLoc is an orthogonal predictor, unlike in the main text Table 1. The grey shaded rows indicate the biologically expected localization. Subscripted values are standard deviations over three seeds, except for natural IDRs (Nat.), which report the standard error of the mean (Appendix I).

P COMPARTMENT MOTIF DETECTION

						RL-SAE		
Compartment	**Motif**	Base	Nat.	SFT	Steering (α)	top-30	private-30	RL-GPS
Nucleolus	cNLS	$0.13_{.02}$	$0.54_{.02}$	$0.37_{.00}$	$0.38_{.01}$ (0.35)	$\mathbf{1.00}_{.00}$	$\underline{1.00}_{.00}$	$1.00_{.00}$
Chromosome	cNLS	$0.13_{.02}$	$0.43_{.02}$	$0.27_{.00}$	$0.82_{.02}$ (0.21)	$\mathbf{1.00}_{.00}$	$\underline{0.96}_{.03}$	$0.64_{.14}$
Speckle	RS repeat	$0.00_{.00}$	$0.08_{.01}$	$0.02_{.00}$	$0.21_{.02}$ (0.20)	$\underline{0.97}_{.04}$	$\mathbf{1.00}_{.01}$	$0.27_{.26}$
Stress gran.	RG/RGG	$0.04_{.01}$	$0.15_{.01}$	$0.09_{.00}$	$0.45_{.01}$ (0.15)	$\mathbf{1.00}_{.00}$	$\underline{0.93}_{.04}$	$0.66_{.52}$
P-body	RG/RGG	$0.04_{.01}$	$0.13_{.02}$	$0.10_{.01}$	$0.04_{.00}$ (0.00)	$\underline{0.34}_{.56}$	$0.00_{.00}$	$\mathbf{0.51}_{.02}$
NPC	GLFG /100 aa	$0.00_{.00}$	$0.29_{.08}$	$0.29_{.01}$	$0.01_{.00}$ (0.20)	$\underline{1.62}_{.45}$	$\mathbf{1.84}_{.05}$	$0.00_{.00}$

Table 10: Fraction of sequences containing the specified biological motif (density per 100 aa for GLFG). Bold indicates the best score per row, underline the second best. The private-30 column targets the compartment's private top-30 features. RL-GPS is included for comparison here because the motifs are orthogonal to ProtGPS. Motif regexes are in Appendix Q. Subscripted values are standard deviations over three seeds, except for natural IDRs (Nat.), which report the standard error of the mean (Appendix I).

Q MOTIF REGULAR EXPRESSIONS

Motif	Regex / rule	Compartment
RS/SR repeat	`(?:RS|SR){3,}`	Nuclear speckle
RG/RGG cluster	≥ 2 of {`RGG`, `RG` not in an `RGG`} within 30 residues	Stress granule, P-body
GLFG repeat	`(?=(GLFG))` (overlapping matches)	Nuclear pore complex
ψKxE	`[VILMF]K.E`	Repression domain
NDSM	`[VILMF]K.E[DE]+`	Repression domain

Table 11: Regular expressions used in Tables 10 and 2.

ELM class	Regex
TRG_NLS_MonoCore_2	`[^DE]((K[RK])|(RK))[KRP][KR][^DE]`
TRG_NLS_MonoExtN_4	`(([PKR].{0,1}[^DE])|([PKR]))((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]`
TRG_NLS_MonoExtC_3	`[^DE]((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))(([PKR])|([^DE][DE]))`
TRG_NLS_Bipartite_1	`[KR][KR].{7,15}[^DE]((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]`

Table 12: Eukaryotic Linear Motif classical-NLS (cNLS) regular expressions used in Table 10.

R TOP-N LOGO GALLERIES

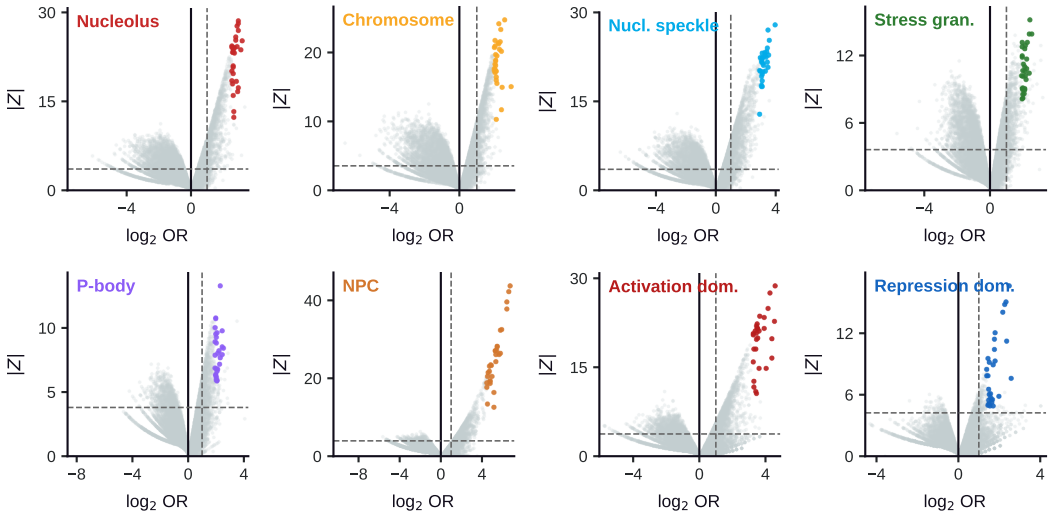

Figure 12: Feature enrichment plots for all eight sequence sets. Feature enrichment plots (hypergeometric $|Z|$ versus $\log_2$ odds ratio (LOR) for all 8 sequence sets. The top-30 features as ranked by LOR are the colored dots. The horizontal dashed line marks the FDR threshold, and the vertical dashed line marks the LOR ≥ 1 threshold.

Logo plots Each of the following figures shows information content sequence logo plots of the top-30 enriched features for a given sequence set (row major). Each logo plot is centered on the feature's top-activating residue position (vertical dashed line) within a given sequence, with the orange color depicting the strength of the feature's activation, and is constructed by averaging across the top 80 sequences by feature activation strength from the corresponding positive sequence sets (pooled for localization sequences).

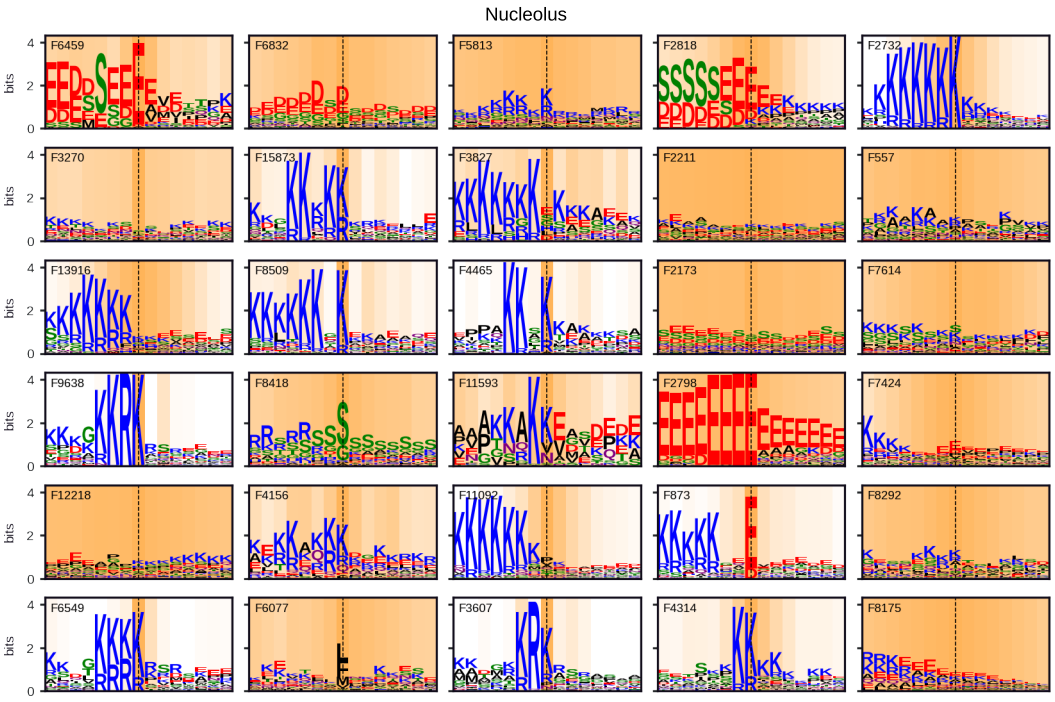

Figure 13: Nucleolus. Logo plots of the top-30 enriched features.

