Biological Graph Priors for Representation Learning in Cellular Microscopy

Reed Naidoo[1,2] **Matt De Vries**[1] **Georgios Skiadas**[1]
Vicky Bousgouni[1] **Ugur Moustafa**[1] **Chris Bakal**[1,2]

[1]Sentinal4D, London, UK [2]The Institute of Cancer Research, London, UK

Abstract

Learning representations of microscopy images that reflect the functional state of the cell remains a central challenge in computational biology. Dominant approaches repurpose self-supervised objectives from natural-image modelling and rely on scale, with no guarantee that the resulting embeddings reflect gene function. We introduce NOVA, a structure-first framework that integrates curated biological interaction graphs into the training objective of a vector-quantised transformer, encouraging image embeddings of related genetic perturbations to lie close together. We show that the discrete bottleneck is central to this design, as quantisation preserves the structure of the representation under this regularisation. Trained on 66K images with a 19M-parameter model and evaluated on an unseen cell line without fine-tuning, NOVA matches or exceeds baselines trained on up to 10M images or with up to 1.5B parameters on three of four gene–gene interaction benchmarks. We further show that the nature of the relational prior matters as much as its presence: soft, probabilistic constraints generalise, while hard, deterministic constraints degrade performance. Model weights are available on HuggingFace: `https://huggingface.co/Sentinal4D/NOVA`

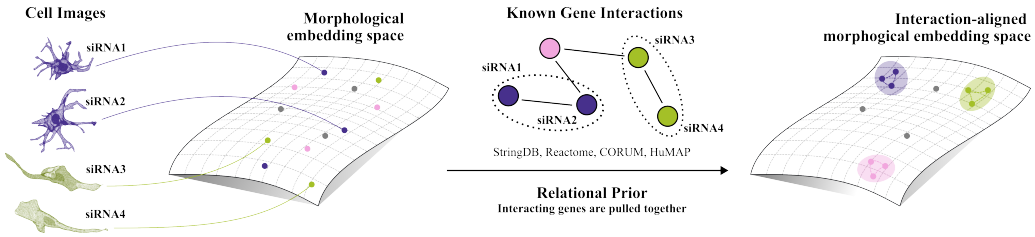


Figure 1: **Linking form to function.** Image-only embeddings organise cells by appearance. NOVA uses known biological interactions as a training prior, pulling interacting genes together in embedding space.

1 Introduction

Silencing a gene often changes a cell's morphology, and microscopy records that change as an image. The goal of representation learning for cellular microscopy is to embed these images so that the geometry of the embedding space reflects the underlying biology: cells with functionally related gene knockdowns should lie close together. Achieving this is a central problem for the field, drawing on methods from both computer vision and computational biology [Przybyla and Gilbert, 2022, Kraus et al., 2024, Noutahi et al., 2025]. In response, numerous deep learning approaches have been proposed to extract biologically relevant information directly from images [Kraus et al., 2024, Lu et al., 2025, Fradkin et al., 2024, Sanchez-Fernandez et al., 2023, Huang et al., 2025]. The

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dominant trend has been to repurpose architectures and objectives from natural-image modelling, spanning contrastive [Chen et al., 2020, He et al., 2020], generative, and masked-image variants [He et al., 2021], and to compensate for the absence of biological supervision through scale: billion-parameter models trained on tens of millions of images. While this line of work has produced notable benchmarks, it largely optimises for generic visual similarity rather than biological relatedness, and relies on the implicit assumption that biological structure will emerge from sufficient data.

It is increasingly accepted that form follows function [Bakal et al., 2007, Lomakin et al., 2020, Verde et al., 1995]: morphology reflects underlying pathways and molecular mechanisms, and perturbed-cell phenotypes carry recoverable biological signal [De Vries et al., 2025, Caicedo et al., 2017, Carpenter et al., 2006]. Yet current training objectives treat biological evaluation as a post-hoc criterion; testing whether embeddings recover gene–gene or compound–target associations after training, while the models themselves remain agnostic to these relationships during learning. The result is a structural mismatch: models are rewarded for visual fidelity during training and assessed for biological fidelity at evaluation, with no mechanism to align the two.

We argue for a structure-first alternative: make known biological relationships an explicit part of the learning objective. Community resources: protein interaction networks, pathway databases, and functional annotations capture relational priors among genes. Using them during training provides partial supervision that anchors representation geometry in biologically meaningful topology without requiring exhaustive labels. This is distinct from supervised classification: no target variable is predicted, and no discrete label is assigned to any training example. Instead, interaction graphs impose a soft geometric constraint: related perturbations should occupy nearby regions of latent space, leaving the encoder free to organise the remainder of the manifold according to visual structure. This perspective reframes progress from a problem of scale to one of inductive bias. However, not all priors are equal: we show that the nature of the relational constraint, its specificity and permissiveness, critically determines whether injected structure improves generalisation or induces overfitting.

We introduce **NOVA**, a framework that integrates curated biological graphs into the objective of a vector-quantised transformer trained on microscopy images. The vector-quantised architecture is central to this design rather than incidental: the discrete bottleneck compresses morphological variation into a finite vocabulary of cellular prototypes, creating a natural substrate for relational regularisation. The relational loss operates on pre-quantisation encoder representations, shaping the continuous manifold so that biologically related perturbations are steered toward shared or adjacent codebook regions; quantisation then amplifies this alignment by discretely grouping related genes into shared prototype assignments. Together, relational regularisation and discrete compression form a mutually reinforcing system; one that achieves biologically faithful representations with substantially less data and compute than scale-driven baselines, and without requiring any image-level labels. We propose:

- **Structure-first training.** NOVA injects interaction-network priors directly into the learning objective, aligning morphology with known biological relationships during training rather than only at evaluation. This partial supervision is architecturally lightweight and requires no image-level annotation beyond perturbation identity.
- **Efficiency through structure.** By leveraging relational priors, NOVA matches or exceeds purely visual baselines across three of four gene–gene interaction benchmarks; demonstrating that inductive bias is a viable and complementary alternative to scale, not a replacement for it.
- **Constraint characterisation.** We systematically analyse how prior "hardness," from permissive probabilistic networks to strict causal graphs, affects embedding geometry and transfer, revealing that soft constraints generalise while hard constraints overfit. This finding has direct implications for how biological knowledge graphs should be selected and used in representation learning pipelines beyond NOVA.
- **A designed architecture for biological microscopy.** The vector-quantised transformer is adapted and motivated specifically for this setting: channel-agnostic projection handles variable fluorescence channel configurations, and the discrete bottleneck interacts directly with the relational objective to amplify biological clustering. The resulting model is 19M parameters, trained on 66K images, and releases pre-trained weights to facilitate reproducible research.

2 Related Work

2.1 Benchmarking Biological Recall from Microscopy Images

Evaluating whether microscopy embeddings capture known biological relationships has become a standard criterion in representation learning for cell imaging. The **RxRx3-core** benchmark [Kraus et al., 2025] provides a large-scale, standardised evaluation framework comprising 222,601 Cell Painting images [Bray et al., 2016b] spanning 736 CRISPR gene knockouts and 1,674 small-molecule perturbations. Models are evaluated by computing pairwise cosine similarities between averaged gene embeddings; the fraction of the most and least similar 5% of pairs that correspond to known biological relationships, derived from Reactome [Gillespie et al., 2021], hu.MAP [Drew et al., 2017], SIGNOR [Perfetto et al., 2015], StringDB [Szklarczyk et al., 2020], and CORUM [Tsitsiridis et al., 2022], defines the biological recall. RxRx3-core functions as a dedicated evaluation benchmark: all methods in Table 1, including NOVA, are evaluated on its full gene set without any training on its images or experimental conditions. The **EFAAR** framework [Celik et al., 2024] provides a complementary methodology for constructing perturbative maps, with embedding alignment strategies that mitigate batch and plate-level artefacts.

