MISLOCUS: A SINGLE-CELL IMAGING BENCHMARK FOR PROTEIN MISLOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Genetic variants that alter the location of proteins within cells can disrupt protein function and contribute to disease. Microscopy can detect protein mislocalization, but benchmarking image representations to automate this task requires data for reference (healthy) proteins plus protein variants, with a replicate structure that separates biological signal from technical variation. We introduce **MisLocus**, a single-cell imaging benchmark for variant-induced protein mislocalization. It contains $\sim$3.3 million single-cell images, covering 250 reference proteins and 1,291 variants, with experimental replicates spanning imaging wells, plates, and batches. We evaluate nine image representations from CellProfiler, Cytoself, SubCell, and MorphEm, including handcrafted features, pretrained learned representations, and models retrained or fine-tuned on MisLocus. Representations are assessed along three complementary axes: detecting variant-induced mislocalizations, predicting each variant's clinical pathogenicity based on mislocalization phenotype, and discriminating annotated reference protein localization patterns. We find no single representation dominates all evaluation axes. Engineered features, paired with supervised models trained to predict localization shifts, detect variant-induced mislocalization competitively with the best learned representations, whereas learned representations better organize proteins according to shared localization patterns. MisLocus establishes a controlled and reproducible framework for benchmarking single-cell microscopy representations on protein mislocalization phenotyping.

1 INTRODUCTION

The location of a protein within the cell is typically critical for proper cellular function: it determines its immediate environment, the interaction partners it can access, and the processes it can influence (Sigaeva et al., 2026). Genetic variants ("alleles") in the protein-coding regions of genes ("coding variants") can change the native spatial distribution of the original protein, producing a directly observable phenotype of mislocalization. Such mislocalization has been reported across cancer (Wang & Li, 2014), neurodegeneration (Suk & Rousseaux, 2020), autoimmunity (Bawadekar et al., 2015; Gugliesi et al., 2013), and inherited disorders, including variant-driven mislocalization in dilated cardiomyopathy and cystic fibrosis (Dornan & Simpson, 2024; Kartner et al., 1992; Kornienko et al., 2023; Lacoste et al., 2024), and several successful drugs reverse the incorrect localizations, motivating efforts to discover more (Hung & Link, 2011; Wang & Li, 2014; Van Goor et al., 2011).

Microscopy imaging makes mislocalization phenotypes measurable at scale by fluorescently labeling the protein to capture the shift in the subcellular distribution of a variant protein relative to its healthy ("reference") counterpart. Recent studies show that mislocalization occurs often in disease. For example, Lacoste et al. (2024) profiled $3{,}448$ genetic variants across $1{,}269$ proteins and reported mislocalization in 16% of pathogenic (disease-causing) variants. Existing analysis of images relies on manual observation or handcrafted cell-morphology profiles, using profile distances or supervised classifiers trained to distinguish each variant protein from its reference (Lacoste et al., 2024; Pendyala et al., 2025; Tejura et al., 2026). Meanwhile, microscopy image analysis has moved towards learned representations, including self-supervised encoders and foundation models that embed spatial and morphological structure at scale (Agrawal et al., 2026; Gupta et al., 2025; Kobayashi et al., 2022). However, these models have mainly been used for categorizing healthy protein localizations or general morphological profiling tasks, leaving unresolved whether they better capture variant-induced localization changes than handcrafted features.

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

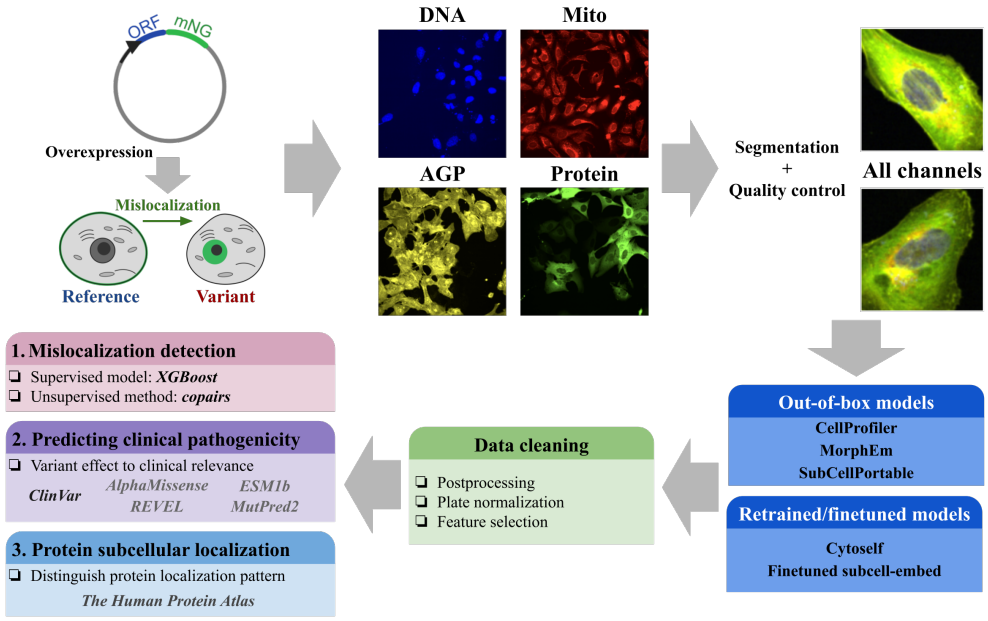

Figure 1: **Overview of the MisLocus benchmark for detecting variant-induced protein mislocalization from multi-channel cell images.** MisLocus comprises mNeonGreen-tagged reference or variant proteins imaged with the GFP protein channel and three cellular-context channels (DNA, Mito and AGP). After segmentation and quality control, high-quality single-cell image crops containing all channels are used as input to several image-representation models. We benchmark out-of-the-box representations, and retrained or fine-tuned deep learning models. Extracted image features are processed through a standardized data-cleaning pipeline involving postprocessing, plate normalization, and feature selection. The resulting representations are evaluated in three complementary settings: mislocalization detection, where supervised XGBoost classifiers and unsupervised copairs analyses distinguish reference and variant profiles; predicting clinical pathogenicity, where imaging phenotypes are compared against clinical variant-effect classfication and existing AI variant-effect predictors, as well as distinguish different protein localization patterns.

Variant-induced mislocalization poses a distinct representation-learning problem from protein localization. The latter—categorizing the absolute location of a protein, for example in the nucleus, cytoplasm, or membrane—is typically easier than comparing reference and variant to each other, which often display highly similar localization. These subtle changes can be confounded by technical variation, including well-position effects (systematic biases tied to a cell's position on the multi-well imaging plate). Therefore, evaluating each image representation's capability to detect protein mislocalization requires paired reference-variant imaging, replication across wells, plates, and batches, and standardized tasks distinguishing localization shifts from technical variation.

Existing resources do not jointly satisfy the above requirements. Large-scale variant-imaging studies demonstrate the relevance of mislocalization, but we cannot compare alternative image representations due to their lack of controlled-benchmark structure (Lacoste et al., 2024). Reference-localization resources such as OpenCell (Cho et al., 2021) and the Human Protein Atlas (Thul et al., 2017; Uhlén et al., 2015) support localization modeling but lack images for genetic variants. General evaluations of learned microscopy representations assess localization or transfer across imaging settings, but do not directly test whether embeddings capture spatial changes (Agrawal et al., 2026; Gupta et al., 2025; Kobayashi et al., 2022; Moutakanni et al., 2025).

We introduce **MisLocus**, a controlled single-cell imaging benchmark for variant-induced protein mislocalization derived from the public VarChAMP image collection (Tejura et al., 2026). Our contribution is the quality-controlled single-cell image crops, harmonized annotations and representations, replicate-aware splits, and evaluation harness rather than the underlying image acquisition. MisLocus contains 3,332,309 high-quality single-cell images across protein and morphology channels, including 3,002,837 images from the biological cohort of 250 reference proteins and 1,291 protein variants

after excluding proteins used as assay controls. Its paired reference-variant design and replicated measurements enable standardized evaluation of image representations along three complementary axes: detecting variant-reference localization shifts, testing mislocalization phenotype as a predictor for a variant's pathogenicity, and preserving annotated reference-localization structure.

Within this framework, we benchmark nine feature sets spanning four representation families: handcrafted CellProfiler descriptors (Stirling et al., 2021); the VQ-VAE-based Cytoself encoder (Kobayashi et al., 2022); the ViT-based SubCell model under frozen and fine-tuned regimes (Gupta et al., 2025); and MorphEm, a channel-agnostic Vision Transformer pretrained on heterogeneous multi-study microscopy data (Agrawal et al., 2026).

MisLocus reveals that the three evaluation axes are related but not redundant. Imaging phenotype scores are often correlated across representations, indicating shared mislocalization signal, but different representations are sensitive to different components of the phenotype. CellProfiler-based detection relies heavily on intensity-related measurements, which can entangle protein abundance, imaging artifacts, and localization changes, while retaining direct interpretability for describing shifts in spatial distribution. In contrast, learned representations better preserve localization structure and more effectively group proteins with shared subcellular patterns. Together, these results establish MisLocus as a benchmark for testing whether microscopy representations capture the impact of genetic variants on localization, and as a resource for developing more robust and interpretable methods for variant-effect characterization using imaging.

2 Related Work and Background

Image representations for protein localization. Protein localization heavily influences protein function, thus measuring it at scale has been a long-standing goal. Early large-scale efforts relied on expert visual inspection or manually defined localization categories, which are effective for assigning proteins to coarse compartments but are poorly suited for detecting subtle or heterogeneous changes in spatial distribution across single cells. Modern image representations address this limitation by learning features directly from microscopy images, reducing dependence on manual annotations and enabling fine-grained comparisons between cellular states.

Existing representations for protein localization can be broadly grouped into two families. *Protein-localization-specific models* are trained to capture the spatial distribution of tagged proteins: Cytoself learns self-supervised representations of tagged proteins (alongside a nucleus image) using a VQ-VAE architecture and a protein-identity pretext task (Kobayashi et al., 2022), while SubCell uses Vision Transformers trained on Human Protein Atlas images using the protein channel together with DNA, microtubule, and endoplasmic-reticulum channels (Gupta et al., 2025). *General-purpose microscopy representations* aim to capture broadly transferable cellular morphology across assays, staining panels, and biological tasks. CellProfiler extracts hand-engineered, interpretable per-cell measurements of intensity, texture, shape, and spatial organization and remains a strong baseline for image-based profiling experiments such as Cell Painting (Bray et al., 2016; Chandrasekaran et al., 2024; Stirling et al., 2021). By contrast, channel-agnostic foundation models, such as MorphEm, are pretrained on heterogeneous microscopy collections to support diverse biological tasks (Agrawal et al., 2026).

These methods provide complementary views of cellular images: handcrafted measurements are interpretable, protein-localization encoders are specialized for subcellular spatial patterns, and foundation models target broader transfer across biological imaging tasks. All have been primarily evaluated on detecting reference-protein localization, where they show decent performance (Gupta et al., 2025; Isselmann et al., 2026; Kobayashi et al., 2022). However, it remains unclear how these representations perform when detecting whether a genetic variant has shifted the distribution of that protein relative to its corresponding reference protein.

Variant-induced mislocalization as a representation benchmark. Genetic variants often produce measurable and disease-relevant changes in protein localization; thus far quantification is by manual observation, CellProfiler-derived features, and supervised classifiers to detect image-level differences between reference and variant proteins (Lacoste et al., 2024; Pendyala et al., 2025; Tejura et al., 2026). A comparison of representation learning strategies has not been performed on this task, which is more challenging than classifying the localization of reference proteins because of the often-subtle relative spatial shift between a variant protein and its paired reference. This gap motivates a benchmark

for this task while explicitly distinguishing variant-associated localization changes from technical variation introduced by well position, plate identity, and experimental batch.

3 MISLOCUS DATASET AND EVALUATION

We introduce MisLocus, a standardized single-cell imaging benchmark for evaluating whether cell image representations can detect variant-induced changes in protein localization. MisLocus is derived from six U2OS imaging batches in the existing public VarChAMP collection (`cpg0020-varchamp`) (Tejura et al., 2026). In the source assay, cells overexpress an mNeonGreen-tagged reference or variant protein and were imaged using an adapted multiplexed Cell Painting protocol, capturing the tagged protein together with channels for nuclear DNA, actin/Golgi/plasma-membrane structure, and mitochondria. We newly contribute the quality-controlled crop cohort, harmonized annotations and representations, replicate-aware splits, and evaluation harness; the underlying image acquisition and CellProfiler measurements are inherited from the upstream release.

MisLocus is a benchmark-ready resource, enabling reproducible evaluation through released crops and precomputed features, downstream feature preprocessing, replicate-aware data splits for model training and testing, and interfaces for extracting additional cell image representations. Released data products are keyed to a common quality-controlled single-cell manifest. Representation-specific preprocessing can remove a small number of ineligible rows, so each result reports its evaluable comparison set rather than assuming identical denominators across methods.

3.1 DATASET OVERVIEW

The benchmark uses six imaging batches, organized as three biological replicate batch pairs, each screened twice with identical experimental setups. Each batch contains arrayed 384-well plates, where each well corresponds to a single reference protein or variant thereof. The imaging hierarchy is batch $\rightarrow$ plate $\rightarrow$ well $\rightarrow$ site $\rightarrow$ single cell; MisLocus provides metadata linking these layers plus the identity of the reference or variant in each well. We use this hierarchy to define both the single-cell representation units and the replicate-aware evaluation splits.