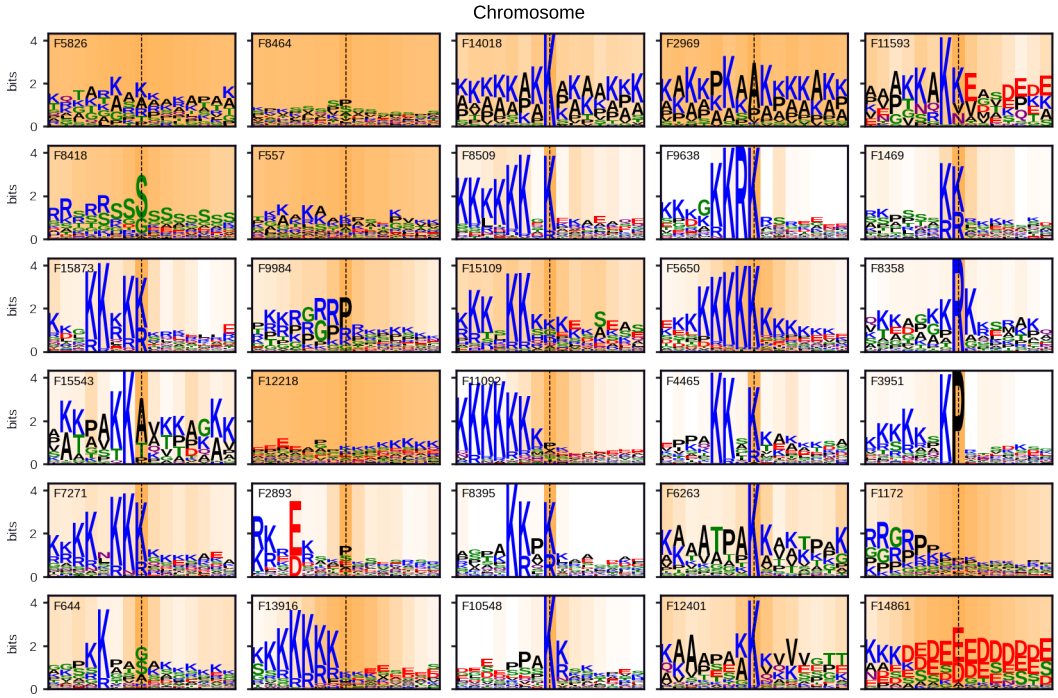

Figure 14: Chromosome. Logo plots of the top-30 enriched features.

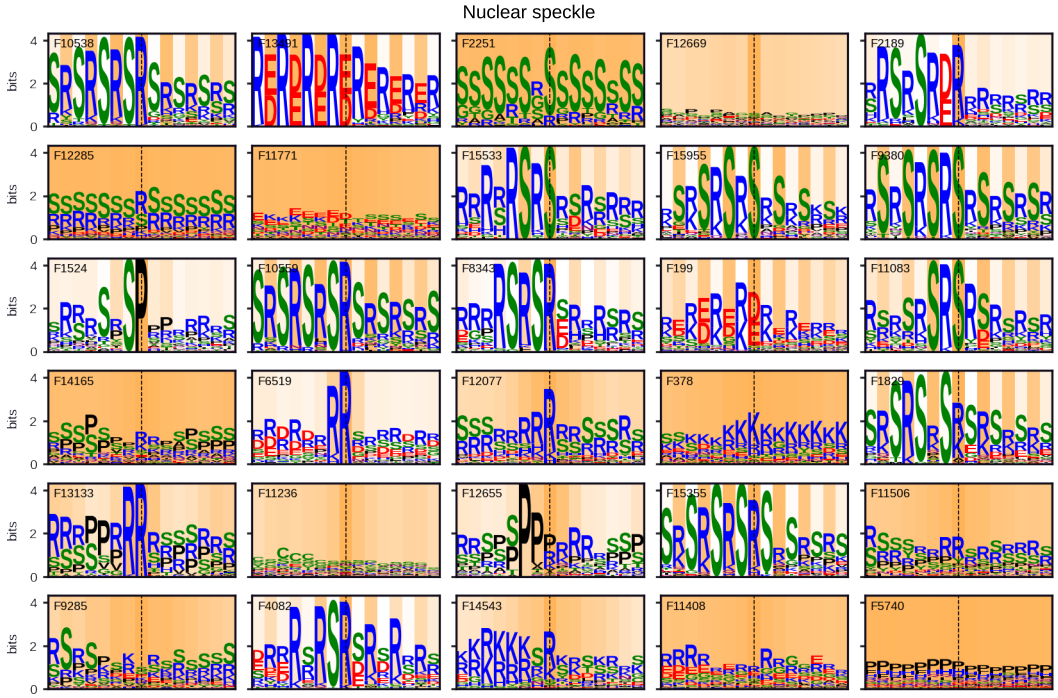

Figure 15: Nuclear speckle. Logo plots of the top-30 enriched features.

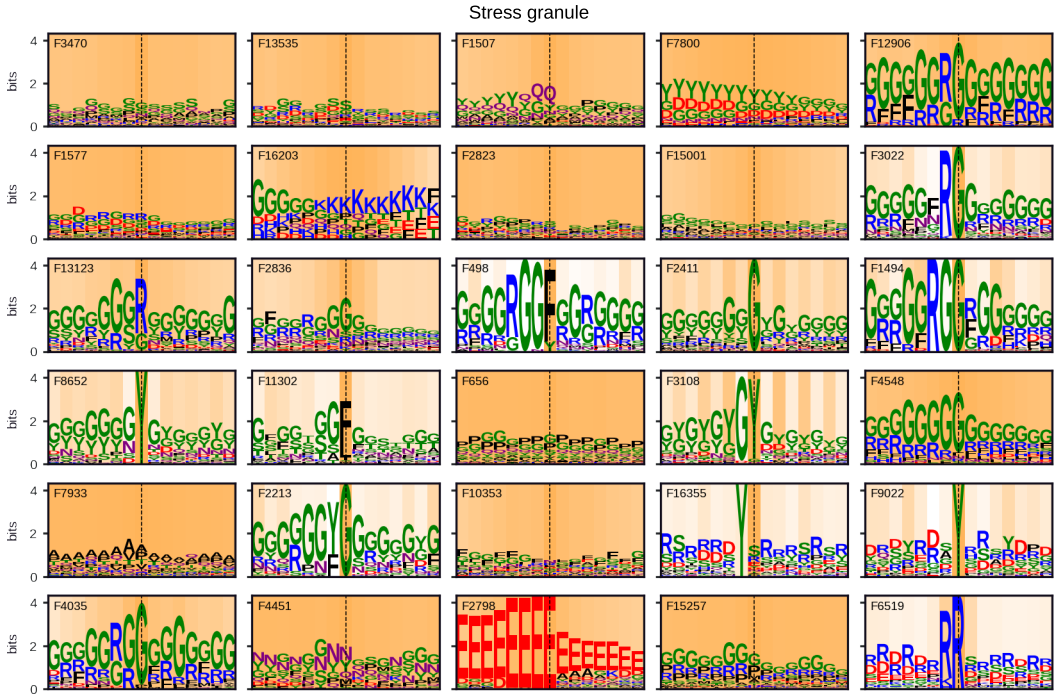

Figure 16: Stress granule. Logo plots of the top-30 enriched features.