2.2 Metadata- and Knowledge-Supervised Microscopy Models

Self-supervised learning [Chen et al., 2020, He et al., 2020, 2021, Oquab et al., 2024, Siméoni et al., 2025] has been widely applied to cellular microscopy, though augmentation-based methods can distort biologically meaningful structures [Kim et al., 2025, Kraus et al., 2024], and masked image modelling typically requires tens of millions of images to achieve strong biological recall [Kraus et al., 2024]. A growing body of work incorporates partial supervision from paired metadata into the training objective without requiring exhaustive annotation [Radford et al., 2021, Wang et al., 2022]. Methods such as MolPhenix [Fradkin et al., 2024], CLOOME [Sanchez-Fernandez et al., 2023], and CellCLIP [Lu et al., 2025] use treatment identity, molecular structure, or textual metadata as soft supervision signals. The Cross-Well Aligned Masked Siamese Network (CWA-MSN) [Huang et al., 2025] enforces semantic consistency by aligning embeddings of cells subjected to the same perturbation across different experimental wells, a form of metadata-derived geometric constraint that improves biological fidelity without image-level labels. Most closely related to our work, Chen et al. [2025] also regularise embeddings with a perturbation graph, but as one component of a disentanglement framework built on continuous encoders, assessed by perturbation classification accuracy after few-shot fine-tuning on held-out cell lines.

NOVA extends this line of work in a specific direction: rather than constraining embeddings by experimental co-occurrence or treatment identity, it uses interaction-network topology as the geometric prior. Metadata alignment (as in CWA-MSN) enforces within-perturbation consistency but provides no signal about relationships between perturbations. Gene interaction graphs encode precisely this cross-perturbation relational structure, enabling the model to learn that two distinct knockdowns should occupy nearby regions of latent space, not because they appeared in the same well, but because their gene products interact. Unlike Chen et al. [2025], NOVA uses curated interaction networks alone, couples the constraint to a discrete bottleneck, and is evaluated zero-shot on biological recall. The relational prior constitutes partial supervision, not supervised classification: no target variable is predicted, no discrete label is assigned, and the constraint is a soft geometric regulariser that leaves the encoder free to organise the remainder of the manifold according to visual structure.

2.3 Graph-Regularised Representation Learning

Regularising learned representations with a graph has a long history in machine learning. Graph-based semi-supervised learning [Zhu et al., 2003] and manifold regularisation [Belkin et al., 2006] encourage solutions to vary smoothly over connected nodes. External knowledge graphs have also been used to guide visual recognition [Marino et al., 2017]. NOVA applies this principle in cellular microscopy: curated gene interaction networks provide the graph, and the regulariser shapes a morphological embedding space rather than a set of predictions.

3 Methodology

3.1 Problem Statement

Learning biologically meaningful representations from microscopy images presents a dual challenge: models must capture local morphological structure while maintaining consistency with global biological relationships across perturbations. Existing self-supervised frameworks address the first challenge through reconstruction or contrastive objectives, but treat the second as an emergent property of scale. We argue this is structurally insufficient: biological relationships between perturbations are not fully recoverable from visual similarity alone, particularly for gene pairs whose functional interaction produces divergent rather than convergent morphology.

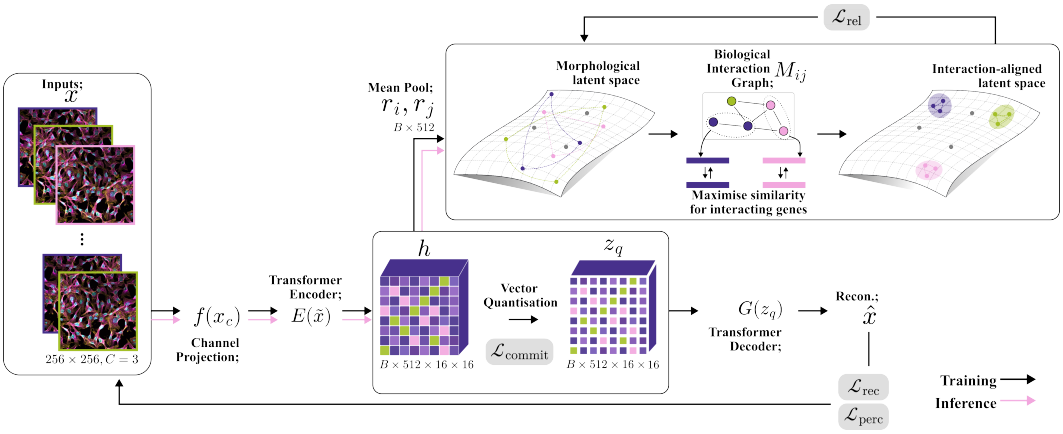


Figure 2: **NOVA: structure-first representation learning for cellular microscopy.** Input multi-channel fluorescence images $x \in \mathbb{R}^{C \times H \times W}$ are aggregated via a channel-agnostic projection f_{proj} and encoded by a transformer encoder E into a spatial latent feature map $h \in \mathbb{R}^{B \times D \times H' \times W'}$. Spatial mean-pooling of h yields perturbation-level representations $r_i, r_j \in \mathbb{R}^D$ (morphological manifold), which are regularised by a curated biological interaction graph M_{ij} that maximises cosine similarity between embeddings of interacting gene pairs (interaction-aligned manifold). In parallel, h is discretised via vector quantisation into z_q using a learned codebook Z, and reconstructed by transformer decoder G. The total objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}$ jointly optimises reconstruction fidelity and biological alignment. At inference, M_{ij} is absent: the encoder alone maps unseen perturbations to embeddings evaluated by pairwise cosine similarity.

In NOVA, we formulate representation learning as a joint visual–relational problem. The model combines a vector-quantised transformer (VQT) that encodes cellular morphology with a graph-based regularisation term that constrains latent geometry according to curated gene–gene networks (Figure 2). The following sections describe the model architecture, relational objective, and dataset construction in detail.

3.2 Vector-Quantised Transformer for Cellular Representation Learning

Our base model follows a vector-quantised autoencoding formulation that reconstructs microscopy images from compact discrete representations. Let $x \in \mathbb{R}^{C \times H \times W}$ denote a multi-channel microscopy image, where C represents the number of fluorescence channels. Each channel is first projected individually through a shallow convolutional network with an intermediate channel expansion:

$$\tilde{x}_c = f_{\text{proj}}(x_c), \quad \tilde{x} = \frac{1}{C}\sum_{c=1}^{C} \tilde{x}_c, \tag{1}$$

where f_{proj} comprises two convolutional layers ($1 \to 8 \to 1$ channels) with group normalisation and ReLU activation applied after the first layer only. This bottleneck structure enables channel-agnostic training by projecting each fluorescence channel to a common representation before aggregation.

The aggregated projection $\tilde{x}$ is tokenised into non-overlapping 16×16 image patches, linearly embedded, and processed by a transformer encoder:

$$h = E(\tilde{x}), \tag{2}$$

which outputs a spatial latent feature map $h \in \mathbb{R}^{D \times H' \times W'}$, where $H' = H/16$ and $W' = W/16$.

To obtain a discretised representation, the encoder output is passed through a vector quantisation module [Esser et al., 2021] with a learned codebook $Z = \{z_k\}_{k=1}^{K}$, where each codebook entry $z_k \in \mathbb{R}^D$ represents a prototypical latent pattern. Quantisation is applied independently at each spatial location:

$$z_q^{(i,j)} = \arg\min_{z_k \in Z} \|h^{(i,j)} - z_k\|_2, \tag{3}$$

producing a grid of discrete codes $z_q \in \mathbb{R}^{D \times H' \times W'}$. This discrete bottleneck constrains representations to be expressed as compositions of morphological prototypes, promoting stability during training and encouraging semantically consistent latent codes that capture recurring cellular structures [Naidoo et al., 2025].

The quantised embeddings are passed to a transformer decoder that reconstructs the original image:

$$\hat{x} = G(z_q), \tag{4}$$

where G applies self-attention over quantised patch tokens followed by a linear projection to pixel space. The decoder mirrors the encoder structure in depth and width.

Loss formulation. The reconstruction process is optimised using a combination of L1 reconstruction, perceptual [Johnson et al., 2016], and codebook losses:

$$\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|_1, \tag{5}$$

$$\mathcal{L}_{\text{perc}} = \text{LPIPS}(x, \hat{x}), \tag{6}$$

$$\mathcal{L}_{\text{commit}} = \|\text{sg}[h] - z_q\|_2^2 + \beta\|h - \text{sg}[z_q]\|_2^2, \tag{7}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator and $\beta = 0.25$ weights the encoder commitment term. The first term updates the codebook toward encoder outputs, while the second encourages encoder outputs to remain close to their assigned codes. Together, these terms promote stable codebook utilisation and faithful image reconstruction.