The manifest indexes the QC-filtered single-cell crops and provides the identifiers used to join each representation's profiles. The crops retain raw pixel intensities; "raw" here describes the intensities, not a separate pre-QC crop cohort. This design gives every representation the same starting cohort and quality-control decisions, while allowing documented representation-specific preprocessing and row filtering. We extract one four-color-channel crop per retained cell, centered on the segmented nucleus. The full manifest contains 3,332,309 QC-passed cells. Detailed single-cell quality-control thresholds, crop extraction rules, and batch-level counts are provided (Appendix A.1).

3.2 BENCHMARK STRUCTURE

MisLocus is organized around paired reference proteins and their variants. Across the designed platemaps, the original VarChAMP dataset covers 1,623 tested proteins (excluding proteins used as experimental controls), including 281 reference proteins and 1,342 variants spanning 310 proteins (some variants lack a paired reference protein and vice versa, because their partner was dropped due to sequence truncation or sample contamination; these cells are useful for representation learning but are excluded from paired mislocalization evaluation). After imaging and single-cell quality control, the released crop manifest retains 1,541 tested proteins, including 250 reference proteins and 1,291 variants, across 282 genes. Of these, 125 proteins retain both a reference and at least one variant, yielding 1,055 paired reference-variant pairs for mislocalization analysis (Table 3).

3.3 REPLICATE STRUCTURE AND BENCHMARK SPLITS

MisLocus is designed to separate biological signal from technical variation. Each unique platemap is replicated across physical plates, denoted by technical replicates T1-T4. These technical replicates contain the same proteins in the same well positions but correspond to independent (though in theory identical) physical preparations, staining, and imaging processes. On each imaging plate, there are always six controls, consisting of five reference proteins and one variant protein, each screened across four different well positions; these are used to assess well-position effects and technical variation.

For representation models trained on MisLocus images, we use a deterministic technical-replicate split: T1 and T2 are used for training, T3 for validation, and T4 for testing. This split ensures that the same protein identities are observed across splits, while cells from the same physical plate never appear in both training and test sets. As a result, held-out performance measures robustness to technical replicate variation rather than generalization to unseen protein identities. Benchmark evaluation uses fixed folds derived from the same manifest and replicate structure. Representations that do not require re-training or fine-tuning, such as CellProfiler profiles or frozen pretrained image encoders, are not trained on images but are still evaluated using the same downstream cells, controls, and aggregation rules. The full split specification by batch and platemap is provided in Appendix A.3.2.

3.4 Clinical and computational variant annotation

We annotated each variant using two complementary external resources: ClinVar (Landrum et al., 2018) (January 2025 release) and dbNSFP (Liu et al., 2020) (v5.0a). ClinVar provides clinician- and expert-submitted assertions of clinical significance for human genetic variants. For each assayed variant, we queried ClinVar and aggregated these assertions into binary pathogenicity labels by merging pathogenic and likely pathogenic variants into the pathogenic category and similarly for the benign variants. To complement these clinically curated labels with model-based functional priors, we used dbNSFP to extract per-variant scores from four widely used computational variant-effect predictors (AlphaMissense (Cheng et al., 2023), REVEL (Ioannidis et al., 2016), MutPred2 (Pejaver et al., 2020), and ESM1b (Brandes et al., 2023)). For benchmarking we use each predictor's binary pathogenic/benign call (the continuous scores are also released for completeness). This allows correlation analyses between imaging-derived phenotype scores and pathogenicity predictions, enabling us to evaluate whether microscopy-based representations capture consistent variant effects.

3.5 Evaluation tasks

MisLocus evaluates image representations along three complementary axes: **(1) reference–variant mislocalization detection**, testing whether variant proteins separate from their paired references; **(2) clinical agreement**, testing whether per-variant scores track ClinVar annotations and sequence-based variant effect predictors; and **(3) reference-localization structure**, testing whether reference proteins organize by annotated subcellular localization. Performance on these axes can diverge because they probe different biological tasks and also use overlapping but non-identical subsets of the shared single-cell manifest. Implementation details, including model settings, control-null construction, and per-task cell selection, are provided (Appendix A.3.1).

(1) Reference–variant mislocalization detection. The inventory contains 1,055 potential paired reference–variant comparisons. After task-specific availability and classifier-quality filters, the headline table evaluates 931 pairs eligible for supervised XGBoost and 993 pairs eligible for unsupervised retrieval across all the representations. We evaluate this task using both supervised and unsupervised imaging phenotype scores. In the *supervised* setting, we train a separate XGBoost classifier for each evaluable reference–variant pair to distinguish single cells expressing the variant protein from cells expressing the paired reference protein, and use the held-out AUROC as the supervised imaging phenotype score. In the *unsupervised* setting, we score each evaluable pair using the copairs normalized mean average precision (mAP) metric (Kalinin et al., 2025): variant images are used as queries, images are ranked by cosine similarity in the embedding space, and the score measures whether same-variant images are retrieved ahead of paired-reference images. We report normalized mAP, where 0 corresponds to the random-retrieval baseline and 1 to perfect separation. For hit-rate analyses, both supervised and unsupervised scores are calibrated against the 95th percentile of the corresponding batch-, representation-, and channel-matched control-null distribution. Unsupervised hits must additionally pass statistical significance testing of the `mAP_norm` retrieval score after Benjamini–Hochberg (BH) correction. Details of eligibility filters, null construction, replicate aggregation, and hit calling are provided (Appendix A.3.3).

(2) Clinical agreement. This axis evaluates whether per-variant mislocalization scores reflect clinical annotations for that variant in ClinVar (Landrum et al., 2018) and sequence-based variant effect predictors. Higher imaging phenotype scores should be more likely to be annotated or

Table 1: **Mislocalization hit rates and ClinVar pathogenic enrichment across representation settings.** Each row reports one representation and channel setting. Results include per-variant phenotype scores for supervised XGBoost AUROC (XGB) and for unsupervised copairs `mAP_norm` scoring (PA). *Hits/Hit rate* count variants passing calibrated hit criteria among XGB-eligible or PA-eligible pairs. Precision, F_1, and Fisher p_{BH} compare binary imaging hit calls with ClinVar variant classifications of pathogenic or benign. Full eligibility and hit-calling procedures are provided (Appendix A.3.3).

		Mislocalization hits & Imaging phenotype scores							Mislocalization hits vs ClinVar classifications					
		XGB				PA			XGB			PA		
Representation	Channel	Hits	Hit rate	AUROC	Null p_{95}	Hits	Hit rate	$\mathrm{mAP}_{\mathrm{norm}}$	Precision	F_1	Fisher p_{BH}	Precision	F_1	Fisher p_{BH}
SC ft/ViT	EMBED	234	25.1%	0.809	0.940	357	36.0%	0.381	84.5%	0.488	0.017	85.6%	0.577	1e-4
SC ft/MAE	EMBED	246	26.4%	0.802	0.934	338	34.0%	0.385	78.6%	0.474	0.285	87.2%	0.552	1e-4
SC rbg/MAE	EMBED	201	21.6%	0.740	0.909	319	32.1%	0.301	84.3%	0.417	0.034	87.7%	0.512	1e-4
SC rbg/ViT	EMBED	173	18.6%	0.740	0.918	309	31.1%	0.284	82.4%	0.354	0.126	86.4%	0.469	1e-3
MorphEM	ALL	146	15.7%	0.759	0.948	183	18.4%	0.170	76.7%	0.278	0.497	83.7%	0.327	0.050
MorphEM	GFP	321	34.5%	0.713	0.824	277	27.9%	0.298	88.2%	0.563	1e-4	89.0%	0.517	1e-5
Cytoself	combined	348	37.4%	0.695	0.789	203	20.4%	0.134	85.7%	0.584	0.001	89.4%	0.375	1e-3
CellProfiler	ALL	274	29.4%	0.796	0.952	36	3.6%	0.003	76.2%	0.473	0.497	94.1%	0.086	0.056
CellProfiler	GFP	190	20.4%	0.798	0.957	48	4.8%	0.036	85.4%	0.397	0.028	91.3%	0.111	0.056

predicted as pathogenic (disease-causing) than benign (harmless). Disagreement on variants with strong imaging phenotypes points to inconsequential localization changes, a false negative for the sequence-only models, or potential inaccurate annotations in ClinVar.

(3) Reference-localization structure. This axis evaluates whether a representation organizes reference proteins by annotated subcellular localization. It is computed on reference cells only, pooled across batches and aggregated to one consensus profile per protein, and uses Human Protein Atlas (HPA) (Thul et al., 2017) location annotations as coarse compartment labels. Proteins sharing a location label should be closer in feature space than proteins that do not. We use unsupervised retrieval rather than a supervised HPA-label classifier because location annotations are sparse and multi-label per protein, leaving too few clean per-class examples to train per-location classifiers. This axis tests whether the representation captures compartment-level spatial distributions of different proteins.

4 Results

We evaluated image representations along the three complementary axes described above. Unless otherwise stated, results are computed on the held-out T4 plate within each imaging batch and then aggregated across biological-replicate batch pairs as described in section 3.5.

4.1 Supervised and unsupervised scoring reveal distinct representation strengths

We first quantified each representation's sensitivity to variant-induced mislocalization. For every reference–variant pair, we computed two imaging phenotype scores: held-out AUROC from a supervised XGBoost classifier and `mAP_norm` from the unsupervised copairs retrieval framework, which captures phenotypic activity (PA). These scores capture complementary properties of a representation: supervised discriminability versus native metric-space organization. Continuous scores were converted into binary mislocalization hit calls using calibrated control-null thresholds (Appendix A.3.3). To benchmark against manual annotations, we compared both imaging phenotype scores using mislocalization labels from Lacoste et al. (2024) for the small subset of overlapping variants; the imaging phenotype scores are consistently higher for mislocalized variants across representations (Supplementary Fig. 8).

Under supervised XGBoost scoring, hit rates vary substantially across representations but do not simply track the mean AUROC. Cytoself (37.4%) and MorphEM-GFP (34.5%) yield the highest supervised hit rates, followed by CellProfiler-ALL (29.4%) and the fine-tuned SubCell models (25.1–26.4%). However, several representations with strong mean AUROC have lower hit rates because their matched control-null thresholds are also high: for example, SC ft/ViT has the highest

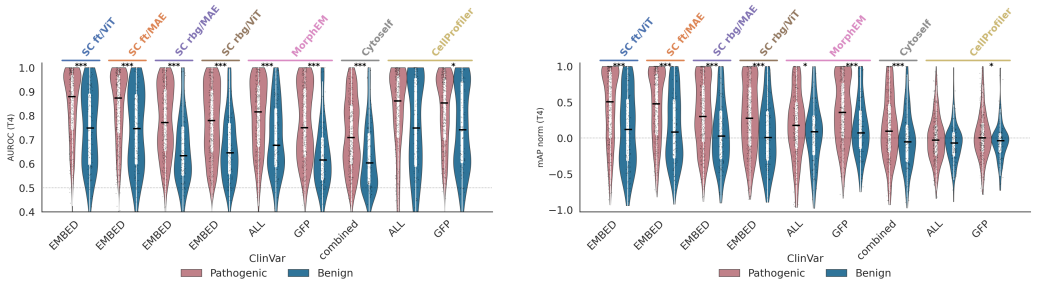

Figure 2: **Per-variant imaging phenotype scores by ClinVar label.** Violin distributions compare pathogenic and benign variants across representation/channel settings for supervised held-out XGBoost AUROC (left) and unsupervised copairs `mAP_norm` (right). Stars mark BH-corrected Mann–Whitney U tests (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

mean AUROC (0.809) but a high average null threshold (0.940), while Cytoself has a lower mean AUROC (0.695) but a lower null threshold (0.789) (Table 1). We believe supervised hit calls are the more relevant metric than mean AUROC because they reflect both variant–reference separability and each representation's sensitivity to technical well-to-well variation. Continuous AUROC scores are strongly correlated across representations (Appendix Fig. 7), while calibrated hit rates diverge after control-null thresholding. For CellProfiler, XGBoost feature importance analysis further shows that high control-null AUROC is largely driven by GFP intensity-related features, indicating that some supervised CellProfiler signal aligns with technical intensity variation rather than localization-specific morphology (Supplementary Fig. 6; Appendix A.4.3).

The unsupervised `mAP_norm` results show a clearer advantage for learned embeddings. All four SubCell settings produce the highest unsupervised hit rates (31.1–36.0%), with fine-tuned SubCell also achieving the highest mean `mAP_norm` scores (0.381–0.385). MorphEM-GFP is competitive (27.9% hits; `mAP_norm`=0.298), whereas Cytoself drops to 20.4% hits and `mAP_norm`=0.134. In contrast, CellProfiler performs poorly under unsupervised retrieval, with only 3.6–4.8% hits and near-zero mean `mAP_norm`. This supervised–unsupervised asymmetry suggests that CellProfiler contains discriminative variant information that XGBoost can extract, but that this signal is poorly aligned with the full CellProfiler distance geometry used for nearest-neighbor retrieval as in `mAP_norm`.

Consistent with this interpretation, restricting CellProfiler to top XGBoost-ranked features recovers much of the lost unsupervised signal (Supplementary Fig. 9). Thus, the weak unsupervised CellProfiler result does not imply an absence of variant-relevant information; rather, it indicates that this information is diluted by nuisance dimensions in the full engineered feature space. By contrast, SubCell embeddings, and to a lesser extent MorphEM-GFP, better organize variant-induced phenotypes in their native embedding geometry, enabling stronger retrieval without per-variant classifier training.

4.2 Imaging phenotype scores separate pathogenic from benign variants

ClinVar. We next tested whether imaging phenotypes align with clinical variant annotations from ClinVar. As expected, pathogenic variants generally show higher imaging phenotype scores than benign variants across most learned representations, for both supervised XGBoost AUROC and unsupervised `mAP_norm`, with CellProfiler-ALL as the exception (Figure 2). Table 1 shows similar trends using binary hit calls to evaluate precision, F_1, and Fisher enrichment against the binary classifications. CellProfiler again reveals a supervised–unsupervised asymmetry: pathogenic and benign variants separate under XGBoost, indicating that engineered features contain clinically relevant signal, but the very limited number of unsupervised mislocalization hits leads to a very low F_1 score when comparing with pathogenic and benign labels.