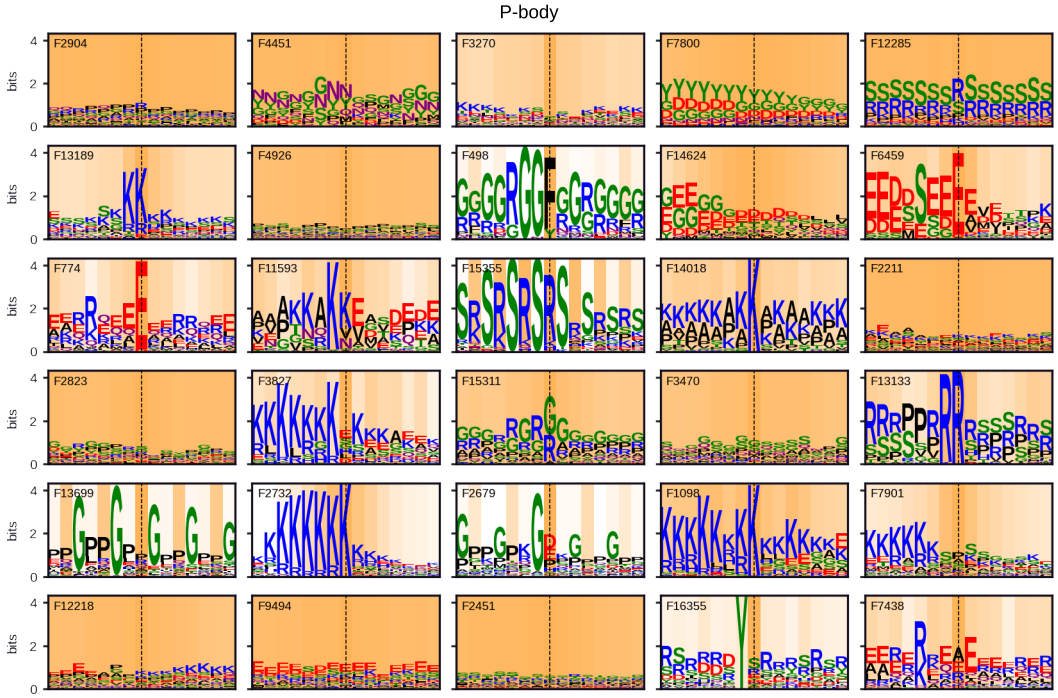

Figure 17: P-body. Logo plots of the top-30 enriched features.

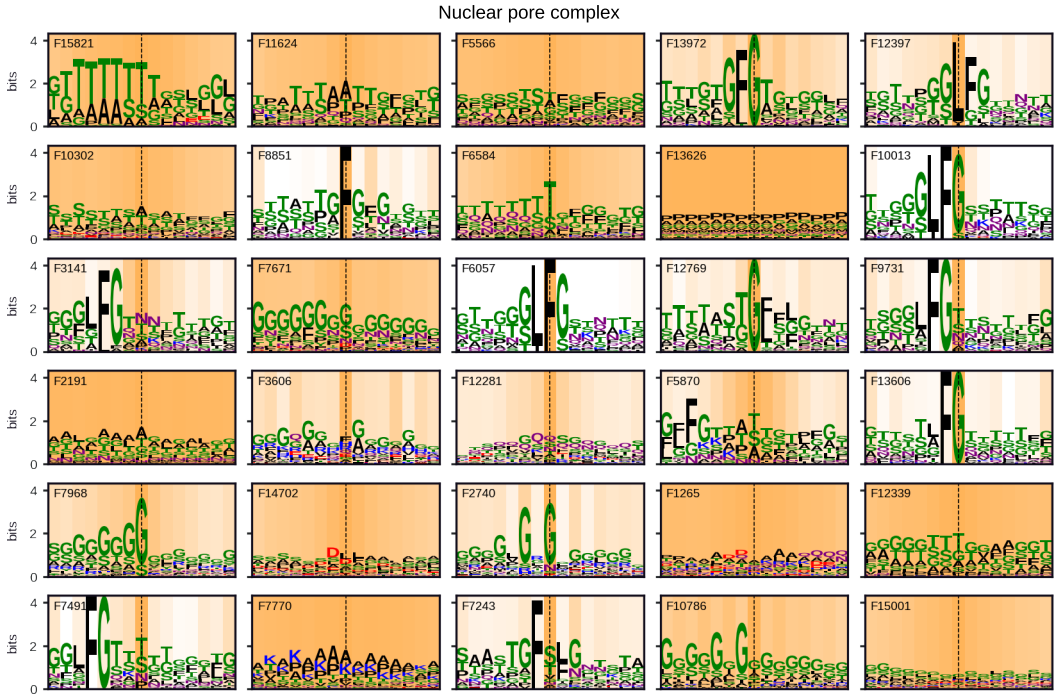

Figure 18: Nuclear pore complex. Logo plots of the top-30 enriched features.

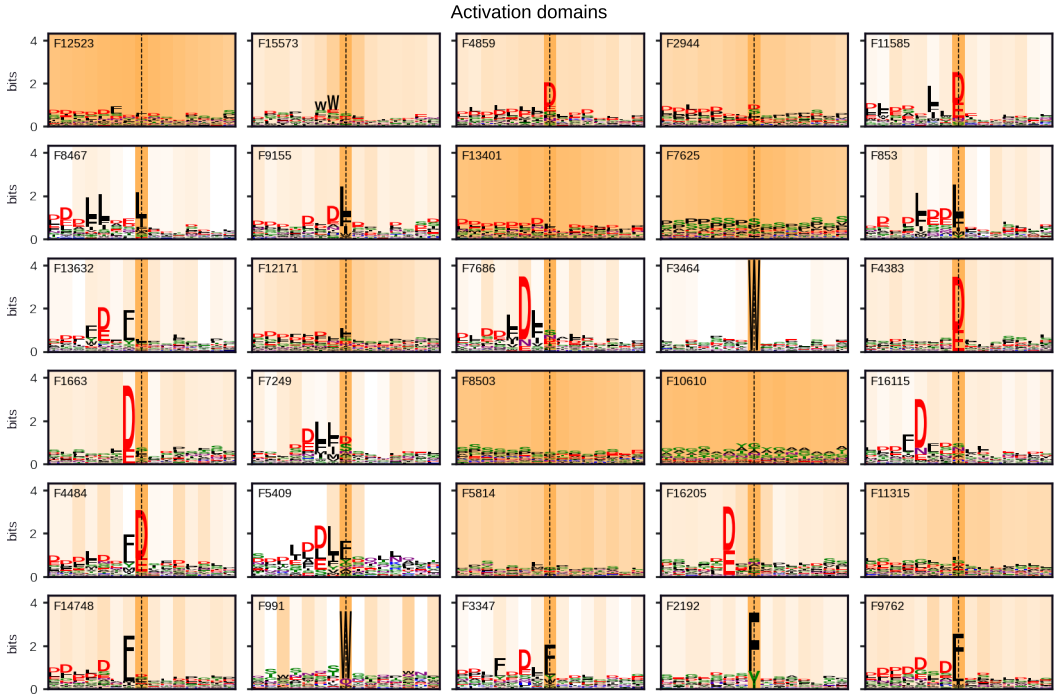

Figure 19: Activation domains. Logo plots of the top-30 enriched features.

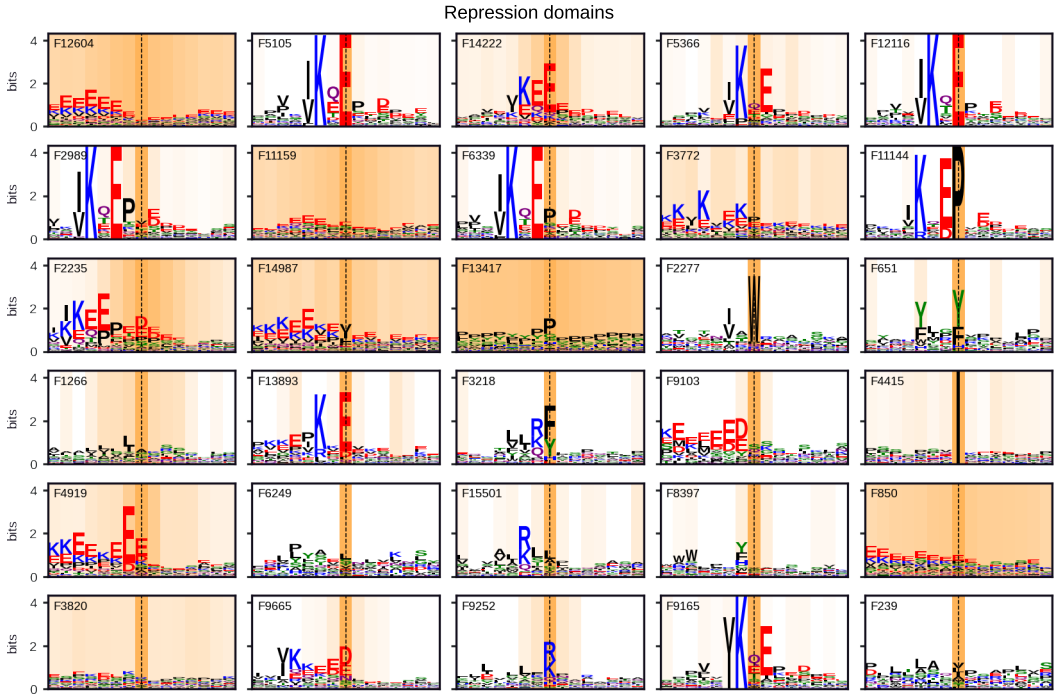

Figure 20: Repression domains. Logo plots of the top-30 enriched features.