3.3 Graph-Regularised Relational Consistency

While the vector-quantised transformer learns to encode morphological structure, its representations are initially aligned with visual similarity without direct supervision on biological context. To incorporate known gene–gene relationships, we introduce a relational loss that anchors embeddings of biologically related perturbations within shared regions of latent space, encouraging relational consistency across known interaction networks.

Let $\mathcal{G} = (V, E)$ denote a curated biological interaction graph, where nodes correspond to gene perturbations and edges represent known associations. For each image x_i in a mini-batch, the encoder produces a latent feature map h_i, which is spatially mean-pooled to obtain a perturbation-level representation vector $r_i \in \mathbb{R}^D$. Each training sample is associated with a perturbation label g_i (the target gene of an siRNA knockdown), provided alongside the image and used to query the interaction graph during loss computation.

Given perturbation labels g_i, g_j associated with samples i and j, we construct a binary relation mask:

$$M_{ij} = \begin{cases} 1, & \text{if } (g_i, g_j) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The relational regularisation term encourages embeddings of related perturbations to exhibit high cosine similarity:

$$\mathcal{L}_{\text{rel}} = \frac{1}{|E_b|} \sum_{(i,j) \in E_b} \big(1 - \cos(r_i, r_j)\big), \tag{9}$$

where E_b denotes the subset of graph edges present within the current mini-batch. This objective transfers prior biological structure from curated interaction networks into the learned latent geometry, constraining representations to preserve relationships among functionally related genes.

Batch sampling considerations. Effective relational learning requires biologically connected perturbations to co-occur within mini-batches with sufficient frequency. At a global batch size of B=512B, P=707 perturbations with average node degree $\bar{d} \approx 10$, this yields approximately 5 positive pairs per batch: sparse but stable, accumulating consistently across thousands of training steps. When no related pairs are present, $\mathcal{L}_{\text{rel}}$ defaults to zero.

Total objective. The overall training objective combines reconstruction, perceptual, commitment, and relational terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}} + \lambda_{\text{rel}}^{\text{eff}}\mathcal{L}_{\text{rel}}, \tag{10}$$

where $\lambda_{\text{rel}}^{\text{eff}} = 50$ controls the influence of biological alignment. In practice, we apply a scaling factor of 10^3 to $\mathcal{L}_{\text{rel}}$ before weighting by $\lambda_{\text{rel}} = 0.05$ to match the magnitude of reconstruction losses. This lightweight inductive prior encourages cells with related perturbations to occupy neighbouring regions in latent space, improving biological faithfulness without requiring additional supervision or large-scale architectural changes.

3.4 Training Data and Preprocessing

NOVA was trained on an in-house collection of 2D fluorescence microscopy images obtained from SHEP neuroblastoma cells subjected to 707 distinct siRNA-mediated gene knockdowns. Cells were fixed and stained for α-tubulin (microtubules), F-actin (cytoskeleton), and Hoechst (DNA). Imaging was performed using a $40\times$ water-immersion objective, yielding three fluorescence channels. Images were normalised per-channel to the range $[-1, 1]$ using min-max scaling, centre-cropped to 256×256 pixels, and augmented with random horizontal flips to preserve morphological integrity while increasing sample diversity. The final training corpus contained $\sim$66K microscopy images.

3.5 Training Configuration

All models were trained using PyTorch Lightning on 8 NVIDIA H100 GPUs (80 GB HBM3) for approximately 80 GPU hours. Optimisation used AdamW ($\beta_1 = 0.5$, $\beta_2 = 0.9$) with a base learning rate of 4.8×10^{-4}, scaled linearly with effective batch size. Training used full precision (FP32) with gradient clipping (max norm 1.0) and a per-GPU batch size of 64 (global batch size 512). The encoder and decoder each comprised $L = 6$ transformer layers with 8 attention heads and hidden dimension $D = 512$. The vector quantiser contained $K = 2048$ learned codebook entries, updated via exponential moving average with decay $\gamma = 0.99$. Loss weights were set to $\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{commit}} = 0.25$, and $\lambda_{\text{rel}}^{\text{eff}} = 50$. The relational loss was computed only when at least one valid gene–gene pair was present within a batch; otherwise, training proceeded with the reconstruction objective alone.

4 Results

4.1 Recalling known gene-gene interactions (RxRx3-core)

We evaluate NOVA on the RxRx3-core gene–gene interaction benchmark and report biological recall against four curated relational resources (Table 1). All embeddings are evaluated following PCA Center-Scaling (PCA-CS), normalised to the `EMPTY_Control` condition as defined in the RxRx3-core protocol. NOVA is trained on SHEP neuroblastoma cells under siRNA knockdown, a dataset entirely disjoint from RxRx3-core in cell line, imaging platform, and experimental modality. The interaction graph is absent at inference, with the encoder alone mapping perturbation images to embeddings.

NOVA achieves the strongest results on CORUM, HuMAP, and StringDB, with competitive performance on Reactome. Relative to CWA-MSN, the strongest purely visual baseline, NOVA improves on CORUM (+0.013, +3.4% relative), HuMAP (+0.016, +3.6%), and StringDB (+0.019, +5.8%). Performance on Reactome is addressed directly by the constraint hardness analysis in Section 4.4: Reactome's rigid pathway-membership structure makes it the hardest benchmark for any model regularised by softer interaction signals, and its underperformance is consistent with, rather than anomalous to, the framework we propose.

To disentangle the relational objective from dataset-specific effects, Table 1 includes NOVA-no-prior trained on both the in-house SHEP dataset and the public Cell Painting dataset from Bray et al.

Table 1: **Biological recall on RxRx3-core.** Recall (↑) against curated gene–gene resources. Best results per column are bolded.

	Training Settings		Relational Database			
Method	**Images**	**Params**	**Dataset**	**CORUM ↑**	**hu.MAP ↑**	**Reactome ↑**	**StringDB ↑**
MolPhenix [Fradkin et al., 2024]	200K	36M	[Bray et al., 2016a]	0.262	0.306	0.142	0.241
CLOOME [Sanchez-Fernandez et al., 2023]	200K	25M	[Bray et al., 2016a]	0.328	0.406	0.135	0.278
OpenPhenom [Kraus et al., 2024]	10M	10M	[Kraus et al., 2024]	0.326	0.388	0.155	0.297
CellCLIP [Lu et al., 2025]	200K	1.5B	[Bray et al., 2016a]	0.354	0.416	0.145	0.307
CWA-MSN [Huang et al., 2025]	200K	22M	[Bray et al., 2016a]	0.386	0.447	**0.158**	0.327
NOVA-no-prior	200K	19M	[Bray et al., 2016a]	0.361	0.437	0.134	0.305
NOVA-no-prior	65K	19M	In-House	0.377	0.426	0.152	0.316
NOVA	**65K**	**19M**	**In-House**	**0.400**	**0.463**	0.148	**0.346**

[2016a] under identical settings. Gains from the relational prior are consistent across both training domains, isolating graph regularisation as the source of improvement independently of dataset choice. Averaged across all four databases, NOVA attains a macro-mean recall of 0.339, exceeding the strongest open-source baseline CellCLIP (0.306) by +0.033 absolute (10.8% relative). Pre-trained weights are released on HuggingFace to facilitate reproducible research.

4.2 Relational Priors Exhibit Unequal Gains in Functional Recall

While relational supervision improves biological recall overall, gains vary substantially across relational resources. StringDB-derived priors consistently yield the strongest improvements in functional recall, often surpassing more curated resources such as CORUM, HuMAP, and Reactome (Figure 3).

Notably, this dominance is counterintuitive: StringDB encodes noisier, less specific interactions than curated complex- or pathway-level databases. One might expect that higher-quality, more biologically precise priors would translate to stronger downstream recall. Instead, we observe the opposite.

These results suggest that the effectiveness of a relational prior is not determined solely by biological specificity or curation depth. Rather, they point to a more fundamental interaction between (i) the structure of the relational graph and (ii) the geometry of the embedding space learned from microscopy data.