Variant effect predictors. On the same pathogenic-versus-benign task, we tested predictor-derived labels from AlphaMissense, ESM-1b, REVEL, and MutPred2. Under supervised XGBoost scoring, imaging hit calls are enriched for predicted pathogenic variants across most representation–predictor combinations. Cytoself and MorphEM-GFP show the strongest overall agreement, while fine-tuned

Table 2: **Agreement between imaging hit calls and four sequence-based variant-effect predictors.** Rows are representation/channel pairs; columns report supervised XGBoost (XGB) and unsupervised copairs (PA) results. For each binarized predictor, F_1 measures agreement with pathogenic labels and p_{BH} is the BH-corrected one-sided Fisher enrichment p-value (capped at 10^{-8}).

		XGB								PA							
		AlphaMissense		ESM-1b		REVEL		MutPred2		AlphaMissense		ESM-1b		REVEL		MutPred2	
Representation	Channel	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}
SC ft/ViT	EMBED	0.546	1e-8	0.508	1e-6	0.485	1e-4	0.541	1e-4	**0.584**	1e-8	**0.555**	1e-5	**0.549**	1e-6	**0.598**	1e-4
SC ft/MAE	EMBED	0.524	1e-6	0.495	1e-4	0.483	1e-3	0.539	0.001	0.563	1e-8	0.531	1e-4	0.516	1e-5	0.573	1e-4
SC rbg/MAE	EMBED	0.481	1e-7	0.440	1e-5	0.407	1e-3	0.474	1e-3	0.536	1e-8	0.488	1e-4	0.445	1e-3	0.535	1e-3
SC rbg/ViT	EMBED	0.425	1e-6	0.395	1e-4	0.354	0.008	0.423	0.003	0.505	1e-7	0.460	0.002	0.421	0.012	0.509	0.002
MorphEM	ALL	0.371	1e-5	0.354	1e-4	0.282	0.058	0.374	1e-3	0.392	1e-4	0.346	0.036	0.331	0.003	0.403	1e-3
MorphEM	GFP	**0.596**	1e-8	0.569	1e-7	0.555	1e-6	0.592	1e-5	0.554	1e-8	0.507	1e-6	0.466	1e-5	0.551	1e-6
Cytoself	combined	0.595	1e-6	**0.590**	1e-6	**0.565**	1e-3	**0.596**	0.001	0.377	1e-4	0.343	0.002	0.291	0.157	0.357	0.052
CellProfiler	ALL	0.499	0.002	0.524	1e-4	0.474	0.008	0.502	0.036	0.079	0.293	0.093	0.114	0.107	0.024	0.104	0.170
CellProfiler	GFP	0.444	1e-6	0.432	1e-6	0.375	0.004	0.432	0.007	0.122	0.153	0.121	0.114	0.134	0.024	0.137	0.122

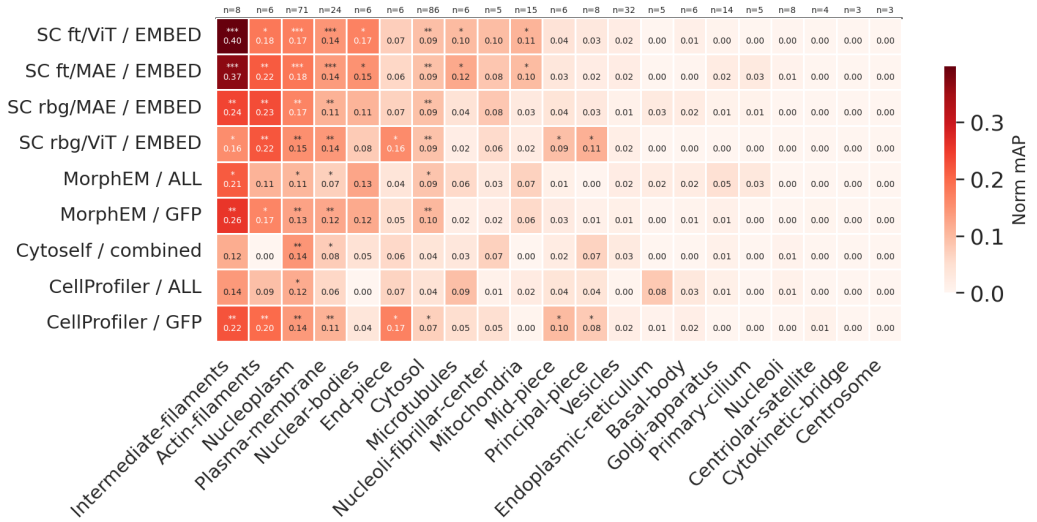

Figure 3: **Reference-protein similarity by annotated location.** Per-location normalized mAP measures whether each representation places profiles of reference proteins with the same HPA location close together. Rows are representation/channel pairs; columns are HPA locations with at least three proteins, sorted by mean signal (counts above columns). Benjamini–Hochberg-corrected significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

SubCell and CellProfiler-ALL are also competitive, indicating that supervised imaging phenotypes capture variant effects broadly concordant with sequence-based pathogenicity estimates (Table 2).

For unsupervised `mAP_norm` scoring, agreement shifts toward learned embeddings with stronger native metric geometry. Fine-tuned SubCell achieves the highest unsupervised F_1 values across all four predictors, and MorphEM-GFP also remains competitive. CellProfiler drops sharply under unsupervised scoring. Together, these results mirror the mislocalization hit-rate analysis: supervised classifiers can recover pathogenicity-relevant signal from several feature spaces, whereas unsupervised retrieval most strongly benefits from learned embeddings that organize variant-induced phenotypes directly in distance space.

4.3 Learned embeddings better preserve localization structure

As a third evaluation axis, we asked whether each representation organizes reference proteins according to annotated subcellular localization. We restricted this analysis to reference proteins because many have curated Human Protein Atlas (HPA) (Thul et al., 2017) subcellular localization labels, while variant localization is rarely annotated. After filtering HPA annotations by confidence,

we computed per-location mAP_norm, measuring whether reference proteins annotated to the same compartment are retrieved near one another in feature space. We used unsupervised retrieval rather than supervised HPA-label prediction because HPA labels describe endogenous protein localization, whereas MisLocus images are generated from an overexpression assay that can change or add localizations. In addition, HPA annotations are multi-label rather than mutually exclusive, and some compartments contain few reference proteins. A retrieval-based metric therefore provides a more appropriate test: it asks whether proteins sharing a curated localization occupy nearby regions of the representation space, without forcing each protein into a single predicted compartment class.

All representation families recover some HPA-consistent localization structure, but learned embeddings show stronger and more coherent retrieval patterns than CellProfiler features (Fig. 3). Fine-tuned SubCell embeddings give the clearest signal across several cytoskeletal, nuclear, and membrane-associated locations, with portable SubCell and MorphEM-GFP also recovering multiple major locations. CellProfiler shows lower overall mAP_norm scores, except for a few strongly distinctive localizations. Across representations, broad cytoskeletal and nuclear/plasma-membrane locations are easiest to recover, whereas smaller or more heterogeneous classes with few annotated proteins show weak retrieval. Thus, the HPA analysis provides an orthogonal check that most representations encode coarse localization information, while learned embeddings—especially fine-tuned SubCell models—better organize reference proteins by subcellular structure.

5 Discussion, Limitations, and Conclusion

MisLocus provides a unified benchmark for evaluating image representations on variant-induced protein mislocalization. By combining paired reference–variant imaging, replicate-aware control nulls, clinical annotations, and reference-protein localization labels, it supports evaluation across three complementary axes: reference–variant mislocalization detection, agreement with clinical pathogenicity annotations, and preservation of annotated localization patterns. Across these axes, learned embeddings generally provide stronger native metric geometry than engineered CellProfiler features, especially for unsupervised retrieval and reference-localization structure. However, this advantage is not uniform: CellProfiler remains competitive under supervised XGBoost scoring and provides explicit, interpretable measurements.

Several limitations define the scope of these conclusions. Available annotations are imperfect proxies for biological ground truth: ClinVar labels are incomplete and clinically biased, sequence-based predictors are not independent experimental labels, manual mislocalization annotations cover only a small overlapping subset, and HPA labels describe natural localization rather than localization in the MisLocus overexpression assay. In addition, the current benchmark evaluates performance within the available imaging cohorts, cell lines, microscope setting, and experimental design; cross-cell-line, cross-instrument, and cross-assay generalization remain future work. Finally, supervised and unsupervised scores answer different questions: XGBoost can recover discriminative signals from selected feature combinations, whereas mAP_norm retrieval tests whether phenotypes are already organized in the representation's native distance geometry.

MisLocus is therefore not intended to declare a single universally best representation, but to expose how representations behave under controlled, biologically motivated evaluation tasks. We release the curated cohort, metadata, bundled precomputed feature sets, and evaluation harness so that future representations can be evaluated under the same cell inclusion criteria, replicate structure, control-null calibration, and clinically and biologically relevant tasks. As additional imaging data, variant annotations, and representation models become available, MisLocus can be expanded to better characterize how genetic variation reshapes protein localization in cells.

AI use statement

We used generative AI for manuscript and reference editing, code-inspection and editing support, and numerical and provenance checks. The authors verified all outputs. We did not use AI for data generation, developing mathematical theory or proofs, proposing hypotheses, study and experiment design, method implementation, translation, data cleaning, qualitative analysis, or result interpretation. The authors take responsibility for all the content.

Reproducibility statement

The anonymous analysis code is available at https://anonymous.4open.science/r/MisLocus/, and the versioned dataset bundle is at https://anonymous-hf.com/a/vqlxvun682t7/. The release includes per-batch QC manifests, four-channel single-cell image crops, preprocessed feature matrices for all the tested representations, evaluation code, and curated annotations. All the released imaging profiles can be joined to the manifest. A new method can therefore compute embeddings for the released crop cohort, write them in the expected schema, and reuse the common downstream aggregation and evaluation pipeline without redefining cells, controls, or train/test splits. The source TIFFs of cell images are part of the public VarChAMP Cell Painting Gallery collection cpg0020-varchamp and are referenced by URI in the released manifests.

References

Vidit Agrawal, John Peters, Tyler N. Thompson, Mohammad Vali Sanian, Chau Pham, Nikita Moshkov, Arshad Kazi, Aditya Pillai, Jack Freeman, Byunguk Kang, Samouil L. Farhi, Ernest Fraenkel, Ron Stewart, Lassi Paavolainen, Bryan A. Plummer, and Juan C. Caicedo. CHAMMI-75: Pre-training multi-channel models with heterogeneous microscopy images. arXiv, March 2026. URL https://arxiv.org/abs/2512.20833v2. Preprint, version 2.

Mandar Bawadekar, Marco De Andrea, Marisa Gariglio, and Santo Landolfo. Mislocalization of the interferon inducible protein IFI16 by environmental insults: Implications in autoimmunity. *Cytokine & Growth Factor Reviews*, 26(2):213–219, April 2015. ISSN 13596101. doi: 10.1016/j.cytogfr.2014.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1359610114001312.

Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55 (9):1512–1522, September 2023. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-023-01465-0. URL https://www.nature.com/articles/s41588-023-01465-0.

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 2016. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2016.105. URL https://www.nature.com/articles/nprot.2016.105.

Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G. Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, Alexandr A. Kalinin, John Arevalo, Marzieh Haghighi, Juan C. Caicedo, Daniel Kuhn, Desiree Hernandez, James Berstler, Hamdah Shafqat-Abbasi, David E. Root, Susanne E. Swalley, Sakshi Garg, Shantanu Singh, and Anne E. Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, June 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02241-6. URL https://www.nature.com/articles/s41592-024-02241-6.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco California USA, August 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492, September 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adg7492. URL https://www.science.org/doi/10.1126/science.adg7492.

Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y.S. Kim,

Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loic A. Royer, Matthias Mann, and Manuel D. Leonetti. OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization. *bioRxiv*, 2021. doi: 10.1101/2021.03.29.437450. URL https://www.biorxiv.org/content/early/2021/12/08/2021.03.29.437450.

Lucy G. Dornan and Jeremy C. Simpson. Lost in translation: Illuminating protein mislocalization through high-content screening microscopy. *Cell Genomics*, 4(11):100695, November 2024. ISSN 2666979X. doi: 10.1016/j.xgen.2024.100695. URL https://linkinghub.elsevier.com/retrieve/pii/S2666979X24003240.

Francesca Gugliesi, Mandar Bawadekar, Marco De Andrea, Valentina Dell'Oste, Valeria Caneparo, Angela Tincani, Marisa Gariglio, and Santo Landolfo. Nuclear DNA Sensor IFI16 as Circulating Protein in Autoimmune Diseases Is a Signal of Damage that Impairs Endothelial Cells through High-Affinity Membrane Binding. *PLoS ONE*, 8(5):e63045, May 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063045. URL https://dx.plos.org/10.1371/journal.pone.0063045.

Ankit Gupta, Zoe Wefers, Konstantin Kahnert, Jan N. Hansen, Mohini K. Misra, Will Leineweber, Anthony Cesnik, Dan Lu, Ulrika Axelsson, Frederic Balllllosera, Russ B. Altman, Theofanis Karaletsos, and Emma Lundberg. SubCell: Proteome-aware vision foundation models for microscopy capture single-cell biology. bioRxiv, October 2025. URL https://www.biorxiv.org/content/10.1101/2024.12.06.627299v2. Preprint, version 2.