S EXAMPLE SEQUENCES

Here we show five randomly chosen example sequences per condition.

NUCLEOLUS

RL-SAE, Nucleolus (N=30)

KKKRKKGTKRLKDAEKTAQKKKKMKQKGPQGAEDSSTSEEESEEEGQKDEDEGDDFFVGHEKTEPKTDKMPKHKKHKNQMQTTLHANGDSD
SEDEEDDIVLQVTSGDIA

KKKRKKEKKGGRETKTAKQRKKRKQADEKGERDSSSEEEEIEEEGPGPVGNDNDDFFQNKEEASKQQRPKKQAAPHKKAKMQHHSTASNPL
TDSDKSEDEDEDEFDDDIVL

KKKRKKGTKAADQLNEKKVSKKRKEKQKAAPMTEDDSQKSSSEESEEEKEDDGEDKVKAEDDGDDFFVGTNHKEQPTLKKHKTGKKHTVQN
PNHNSDSDESDEEDEIVLML

EKVKKKRKKGSKAAVEKASKRRKQSKQTKEPVESESSSEEEEEEEEDAGDSDDFFVGEDTSRDAPMKLKKNKTKVQPHHKGAKPQQQNHPQ
PVAG

AKPKNKKRKKEPQGKKGAKKEKQAKAHGAGSEEESGSSSEEEQEQDTGGDDGDDFFTDKRSDQLPQQPTKKMAKKHKSATKDHNQQTAPHR
IPDSSDEEEDDEDDILL

RL-GPS, Nucleolus

AATDDEEDRTAKGHPTGVQAQKSDQKGPKKTLRKPKKKGRKSRKKKRQKTEDPDLVLDREPTSTSTNEEESDREEDQKDTSEENETDQHRS
QITKQNLN

VKRESKAAFSTTATTQKEGPPENRKKTAKVKKVKKKRRKESKKQEGKVEPSDVASKDDIDDSQTVKNARSNQDPDDDAESNVEAPSMRNRS
RDGVEKQTHPT

LQQQGPGPGSAPFRPPDPKSTTPKEKGKPRRKAAKKRQKRRRKKEKRKKRALEAENSDEAGGLAVSVESESQHEVEGSSSLDSAEEADTEK
TPQRRPGQPPVN

KANAAGTESKSVSENKTRQKRDRKEKKRKRKNAEKKQPQDTTTQDANESSETSEDEEEEVERPNDDAPQQSRPEELPNAPQDDRVHEFQDT
PPGDKRVRTGQ

DKSGKSTAEVNRSSSKKRKKHKKTKKKATRKRRELKEDQVEDQSTNTREPDQRSEDADGDDSEPETSENPDTTADEQRGTEHPKSQNSHED
RFKGHPHNNLL

Activation steering (N=30, α=0.35)

VENTKAKPVNKRKTEGDEEEGETQASKASAKKAKVEPAKEGSSKAAIKEEPAKEVKSEVKEEKKEEPVKKAASKEKKVEE

TRLPVTSGSVSPQKTPLVSKRETGEDNEKGIPSLEEVAQRAQIADKKKTAKASKKKKDAEKKKEAADMSGEKLQRKVAREEEAKKQAAEKA
AKAKAEK

TARPPEEGEAPAGGEAPAGEGKAEGGVPSSGTSADAAEGSAKAAKGEPIGKKRAAAADEGGAAKKARKEEAGGDDKKSAAKAESAGSGKPA
AGGGQGKARAGGGKAAAAGKKRKAEE

FSPTRDEAAGESSQKPEKRKKLREAKEDGDPGLEQGDENDQEQEESDKELAKQSDDVEEAEENGEVEDEEHREQEGKEKPKKESSKSEEAQ
ENGKEAEKAKKDENN

TLTEVHTEKFQKDHKQQKVEQSTDQQHVSDTQKAAQQNVKATEEVEPRKAAPQAESDSVVESEKEKTPPAANTKPKKKTKK

Supervised fine-tuning

VTRPAGTSEGTILHQRALSSPKPVHRHGLSSLPQVVFTTAESTSKWEDVPLDRGSYLDDLNMEALSIKQDISGNDVSASLLPGERR

LTYTVGLRRGFASSPQSWDRDDAAIRRAGSDIYSRISRGMPLPDKGRDEAPGWK

MAAEKQVPGGGGGGGSGGGGGSGGGGSGGGRGAGGEENKENERPSAGSKANKEFGDSLSLEIL

MAAPLMDQRARGQGDVHRGKLKEPLNLLQGRHRGQRRNRGDGGGRQPLPGLRRALPQRGACPLPALPDRGVNMAQRPMEQPMMMMM

YKSLRRKPLIPLVSEPTFKDGFSECDSKDSELGKILQRDDCPPEGMMGCSASGELGSPPESCYHSPSNGNKPRDSASMDGPKYRGSGDSGT
ESDGGTKNKVATWPRQQPKHDPTQPSSDLGSGQP

Natural nucleolus IDRs

MTRILTAFKVVRTLKTGFGFTNVTAHQKWKFSRPGIRLLSVKAQTAHIV

MNRSRAIVQRGRVLPPPAPLDTTNLAGRRTLQGRAKMASV

MPTVISASVAPRTAAEPRSPGPVPHPAQSKATEAGGGNPSGIYS

MEPQPGGARSCRRGAPGGACELGPAAEAAPMSLAIHSTTGTRYDLAVPPDETVEGLRKRLSQRLKVPKERLALLHKDTRLSSGKLQEFGVG
DGSKLTLVPTVEAGLMSQASRPEQS

MPGANYRAGAGAGAGARRPRGARDREEDGGGLEPAAVARDLLRGTSNMSFEELLELQSQVGTKTYKQLVAGNSPKKQASRPPIQNACVADK
HRPL

CHROMOSOME

RL-SAE, Chromosome (N=30)

KSRDEDKPQKISKKKSKTPNKKAKKAPRKPPGKRARKTLKKTAEASENMELDCGNTQDEFEKISAEDGPVKEEHPVKKRKRPPKLNSTPAK
VHQTSDNKDVAL

EDSSPKKTKKKKKSPVKKAKGPRKAGRPKKATRKTTEKLISTLESEMEENGLREESMNNEDD

EVENDKSPKKKKKSKKPPSKKVKGRPAKTRAKRAARKTQELIHEENSLNSEDQEENLDQNDDVNEPPAKRKRKAQELSKSHSTKSDTPIKK
TKGAPCTRCN

DDPRTKETVVNKKKKNKKPPAKKAKSGRPRKAPRKTARKMQDTSDLDCEEQSGEESASEVPAQKRRKRKPQESADSSVVKQKRQPTSKVSQ
KCFEVNTD

SKSGEKNDATIEKTKKKKKVPSRKTKKPPVKKIARKTARKKMASKPSDNEDSENEVVPSIEENDLSVPQKKRKRKPPQPGDVVECSLTQNN
KRKRGVAG

RL-GPS, Chromosome

KSSSSGSPESTTPQRRRSQKSSTCLTQKKVNLHNRSAVESSRICNKRSQRSSCAEDFEDLSGSASGTAVSDEDSTDKSDDFESNGQKASRS
LKEGSLASNKRHS

EQRISELTLNSKPSKGNSCKASAGKKLYSCTVSSSRRSTNLTGSKQSKHARESP

DKLQKQQADARSASTCKEQSCSVTPLKKSQSLRKSDSPEIQNNSCCLDSDFNSSPTKRQKLICSVENHPSVELVTGSELKKTECLDSDKRN
GQPRSVT

SKDTGKKASLCTSSILSEAKSLQGSRASSLTNGKSQPSEHNDSKSLTVTPQKTTSDKQTYCDSENTLIQRSSKKDPSTQPHPSNKEDPASD
EETDASPNKRKRKPRQ

RDQGSSCRKSKRTSSTCCKSRKSIQNSQEGFPLKKPTGKKKPCRKSSVDYTEESELEAAETVEGADQLSISKASRRAKSPQDTSGIENTEE
TSDS

Activation steering (N=30, α=0.21)