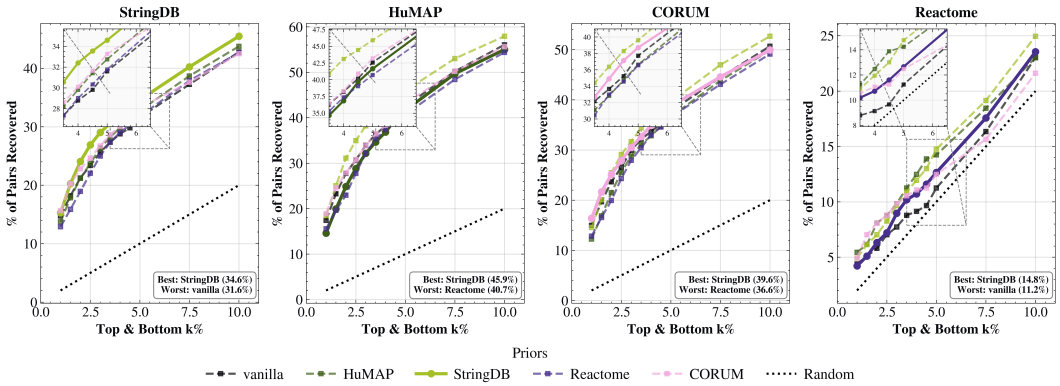


Figure 3: **Enrichment of Known Biological Relationships by Prior Type.** Recovery of known gene-gene relationships from four reference databases (StringDB, HuMAP, CORUM, Reactome) when examining the top and bottom k% of RxRx3-core gene pairs ranked by cosine similarity. Dashed lines indicate non-matching priors; the dotted black line represents random expectation (2k% recovery, since both extremes are examined). "Vanilla" is NOVA trained without a relational prior.

4.3 Soft Constraints and Relational Scale Enable Generalisation

To understand why some relational priors generalise better than others, we examine the interplay between constraint strength and graph size (Figure 4). Stronger constraints produce greater embedding separation between related and unrelated gene pairs (Figure 4A). While moderate separation is desirable, it reflects genuine biological structure in the embedding, excessive separation signals over-regularisation: the embedding space has collapsed around training relationships at the expense of flexibility. Figure 4B shows this directly: higher embedding separation correlates with worse transfer performance. Models trained with hard constraints (SIGNOR, Reactome) achieve strong in-distribution separation but fail to generalise; soft-constraint models (StringDB) maintain moderate separation and transfer effectively.

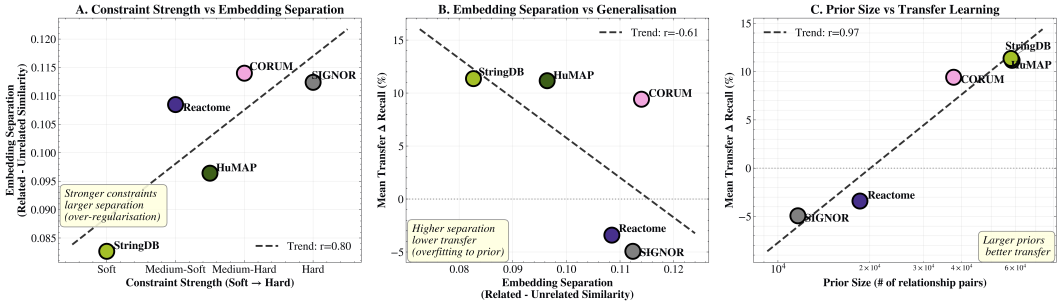


Figure 4: **Effective relational priors balance constraint softness with graph coverage.** (A) Stronger relational constraints increase embedding separation between related and unrelated gene pairs. (B) Higher separation correlates with decreased transfer performance, indicating over-regularisation. (C) Prior size shows strong positive correlation with transfer ($r = 0.97$), but size alone is insufficient: priors of similar scale (e.g., CORUM vs. StringDB) diverge based on constraint type. The optimal regime combines *soft* constraints with *broad* coverage. Models trained on SHEP, evaluated on RxRx3-core.

Prior size provides a complementary effect (Figure 4C). Larger graphs correlate strongly with better transfer ($r = 0.97$), but size alone is insufficient: CORUM and Reactome have comparable scale but diverge sharply in transfer performance. The critical factor is constraint type: StringDB's probabilistic, functional edges act as soft regularisation, while Reactome's physical pathway annotations impose rigid structure that the encoder cannot partially satisfy. Together these results establish that effective relational supervision requires both soft constraints and sufficient coverage. Large graphs with soft semantics occupy the optimal regime; small graphs with hard constraints induce over-regularisation regardless of biological accuracy.

4.4 Sources of Constraint Hardness

What determines whether a relational prior imposes soft or hard constraints? We identify three contributing factors, summarised in Appendix 6:

Graph size and redundancy. Larger graphs with overlapping relationships act as ensemble regularisation, ensuring no single edge dominates the gradient signal. StringDB and HuMAP (58,000 pairs each) achieve the highest transfer gains (+11.37% and +11.17% respectively).

Semantic specificity. Priors encoding deterministic, physically interacting relationships impose hard constraints requiring precise geometric conditions. SIGNOR produces the worst individual transfer (-4.93%) despite moderate size, due to its causal semantics. Probabilistic priors like StringDB instead permit slack, allowing constraints to be satisfied approximately.

Interaction directionality. Directed graphs enforce asymmetric relationships that constrain embedding geometry more rigidly than undirected networks. This contributes to degradation in both Reactome and SIGNOR, where rigid enforcement of directed signalling flow hinders generalisation to unseen biological contexts.

The ensemble paradox. Combining all priors (184,428 pairs) produces the worst transfer performance (-8.57%), despite 3× the coverage of any individual prior. Constraint quality dominates

quantity: mixing soft probabilistic associations with hard causal constraints dilutes beneficial regularisation while amplifying geometric distortions. StringDB alone outperforms the combined prior by nearly 20%.

4.5 Relational Priors Recover Divergent Interactors

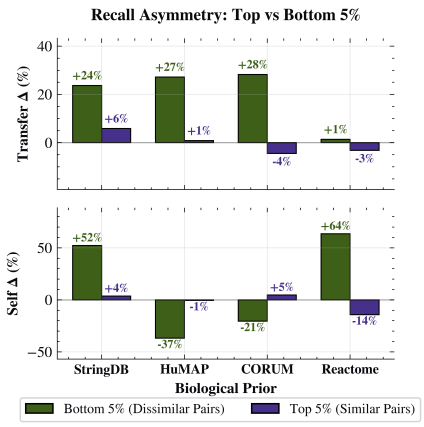


Figure 5: **Relational priors recover phenotypically divergent interactors.** Recall Δ (relative to vanilla baseline) decomposed by embedding similarity. The "Bottom 5%" represents gene pairs with ground-truth functional relationships that are initially phenotypically dissimilar.

To investigate the geometric shifts driving improved recall, we decompose performance gains by initial embedding similarity (Figure 5). Across all transferable priors, gains are heavily concentrated in the Bottom 5% (gene pairs with ground-truth functional relationships but the lowest phenotypic similarity in the baseline embedding space) with improvements exceeding +24%, against marginal gains of +1% to +6% in the Top 5%. This asymmetry identifies the primary function of relational supervision: not reinforcing obvious phenotypic clusters, but correcting for phenotypic divergence that purely visual approaches cannot resolve. This finding also speaks directly to whether the purely attractive loss risks collapsing antagonistic or inhibitory relationships. Such pairs, functionally linked but morphologically divergent, appear precisely in the Bottom 5%. Gains concentrated here indicate the relational loss is recovering these difficult cases rather than homogenising them. Hard constraints (CORUM, Reactome) tell the opposite story: degradation in Self recall for dissimilar pairs (21% to 37%) confirms that over-regularisation forces alignment at the expense of local biological variation. StringDB is the only prior to maintain positive gains across both extremes, reinforcing that soft constraints provide the flexibility to bridge functional gaps without distorting the broader manifold.

5 Discussion and Conclusion

We introduced NOVA, a framework that integrates biological interaction networks into the training objective of a vector-quantised transformer for microscopy representation learning. Trained on 66K images with a 19M parameter model, NOVA matches or exceeds purely visual baselines trained with substantially more data and compute on three of four gene–gene interaction benchmarks; demonstrating that relational inductive bias is a viable and complementary alternative to scale.