Mien-Chie Hung and Wolfgang Link. Protein localization in disease and therapy. *Journal of Cell Science*, 124(20):3381–3392, October 2011. ISSN 1477-9137, 0021-9533. doi: 10.1242/jcs.089110. URL https://journals.biologists.com/jcs/article/124/20/3381/31946/Protein-localization-in-disease-and-therapy.

Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, Lisa A. Cannon-Albright, Craig C. Teerlink, Janet L. Stanford, William B. Isaacs, Jianfeng Xu, Kathleen A. Cooney, Ethan M. Lange, Johanna Schleutker, John D. Carpten, Isaac J. Powell, Olivier Cussenot, Geraldine Cancel-Tassin, Graham G. Giles, Robert J. MacInnis, Christiane Maier, Chih-Lin Hsieh, Fredrik Wiklund, William J. Catalona, William D. Foulkes, Diptasri Mandal, Rosalind A. Eeles, Zsofia Kote-Jarai, Carlos D. Bustamante, Daniel J. Schaid, Trevor Hastie, Elaine A. Ostrander, Joan E. Bailey-Wilson, Predrag Radivojac, Stephen N. Thibodeau, Alice S. Whittemore, and Weiva Sieh. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, 99(4):877–885, October 2016. ISSN 00029297. doi: 10.1016/j.ajhg.2016.08.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929716303706.

Ben Isselmann, Dilara Göksu, Heinz Neumann, and Andreas Weinmann. Using Deep Learning Models Pretrained by Self-Supervised Learning for Protein Localization. arXiv, September 2026. URL https://arxiv.org/abs/2604.10970v2. Preprint, version 2.

Alexandr A. Kalinin, John Arevalo, Erik Serrano, Loan Vulliard, Hillary Tsang, Michael Bornholdt, Alán F. Muñoz, Suganya Sivagurunathan, Bartek Rajwa, Anne E. Carpenter, Gregory P. Way, and Shantanu Singh. A versatile information retrieval framework for evaluating profile strength and similarity. *Nature Communications*, 16(1):5181, June 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60306-2. URL https://www.nature.com/articles/s41467-025-60306-2.

Norbert Kartner, Olga Augustinas, Timothy J. Jensen, A. Leonard Naismith, and John R. Riordan. Mislocalization of ΔF508 CFTR in cystic fibrosis sweat gland. *Nature Genetics*, 1(5):321–327, August 1992. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng0892-321. URL https://www.nature.com/articles/ng0892-321.

Hirofumi Kobayashi, Keith C. Cheveralls, Manuel D. Leonetti, and Loic A. Royer. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature Methods*,

19(8):995–1003, August 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01541-z. URL https://www.nature.com/articles/s41592-022-01541-z.

Julia Kornienko, Marta Rodríguez-Martínez, Kai Fenzl, Florian Hinze, Daniel Schraivogel, Markus Grosch, Brigit Tunaj, Dominik Lindenhofer, Laura Schraft, Moritz Kueblbeck, Eric Smith, Chad Mao, Emily Brown, Anjali Owens, Ardan M. Saguner, Benjamin Meder, Victoria Parikh, Michael Gotthardt, and Lars M. Steinmetz. Mislocalization of pathogenic RBM20 variants in dilated cardiomyopathy is caused by loss-of-interaction with Transportin-3. *Nature Communications*, 14(1):4312, July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39965-6. URL https://www.nature.com/articles/s41467-023-39965-6.

Jessica Lacoste, Marzieh Haghighi, Shahan Haider, Chloe Reno, Zhen-Yuan Lin, Dmitri Segal, Wesley Wei Qian, Xueting Xiong, Tanisha Teelucksingh, Esteban Miglietta, Hamdah Shafqat-Abbasi, Pearl V. Ryder, Rebecca Senft, Beth A. Cimini, Ryan R. Murray, Chantal Nyirakanani, Tong Hao, Gregory G. McClain, Frederick P. Roth, Michael A. Calderwood, David E. Hill, Marc Vidal, S. Stephen Yi, Nidhi Sahni, Jian Peng, Anne-Claude Gingras, Shantanu Singh, Anne E. Carpenter, and Mikko Taipale. Pervasive mislocalization of pathogenic coding variants underlying human disorders. *Cell*, 187(23):6725–6741.e13, November 2024. ISSN 00928674. doi: 10.1016/j.cell.2024.09.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867424010213.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46 (D1):D1062–D1067, January 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx1153. URL http://academic.oup.com/nar/article/46/D1/D1062/4641904.

Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, 12(1):103, December 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00803-9. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9. URL https://www.nature.com/articles/s42256-019-0138-9.

Théo Moutakanni, Camille Couprie, Seungeun Yi, Michael Doron, Zitong S. Chen, Nikita Moshkov, Elouan Gardes, Mathilde Caron, Hugo Touvron, Armand Joulin, Piotr Bojanowski, Wolfgang M. Pernice, and Juan C. Caicedo. Cell-DINO: Self-supervised image-based embeddings for cell fluorescent microscopy. *PLOS Computational Biology*, 21(12):e1013828, December 2025. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1013828. URL https://dx.plos.org/10.1371/journal.pcbi.1013828.

Vikas Pejaver, Jorge Urresti, Jose Lugo-Martinez, Kymberleigh A. Pagel, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, David N. Cooper, Jonathan Sebat, Lilia M. Iakoucheva, Sean D. Mooney, and Predrag Radivojac. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, 11(1):5918, November 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19669-x. URL https://www.nature.com/articles/s41467-020-19669-x.

Sriram Pendyala, Katie Partington, Nicholas Bradley, Abbye E McEwen, Gwenneth Straub, Hyeon-Jin Kim, Shawn Fayer, Daniel Lee Holmes, Katherine A Sitko, Allyssa J Vandi, Rachel L Powell, Clayton E Friedman, Evan McDermot, Nishka Kishore, Frederick P Roth, Alan F Rubin, Kai-Chun Yang, Lea M Starita, William S Noble, and Douglas M Fowler. Image-based, pooled phenotyping reveals multidimensional, disease-specific variant effects. bioRxiv, July 2025. URL https://www.biorxiv.org/content/10.1101/2025.07.03.663081v1. Preprint, version 1.

Erik Serrano, Srinivas Niranj Chandrasekaran, Dave Bunten, Kenneth I. Brewer, Jenna Tomkinson, Roshan Kern, Michael Bornholdt, Stephen J. Fleming, Ruifan Pei, John Arevalo, Hillary Tsang, Vincent Rubinetti, Callum Tromans-Coia, Tim Becker, Erin Weisbart, Charlotte Bunne, Alexandr A. Kalinin, Rebecca Senft, Stephen J. Taylor, Nasim Jamali, Adeniyi Adeboye, Hamdah Shafqat Abbasi, Allen Goodman, Juan C. Caicedo, Anne E. Carpenter, Beth A. Cimini, Shantanu Singh, and Gregory P. Way. Reproducible image-based profiling with Pycytominer. *Nature Methods*, 22 (4):677–680, April 2025. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-025-02611-8. URL https://www.nature.com/articles/s41592-025-02611-8.

Alina Sigaeva, Charlotte Hutchings, Anthony Cesnik, Kathryn S. Lilley, and Emma Lundberg. Subcellular localization as a driver of protein function. *Nature Reviews Molecular Cell Biology*, 27 (7):493–510, July 2026. ISSN 1471-0072, 1471-0080. doi: 10.1038/s41580-026-00947-3. URL https://www.nature.com/articles/s41580-026-00947-3.

David R. Stirling, Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, September 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04344-9. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04344-9.

Terry R. Suk and Maxime W. C. Rousseaux. The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Molecular Neurodegeneration*, 15(1):45, August 2020. ISSN 1750-1326. doi: 10.1186/s13024-020-00397-1. URL https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-020-00397-1.

Malvika Tejura, Yile Chen, Abbye E. McEwen, Ross Stewart, Yuriy Sverchkov, Florent Laval, Ivan Woo, Daniel Zeiberg, Runxi Shen, Shawn Fayer, Jeremy Stone, Nahum Smith, Silvia Casadei, Ziyu R. Wang, Matthew W. Snyder, Benjamin J. Capodanno, Pankhuri Gupta, Mariam Benazouz, Shantanu Jain, Sarah Heidl, Lara Muffley, Shengcheng Dong, Benjamin C. Hitz, Idan Gabdank, Khine Lin, Estelle Y. Da, Sabrina Best, Sally Grindstaff, David Reinhart, Leslie Rodriguez-Salas, Obsa Seid, Allyssa J. Vandi, Cameron Wenman, Melinda K. Wheelock, Sriram Pendyala, Dan Holmes, Alicia Xu, Airi Hosokai, Maxime Tixhon, Chloe Reno, Jessica D. Ewald, Kerstin Spirohn-Fitzgerald, Tanisha Teelucksingh, Tong Hao, Zitong S. Chen, Marzieh Haghighi, Ahmad Kamal Hamid, Esteban A. Miglietta, Erin Weisbart, Georges Coppin, Luke Lambourne, Marinella Gebbia, Atina G. Coté, Warren Van Loggerenberg, Kirby M. Fawcett, Robert D. Steiner, Jill M. Johnsen, Andrew B. Stergachis, Lilia M. Iakoucheva, Shantanu Singh, Beth A. Cimini, Frederick P. Roth, Richard G. James, the IGVF Coding Variants Focus Group, Marc Vidal, Mikko Taipale, Anne E. Carpenter, Michael A. Calderwood, Mark Craven, Vikas Pejaver, Alan F. Rubin, Predrag Radivojac, Douglas M. Fowler, and Lea M. Starita. A scalable approach to resolving variants of uncertain significance. bioRxiv, February 2026. URL http://biorxiv.org/lookup/doi/10.64898/2026.02.14.705848. Preprint.

Peter J. Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M. Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M. Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M. Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P. Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle Von Feilitzen, Kathryn S. Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, May 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal3321. URL https://www.science.org/doi/10.1126/science.aal3321.

Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle Von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar Von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, January 2015. ISSN 0036-8075, 1095-9203. doi:

10.1126/science.1260419. URL https://www.science.org/doi/10.1126/science.1260419.

Fredrick Van Goor, Sabine Hadida, Peter D. J. Grootenhuis, Bill Burton, Jeffrey H. Stack, Kimberly S. Straley, Caroline J. Decker, Mark Miller, Jason McCartney, Eric R. Olson, Jeffrey J. Wine, Ray A. Frizzell, Melissa Ashlock, and Paul A. Negulescu. Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proceedings of the National Academy of Sciences*, 108(46):18843–18848, November 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1105787108. URL https://pnas.org/doi/full/10.1073/pnas.1105787108.

Xiaohong Wang and Shulin Li. Protein mislocalization: Mechanisms, functions and clinical applications in cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1846(1): 13–25, August 2014. ISSN 0304419X. doi: 10.1016/j.bbcan.2014.03.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0304419X14000274.

A APPENDIX

This appendix provides the methodological and experimental details needed to reproduce the MisLocus benchmark and to evaluate future image representations on the same single-cell cohort. The workflow is built around a shared evaluation substrate: all representations are applied to the same quality-controlled single-cell crops, matched reference–variant comparisons, control-null comparisons for estimating technical variation, and replicate-aware evaluation splits. Thus, differences in reported performance primarily reflect the representation rather than upstream filtering, cell selection, or task-specific data partitioning.

We first describe dataset curation, site-level image quality control, single-cell quality control, crop extraction, and allele inventory construction (Appendix A.1). We then describe the extractions of image representations, model retraining and finetuning, and embedding preprocessing applied before downstream evaluation (Appendix A.2), followed by the replicate-aware supervised and unsupervised evaluation protocols (Appendix A.3). Finally, we provide supplementary analyses supporting the main results, including control-null calibration, CellProfiler feature-attribution diagnostics, representation-agreement analyses, manual-label comparisons, and additional clinical-annotation analyses (Appendix A.4).

A.1 DATASET CURATION AND SINGLE-CELL CROP PREPROCESSING

A.1.1 DATASET CURATION

Source imaging. The benchmark derives from six previously acquired and publicly released VarChAMP imaging batches (Cell Painting Gallery collection `cpg0020-varchamp`) (Tejura et al., 2026), organized as three biological-replicate pairs (B7/B8, B13/B14, B15/B16). Batches within each pair use the same experimental setup and plate design but were screened at separate times, enabling robustness to biological-replicate and acquisition-date variation to be assessed. The source batches were imaged in U2OS cells on a spinning-disk Opera Phenix at $20\times$/NA1.0 ($0.598\ \mu$m/pixel). Each batch contains arrayed 384-well plates, with each well corresponding to cells expressing one mNeonGreen-tagged reference or variant protein. All six batches use the same four imaging channels: DNA, GFP (protein), AGP, and Mito, with nine imaging sites per well. MisLocus newly applies the quality-control, crop-extraction, harmonization, and benchmark procedures below to these source images.

The imaging hierarchy is batch $\rightarrow$ plate $\rightarrow$ well $\rightarrow$ site $\rightarrow$ single cell. Each retained single-cell record is assigned metadata linking it to its batch, physical plate, well, imaging site, object identifier, protein identity, and control category. This metadata defines the shared cell manifest used by all image crops, CellProfiler profiles, and learned embeddings.

Protein inventory and paired structure. MisLocus is organized around matched reference and variant proteins. We define the gene symbol for a protein by parsing the field `gene_variant`: entries without an underscore are treated as reference proteins, whereas entries with an underscore are treated as coding variants. Counts reported for the benchmark exclude technical, negative, and positive controls. Across the designed platemaps, the non-control portion of the dataset contains 1,623 tested proteins, including 281 reference proteins and 1,342 variants of 310 proteins. After imaging and single-cell quality control, the released crop manifest retains 1,541 tested proteins, including 250 reference proteins and 1,291 variants of 282 genes. Among these, 125 genes retain both a reference protein and at least one variant protein, yielding 1,055 paired reference–variant comparisons for mislocalization analysis.