VAATAERLSKTPKKDRDMKRIIEAVIKKKKEKKKAAQKAEYEKIKKRAAALKKKKKKIGKKKR

RPSVPADHGALATALERSRAAVRPQPMKRVSRAANAASVAARALVKAADLMSVRRGVAAVKASAAPKATATKAPAVKATAAKATAAKPTAK
TKSPAKRKATPATSRTRRRKSTDDAPAASPRKRAAKGPAKAARKRASKR

MSSQPLSRRRPHWLRPLLRLAVAVVVLFTVAWLKGPDVLSPNDTALKAVASTTSSEKPLSTPVLMEKGPGKKSKDRTSSQANSGERSTSTS
KRWTQTEDKTASEKGEKVERPEKKKKKVREKKKA

LSVSGTMTGSSASAAGSEAPAPEQKEPKKVSGSSKKKKPAASKKKAAGSKEAKKRRKTRKKSAKKTKKATAKKAAKAAKTVKKKTKASKSK
KASKAYKSKKKTKKTAKKAKKAKKSSKKKVTKSKKKTKASKSKKKAK

MRRGLKRNWDGYINKRHREKRRLDARALAGGNIDLKGLKSKAGKAGKKRAAEKKPAAKGKAGKPPAKKPAKKPAAKKPAKKPAAKKPAAKK
PAAKKAAKKPAKKPAAKKPAAKKPAAKKAAKKPAAKKTKKK

Supervised fine-tuning

MVKQSQHCPGTLGQESGLENHRDRAATKKTRGNSFERHSPLSNSPHKEPFLKHHHTGRKGERRSPKRRPGAPPVKNEPLTPVITGNHGKSV

QPLNPKPKPKKEKTEQDKRRIEAEINTLTVESPWRAPPDLSPRPHPQNMPPKARLMRWGEFVEDPLLKQEVLVPGKKKTPP

CSPAPSGAALAAGDRGPDPAPAVSDTGSQGGQPGPAAKTASPSASPSPKAPAKGA

MSAQKDCEFWDQHEAKGTPQARRDDSYSPAEDASSKLVRKQTGNEEAGSYKAHRPLGKQHKLVTKPLG

KPHSPKHARRTRSRHDQEPEAMVADSGSEDQHSVASDANRGREILAFSRPAAAAAPPPPQSQQQKSSRSGPNHHHNSGGGGSSRRKPTNLS
PATQQAILSMLEDNQDPVIKGHGKEKNCA

Natural chromosome IDRs

ETDIFKKKKRKGQEDRKSKKKSAPSILSNGRIGQVGKFKNGTLILSPVDIKKINSSRVAK

ESIYKTEPRPEAKTMPNSSTPTSQINTIDSRTLSDVSNISNINASCVSESICNSLNSKLEPTLENL

TVAREATELNEDGSQVKRTGGAEDGAPAPFI

MPRKGTQPSTARRREEGPPPPSPDGASSDAEPEPPSGRTESPATAAETASEELDNRSLEEILNSIPPPPPPPAMTNE

PSLEEVEGHVADGSATEMGTTKKAVKPKKKHLHRTIEQNINNLNVSEADRKCE

NUCLEAR SPECKLE

RL-SAE, Nuclear speckle (N=30)

LSRRSTASRSRSRSRSRSRSYSRSRSPRRDRDKSPAKDKSRSKSPAKSSSRTPERDSHERDLDKHGHKDGDRNHCNCEHERNHAPAVEQNMCTNIIQLTTF

DTARSRSRSVSRSRSGSRSRSYSRSRSPRRDRKHEAASRSRDRSPSKSPAKSLNQSKEHSREKDRDKNCDCGHERNNCHGAECNADADYVLDKQMVFATFPP

IQTRSRSRSRSRSGSRSRSRSYSRSRSPRRERDRSPGKERSRSKSPANAAASHSRDKDRDKHREKDHHEHKNEWCEHGCNCGNEKPAQIANPNMGMPFI

RTRTKNSHDRARSRSSSRSKSRSRSRSYSRSRSPRRERRDKDKSQTRSPSQDKGRSRSRDHEHDYQKDGKDCNCRETDKDCSCGTDQNMIINTGPFLIGEFP

ARRSRSRSRQRDRSRSRSRSRSYSRSRSRSYSRSRSPRRSKDKDRERGASPSRGKEKRSVTRDSSRENHGKDRDHECADCENHSCDHNVDIDTCMMQFFTMQTF

RL-GPS, Nuclear speckle

ADPKHAKKRKPTDHRTYSRHSRTSKHRSSSRSRSRSGSHRRKKSRDRDKRRAEERKGSPPRTPATHVHQHDNHYGGAAGVKESANMAEMGDHVGISKIPQH

SHHNHKKQQRDSRHPTSHDRRHNRSRSHSRSRSRPRDSRDRKSKKHTSSRRRSRSHDRDDGKKSKDPSANAPVDIVPTEMQPEDEMRLHHHDNIDLPPEIQ

RAGHHKPPSSHQQRHRPRHYPQAHRKSTAASDRSRSESQRSRRHRSPDRGHHREKRKRTPQREGTQEKENKPHKEPANSMGNPQPNQHDQPNQDDHHDMAEGI

DHHHHHPNRTPSTSRSRRHSNRHRSAKHSRSHRAHDKGRHPPNRKPRDYPPVEKGHRREAVAEAHGPGAIGNSGHNPLDSQHQDMHEMKRDHHQGKMESHDLP

GAAHRHQHIKEKRKTIHPAHAHRKSASRRSTSKHDRHRSRRDRTKKSSSRREEPRKQRSAARDATQDRRTPAPIQEQTTKAPEMNPHDHNEPSRNVHEDLDHHHGQ

Activation steering (N=30, α=0.20)

LMGSASSRQGHSEDSAFAAERQGTRNRKRGRRDRDHCSDNERRVNERRSDDYGSRRRSRRRCRSRRENSSRSRSRRSSRSRSRGRRSRSRDRRRGRRSRRTHSRSRRSRSRSRRRSRRSRRSRSRRRGSRSRSRRKR

TGDEPRDTDANLALSPQPAAANTEDGPWVRMELHPYMARNYSQCIFRDDQGAARRRR

GKAKDDPETTPAPKQSKPRPPKGKRNPRSRRKPKSAKTKSSPSRKRKSKRR

SPQRSTETLKQTEAPYSSKRQRNDSTASDSSDTPIKRLKGNDNDEESPDAS

INPAGGSSRGGSSRRAAVESLILDSRIMGGHGKHHRSGSRGHRSRSRGSRSGRDRSRDRDRDRDRERDRVDRTRSRSRRSRSR

Supervised fine-tuning

MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDRDYMDTLPPTVGDDV

MPVLSARRRELRDPLRSLSTPVVATCLALGIVLLAGCTAAGSDTAAGDAQARAGGAAADTPEAP

MQSNKTFNLEKQNHTPRKHHQHHHQQQHHQQQQQQPPPPPIPANGQQASSQNEGLTIDLKN

MASGPHSTATAAAAASSAAPSAGGSSSGTTTTTTTTTGGILIGDRLYSEVSLTIDHSLIPEERLSPTPSMQDGLD

MLGKDVLNPDFQKCGRLPATEDSSLHQEDSNPLTLRTSRPTGCRYSSQSHHGIYGRGTRGIKGKEPQGFYLSFREN

Natural nuclear speckle IDRs

LLEGAPQEEAEAAPEGNVGAEVVGEPRANIPCK

RQEEERRRREEEMRRQQEEMMRRQQEGFKGTFPDAREQEIRMGQMAMGGAMGINNRGAMPPAPVPAGTPAPPGPATMMPDGTLGLTPPTTERFGQAATMEGIGAIGGTPPAFNRAAPGAEFAPNKRRRY

LSTIKSSVFSLDGGSSPVEPDLAVAGIHSLPSTSVTPHSPSSPVGSVLLQDTKPTFEMQQPSPPIPPVHPD

YGRRRKKHRSRSRSRERRSRSRDRGRGGGGGGGGGGGGGRERDRRRSRDRERSGRF

FPRSGRGTGRGGGGGGGGGAPRGRYGPPSRRSENR

STRESS GRANULE

RL-SAE, Stress granule (N=30)