Constraint hardness matters. The nature of the relational prior critically determines generalisation. Soft constraints from probabilistic functional networks (StringDB) consistently outperform hard constraints from causal graphs (SIGNOR), and naive ensembling performs worst of all: a consequence that may be reflective of the many-to-many mapping between genotype and phenotype, where genes in the same pathway regularly produce divergent morphologies. Permissive priors better accommodate this biological uncertainty.

Limitations. Evaluation is restricted to genetic perturbations in a single neuroblastoma cell line; generalisation to compound perturbations, alternative cell types, and imaging platforms requires further validation. For sparsely annotated organisms or novel perturbation types, interaction graph coverage limits the relational signal, though maturing compound-target databases represent a natural extension of the framework.

Outlook. Structure-aware objectives offer a complementary path to scale, leveraging community-curated knowledge to achieve strong performance under constrained resources. Future work could extend NOVA to hierarchical priors, incorporate contrastive constraints for inhibitory relationships, or apply the graph-regularised objective as a fine-tuning strategy atop pretrained foundation models.

Acknowledgments and Disclosure of Funding

This is a summary of independent research supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and The Institute of Cancer Research. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The authors also thank StratMedPaediatrics2, funded by Cancer Research UK (CRCEMA-Jul23/100001).

References

Chris Bakal, John Aach, George Church, and Norbert Perrimon. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science*, 316(5832):1753–1756, 2007. doi: 10.1126/science.1140324. URL https://www.science.org/doi/abs/10.1126/science.1140324.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7 (85):2399–2434, 2006. URL http://jmlr.org/papers/v7/belkin06a.html.

Mark A. Bray, Sigrun M. Gustafsdottir, Vebjørn Ljosa, Shantanu Singh, Katherine L. Sokolnicki, Joshua A. Bittker, Nicole E. Bodycombe, V. Dančík, Thomas P. Hasaka, Christopher S. Hon, Matthew M. Kemp, Kun Li, Das Walpita, Mathias J. Wawer, Todd R. Golub, Stuart L. Schreiber, Paul A. Clemons, Alykhan F. Shamji, and Anne E. Carpenter. Supporting data for "a dataset of images and morphological profiles of 30,000 small-molecule treatments using the cell painting assay", 2016a. URL https://doi.org/10.5524/100200.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, September 2016b.

Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Jose Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9): 849–863, September 2017.

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, October 2006.

Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta M. Fay, Berton A. Earnshaw, and Imran S. Haque. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLOS Computational Biology*, 20(10):1–24, 10 2024. doi: 10.1371/journal.pcbi.1012463. URL https://doi.org/10.1371/journal.pcbi.1012463.

Jiayuan Chen, Thai-Hoang Pham, Yuanlong Wang, and Ping Zhang. Integrating biological knowledge for robust microscopy image profiling on de novo cell lines. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22846–22856, 2025. doi: 10.1109/ICCV51701.2025.02121.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Matt De Vries, Lucas G Dent, Nathan Curry, Leo Rowe-Brown, Vicky Bousgouni, Olga Fourkioti, Reed Naidoo, Hugh Sparks, Adam Tyson, Chris Dunsby, and Chris Bakal. Geometric deep learning and multiple-instance learning for 3D cell-shape profiling. *Cell Systems*, 16(3), March 2025.

Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeson, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology*, 13(6): MSB167490, June 2017.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. URL https://arxiv.org/abs/2012.09841.

Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 110667–110701. Curran Associates, Inc., 2024. doi: 10.52202/079017-3513. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c793577b644268259b1416464a6cdb8c-Paper-Conference.pdf.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1028. URL https://doi.org/10.1093/nar/gkab1028.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. URL https://arxiv.org/abs/1911.05722.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Pin-Jui Huang, Yu-Hsuan Liao, SooHeon Kim, NoSeong Park, JongBae Park, and DongMyung Shin. Efficient cell painting image representation learning via cross-well aligned masked siamese network, 2025. URL https://arxiv.org/abs/2509.19896.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. URL https://arxiv.org/abs/1603.08155.

Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports*, 15(1):4876, February 2025.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11757–11768, June 2024.

Oren Kraus, Federico Comitani, John Urbanik, Kian Kenyon-Dean, Lakshmanan Arumugam, Saber Saberian, Cas Wognum, Safiye Celik, and Imran S. Haque. Rxrx3-core: Benchmarking drug-target interactions in high-content microscopy, 2025. URL https://arxiv.org/abs/2503.20158.

A. J. Lomakin, C. J. Cattin, D. Cuvelier, Z. Alraies, M. Molina, G. P. F. Nader, N. Srivastava, P. J. Sáez, J. M. Garcia-Arcos, I. Y. Zhitnyak, A. Bhargava, M. K. Driscoll, E. S. Welf, R. Fiolka, R. J. Petrie, N. S. De Silva, J. M. González-Granado, N. Manel, A. M. Lennon-Duménil, D. J. Müller, and M. Piel. The nucleus acts as a ruler tailoring cell responses to spatial constraints. *Science*, 370 (6514):eaba2894, 2020. doi: 10.1126/science.aba2894. URL https://www.science.org/doi/abs/10.1126/science.aba2894.

MingYu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP - learning perturbation effects in cell painting via text-guided contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=W3JnXa9mW2.

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The More You Know: Using Knowledge Graphs for Image Classification . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–28, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.10. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.10.

Reed Naidoo, Matt De Vries, Olga Fourkioti, Vicky Bousgouni, Mar Arias-Garcia, Maria Portillo-Malumbres, and Chris Bakal. Morphologically intelligent perturbation prediction with form, 2025. URL https://arxiv.org/abs/2510.21337.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, Emmanuel Bengio, Dominique Beaini, Christopher Gibson, Daniel Cohen, and Berton Earnshaw. Virtual cells: Predict, explain, discover, 2025. URL https://arxiv.org/abs/2505.14613.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, Daniele Peluso, Lucia Lisa Petrilli, Stefano Pirrò, Daniela Posca, Elena Santonico, Alessandra Silvestri, Filomena Spada, Luisa Castagnoli, and Gianni Cesareni. Signor: a database of causal relationships between biological entities. *Nucleic Acids Research*, 44(D1):D548–D554, 10 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1048. URL https://doi.org/10.1093/nar/gkv1048.

Laralynne Przybyla and Luke A. Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, Nov 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42328-w. URL https://doi.org/10.1038/s41467-023-42328-w.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1074. URL https://doi.org/10.1093/nar/gkaa1074.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 51(D1):D539–D545, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1015. URL https://doi.org/10.1093/nar/gkac1015.