The distinction between retained proteins and paired proteins is important for reproducibility. Some variant proteins remain in the manifest but lack a retained reference protein after quality control; these "orphan" variants are useful for representation learning and embedding extraction but are excluded from paired reference–variant evaluation. Conversely, reference-only genes provide additional reference-protein diversity for representation learning and reference-localization analyses but do not define variant-pair tasks.

Control protein curation. In addition to experimental reference and variant proteins, each batch contains control conditions used to quantify technical variation. Technical controls (TC) include

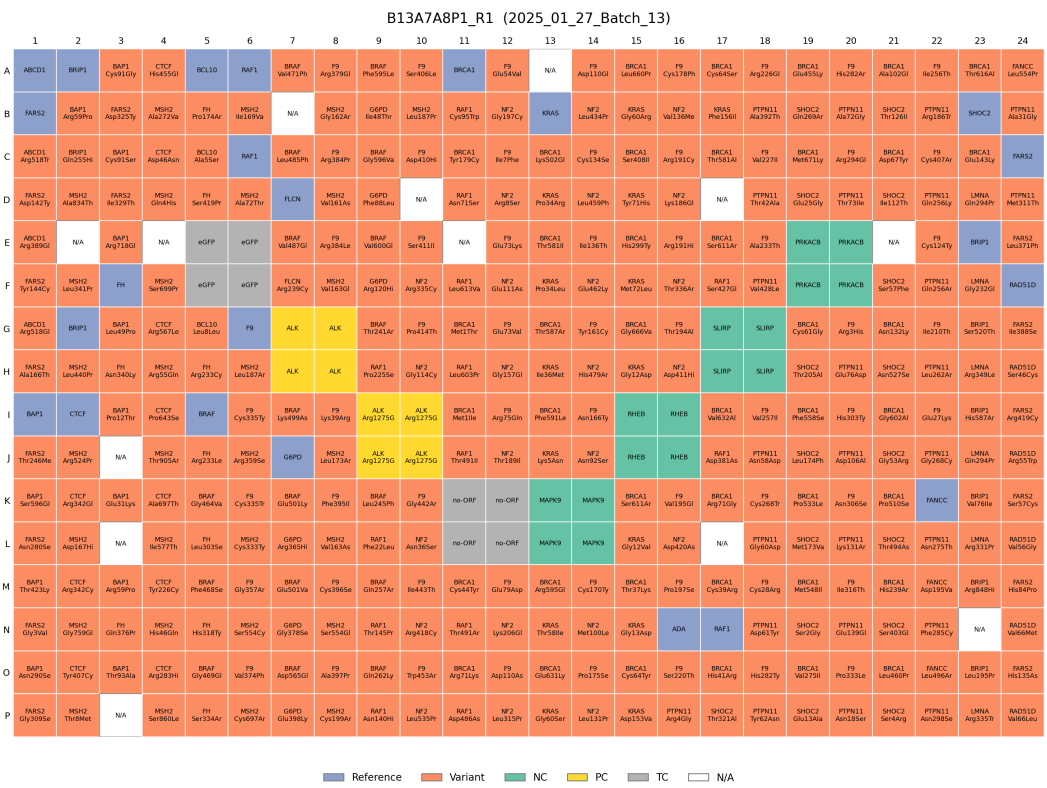

Figure 4: **Example 384-well plate layout for a MisLocus imaging plate.** The layout of an imaging plate from Batch 13, with columns numbered 1–24 and rows A–P. Each well contains cells expressing one mNeonGreen-tagged protein and is colored by assay role: reference or variant proteins, negative controls (NC), positive controls (PC), technical controls (TC), and unassigned wells (N/A).

fluorescent-protein-only or empty-vector conditions that are not part of the paired protein inventory. Negative-control (NC) proteins include RHEB, MAPK9, PRKACB, and SLIRP, which are reference proteins screened in multiple wells per plate. Positive controls (PC) include the ALK reference protein and the ALK Arg1275Gln variant protein, a control pair expected to produce a reproducible localization-related phenotype. The biological inventory also excludes observations annotated cNC (complementary NC) or cPC (complementary PC), and recurring identities of these controls are excluded even when an observation carries another label. They remain in the release manifest for training, testing and technical-null calibration, but are not counted in the biological cohort for detecting mislocalization.

The control wells are used to estimate well-position and plate-level technical effects. Each protein was screened with four replicates on every 384-well imaging plate in a quadrant square (Figure 4), and these replicates are compared against one another to construct empirical control-null distributions for downstream hit calling (Appendix A.3.3). Because these comparisons involve the same protein measured in different wells or plates, high discriminatory performance in this setting reflects technical or well-position variation rather than a variant-induced localization change.

A.1.2 Image quality control

Site-level image quality control. Failed acquisitions are screened out at the site level before any cell-level analysis. For every field of view we compute nine image-quality metrics that capture distinct failure modes: focus score (normalized variance), signal-to-noise ratio defined as the difference between the 95th-percentile and median pixel intensities divided by the standard deviation of background pixels, Shannon entropy, coefficient of variation, foreground-to-background ratio (mean of pixels above the 75th percentile divided by mean of pixels below the 25th percentile), the slope of the log-log power spectrum, a tile-averaged local focus score, and the percentages of saturated and

minimal pixels. For each plate-channel combination we apply a robust median-absolute-deviation (MAD) outlier rule at five MAD: a channel is flagged when at least two of its nine metrics fall outside the band, a site is flagged whenever any of its four channels is flagged, and a well is dropped entirely when at least three of its nine sites are flagged; otherwise only the individual flagged sites are removed. Empty or unreadable images are auto-flagged. The thresholds were chosen to remove obvious focus and contamination failures without rejecting plate-to-plate biological variability.

A.1.3 SINGLE-CELL QUALITY CONTROL

Cell-level quality control. Per-cell records are filtered through cell-level quality-control gates applied in fixed order. We first restrict the data to wells with valid protein annotations and propagate the site-level image-quality verdicts above. We then retain only cells whose nucleus-to-cell area ratio falls in $[0.08, 0.40]$, which removes debris on the low end and merged segmentation errors on the high end. The nuclear centroid must also leave room for a complete 128×128 crop within the imaging field. Within each well, we discard cells whose mean protein-channel intensity deviates from the well median by more than five MADs; this per-well outlier filter is computed only on cells from good-quality sites so that failed sites do not distort the well-level median. Finally, proteins with fewer than 100 retained cells after the preceding filters are dropped entirely.

Although the gating criteria use CellProfiler-derived measurements, including compartment areas, centroid coordinates, and mean protein-channel intensity, the resulting per-cell manifest is shared across every representation in the benchmark. Thus, CellProfiler features, raw image crops, and learned embeddings are all evaluated on the same QC-approved cells.

Single-cell crop extraction. For image-based representations, we extract one 128×128 four-channel crop around each retained nuclear centroid directly from the raw 16-bit TIFFs. Crops are stored as `uint16` NumPy arrays without per-image normalization, illumination correction, background subtraction, or model-specific preprocessing at extraction time. This preserves the original pixel intensities and allows each representation family to apply its own expected preprocessing. For example, Cytoself consumes native-scale crops with its own channel construction, whereas SubCell-style models apply their own resizing and normalization in the dataloader. Avoiding extraction-time normalization prevents irreversible preprocessing choices from favoring one representation family over another.

Table 3: Reference and variant inventory before and after single-cell quality control. Counts exclude technical, negative, and positive controls.

Quantity	Designed platemaps	QC-passed manifest
Tested proteins	1,623	1,541
Reference proteins	281	250
Variant proteins	1,342	1,291
Unique genes	310	282
Paired reference proteins	135	125
Paired variant proteins	1,157	1,055
Median variants per paired reference protein	4.0	4.0
Maximum variants per paired reference protein	164	161

A.2 IMAGE REPRESENTATIONS AND EMBEDDING PREPROCESSING

MisLocus evaluates five representation groups spanning hand-engineered image features, self-supervised representations trained on the benchmark cohort, frozen pretrained microscopy encoders, cohort-fine-tuned encoders, and a frozen microscopy foundation-model baseline. Each representation ultimately enters the downstream benchmark as a per-cell feature matrix keyed to the shared QC-approved cell manifest described in Appendix A.1. This common cell set ensures that performance differences reflect the representation and its feature geometry rather than differences in cell selection, crop extraction, or allele pairing.

The representation groups are: CellProfiler hand-engineered features; Cytoself trained from scratch on MisLocus crops; SubCell-portable frozen pretrained models; SubCell-embed models fine-tuned

on MisLocus crops; and MorphEM, a frozen microscopy foundation-model baseline. For all learned representations using cell-crop images, the starting input is the same 128×128 four-channel `uint16` crop centered on the segmented nucleus by CellProfiler. Model-specific preprocessing is applied only after crop extraction, so no irreversible crop-level normalization favors one representation family over another.

A.2.1 SHARED TRAINING SPLIT FOR RETRAINED REPRESENTATIONS

Only Cytoself and the fine-tuned SubCell models are trained or fine-tuned on MisLocus images. For these models, we use the deterministic technical-replicate split defined by the physical replicate plates within each platemap: T1 and T2 are used for representation training, T3 for validation and model selection, and T4 is held out from representation learning. The split is applied independently within each platemap, so the same allele identities appear across train, validation, and test plates, but cells from the held-out physical replicate are never used to fit the representation.

This split is distinct from the downstream supervised evaluation described in Appendix A.3.1. In representation training, T1/T2/T3/T4 define which images are used to learn the encoder. In downstream evaluation, all representations—including frozen encoders and CellProfiler features—are scored on the same T4-held-out reference–variant tasks. Thus, retrained and fine-tuned representations are evaluated on cells that were not used during training, while frozen representations are evaluated on the identical held-out cells for consistency.

A.2.2 CELLPROFILER

Image analysis and raw features. CellProfiler provides an interpretable hand-engineered feature baseline. We use CellProfiler profiles generated with the upstream VarChAMP image-analysis workflow (Lacoste et al., 2024; Tejura et al., 2026), rather than presenting feature extraction as newly performed for MisLocus. In that workflow, illumination variation is corrected before segmentation. Nuclei are identified from the DNA channel, and whole-cell boundaries are estimated from the AGP channel using watershed-based secondary-object segmentation. Features are then measured across the three segmented compartments—Cells, Cytoplasm, and Nuclei—and across the four imaging channels: DNA, GFP/protein, AGP, and Mito.

The raw CellProfiler profile contains standard morphology and intensity measurements, including fluorescence intensity, texture, granularity, radial distribution, neighbor, and shape/location features. Joining the compartment-level outputs and metadata yields $3{,}194$ raw CellProfiler feature columns per cell before benchmark-specific preprocessing.

Profile filtering and normalization. CellProfiler profiles are first restricted to the shared QC-approved cell manifest so that the engineered-feature baseline is evaluated on the same cells as all crop-based representations. Wells with fewer than 20 retained cells are removed. Non-finite values are treated as missing; feature columns with more than 100 missing rows are dropped, and cells with remaining missing values in retained features are removed.

For each batch, per-plate medians and median absolute deviations are computed for every feature. Features are retained only if they show non-zero robust variation and sufficient absolute coefficient of variation on every plate in the batch. Each retained feature is then normalized within plate as

$$x' = \frac{x - \mathrm{median}_{\mathrm{plate}}}{\mathrm{MAD}_{\mathrm{plate}}},$$

and extreme normalized values are capped at ± 100 MAD units. This plate-wise RobustMAD normalization reduces systematic staining, illumination, and acquisition differences while preserving within-plate biological variation (Serrano et al., 2025).

Feature selection. After normalization, we apply feature selection to remove measurements that are uninformative, unstable, or likely to encode nuisance variation. Neighbor features are removed because they primarily capture local cell density and spatial arrangement within a field of view, which can act as well- or site-level confounders. Standard problematic CellProfiler features, including known unstable correlation and high-frequency granularity measurements, are also removed. Near-zero-variance features are discarded.

Correlation filtering is performed in a channel-aware manner. GFP/protein-channel features are decorrelated separately from non-GFP features, and the retained sets are then combined. This prevents protein-channel features from being removed simply because they correlate with morphology features from other channels, preserving diversity in the feature subset most directly tied to protein localization. The resulting processed CellProfiler profile contains approximately $1{,}060$ features per batch (down from $3{,}194$ raw columns), reflecting both feature selection and the channel-aware correlation filter.

Feature-channel groups. For downstream benchmarking, CellProfiler features are organized into biologically interpretable feature groups. The GFP group contains protein-channel features; DNA, AGP, and Mito groups contain features specific to their corresponding stains; the Morph group contains non-protein morphology/context features; and the ALL group contains the full retained feature set when used for combined-channel analyses. For stain-specific groups other than GFP, features that reference the GFP/protein channel are excluded so that non-protein-channel comparisons cannot directly use protein-channel intensity or texture information.

A.2.3 CYTOSELF

Model rationale. Cytoself (Kobayashi et al., 2022) is a self-supervised VQ-VAE-based representation model originally developed for protein-localization imaging. We train Cytoself from scratch on MisLocus rather than transferring published OpenCell weights, because the original OpenCell supervision relies on stable per-gene localization labels. In MisLocus, the central biological question is whether different alleles of the same gene differ from one another, so a representation trained only to collapse all cells from the same gene would be misaligned with the variant-effect task. We therefore use the allele identifier as the classification pretext label during Cytoself training.