SQSHSVTGRGGFGRGRGFDSGYGGRGTSRGGAGYGYSSRGDGSYGSRRNDGYGSYGNQSYGYNSQSSRRFGNAQGGPQSMHAMRIPMPSLH
IPIPPVAIPNEIAVPVLE

QSSTATSSMRGSGGRGFRGRGFDSGYGGRGGSSRGGGGYGYSSRRFDGGYGNGGYGNRDAYSNQQQRFGQSQQNRQARGMAMQIAPPQVRM
PIPVAAAVAVPVLEVECNV

ISSASTTSTMAAGRIGSPRMGGRGGFRGGFDSGYGSRGGSGYGYSSRRPDGYGSYGNRRGDGYANGYGNDRFNNQQNFGSRSQNMRSMVGV
PMQNIHVPPMGIAAPAPVITTLPTPVE

NPSASSTMSTADLGPRGGLRGGRGFGRGFDSGYGSRGGGGYGGYSGGRRFDNGYGNGHSAYGNAERRFGNGYSQQQQGGISFNHAQPIRMP
PRVPQTPAGVPIVVNPPTRVIEQ

VSPSASTQSMTGTGRGGFRGRGFDSGYGGRGGSGYGYSSRRAGGDFSGGYSSRRSDGYGNYNQNRHNRDNYGAQQQQVRMPRSTAMGGSMP
VLMMPPVIAPALKME

RL-GPS, Stress granule

DQRTPGQEVAQVRAPGRDEADVKGRDARRNVDASEKRGALKGDKSYEKYEREGAKQGRNNGNGTKKGGEDTDGEDSEQVAGNSNGMNGNGN
GNLDGGQTGNMVGN

TAADEPAAQQTDAGRQLMGQHRGQLSNQRGPFAQANSGNAAPIGRVDNNRRGNGRGRGRRDNGNRGFRNGMPRMEGISTPVSNQPDIGAGG
YNIGSPTGNLLQSS

SLVSPAATTGKPDVVAEQPGSQNPRAINPPSGLRGDQQPANGKLNSEGNRGNGNRNRGGRNGWRRGGGWNNRNTYGGAEQMNNQTTGNNMG
PTRVDIGPDTGLGA

TSQAGNGVNPAMMRGQIPPNRIANTSQGNSRSDNRRNEGRDGLGRGNFRRGNNRGNGWRDSQGENIPVPSFAPQNISPSLSQRNATFGQIS
APTSSTNFN

ENQQSAPTQAQPNSVNGVSQRDAPKTQDPAAVSQNTRFGDGLGWDMNQSQNSSNGYGRNRNNRGGRGNRGNNAGFNGPNRMSGNNTVTMVS
QITGTPSSSYN

Activation steering (N=30, α=0.15)

LDYSTYIPKYNTTNQNWLEKRRAEQTNSNEFQIRRKLAYINMFDNNGPGSFDILSNTPNN

MKSRNRLEINLLTGQNSSVFIQRVQRAVANASARGEYSDDAELGRFLLDGKLSANGDGAG

QPNCRVPVPVPVHPPPPRGGQPGGPPPCPGPCMGPPGPVPVPVPVPCAVPRRGPGPGPVPAPGPRKS

EVARYSPGEQSADSVKSFENRQPAVNEAVPSHTSASLNPALNSAFKRPAGATSSQSFRVPSMGDIGRIHPGF

GKAKDDPETTPAPEQSEPRPPRGPRNPRGRGGPRGGFGGSGPSRGGGGGGGGGGMRSGPGGFFGFGGGGGRGYGRGGGYGRGGGGYGRGGG
RGGGGGGGGRW

Supervised fine-tuning

MQQMGQGGMPGMGGGMPGMGGGPGRGAKGKQQPRKKKKKKGFGNLPKNFENSGPQGGQQSGGLGEDLPPGLGGGEKKQEGGLGGLPPGMEL
PKGQ

DDANLGNDPPVKRKSQSWTSKILRKKNGVISKKIFIEQEEETPLPEVNQL

MATDSWALAVDEQEAAVKSMTNLQIKEEKVKADTNGIIKTSTTAEKTDEEEKEDRAAQS

QLIPVTPRTTRVNLIEVSGKDNEEALLNLEAPSDEEGSVPASSFSQVNCVDVKEEWQELSELSKACR

HRHNVSSFQNGWAVHSTDLYDLCRELHGIVRVYAVQTGQKPPNALLSQKLSPP

Natural stress granule IDRs

QRKLRESQEGREVQRQEAGELRRSLGEGAKEREALRRSNEELRSAVKKAESERIS

KVKMKQPKGSLQAHDTSSLPTVIMRNKPSQPKERPRSAVLLVDETATTPIFANRRSQQSVSLSKSVSIQNITGVGNDENMSNTWKFLSHST
DSLNKISKVNESTESLTDEGVGTDMNEGQLLGDFEIESKQL

MGLCLAGRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQAIAENR

MPLGLLWLGLALLGALHAQAQDSTSDLIPAPPLSK

MACRPRSPPRHQSRCDGDASPPSPARWSLGRKRRADGRRWRPEDAEEAEHRGAERRPESFTTPEGPKPRSRCSDWASAVEEDEMRTRVNKE
MARYKRKLLINDFGRERKSSSGSSDSKESMSTVPADFETDESV

P-BODY

RL-SAE, P-body (N=30)

NADEDKTKKEKKKKRKKRKDKEESSSDEEDEEIDEDLPQPHQVKKSKKKKRKHKEGADDSYDDDYATPKVNHAGSKSKGSNIPLPGQGAMP
PPGQVSKNTPRHASIEGP

DADEKKKKKRKTRDDDESSSEEEEVDEELPQPHQIKKKKKKRKHKAYDDDDYDYDDNYDDGSKKNKQQSETPLPGQGSYVPPHTKKIPAGD
GSPVTPGAPTGHGNT

DADEKKKKKRKKRKDKEQDESSESEEEEIDEDLPQPHQLPHPKKKKKKRKREDGEDSDNESDSDSSDLDNQEEDNLNPGSIPLPGQGKMKV
PSGHPHNQDPPPSDSGGP

DADEKKKKKRKKRKDKDEEESSSDEEEDEIDESDLPQPHQMKKKKKKKHKQDDDYDDDEYDERDYRGKTSGGVSLPGQGNRLPGPPKPGTH
LSATGNRTNSTASEPQGAV

DADEKKKKKRKKRKDKEDEESSGEKDEEEEIDEDLPQPHQIKKKHKKKKHKDAQEEISDDYDDDYDDQQTTGGKNKQEDLPLPGQGGLPVP
GQGAQVPTPDLPNKNLS

RL-GPS, P-body

TAGNAGKTDSPNPASKVDDNGKGIIGATLEKPKEAVSDIQGEKLNTNDRNQLRPGIGNEPVGGSLGRNTESGPNGGRQGDTLGGTGIRGSN
VTQGGNGNT

EKKDGKDNDDKDKKSDDNKGSNGKKVDSLKIPNKITVQSDTIGQSTNGVQNPASSVIGEAQIKQENQINAGEIKAKNGRQGYNTGRGGTVS
SGLGGGPGS

KGQSIKDVSDDYKINGLGPIPGKNDDKFQVGNPEGKPLVIAGTKIKGSDGGGHDISSPRQPGGGNVKYRLGSTQPSSHGQHGSRGKSQTGG
HGAVNSTSGRGI

QGVKDTMNLVKGGDLNVGVPTGGKSVVKGDQARKNKQTPINVSPSDTGQVRAASAPRTAGNVGRGEGQYKDQRNVIDAGSRGGDGRPGPKG
QNGPYAGGD

PHKDSDSRSDVKNVGVEGNSKVVPSGGLNRFKGSSNILTPTNEQIAGIGSGRNAGNKAVEGRTQGSAYSNVGVPGGGRSGTGYSGGGVGRG
SYPSQGSRGQSA

Activation steering (N=30, α=0.00)