F Verde, J Mata, and P Nurse. Fission yeast cell morphogenesis: identification of new genes and analysis of their role during the cell cycle. *J Cell Biol*, 131(6 Pt 1):1529–1538, December 1995.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *CoRR*, abs/2210.10163, 2022. doi: 10.48550/ARXIV.2210.10163. URL https://doi.org/10.48550/arXiv.2210.10163.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 912–919. AAAI Press, 2003. ISBN 1577351894.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction contributions accurately state that NOVA matches or exceeds purely visual baselines across three of four gene-gene interaction benchmarks, and explicitly scope the evaluation to genetic perturbations in a single cell line. Aspirational directions (compound generalisation, foundation model fine-tuning) are clearly framed as future work in the Discussion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed explicitly in Section 5 (Discussion and Conclusion). These include dependence on interaction graph coverage, restriction to genetic perturbations in a single neuroblastoma cell line, and the absence of validation on compound perturbations, alternative cell types, and imaging platforms.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include formal theoretical results or proofs. Contributions are empirical and methodological; all equations are design choices or loss formulations rather than theorems requiring proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full architecture details, hyperparameters, loss weights, optimiser settings, and hardware configuration are reported in Section 3.5 and Appendix A. The training algorithm is detailed in Appendix F. The evaluation protocol follows the publicly available RxRx3-core benchmark. Pre-trained model weights will be released on HuggingFace to enable direct reproduction of evaluation results without requiring access to the proprietary SHEP training data; we additionally provide results on the public Bray Cell Painting dataset under identical settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The primary training dataset (SHEP neuroblastoma siRNA screen) is proprietary and cannot be publicly released. However, pre-trained model weights will be made available on HuggingFace, and we provide results for NOVA-no-prior trained on the publicly available Bray Cell Painting dataset under identical settings, providing a fully reproducible baseline. Evaluation is conducted on the public RxRx3-core benchmark following its standard protocol.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training configuration is fully specified in Section 3.5 and Appendix A, including optimiser, learning rate, batch size, loss weights, and hardware. Evaluation follows the RxRx3-core protocol (Section 4.1 and Appendix B–C); no training is performed on evaluation data. The batch correction procedure applied to embeddings at evaluation time is detailed in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported. The RxRx3-core benchmark is a fixed evaluation set with deterministic recall computation given fixed embeddings; variance across runs is driven by training stochasticity rather than evaluation noise. This is consistent with practice across all baseline methods compared in Table 1, none of which report error bars on this benchmark.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.5 reports that all models were trained on 8 NVIDIA H100 GPUs (80GB HBM3) for approximately 80 GPU hours using PyTorch Lightning with DDP. Additional hardware details are provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research involves computational analysis of fixed, anonymised cell imaging data and does not involve human subjects, sensitive personal data, or dual-use concerns. All biological databases used are publicly available. The work conforms with the NeurIPS Code of Ethics in all respects.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Discussion notes positive impacts: improved biological representation learning under constrained resources democratises access to high-quality morphological embeddings for smaller research groups and under-resourced organisms. No significant negative societal impacts are foreseen; the method processes fixed cell images for basic research and has no direct path to harmful applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released model encodes cellular morphology from fixed fluorescence microscopy images and poses no meaningful risk of misuse. It is not a generative model, does not process personal data, and has no plausible dual-use pathway.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (RxRx3-core, Bray Cell Painting, StringDB, HuMAP, CORUM, Reactome, SIGNOR) and baseline models are cited with their original papers. The VQ-transformer architecture builds on the publicly available implementation from Esser et al. (2021), which is properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Pre-trained NOVA model weights will be released on HuggingFace with documentation covering architecture, training configuration, intended use, and evaluation protocol. The proprietary SHEP training dataset cannot be released; its collection protocol is described in Section 3.4 and Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All data are fixed fluorescence microscopy images of cell lines.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects. IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methods. Any use was limited to writing and editing assistance, which does not require declaration under NeurIPS policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Baseline Methods

To contextualise NOVA's performance in Table 1, we describe each baseline method and training dataset below. Baselines vary along three axes relevant to the comparison: learning objective (purely visual reconstruction versus metadata-supervised contrastive), supervision signal (none, compound identity, or natural language), and training scale. Understanding these differences is essential for interpreting the results: NOVA is the only method that uses gene interaction graphs as a geometric constraint during training.

1. **MolPhenix, [Fradkin et al., 2024]**: learns a joint latent space between molecular structures (compounds) and cell painting microscopy images using contrastive learning, building on a pre-trained MAE image encoder. The supervision signal is compound identity: image embeddings are aligned with molecular structure embeddings, rather than any gene interaction information. It is a multimodal compound-phenotype model, not a purely visual one.
2. **CLOOME, [Sanchez-Fernandez et al., 2023]**: applies a CLIP-style contrastive objective to align microscopy image embeddings with chemical structure representations, enabling bioimaging databases to be queried by chemical structure. Like MolPhenix, its supervision signal is compound-image pairing rather than gene interaction structure.
3. **OpenPhenom, [Kraus et al., 2024]**: is a channel-agnostic masked autoencoder (CA-MAE) trained on a large proprietary, in-house dataset. It is fully self-supervised, no biological metadata of any kind is used during training, and represents the scale-driven ceiling of purely visual representation learning for microscopy.
4. **CellClip, [Lu et al., 2025]**: aligns cell painting image embeddings with natural language descriptions of perturbations using a cross modal contrastive objective. It introduces a *CrossChannelFormer* architecture to handle the multi-channel structure of cell painting images. Supervision comes from textual perturbation descriptions, not gene interaction graphs.
5. **CWA-MSN, [Huang et al., 2025]**: is the strongest purely-visual baseline and the most direct point of comparison for NOVA. It is a masked Siamese network that aligns embeddings of cells subjected to the same perturbation across different wells, enforcing within-perturbation consistency as a form of metadata-derived geometric constraint. Crucially, like NOVA, it uses no image-level class labels; but its constraint is within-perturbation (same gene, same well type), whereas NOVA's is cross-perturbation (different genes, known to interact).

A.1 Datasets

A.1.1 SHEP In-House Dataset (NOVA training)

The primary training dataset for NOVA is a proprietary fluorescence microscopy screen of SHEP human neuroblastoma cells under 707 distinct siRNA-mediated gene knockdowns. siRNA (small interfering RNA) is a reagent that silences a specific target gene, producing a cellular phenotype that can be captured by imaging. Cells were fixed and stained for three structural compartments; the nucleus (Hoechst), microtubules (α-tubulin), and the actin cytoskeleton (F-actin), yielding three fluorescence channels per image. All imaging was performed using a $40\times$ water-immersion objective. Images were centre-cropped to 256×256 pixels and normalised per-channel; the final corpus contains approximately 66K images. The SHEP dataset is entirely disjoint from RxRx3-core in cell line, imaging platform, perturbation modality, and channel configuration.

A.1.2 Bray Cell Painting Dataset (NOVA-no-prior training)

The JUMP-CP dataset [Bray et al., 2016a] is a large public Cell Painting screen comprising approximately 919K five-channel fields of view covering 30,600 small-molecule compounds, imaged in U2OS human bone osteosarcoma cells. Cell Painting uses a standardised panel of six fluorescent dyes, across five imaging channels, revealing eight broadly relevant cellular components including the nucleus, endoplasmic reticulum, nucleoli, actin cytoskeleton, Golgi apparatus, and mitochondria [Bray et al., 2016b]. The dataset is the standard training corpus in the field and is used by four of the five baselines in Table 1. In NOVA, it is used exclusively to train the NOVA-no-prior ablation under identical settings to the in-house variant, allowing us to isolate graph regularisation as the source of improvement independently of dataset choice (Table 1). NOVA itself is not trained on this dataset.

Table 2: Dataset comparison across training and evaluation corpora used in this work.

	SHEP (In-House)	Bray Cell Painting	RxRx3-core
Reference	This work	Bray et al., 2016	Kraus et al., 2025
Perturbation type	siRNA knockdown	Small molecules	CRISPR KO + compounds
Perturbation count	707 genes	∼30,600 compounds	736 KO + 1,674 compounds
Cell line	SHEP neuroblastoma	U2OS	HUVEC
Channels	3 (Hoechst, α-tubulin, F-actin)	5 (Cell Painting panel)	6 (modified Cell Painting)
Images	∼66K	∼919K fields of view	222,601 images
Public	No	Yes	Yes
Role in NOVA	NOVA training	NOVA-no-prior training	Evaluation only

A.1.3 RxRx3-core (Evaluation Only)

RxRx3-core [Kraus et al., 2025] is a publicly available benchmark for evaluating representation learning models on high-content microscopy data. It is a curated, compressed subset of the larger RxRx3 dataset [Kraus et al., 2024], comprising 222,601 images of six-channel fluorescence microscopy images of HUVEC human umbilical vein endothelial cells, spanning 736 CRISPR gene knockouts and 1,674 small-molecule compounds at eight concentrations each. Ground-truth biological relationships are derived from five orthogonal curated databases: StringDB [Szklarczyk et al., 2020], HuMAP [Drew et al., 2017], CORUM [Tsitsiridis et al., 2022], Reactome [Gillespie et al., 2021], and SIGNOR [Perfetto et al., 2015]. Models are evaluated by computing pairwise cosine similarities between mean-aggregated perturbation embeddings and measuring what fraction of known interacting gene pairs appear in the most and least similar 5% of pairs; a threshold-free recall metric that directly tests whether the embedding space reflects biological relational structure. All methods in Table 1, including NOVA, are evaluated on RxRx3-core in a zero-shot setting: no method is trained on RxRx3-core images or experimental conditions.

B NOVA Training Algorithm

Algorithm 1 outlines a single training step of NOVA. For each mini-batch, images are first projected through a channel-agnostic convolutional block and encoded into a spatial latent map. A perturbation-level representation is extracted via spatial mean-pooling *prior* to quantisation: this is the representation the relational loss acts on. The encoder output is then discretised via nearest-codebook-entry assignment, decoded to reconstruct the input, and the full objective is computed. The relational term is only active when at least one positive gene pair is present in the batch; otherwise training proceeds on the reconstruction objective alone.