Input construction. Following the Cytoself approach (Kobayashi et al., 2022), we train on native-scale 128×128 crops using GFP/protein and DNA fluorescence. Our implementation also computes a nuclear-distance image by Otsu-thresholding the DNA image and scaling the Euclidean distance transform of the resulting mask by 0.01. GFP, DNA, and this derived image are stacked as a three-channel model input; only the first two are acquired fluorescence channels. This is an adaptation of Cytoself's nuclear-fiducial preprocessing: our distance image is nonnegative and zero outside the mask, whereas the original paper describes a signed nuclear-distance channel.

No per-crop z-scoring, min–max normalization, logarithmic transformation, background subtraction, or illumination correction is applied before the Cytoself encoder. Data augmentation is restricted to rotations by multiples of 90° and horizontal/vertical flips. These transformations provide minimal augmentation while preserve the localization phenotype.

Training and architecture. Cytoself is trained on the six benchmark batches using the shared technical-replicate split: T1 and T2 for training, T3 for validation, and T4 as held out test set. The model uses a two-stage VQ-VAE architecture with hierarchical encoders and vector-quantization layers. The second-stage vector code captures coarse spatial organization, whereas the first-stage codebook-index histogram captures finer-grained texture and local pattern information. Training uses Adam optimization with a base learning rate of 10^{-3}, batch size 64, a maximum of 20 epochs, early stopping on the T3 validation split, and learning-rate reduction on validation plateau. In the final training run, the model converged before the maximum epoch count.

Embedding extraction and feature channels. After training, the encoder is frozen and applied to QC-approved cells from all six benchmark batches. We extract two complementary feature types per cell: the flattened second-stage vector code, yielding a $1{,}024$-dimensional `global` representation, and the first-stage codebook-index histogram, yielding a 512-dimensional `spectrum` representation. Their concatenation gives the $1{,}536$-dimensional `combined` representation. These three Cytoself feature channels expose coarse spatial structure, fine-grained codebook usage, and their joint representation, respectively.

A.2.4 SUBCELL-PORTABLE

Model variants. SubCell is a Vision Transformer representation model trained on Human Protein Atlas microscopy images using masked-image modeling and protein-supervised contrastive objec-

tives (Gupta et al., 2025). We evaluate frozen SubCell-portable checkpoints directly on MisLocus crops, without updating their weights. The portable variants differ by input-channel configuration and pretraining objective. We evaluate SubCell models, including both ViT-style and MAE-style pretrained checkpoints where available. Each model produces a single $1{,}536$-dimensional embedding per cell through an attention-pooling head.

Channel mapping. SubCell expects channels in the RYBG (red, yellow, blue and green) convention used during HPA pretraining: red for microtubules, yellow for endoplasmic reticulum, blue for nuclei, and green for the protein channel. MisLocus does not contain the exact HPA marker set, so we map the available stains to the closest functional analogues: AGP to the red/cytoskeletal slot, Mito to the yellow/organellar slot, DNA to the blue/nuclear slot, and GFP to the green/protein slot. The same channel mapping is used for frozen SubCell-portable inference and for fine-tuned SubCell, ensuring that the two settings differ by weight adaptation rather than input-channel semantics.

Inference preprocessing. SubCell was trained at a finer physical pixel size than MisLocus. To preserve the physical scale expected by the pretrained patch embeddings, each 128×128 MisLocus crop is first cast to floating point and resized by the MisLocus-to-HPA pixel-size ratio, producing a larger image at approximately the HPA physical sampling scale. A centered 448×448 crop is then taken to match the SubCell training resolution. Intensities are min–max normalized across the supplied channels and spatial dimensions before inference. This preprocessing keeps the model input resolution and patch-level physical scale aligned with the HPA-pretrained checkpoint, rather than simply resizing the 128×128 crop directly to 448×448.

Embedding extraction. Frozen SubCell-portable encoders are applied to all QC-approved cells in the six benchmark batches. For the SubCell-portable, we only extracted the embeddings using the three-channel (rgb) variants and the corresponding subset of channels, because the original paper reported that dropping the yellow channel (ER) led to better results in protein localization prediction tasks (Gupta et al., 2025). For each portable configuration, the benchmark stores one $1{,}536$-dimensional embedding per retained cell. These embeddings are then passed through the shared learned-embedding preprocessing described below before downstream evaluation, where applicable.

A.2.5 SubCell-embed fine-tuned models

Motivation. Frozen SubCell-portable models provide a strong pretrained baseline but were not exposed to MisLocus imaging conditions during training. We therefore also fine-tuned a full four-channel SubCell models in both ViT and MAE configurations on MisLocus crops while preserving the same channel mapping and HPA-scale inference preprocessing used for the portable setting. This design allows the benchmark to compare frozen transfer against cohort-adapted embeddings under matched input conventions.

Training data and labels. Fine-tuned SubCell uses the same six-batch training manifest and technical-replicate split as Cytoself: T1 and T2 for training, T3 for validation, and T4 held out. During fine-tuning, cells are grouped by their respective protein (reference or variant) labels for the protein-supervised contrastive objective. This encourages the model to learn protein-level localization structure while still allowing downstream evaluation to test whether variant alleles separate from their references.

Model objectives. We fine-tune four-channel MAE and ViT SubCell models, initializing each encoder and gated-attention pool from its corresponding HPA-pretrained weights. The MAE model combines reconstruction of masked, patch-normalized image targets, cell-level contrastive learning, and allele-supervised contrastive learning. The ViT model uses allele-supervised contrastive learning without reconstruction or cell-level contrastive loss. Both models also train an auxiliary allele-classification probe with cross-entropy loss. The probe receives detached pooled features, so its loss does not directly backpropagate through the encoder or pool. The decoder, contrastive projectors, and probe are initialized anew. Embeddings are the $1{,}536$-dimensional outputs of the two-head gated-attention pool.

Augmentation and optimization. Two views of each cell are generated with independently sampled horizontal and vertical flips and affine or perspective transformations. Intensity augmentation may remove a non-GFP channel, rescale GFP, apply per-channel brightness and contrast jitter, blur or sharpen, add Gaussian noise, and erase image regions. MAE applies these intensity transformations only to the second view; ViT applies them to both. Images are not clipped or renormalized after augmentation.

Both models use AdamW with a peak learning rate of 5×10^{-5}, a five-pass warmup, and a subsequent linear learning-rate decay over 100 sampling passes (9,700 updates). Weight decay is 0.05 for eligible encoder and decoder parameters, zero for their biases and one-dimensional parameters, and 0.01 for the pool, projectors, and probe. Training uses FP32 tensors and global gradient-norm clipping at 1. MAE masks 25% of image patches in the first view; its second view is unmasked. Neither model uses cell or object masking, and ViT uses no image-patch masking. Every ten passes, checkpoint selection uses macro one-versus-rest average precision for allele identity on a fixed T3 validation subset. The selected seed-42 checkpoints were chosen before T4 extraction.

Embedding extraction. The selected encoder and gated-attention pool for each model are frozen and applied to QC-approved cells from all six batches, including T1–T4. Each four-channel crop is ordered AGP, Mito, DNA, GFP; its native 128×128 image is bilinearly resized to 955×955, center-cropped to 448×448, and min–max normalized jointly across channels and spatial positions for each cell. Inference uses neither stochastic augmentation nor image-patch masking. Each model exports one $1{,}536$-dimensional FP32 embedding per cell, covering 3,332,309 cells.

A.2.6 MorphEM

MorphEM is included as a frozen microscopy foundation-model baseline (Agrawal et al., 2026). Unlike Cytoself and fine-tuned SubCell, MorphEM is not trained or fine-tuned on MisLocus images. The frozen encoder is applied to the shared single-cell crops to produce per-cell embeddings for downstream evaluation. We evaluate a GFP-only representation, which focuses directly on the tagged protein channel, and an ALL-channel representation, which incorporates the protein channel together with cellular context channels. Reporting both regimes lets us distinguish the contribution of the protein channel itself from that of the surrounding cellular context, and provides parity with CellProfiler, which exposes analogous GFP-only and all-channel feature groups; the GFP-only regime is the headline setting because it isolates protein-localization signal from morphology features that vary independently across alleles. Because MorphEM is frozen, any performance differences relative to fine-tuned SubCell reflect the suitability of the pretrained representation and input adaptation rather than MisLocus-specific representation learning.

A.2.7 Shared preprocessing for learned embeddings

After raw embeddings are extracted, Cytoself, SubCell-portable, fine-tuned SubCell, and MorphEM embeddings are processed through a shared embedding-preprocessing pipeline before downstream evaluation. This preprocessing is parallel in spirit to CellProfiler normalization but is adapted to the statistical properties of learned embeddings, which lack explicit CellProfiler-style feature names and stain-channel semantics.

First, wells with fewer than 20 retained cells are removed. Non-finite values are treated as missing; unreliable embedding dimensions with excessive missingness are removed, and cells with remaining missing values are dropped. Per-plate medians and MADs are then computed for every embedding dimension, and each dimension is RobustMAD-normalized within plate. Unlike CellProfiler feature preprocessing, learned embeddings are not filtered by median absolute deviation or coefficient of variation, because sparse embedding dimensions—especially Cytoself codebook-histogram features—can have median and MAD equal to zero while still carrying real signal. Instead, dead dimensions are removed using a batch-wide standard-deviation threshold. Normalized values are clipped to ± 100, and redundant embedding dimensions are removed with a single-pass correlation filter across all embedding dimensions.

The same preprocessing logic is applied to all learned embeddings so that downstream differences are not driven by representation-specific normalization choices. For sensitivity analyses, we also retain unprocessed embedding variants that bypass plate normalization and decorrelation. These

variants are used to confirm that the major benchmark conclusions are not artifacts of the embedding preprocessing step.

A.2.8 REPRESENTATION OUTPUTS USED FOR DOWNSTREAM EVALUATION

The output of this section is one processed per-cell feature matrix for each representation and channel used in the benchmark. CellProfiler contributes interpretable stain-specific and morphology feature groups. Cytoself contributes `global`, `spectrum`, and `combined` embeddings. SubCell-portable contributes one embedding channel for each frozen channel/objective configuration. Fine-tuned SubCell contributes one embedding channel for each fine-tuned MAE-style or ViT-style model. MorphEM contributes GFP-only and all-channel embeddings. Each feature matrix is indexed by the shared cell identifiers and is evaluated using the common reference–variant, clinical-agreement, and HPA localization protocols described in Appendix A.3.1.

A.2.9 COMPUTE RESOURCES.

We report compute requirements for reproducing the benchmark starting from QC-approved single-cell crops. Representation training and embedding extraction were run on a single compute node with 4 NVIDIA H100 NVL GPUs, each with 93 GiB of GPU memory. Cytoself was trained from scratch on the T1+T2 training subset, comprising approximately 1.66 million single-cell crops, using batch size 64 for at most 20 epochs. Training converged at epoch 10 and required approximately 7 hours on one GPU.

The two SubCell fine-tuning variants, corresponding to the ViT-style and MAE-style objectives, were each trained with two-GPU distributed data parallelism. Each fine-tuning run completed in approximately 3.8 hours; the two variants were trained simultaneously on disjoint GPU pairs, so the combined wall-clock time was also approximately 3.8 hours. Frozen SubCell-portable embedding extraction was run on the shared 3,332,309-cell QC cohort across all six benchmark batches for each evaluated checkpoint and input configuration. Using FP16 inference at 448×448 resolution with batch size 128, these extraction runs required approximately two hours when sharded across all four GPUs. This timing includes repeated evaluation of the same crop under multiple checkpoint and configuration settings.

The downstream benchmark was also GPU-accelerated. XGBoost classifiers were trained using fold-level parallelism across the four GPUs. A single sweep for one representation and one batch required 1–25 minutes depending on the number of QC-passing cells in that batch, with a median runtime of approximately 14 minutes. The complete supervised benchmark across nine feature sets and six batches required approximately 10–12 hours of wall-clock time. Reproduction of the reported representations and benchmark summary outputs, starting from QC-approved single-cell crops, requires approximately 1–2 days on the node described above. The released pipeline ends at those summary outputs; the manuscript-specific figure assembly script is not part of the anonymous companion repository. CellProfiler feature extraction was inherited from the upstream imaging data release and was not rerun as part of MisLocus.

A.3 EVALUATION PROTOCOLS

A.3.1 OVERVIEW OF EVALUATION PROTOCOLS

This section describes the shared evaluation protocols used for all image representations. The evaluations follow the three axes introduced in Section 3.5: variant protein mislocalization detection, clinical agreement of imaging-derived phenotype scores, and reference-protein localization structure. All evaluations consume the same quality-controlled single-cell manifest, curated reference–variant pairs, control proteins, and replicate-aware training and testing splits. Thus, differences across rows in the benchmark reflect the image representation and feature channel rather than differences in cell selection, allele pairing, or split construction.

For each representation, evaluations are run independently for each feature channel exposed by that representation. For CellProfiler, these channels correspond to features from only the protein channel and across all four channels. For learned representations, channels correspond to embedding heads or input-channel configurations described in Appendix A.2.

A.3.2 HELD-OUT TECHNICAL-REPLICATE EVALUATION

The primary goal of MisLocus is to benchmark whether image representations capture reproducible variant–reference imaging phenotypes across independently prepared plates, rather than to assess generalization to protein identities unseen during representation training. Therefore, MisLocus uses the same deterministic technical-replicate split throughout the benchmark. Each platemap is replicated across four physical plates, denoted T1–T4, which contain the same allele layout. For representations trained or fine-tuned on MisLocus images, T1 and T2 are used for representation training, T3 for validation, and T4 is never used during representation learning. CellProfiler, frozen SubCell-portable, and MorphEM embeddings do not require MisLocus-specific representation training, but their downstream benchmark scores are still computed on the same T4 plates.