HVNNVTRVKFFNSSDMPSAVKYFNKTTHAVLIDIKNQEYNEYTSKDNKVLIKSVGYVKDGVHKYRIILPDISLNSVLESDADRIFDSYLQF
FPGGHMDAFTKMS

IANNVQASTSARPGRYWVSELHTDEVPEATPAPTSATIMSKDAHYGILRFEWVSNHDLCESNRIQWKDLVSFPGQHFPLMYFQR

MSTLLRGARQSKVCWAILSRCDQYDYTCQQVKVVTHKLDGKKQIAFHSVGTHPDACDCAIYIPIVPSRSDHGGNAIVV

MATATLLNVESPSLRIPKLKGKRGPYHHRLSYQLYNASQCRVCGEKASGYHFSVSLA

QKTSRMYDRLTSVHLKKMCDSCFGVCPVNAVNVQKCLQTLDLDSIGQPEYTPST

Supervised fine-tuning

MADVLDLHEAGGEDFAMDEDGDESIHKLKEKAKKRKGRGFGSEEGSRARMREDYDSVEQDGDEPGPQR

MASCASIDIEDATQHLRDILKLDRPAGGPSAESPRPSSAYNGDLNGLLVPDPLCSGDSTSANKTGLRTMPPINL

APAAAPATAAPAKHAAAPKHRKAQAQMPEAKAAGAKHKKVAKKETAQKAQAAKKHGKKAPAKSTAAKPATKPEAAPAQ

MASCASIDIEDATQHLRDILKLDRPAGGPSAESPRPSSAYNGDLNGLLVPDPLCSGDSTSANKTGLRTMPPINL

PSGTPAPQRYENGAPGGAETGGRFAGANGSVGSWTGQEPAYSGSGSSTGHHTAAPRSSQANLNQSPADYSTVTPQSLQAPTANASN

Natural P-body IDRs

VKPESNDKETEAAYESDIPEELCGHHLPQQSLKSYNDSPDVIVEAQFDGSDSEDGHGITQNVLVDGVKKLSVCVSEKGREDGDAPVTKDET
TCISQDTRALSEK

LEKLFSGPNAANNKKKKIIPQAKGVVNTAVSAAVQAVRGRGRGTLTRGAFVGATAAPGYIAPGYGTPYGYSTAAPAYGLPKRMVLLPVMKF
PTYPVPHYSFF

AGIGPPSEASDSVLMKDAAYPPGPPSNPHVTDTTKKS

MIAPTDSHEEVRSGTSYILPFASRFLSFRADSAHASMVNVPKTRRT

QQIEKEIKEKDSELNQKRPQYIKAKENTSHKIKKLEAAKKSLQNAQKHYKKRKGDMDELEKEMLSVEKARQEFEERMEEESQSQGRDLTLE
EN

NUCLEAR PORE COMPLEX

RL-SAE, Nuclear pore complex (N=30)

QTLTSSTTAATAFTSTSFSFGLGATTNTGLFGNKTNLGLNTGGGLGVNAGGGGSLGLNTGGGGLFGQKPATAASAGYYQAPTFYAANQQHASQIPDVIRASENF

DAPAPSVNTFTAAPATFGFGNSTFGLSSNNTNTGLFGSKTLGLTNGSGGLGVNTGGGLFGAKPGLGVIGSQPTQTGFHAGAGYGVGAQGVPAGYYHEDEQQAESRKDL

AAASTNSFSTFSFNSTNTFGFGTTNNTGLFGSKTNAGSLGVNTGGLFGNKPAVGVGGGIGVNTGTAGIGVGTPTTSAPAPYYHVDRHREIERVEIEE

PKSSPTNSFNTGVSAFSSFSFGSTNFGQSNTGLFGNKTNLGLNSGTGGLGVNTGGGGLGVNTGGGLFGGKPQVPSAGAPAASASQAGYYHPIHHRVEQIEAE

TEPSNFSLVGTGFSTISSSNFGFGSTNTFGLSNNNTGLFGNTGLNVGGGLFGNKPGGGLGVNNTTGGGYGVNAGQTAGGGFYGMDPASAVLREVIANINNDSIETINQ

RL-GPS, Nuclear pore complex

MFSLYTLWSIFFPSFSATWTSIQLSWLNMITSFEFQGIGMSATTASIQLSGVAEFASLSNIFFVAQTMGATTTVEGSQISAGASFQGIESSLNMLAYQNS

TVFLFLGTQANDALGSLTTTLLSYFLFGTPVTLTNFIEYALSLQWTYWAALGPPSSASEQLGHIAFTTNWILLQSLLGAASSAADSVAWFSVNLNALTL

WITAVGSNVGSGTATENWAGWLTSVNGQLSQMGTSSFTTTVTSTGFQPYPFTPFGFTTVENAGSSTATFTQLWWFTGQAELATFASPTAIISYGVSEIQVQ

RTLQTSGTTATAPGWTSWADWTGFIVSGGYSLVTSSFPSSLINGSPTSITAGLSIPSPVWQSLLQSVYSGSPVNVNFQQLPIGTSTDVYFAVYTNDFAQIR

SEIGFTTSSGITNWFLAFFSTSWVTLNSLIGLIFSGSGPSIYPTDSIVGVGNSLVQGLTTATPVPALGNTAFPSVIFGNITSYEGTFQQPAYFMWFMTTA

Activation steering (N=30, α=0.20)

RSTAAAMATAPPSSQVAPLDFNRDGHVDFLLAHAGGGAAAAIARYTNTGSFNAFTALAGGTFGTVASPVPAFYGGTKFGTLGGAGAAAVPAPIS

MAADNQVLPLVDDDAAQQQGVPLLGAANHSAPQQQSAGFGDLFGAGGAAGGSATAQGSGLGGLLGALTGAPANQGGIGGLLGGGAAMGGTMSGGGLFSGAGGLSLGGGLMA

LSEIASGASRSLAPPVGLTGRARPTNAGSSPERRSAPPSTSVAHRSPSGTMAAASPKSSQSLARTTSTRASFASSSNGSSGPAGGNDK

MKKYKITMNGKSYEVEVAEVSKNATAASAQTAPASSAASQSQVAAPVAAAPAAVAAPAAAPAAVAPAAPAANGSSAGAAIS

LTADLQPGLTGNQSGTASTAAGTAAGTGGTSGSTSTGSTTGTSLTGGSSSLSAAANSSTANGFAGNTASTTNTASTATTGTTTASGTGTASGTATATATATAGAATTLGTGAVGLAGLAAFATALLA

Supervised fine-tuning

QNTPRVTTGPTPFSTMPNAAAVAMAATLTQQQQPATGFGTSSGFGCSTTGASTFGFGTTNKPSGWREANDVDFIFFVDLLIIFILYQILYL

MDTEGFGELLQQAEQLAAETEGISELPHVERNLQEIQQAGERLRSRTLTRTSQETADVKASVLLGSRGLDISHISQRLESLSAATTFEPLEPVKD

MAAAEGPVGDGELWQTWLPNHVVFLRLREGLKNQSPTEAEKPASSSLPSSPPPQ

NTPRVTTGPTPFSTMPNAAAVAMAATLTQQQQPATGDKWCNVAELKNRTQLSLFVNKLELSCYCVID

MDTEGFGELLQQAEQLAAETEGISELPHVERNLQEIQQAGERLRSRTLTRTSQETADVKASVLLGSRGLDISHISQRLESLSAATTFEPLEPVKDT

Natural nuclear pore complex IDRs

LQDSDEEEEEHPSKTSTKKLKTAPLPPASQTTPLQMALNGKPAPPPQVEKKGQ

MAAAEGPVGDGELWQTWLPNHVVFLRLREGLKNQSPTEAEKPASSSLPSSPPPQ

MYPDLSELLMVSAPNNPTVSGMSDMDYPLQGPGLL

RFPTTLPPPSNVGPILGSPVYSSSPVPSGSPYPNPSFLGTPSHGIQPPAMSTPVCALGNPATQATNMSC

MDTEGFGELLQQAEQLAAETEGISELPHVERNLQEIQQAGERLRSRTLTRTSQETADVKASVLLGSRGLDISHISQRLESLSAATTFEPLEPV

ACTIVATION DOMAINS

RL-SAE, Activation domains (N=30)