Channel-agnostic projection. Each fluorescence channel is projected independently through a shared convolutional block ($1 \to 8 \to 1$ channels, group normalisation, ReLU) before averaging. This enables the same encoder to process any number of fluorescence channels without architectural modification.

Pre-quantisation relational loss. The relational loss operates on $r_i = \text{MeanPool}(h_i)$: the spatially pooled encoder output *before* the codebook step. This preserves gradient flow directly to the encoder, ensuring the relational signal shapes the continuous manifold rather than operating solely on discrete codes.

Codebook update. The codebook Z is updated via exponential moving average rather than gradient descent. For each entry z_k, running estimates of assignment count N_k and centroid m_k are maintained across batches:

$$N_k \leftarrow \gamma N_k + (1-\gamma)\, n_k, \tag{11}$$

$$m_k \leftarrow \gamma m_k + (1-\gamma) \sum_{i\,:\, z_q = z_k} h_i, \tag{12}$$

$$z_k \leftarrow m_k / N_k, \tag{13}$$

where n_k is the per-batch assignment count and $\gamma = 0.99$. EMA updates stabilise the codebook under the relational loss, preventing the attractive objective from destabilising prototype assignments.

Algorithm 1 NOVA: single training step

Mini-batch $\mathcal{B} = \{(x_i, g_i)\}_{i=1}^{B}$ of images and perturbation labels, biological graph $\mathcal{G} = (V, E)$, encoder E, decoder G, codebook Z, channel projection f_{proj} Updated parameters of E, G, f_{proj}; updated codebook Z

Step 1: channel-agnostic projection and encoding

$\tilde{x}_i \leftarrow \frac{1}{C}\sum_{c=1}^{C} f_{\text{proj}}(x_{i,c})$ for each sample i

$h_i \leftarrow E(\tilde{x}_i) \in \mathbb{R}^{D \times H' \times W'}$ for each sample i

Step 2: pre-quantisation perturbation embedding

$r_i \leftarrow \text{MeanPool}(h_i) \in \mathbb{R}^D$ for each sample i

Step 3: vector quantisation and reconstruction

$z_q \leftarrow \arg\min_{z_k \in Z} \|h_i - z_k\|_2$ (applied at each spatial location)

$\hat{x}_i \leftarrow G(z_q)$ for each sample i

Step 4: reconstruction losses

Compute $\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_i \|x_i - \hat{x}_i\|_1$, $\mathcal{L}_{\text{perc}} = \frac{1}{B}\sum_i \text{LPIPS}(x_i, \hat{x}_i)$, $\mathcal{L}_{\text{commit}}$ (Eq. 7)

Step 5: graph-regularised relational loss

$E_b \leftarrow \{(i,j) : i < j,\ (g_i, g_j) \in E\}$ (positive pairs in this batch)

$$\mathcal{L}_{\text{rel}} \leftarrow \begin{cases} \frac{1}{|E_b|}\sum_{(i,j)\in E_b}(1 - \cos(r_i, r_j)) & \text{if } |E_b| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Step 6: total objective and parameter update

$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}} + \lambda_{\text{rel}}^{\text{eff}}\mathcal{L}_{\text{rel}}$

Update E, G, f_{proj} via $\nabla\mathcal{L}_{\text{total}}$

Update Z via EMA (see below)

Relational loss scaling. The raw cosine distance loss operates in $[0, 2]$ and is typically two to three orders of magnitude smaller than reconstruction losses at initialisation. A fixed scaling factor of 10^3 is applied before multiplying by $\lambda_{\text{rel}} = 0.05$, yielding $\lambda_{\text{rel}}^{\text{eff}} = 50$. This calibration matches the relational term to the order of magnitude of $\mathcal{L}_{\text{rec}}$ at the start of training and is held fixed across all prior configurations.

C Model Hyperparameters

Table 3: Key hyperparameters for NOVA training.

Parameter	Value
Image size	256×256
Patch size	16×16
Channels	3 (Hoechst, α-tubulin, F-actin)
Encoder/Decoder layers	6
Attention heads	8
Embedding dimension (D)	512
Codebook size (K)	2048
Batch size (global)	512
Optimizer	AdamW ($\beta_1 = 0.5$, $\beta_2 = 0.9$)
Base learning rate	4.8×10^{-4}
Learning rate scaling	Linear with batch size
Gradient clipping	1.0 (max norm)
Loss weights	$\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{commit}} = 0.25$, $\lambda_{\text{rel}}^{\text{eff}} = 50$
Relational loss scaling	10^3 (applied before weighting)
Training Time	$\sim$80 GPU hours (converged in $<$100 epochs)
Precision	FP32

C.1 Additional Implementation Notes

- **Framework:** All models were implemented in PyTorch Lightning (v2.x) for modularity and distributed training.
- **Data augmentation:** Only center-cropping and random horizontal flips were used to preserve biological structure.
- **Channel projection:** Each fluorescence channel is projected independently via a $1 \to 8 \to 1$ convolutional block with group normalization and ReLU.
- **Vector quantization:** The codebook is updated using exponential moving average (EMA) with decay $\gamma = 0.99$.
- **Relational graph:** The interaction graph is loaded at initialisation and stored as a Python dictionary mapping gene IDs to sets of neighbours.
- **Batch sampling:** Relational loss is computed only when at least one positive gene pair is present in the batch; otherwise, it is skipped.
- **Checkpointing:** Best models are selected based on validation total loss, with early stopping patience of 50 epochs.
- **Hardware:** Training was performed on 8 NVIDIA H100 GPUs (80GB each) using DDP (Distributed Data Parallel) strategy.
- **Reproducibility:** All random seeds (PyTorch, NumPy, Python) were fixed at the start of each run.

D Zero-shot Recovery of Biological Relationships

To validate that the latent representations learned by NOVA to capture intrinsic biological logic, we conducted a zero-shot recovery task using the RxRx3-core benchmark [Kraus et al., 2025]. This evaluation measures the alignment between the model's embedding space and curated functional hierarchies, such as shared signalling pathways or co-membership in protein complexes.

We utilised five relational datasets to define "ground truth" biological relationships: StringDB [Szklarczyk et al., 2020], HuMAP [Drew et al., 2017] and CORUM [Tsitsiridis et al., 2022], Reactome [Gillespie et al., 2021], and SIGNOR [Perfetto et al., 2015]. For each perturbation i, we generate a mean embedding $\tilde{p}_i$ by calculating the mean of the latent vectors across all replicates, a robust aggregation method following [Kraus et al., 2024]. We then compute the cosine similarity between the aggregated embedding $\tilde{p}_i$ and all other perturbation embeddings $\tilde{p}_j$ within the manifold.

D.1 Recall Evaluation

To assess how effectively the model clusters biologically related entities at the extremes of the similarity distribution, we define Recall@k%.

$$\text{Recall@k\%} = \frac{\text{no. of known relationships recovered in top and bottom k\%}}{\text{Total no. of known relationships}} \tag{14}$$

In this framework (Recall@5%), a completely random embedding space results in a baseline recall of 10%. Values significantly exceeding this baseline indicate that the encoder has successfully mapped morphological features to underlying biological functions.

E Batch Correction of embeddings

To ensure that our morphological embeddings are invariant to experimental noise and technical artifacts (batch effects), we implemented a correction protocol for the RxRx3-core dataset.

1. We fitted a PCA kernel solely on control profiles across all experimental batches.
2. Following the transformation, we applied batch-specific standard scaling. For each discrete batch, a `StandardScaler` was fitted to the `EMPTY_Control` embeddings and subsequently used to normalise all perturbation embeddings within that batch.

F Overlap Between Training and Evaluation

Of the 707 SHEP training perturbations and the 736 RxRx3-core CRISPR knockouts, 101 genes are shared: 14.3% of the training set and 13.7% of the evaluation set. Within these 101 genes, database coverage is StringDB 97, SIGNOR 95, Reactome 87, CORUM 67 and hu.MAP 60. For the shared genes, the model has seen SHEP images of an siRNA knockdown, not the corresponding HUVEC CRISPR phenotype.