All reference–variant detection results are therefore reported on T4 only. In the supervised paradigm, XGBoost classifier (Chen & Guestrin, 2016) is used only as a downstream evaluator. For each reference–variant pair and representation channel, the classifier is trained on the non-T4 replicate plates from the same platemap and evaluated on cells from the matched T4 plate. The reported AUROC asks whether the variant–reference separation learned from replicate plates generalizes to an independently prepared and imaged replicate of the same protein pair.

In the unsupervised paradigm, no classifier is trained. For consistency with the supervised held-out setting, `mAP_norm` is computed using only profiles from T4 plates for each eligible reference–variant pair. The resulting `mAP_norm` asks whether the T4 representation separates variant profiles from their matched reference profiles. Thus, both supervised and unsupervised scores are reported on the same held-out technical replicate, while preserving the distinction between a supervised decision-boundary test and an unsupervised ranking-based test.

A.3.3 REFERENCE-VARIANT MISLOCALIZATION DETECTION

The reference-variant axis evaluates whether each representation detects a reproducible imaging difference between cells expressing a variant protein and cells expressing the matched reference protein. Only variants with a matched reference protein from the same imaging batch after quality control are eligible for this paired evaluation. Variants without a retained reference are excluded from this axis but remain available for representation learning and other non-paired analyses.

Supervised XGBoost imaging phenotype score. For each eligible reference–variant pair, representation channel, batch, and fold, we train a binary XGBoost classifier to distinguish variant from reference single-cell profiles. The classifier uses a logistic objective, 150 boosted trees, learning rate 0.05, the histogram tree method, and `scale_pos_weight` set to the negative-to-positive class ratio in the training fold.

We skip classifiers whose training split contains only one class or whose training class imbalance exceeds $100{:}1$. For per-variant aggregation, we further exclude classifiers with training imbalance greater than $3{:}1$ to reduce the chance that the model learns cell-count or sampling artifacts rather than a localization phenotype. The per-variant supervised imaging phenotype score is the held-out T4 AUROC score for classifier trained on the T1—T3 imaging plates.

Unsupervised `mAP_norm` imaging phenotype score. The unsupervised paradigm evaluates whether the native geometry of a representation places variant profiles closer to replicate profiles of the same variant than to profiles of the matched reference protein. We use the copairs retrieval framework (Kalinin et al., 2025) to compute average precision under user-defined positive and negative pair rules. Before retrieval, single-cell profiles are aggregated to per-field-of-view median profiles, reducing single-cell noise while preserving well and plate structure.

For each variant query, positive pairs are profiles from the same variant protein measured on different plates. Negative pairs are profiles from the reference gene on the same plate, so the negative partners for a variant query are matched reference-protein profiles rather than unrelated proteins. Reference profiles are assigned unique reference indices to prevent copairs from constructing reference–reference positives. Variants without a same-batch retained reference allele are dropped from this analysis to avoid degenerate retrieval scores.

For each representation channel, copairs reports the raw mean average precision, a permutation-normalized score `mAP_norm`, and an empirical p-value. We use $10{,}000$ permutations with `seed=42`, followed by Benjamini–Hochberg correction at $\alpha = 0.05$. The primary unsupervised phenotype score is `mAP_norm`; the BH-corrected retrieval p-value supplies the significance criterion that enters the unsupervised hit definition below.

Control-null calibration and hit calling. A high imaging phenotype score from either the supervised or unsupervised regime does not by itself prove variant-induced mislocalization, because a classifier can also distinguish technical differences between replicate wells. We therefore calibrate hit calls against empirical control-null distributions. For each batch, representation, and feature channel, we apply both the supervised and unsupervised evaluations on the same-allele control comparisons: cells from replicate wells of the same control allele are treated as the two classes. These control–control comparisons preserve plate, well-position, and staining variation but should not contain a variant-induced localization change.

The 95th percentile of the resulting control-null score distribution—computed separately for AUROC and `mAP_norm`—defines the channel-matched hit threshold for that batch and representation. The supervised and unsupervised hit definitions then differ in whether significance testing is also required:

Supervised hit: a reference–variant pair is called a supervised hit in a batch when its held-out AUROC exceeds the 95th-percentile control-null AUROC threshold for that representation channel.

Unsupervised hit: a variant is called an unsupervised hit in a batch when (i) its `mAP_norm` exceeds the 95th-percentile control-null `mAP_norm` threshold for that representation channel, *and* (ii) its BH-corrected `mAP_norm` permutation p-value is below 0.05.

Batch-level hit indicators are then collapsed across biological-replicate pairs using the aggregation rule reported in the main hit-rate table. Reporting both definitions allows the benchmark to distinguish discriminative signal recoverable by a supervised model from signal already organized in the metric geometry of the representation.

A.3.4 Clinical agreement

The clinical-agreement axis asks whether imaging phenotype scores are larger for variants annotated or predicted to be pathogenic than for variants annotated or predicted to be benign. This axis does not define the primary mislocalization hit threshold; instead, it evaluates whether the continuous imaging scores and binary hit calls align with external clinical or computational annotations.

ClinVar grouping and continuous-score comparison. ClinVar annotations are collapsed into a coarse two-class comparison by merging Pathogenic and Likely Pathogenic into a Pathogenic group, and Benign and Likely Benign into a Benign group. Variants with conflicting, uncertain, or missing ClinVar annotations are excluded from this two-class comparison. Pathogenic is treated as the positive class and metrics are computed on the ClinVar-annotated subset only. For each representation and channel, we compare the per-variant supervised AUROC or unsupervised `mAP_norm` between the Pathogenic and Benign groups using a two-sided Mann–Whitney U test. Multiple testing is controlled with Benjamini–Hochberg correction across representation–channel comparisons.

Hit-call enrichment. As a discrete companion to the continuous-score comparison, we treat each representation's binary hit call as a classifier of ClinVar Pathogenic versus Benign status. For each representation and channel, we report Precision, Recall, and F_1 on the ClinVar-annotated subset, along with a one-sided Fisher's exact test for enrichment of hit calls among Pathogenic variants. Fisher p-values are corrected with Benjamini–Hochberg across representation–channel comparisons. This analysis asks whether variants called as imaging hits are enriched for clinically pathogenic annotations.

Variant effect predictor agreement. To extend the clinical-agreement analysis beyond the subset of variants with ClinVar labels, we repeat the same comparison using sequence-based variant-effect predictors extracted from dbNSFP: AlphaMissense, REVEL, MutPred2, and ESM-1b. Each predictor's continuous score is binarized into Pathogenic-versus-Benign groups using the predictor-specific thresholds used in our annotation table. For each predictor, representation, and channel,

we run the same continuous-score comparison and hit-call enrichment analyses used for ClinVar. These analyses test concordance between image-derived localization phenotypes and sequence-based functional priors; they do not treat the sequence-based predictors as ground truth.

Per-location clinical decomposition. As a secondary analysis, we decompose ClinVar agreement by the annotated localization of the reference protein. Each variant is assigned to every HPA location label carried by its reference protein, so variants of multi-localizing proteins contribute to multiple location strata. Within each location, we compare Pathogenic versus Benign variants using the same imaging score and Mann–Whitney framework described above. Locations with fewer than three annotated reference proteins or fewer than three variants in either ClinVar group are excluded. This analysis tests whether clinical agreement is broadly distributed across compartments or driven by a small number of localization classes.

A.3.5 Reference-localization structure

The reference-localization axis evaluates whether each representation organizes reference proteins according to canonical subcellular localization. Unlike the reference-versus-variant axis, this analysis does not use variant alleles. It is computed on reference proteins only and asks whether proteins sharing an HPA location label are closer in representation space than proteins with different labels.

Reference-protein consensus profiles. For each representation and channel, we pool QC-approved cells from reference alleles across the six benchmark batches. Cell-level profiles are first aggregated within wells or fields of view as needed for the representation, and then median-aggregated to obtain one consensus profile per reference protein. For CellProfiler, where feature selection can differ slightly by batch, only feature columns available across the relevant batches are retained before constructing the consensus profiles.

HPA label curation. Reference proteins are annotated with subcellular localization labels from the Human Protein Atlas. We retain proteins with at least one HPA location label above the default reliability threshold used in the benchmark, and require each location to be supported by at least two genes for inclusion. Sensitivity analyses use more permissive and more stringent reliability thresholds. Because HPA annotations describe endogenous protein localization while MisLocus images use overexpression, the labels are treated as coarse compartment-level references rather than exact single-label ground truth.

Multi-label `mAP_norm` retrieval. We evaluate reference-localization structure using multilabel `mAP_norm` retrieval on the protein-by-feature consensus matrix. Positive pairs are different proteins sharing at least one HPA location label, whereas negative pairs are proteins sharing no HPA location label. Average precision is computed for each protein and aggregated by HPA location, yielding one retrieval score per location. A permutation null with $10{,}000$ shuffles and Benjamini–Hochberg correction is used to determine which locations show significant retrieval. The primary HPA heatmap reports `mAP_norm` for locations annotated to at least three reference proteins; locations with two proteins are excluded from the primary visualization because their retrieval scores are dominated by sample-size noise.

A.3.6 Implementation and reproducibility

All permutation analyses use `null_size=10000` and `seed=42`, and all multiple-comparison corrections use Benjamini–Hochberg at $\alpha = 0.05$. Output directories follow the representation and regime naming conventions used throughout the released benchmark pipeline. Released scripts write provenance sidecars containing command-line arguments, timestamp, git commit and dirty state, output file sizes, and Parquet row and column counts. Input paths and wall-clock duration are recorded only when supplied by the calling script; the central provenance log from the original artifact-generation environment is not part of the anonymous companion branch.

A.4 Supplementary analyses

This section provides supplementary analyses supporting the three evaluation axes in the main text. We first expand the reference-variant mislocalization detection results by reporting the com-

bined supervised and unsupervised hit-rate table using a more stringent hit-calling approach, the empirical control-null calibration used for hit calling, CellProfiler feature-attribution diagnostics, cross-representation agreement, and comparison to manually annotated mislocalization labels. We then provide additional clinical-annotation analyses, including top-ranked variant enrichment and sequence-based variant effect predictor agreement. These analyses are intended to document how the main-text conclusions depend on thresholding, representation geometry, and external annotation sources.

A.4.1 MISLOCALIZATION HIT CALLING ACROSS BIOLOGICAL-REPLICATE BATCHES

The main analysis calls a variant a hit when it passes the batch-specific criterion in at least one available batch. Here, a variant must have evaluable scores in both batches of at least one biological-replicate pair and pass the criterion in both; PA also requires within-batch BH significance in both batches. Per-batch scores and thresholds are unchanged, but requiring paired availability reduces the eligible set from 931 to 680 XGBoost variants and from 993 to 912 PA variants. Differences between the any-batch and both-batch tables therefore reflect both eligibility and the stricter hit rule.

Under this stricter definition, hit rates decrease for all representations, as expected, but the decrease is representation-dependent. CellProfiler shows a larger drop than the learned representations, indicating that a substantial fraction of CellProfiler hit calls are batch-specific under the 95th-percentile control-null calibration. This pattern is consistent with the control-null analyses above: CellProfiler contains explicit intensity and morphology measurements that can be highly discriminative across wells or batches, but some of this signal reflects technical variation rather than a reproducible variant-induced localization phenotype. Requiring agreement across both batches therefore disproportionately filters CellProfiler calls (Table 4).

By contrast, learned representations retain a larger fraction of their hits under the both-batch rule, suggesting that their high-scoring variant–reference separations are more often reproducible across biological-replicate batches. We therefore use the any-batch rule in the main text to maximize discovery sensitivity, while reporting the both-batch rule here as a conservative robustness check. Together, the two aggregation rules separate sensitivity from cross-batch reproducibility: any-batch measures whether a representation can detect a variant phenotype in at least one matched screen, whereas both-batch emphasizes phenotypes that are robust to batch-level technical and biological variation.

Table 4: **Mislocalization hit rates under stringent cross-batch aggregation.** For each biological-replicate batch pair, a variant is counted as a hit only if its held-out imaging phenotype score exceeds the 95th-percentile control-null threshold in **both** batches. This stricter rule emphasizes variant–reference separations that are reproducible across independently screened batches. The larger reduction in CellProfiler hit rates indicates that more of its high-AUROC calls are batch-specific, consistent with greater sensitivity to well- and batch-level technical variation.

		Mislocalization hits & Imaging phenotype scores							Mislocalization hits vs ClinVar classifications					
		XGB				PA			XGB			PA		
Representation	Channel	Hits	Hit rate	AUROC	Null p_{95}	Hits	Hit rate	mAP$_{\text{norm}}$	Precision	F_1	Fisher p_{BH}	Precision	F_1	Fisher p_{BH}
SC ft/ViT	EMBED	115	16.9%	0.807	0.930	176	19.3%	0.375	87.0%	0.353	0.025	89.9%	0.423	1e-4
SC ft/MAE	EMBED	122	17.9%	0.800	0.924	170	18.6%	0.379	85.9%	0.357	0.028	89.1%	0.396	1e-3
SC rbg/MAE	EMBED	90	13.2%	0.735	0.896	141	15.5%	0.290	85.7%	0.263	0.078	88.1%	0.338	0.005
SC rbg/ViT	EMBED	86	12.6%	0.736	0.907	130	14.3%	0.277	89.4%	0.264	0.025	91.5%	0.306	1e-3
MorphEM	ALL	70	10.3%	0.756	0.939	65	7.1%	0.164	76.5%	0.170	0.700	86.1%	0.159	0.123
MorphEM	GFP	155	22.8%	0.711	0.841	147	16.1%	0.286	91.2%	0.459	1e-4	86.0%	0.336	0.016
Cytoself	combined	154	22.6%	0.695	0.808	48	5.3%	0.131	90.7%	0.437	1e-4	90.0%	0.141	0.064
CellProfiler	ALL	74	10.9%	0.799	0.948	4	0.4%	0.004	73.7%	0.181	0.700	75.0%	0.017	0.755
CellProfiler	GFP	42	6.2%	0.799	0.954	12	1.3%	0.027	73.9%	0.116	0.700	87.5%	0.039	0.438

A.4.2 CONTROL-NULL DISTRIBUTIONS AND THRESHOLD ROBUSTNESS

Since imaging phenotype scores capture both technical variation and mislocalization effect simultaneously, we defined mislocalization hits relative to an empirical control-null distribution per batch rather than using a fixed AUROC cutoff. For each imaging batch and representation setting, we trained the same held-out XGBoost classifier on paired control–control wells corresponding to replicate measurements of the same reference protein. These comparisons preserve the plate, well, and imaging

structure of the reference–variant task but should not contain variant-induced localization changes. We then used the 95th percentile of the resulting control-null AUROC distribution as the batch- and representation-specific hit threshold, and called a reference–variant pair a hit only when its held-out AUROC exceeded this threshold. This procedure calibrates each representation against its own sensitivity to well-to-well technical variation.