MSDTTNQPDSAPYAADDFMNLDNLDDLFNFDTDMANLTNDDFEDLFFGDMPSDQVPELPLVSPVENEPAETETWGNDWPSDETGS

KPDAPPEIQTTDAMDFNVDFLDISDFNNLSSLSDPDFDIFNSDFFENLTPNTPDIPTPIEDAGRSAETEETWSGNDWGNTE

MPSNPPPTLDDDFNLMGSRDPIDLNELNFDSDLFSDLNFGNLDDIFFDSLLGDKDIPNAAATDQTAPQVSGEEPNGWGGDSWG

GGDAPAPAPVADTMPDDLFLTESDFMGLEFGNLDTDFFTNLAEDDFFTDIEIPNSDIPIADSLPSPASVPQESDDWGSSQWGDT

NQAPVPEPPVLPDTNQAVADDFDTFFSPDLDGFNTDFDLLDTFDNKTPAHQSDDDMNLQESAFAPATDILTDGLNDNSNWGD

Activation steering (N=30, α=0.43)

VDFDPGGVPDLSQVGSFDDLAGAISSFMDPSFTGTSLDDFDLAGLSDSNPTDLAFPNFVLDSNFASTLPDDITGGYGKFGTANFLGNY

VEPHRSDDPAGQMSLEMMAMLGRMSTLSLDGRAGESTDGMSVENFHATGLADDGDSMEWMFYGLPS

MSGFYNSGYDEIPGLDSTLSAASNTGFGSSSGFLNGGFSEAQNFGGLGNPADLTSGIGGLDSTFSAGTTLVASFGDDDPDFLRISENSLTG
TPIVDGNPFLINFVL

MLGLVDDFFFDFDMTEESSLLDGASPDSTASIDLPSNDGTSSPVAVSADPPNQHENLFTNMNETLPNDSTAQFPT

LGPDTDGADDAFTGDEDLVFDAASGLAGDPNIPDVPGLPELFAEDSDEDLDSGGSDPLVPDNGGGFNFDFDPFDDMEPDSFSPDTADFDFT
DTGDFDSLGDFETFG

Supervised fine-tuning

LEVGSYRMVRAASPDPYSTEETVVVSCCSSVQQQLPAQPLPVPLPCDAATFFEGGNTGSLHRTTGGQY

MTLPGEAGAAGLAPPGHAAAFGGPPGGLLLDALPGPYAAAAAGPLGAAPDRFPADLDLDMFSGSLECDVESIILNDFMDSDEMDFNFDSAL
PPPPPGLAGAPPPNQSWVP

ALGRNHLADNCYSNYPFPLTSWPCSFSPSQNSSEPFYQQLPLEPPAAKTGCPPLWPNPAGNLYEEKVHVDFNSYVQSPAYHSPQEDPFLFT
YASHPHQQY

GLSQSQSPGQTGFLSYGTSFSTPQPGQAPYSYQMQGSSFTTSSGIYTGNNSLTNSSGFNSSQQDYPSYPSFGQGQYAQYYNSSPYPAHYM

VQLIPNFNNWKEPITSDPMESMDTEADSGHLMAMLQEQLDAINMQAGKIQNFLSEIGVAMD

Natural activation domains

PEDSLFMFPCGPSGGTSQGYCPGASSGQILMQPPAGNIGTASWSDPCLPELPFPGPFCPQALGHPPGGDGYFPDLFPTPC

GGTPDIPSSGLLSGQAQENPGYPYSDSSSILGENPHIGIDMIDNDQGSSSPSNDEAAMAVIMSLLEADAGLGGPVDFSDLVDFSDLPWPL

MVPSPEAKLTEVDNYHFYSSIPSMEKEVGNCSPHFLNAFEDSFSSILSTEDPNQLTVNSLNSDATVNTDFGDEFYSAFIAEPSISNSMPSP
ATLSHSLSELLNGPIDVSDLSLCKAFNQNHPESTAEFNDSDSGISLNTS

DLGAHVTKQQSHPEQNSVDYCQQLTVSQGPSPELCDQAIAFSDPLSYFTDLSFSAALKEEQRLDGMLLDDTISPFGTDPLLSATSPAVSKE
SSRRSSFSSSFSSDDGDEL

LEAGKLPCEDPELDDDFDAHKALEDDEELAHLGLGVDVAKGDDELGTLENLETNDPHLDDLLNGDEFDLLAYTDPELDTGDKKDIFNEHL

REPRESSION DOMAINS

RL-SAE, Repression domains (N=30)

MKIDPPPPRVRLPESRALSVPATEGKPQLKVKKEEPMIIQVKKEEPLVIKVEKMEDDDNDDIKSKDEPLQDKNGKTEPMHIKV

LSYDDDRRNTLAPGPTESPSGLRSDAKTEIRTEIKEEPPSIKKEEPLVIKVEKMEEDDDMPDKKNNKEVKDEPLQIKVEEDET

PEKAAEAGPDAVPASPPTAQAPLKGQVKTEVKKEELKIKEEPLVIKVEKMEDDDDMDVKNVKDEPLQIKVEEDSSTNAN

DRQGGPGSAPRPGTGSQTHPTPTDKMTEIKMEPKAEVDVKKEEPLVIKVEKMEDDDDMDVKNDDDMKNKSEPLHIKIEEETAE

GLLLGKTDMEHPTPVQQNGTAREAEEIKGPKTEAKTEVKEEPLVVKVEKMEDDDMDVKNEKNQIKEEPLQIKVEEDTFGV

Activation steering (N=30, α=0.35)

MLGCLQTGSANHIQWFSTEICPRLCEEEEVNVAECEDTLEYEQIEDDDKIKFIEEIEKLEKEDTEMEKQLEQQEDLEDNEVFVEEEN

MSELTQKEKKPEKKPVEEINVEKEEVFILDEEYEDEEVEEYDLEPSEEIEEFEEIRIREMIEDEIEKEVDEEFSIEEKEEKESSKDE

LKEFEDLHPHQSKQNCMPHFVEEKYEEILNEVKNEIEEPELKVEDYLEDEEESEEVFEEEEFDESKKKNNS

GFIVEENDVDKEPILKEEDWEGVVNPIDEIEEKQVIVEKEVVKEKKVEKKKKVEIKVKEVPVKEEQVKKETDI

MDSNHTDSKPITIEKKEMFSYNDENKTEEEEEEEEEEEEYDDEESIEEFEEKELLSKKESEMQPFHEKIMQEEEELIIEKKEDE

Supervised fine-tuning

PGPATLAPSRPLLVPKAERLSPPAPSGSERRLSGDLSSMPGPGTLSVRVSPPQPILSRGRPDSNKTENRRITHISAEQKR

NHWTLGYPNVHEITRSTITEMAAAQGLVDARFPFPALPFTTHLFHPKQGAIAHVLPALHKDRPRFDFANLAVAATQEDPPKMGDLSKLSPG
LGSPISGLSKLTPDRKPSR

MDPRGVVPSLWDREHEGKVKNGTCMFVKENNMAEPGSEAPLLSKAGAISQRTEGQREPRDQKTDPQSGAKPLGPPAAASPENTKGKELRIL
SNFGSTS

MEVMDSCKFSPSELFYDSSCLSSPEGEFPEDFEPRELPVFGAHAPTPPPSTIQR

DDMSDDQGTSVQSKSPDLEASLDTLENNCHVGSDIDFRPKLVNGKGPLDNFLRNRIETSIGQSTVIIDLTEDSNEQPDSL

Natural repression domains

QQVSDSTKNGDGTKRPFRQNTHGIQMTSIKKRRSPDDELLYLPVRGRETYEMLLKIKESLELMQYLPQHTIETYRQQQQQQHQHLLQKQTS
IQSPSSYGNSSPPLNKMNS

LELTPVQPELQLELVPAAGGGGAAVPGAPAAVVVAPPGYVVVQELMVLPAVAAPAVVAIPGPAGSAALTPARQRRRRRARDRPTICGECGK
GFSRSTDLV

CDTVFTHKPLLSSHFDQHLLPQRVSVFKCPSCPLLFAQKRTMLEHLKNTHQSGRLEETAGKGAGGALLTPKTEPEELAVSQGGAAPATEES
SSSSEEEEVPSSPEPPRPAKRPRRELGSKGLKGGGGGPGGWTCGLCHSW

NSGREPRTPRTLLSIADILAPRMVPRAPSAPQLPESGPGPTSPLCALEELTSKTFRGLDARALQPSEGRAGPDALGPGPF

HLKKHKAALAAAQYKSSNASETEPNAIKNQTLSASLPLDSTVICSTINKANSVGNGQASQTSQPNYHTKLKKAWLTRHSEEDKNTNKMENS
GNSVSEIIKPCSVNLIASTSSDIQNSVDSKIIVDKYVKD