Because the interaction graph shapes the encoder during training, overlap at the level of graph edges could in principle inflate recall. To test this, we computed recall restricted to the 635 RxRx3-core genes that do not appear in the SHEP training set. None of these genes is a node in the training graph, so no training edge can exist between any pair of them (Table 4).

Table 4: **Recall on genes absent from training.** RxRx3-core recall restricted to the 635 evaluation genes not present in the SHEP training set. Macro-mean is over the four standard references.

Reference	NOVA (StringDB prior)	OpenPhenom	Δ
CORUM	0.419	0.379	$+0.040$
hu.MAP	0.466	0.430	$+0.036$
Reactome	0.165	0.225	-0.060
StringDB	0.405	0.362	$+0.043$
SIGNOR	0.263	0.250	$+0.013$
Macro-mean	0.364	0.349	$+0.015$

NOVA outperforms OpenPhenom on four of five references for genes it never saw during training. Macro-mean recall on these genes (0.364) is higher than on the full evaluation set (0.339), so the overall result does not depend on the overlapping genes. The Reactome exception is consistent with the constraint-hardness analysis in Section 4.4. Gene-level overlap counts between the training graph and each reference database are given in Table 5.

Table 5: **Gene overlap between training and evaluation.** Genes shared between the SHEP training perturbations and the RxRx3-core CRISPR knockouts, and coverage of the shared genes in each reference database.

	Genes
SHEP training perturbations	707
RxRx3-core evaluation knockouts	736
Shared	101
% of training set	14.3%
% of evaluation set	13.7%
Evaluation genes absent from training	635
Shared genes covered by reference	
StringDB	97
SIGNOR	95
Reactome	87
CORUM	67
hu.MAP	60

G Relational Prior Characteristics

Table 6 summarises the size, semantic type, constraint hardness, and cross-database transfer performance of each relational prior used in Section 4.4.

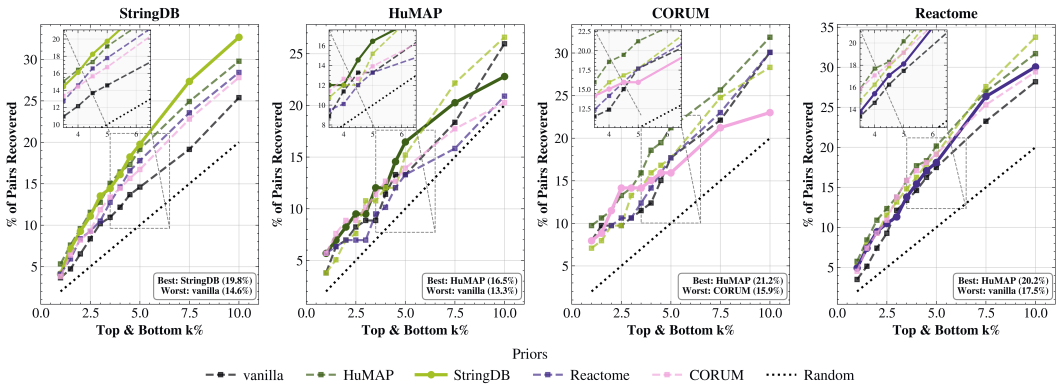


Figure 6: **Recovery of biological relationships across diverse relational priors on SHEP dataset.** Curves represent the percentage of known positive gene-pair relationships recovered within the most extreme (top and bottom) $k\%$ of the cosine similarity distribution. Panels compare performance across five distinct priors: StringDB, HuMAP, CORUM, and Reactome on the SHEP dataset. The dashed line denotes random performance (10% recall at $k = 10\%$). Insets provide a high-resolution view of the $4\% - 6\%$ similarity range. StringDB and HuMAP-guided models consistently exhibit superior relational recall compared to the vanilla baseline, demonstrating that softer, large-scale functional priors better align the latent manifold with established biological hierarchies.

Table 6: **Relational prior characteristics and transfer performance.** Size denotes the number of gene-pair relationships. Transfer Δ indicates mean recall change relative to vanilla (NOVA-no-prior) baseline on non-matching references. Soft constraints with large coverage yield best generalisation.

Prior	**Size**	**Semantics**	**Hardness**	**TransferΔ**
StringDB	$58k$	Functional interac.	Soft	**+11.37%**
HuMAP	$58k$	Protein complexes	Medium	+11.17%
CORUM	$38k$	Protein complexes	Med–Hard	+1.3%
Reactome	$19k$	Pathway membership	Hard	−3.40%
SIGNOR	$12k$	Causal signalling	Hard	−4.93%
Combined	$184k$	All priors	Mixed	−8.57%

H Ablation: Discrete versus Continuous Bottleneck

To isolate the contribution of the vector-quantised bottleneck, we trained a continuous masked autoencoder (MAE) at matched capacity (19M parameters) on the same SHEP corpus, with the same channel-agnostic projection, the same pooled representation r_i and the same relational objective at the same effective weight ($\lambda_{\text{rel}}^{\text{eff}} = 50$). It differs from NOVA only in having no codebook, quantisation or commitment term. Together with NOVA-no-prior, this gives a 2×2 comparison over {discrete, continuous} $\times$ {prior, no prior} (Table 7).

Table 7: **Discrete versus continuous bottleneck.** Recall on RxRx3-core with and without the StringDB relational prior. Chance recall is 0.10.

	Continuous (MAE)			Discrete (VQ)		
Reference	No prior	Prior	Δ	No prior	Prior	Δ
CORUM	0.311	0.145	−53.5%	0.377	0.400	+6.1%
hu.MAP	0.325	0.143	−55.9%	0.426	0.463	+8.7%
Reactome	0.156	0.114	−27.0%	0.152	0.148	−2.6%
StringDB	0.279	0.128	−53.9%	0.316	0.346	+9.5%
Macro-mean	0.268	0.133	−50.5%	0.318	0.339	+6.8%

The same objective moves the two architectures in opposite directions. With quantisation, the prior improves macro-mean recall from 0.318 to 0.339. Without it, recall falls from 0.268 to 0.133, within 0.03 of chance on every reference.

We attribute this to the form of the relational loss. $\mathcal{L}_{\text{rel}}$ is purely attractive, and in a continuous latent space it can be satisfied by contracting the pooled representation globally. This costs little, because r_i is a pooled statistic and not itself on the reconstruction path: the decoder reads the full spatial map. Quantisation removes this route. The commitment term and EMA codebook update tie encoder outputs to a finite set of prototypes, so global contraction collapses codebook utilisation and is penalised through reconstruction. The codebook therefore acts as a variance floor, and the relational gradient can only reassign perturbations between prototypes.

The relational weight was held at NOVA's calibrated value and not re-tuned for the MAE. We therefore do not claim that no continuous model could benefit from the prior under any weighting; only that, at the operating point where the prior helps the quantised model, the two architectures do not behave alike.

I Training Stability Across Seeds

Because the attractive relational loss admits a contraction mode (Appendix H), we report recall across three independent training seeds for every prior configuration (Table 8).

Table 8: **Recall across training seeds.** Mean $\pm$ standard deviation (%) on RxRx3-core over three seeds.

Training prior	CORUM (↑)	hu.MAP (↑)	Reactome (↑)	StringDB (↑)
StringDB	40.7 ± 1.8	45.8 ± 0.7	13.5 ± 0.7	33.7 ± 0.3
hu.MAP	36.6 ± 1.2	42.1 ± 0.8	12.2 ± 1.4	32.4 ± 0.5
CORUM	39.1 ± 0.8	43.4 ± 0.1	11.2 ± 1.0	33.6 ± 0.0
Reactome	35.2 ± 0.2	39.2 ± 0.6	12.3 ± 2.5	31.4 ± 1.0
Combined	34.7 ± 5.7	39.8 ± 6.8	14.1 ± 0.4	30.6 ± 2.0
No prior	37.1 ± 0.5	42.1 ± 0.4	11.5 ± 1.4	31.7 ± 0.3

The gain from the StringDB prior over the no-prior baseline exceeds the pooled standard deviation on all four references ($4.7\times$ on StringDB, $4.6\times$ on hu.MAP, $1.9\times$ on CORUM and $1.3\times$ on Reactome). The individual priors and the no-prior baseline show comparable, low variance. The combined prior has the largest spread (up to 6.8% on hu.MAP), consistent with the ensemble effect in Section 4.4: mixing soft and hard constraints reduces both mean recall and training stability.