Figure 5 illustrates this calibration for imaging Batch 13. CellProfiler features produce broad and high-valued control-null distributions, especially for GFP-channel and all-channel feature sets, yielding conservative thresholds near AUROC ≈ 1. In contrast, the learned embedding has a narrower control-null distribution and a lower empirical threshold, while still assigning high AUROC values to a subset of reference–variant comparisons. Thus, the hit-calling procedure does not assume that all representations have the same technical-null behavior; instead, it normalizes each representation by its own control-null distribution before comparing variant hit rates across models and batches.

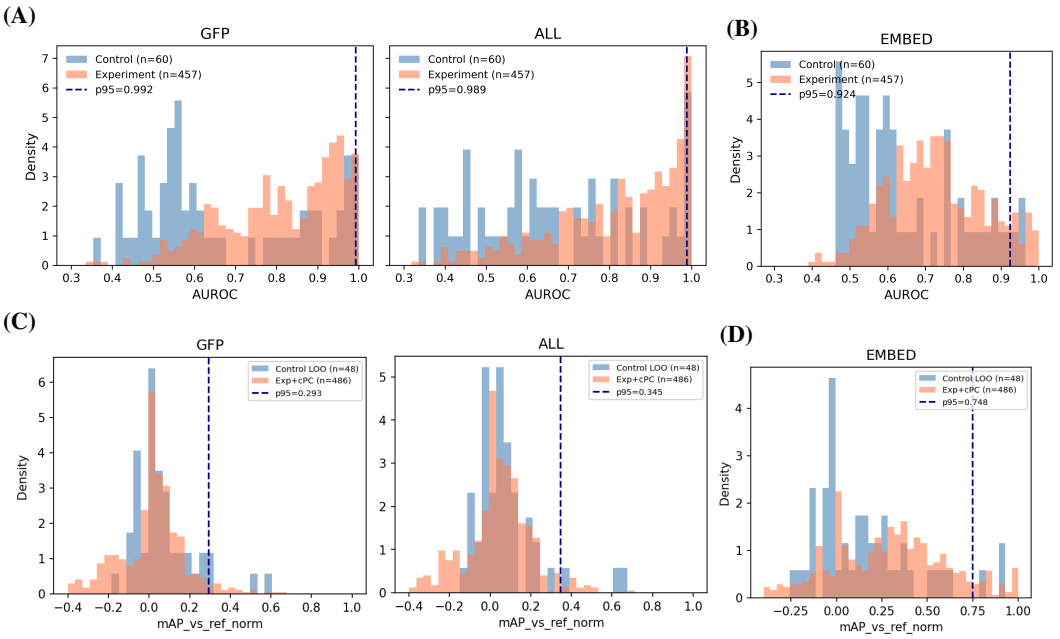

Figure 5: **Control-null imaging phenotype score distributions used for batch-specific hit calling.** For imaging Batch 13, held-out XGBoost classifiers were trained to distinguish paired single-cell profiles. Blue histograms show control–control comparisons between replicate wells of the same protein, which define the empirical technical-null distribution. Orange histograms show reference–variant comparisons used to score variant-induced localization shifts. The dashed vertical line marks the 95th percentile of the corresponding control-null distribution, used as the hit threshold for that batch and representation. **(A)** CellProfiler AUROC distributions using GFP-channel features only and all-channel features. **(B)** AUROC distribution for the SubCellPortable-rgb/mae embedding. **(C)** CellProfiler `mAP_norm` distributions using GFP-channel features only and all-channel features. **(D)** `mAP_norm` distribution for the SubCellPortable-rgb/mae embedding.

A.4.3 CellProfiler control-null signal is dominated by intensity features.

To identify which CellProfiler measurements drive high AUROC values in the control-null setting, we examined the Batch 13 T4-held-out XGBoost classifiers (Chen & Guestrin, 2016) for curated control well pairs and computed TreeSHAP attributions (Lundberg et al., 2020) in these classifiers. Because each classifier distinguishes single cells from two imaging wells of the same protein, any cross-plate discrimination reflects technical or well-position variation rather than variant-induced mislocalization. We aggregated per-feature mean absolute SHAP values across classifiers with AUROC ≥ 0.9, separately for protein-channel-only CellProfiler features (`GFP`) and the full CellProfiler feature set (`ALL`). In both settings, the dominant feature was `Cells_Intensity_MaxIntensityEdge_GFP`, with additional top-ranked features `Cells_Intensity_MinIntensityEdge_GFP` for the

protein channel and `Cells_Intensity_MinIntensityEdge_Mito` for across the channels (Figure 6). Thus, the CellProfiler control-null classifiers that achieve high held-out AUROC are driven primarily by intensity features across channels and related correlation features, consistent with reproducible well-level imaging-intensity confounding factors rather than a true localization phenotype.

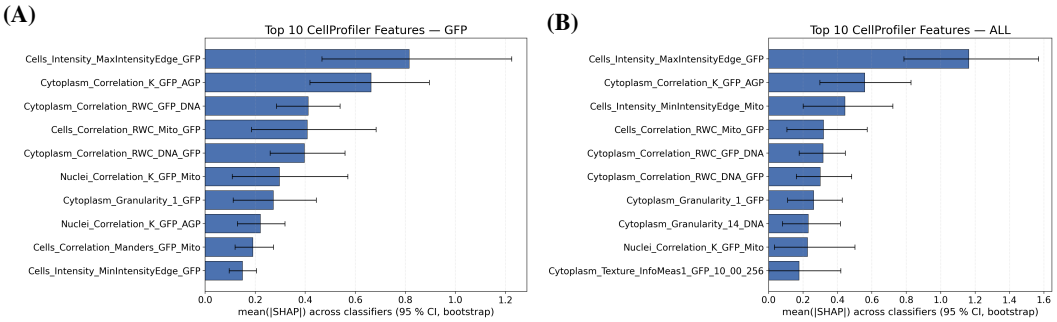

Figure 6: **CellProfiler control-null classifiers are dominated by protein-channel intensity features.** TreeSHAP feature attributions were computed for Batch 13 control well-pair XGBoost classifiers evaluated on the held-out T4 plate. Each classifier compares two wells of the same allele, so high held-out AUROC reflects reproducible technical or well-position variation rather than biological mislocalization. Bars show the top ten CellProfiler features ranked by mean absolute SHAP value across classifiers that achieved held-out AUROC ≥ 0.90; whiskers denote 95% bootstrap confidence intervals over classifiers. **(A)** Protein-channel-only CellProfiler features (GFP channel). **(B)** Full CellProfiler feature set (ALL channels). In both feature spaces, `Cells_Intensity_MaxIntensityEdge_GFP` is the strongest driver, and most top-ranked features directly involve the GFP channel, indicating that the high-AUROC control-null classifiers primarily exploit reproducible protein-channel intensity variation rather than a variant-specific localization signal.

A.4.4 Representation agreement across phenotype scores

To assess whether different image representations recover consistent variant-induced phenotypes, we compared per-variant imaging phenotype scores across representations. For each representation and channel setting, we computed Pearson correlations across matched reference–variant pairs using either supervised XGBoost AUROC or unsupervised `mAP_norm`. The supervised AUROC scores were highly concordant across representations, with especially strong agreement among SubCell-derived embeddings and substantial agreement between learned representations and CellProfiler features (Figure 7, left). This indicates that many high-scoring variant–reference shifts are reproducible across distinct feature families rather than being artifacts of a single representation. Agreement was lower, but still broadly positive, for `mAP_norm` (Figure 7, right), consistent with the greater sensitivity of distance-based retrieval metrics to each representation's metric geometry. Overall, these correlations show that the benchmark identifies a shared imaging phenotype signal across representations, while also highlighting metric- and representation-specific differences in how variant effects are ranked.

A.4.5 Additional mislocalization detection results

As an external sanity check on the learned imaging phenotype scores, we compared per-variant scores with manual mislocalization labels reported by Lacoste et al. (2024). This comparison is necessarily limited to the subset of variants overlapping the manually annotated set, and the labels are not used for model training or threshold selection. We therefore treat this analysis as an orthogonal validation of whether high imaging phenotype scores correspond to visually recognizable localization changes. Across representations, variants labeled as mislocalized shift toward higher supervised AUROC scores than variants not labeled as mislocalized, supporting the interpretation that the benchmark scores capture a biologically meaningful localization phenotype rather than only plate or well-position artifacts (Figure 8).

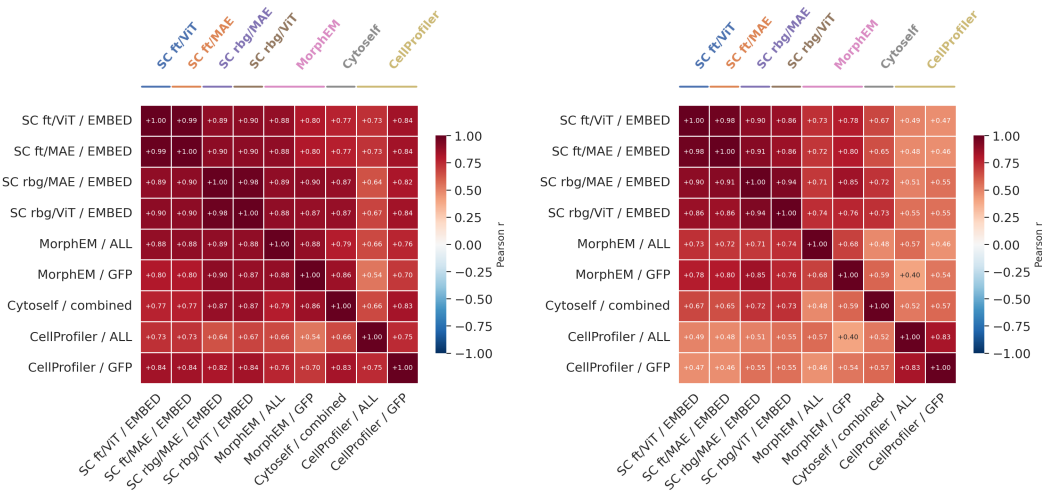

Figure 7: **Variant detection correlation across image representations.** Correlation between representation models and encoded channels for: **Left:** Per-variant XGBoost AUROC for discriminating variant from paired reference single cells. **Right:** Per-variant `mAP_norm` (retrieving same-variant cells over paired reference cells)

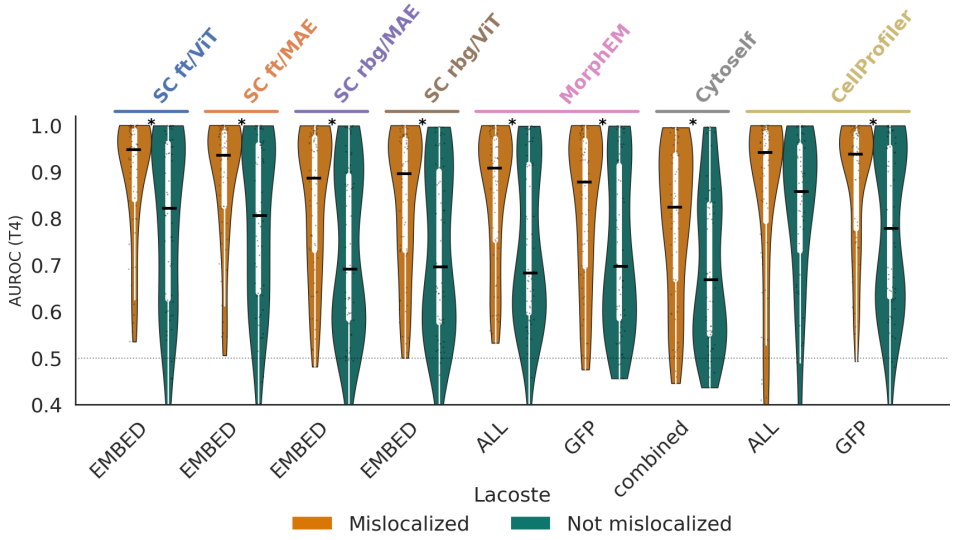

Figure 8: **XGBoost AUROC distributions on T4 split by Lacoste et al. (2024) mislocalization label.** For each (representation, channel) the held-out per-variant XGBoost AUROC is shown as a violin separately for variants labeled as mislocalized (orange) and not mislocalized (green) by manual annotations; the black bar marks the median. A star above the violin pair indicates a Mann–Whitney U test with BH-corrected $p < 0.05$ between the two groups. Across deep representations, Lacoste-positive variants consistently shift to higher AUROC than Lacoste-negative ones, indicating that imaging-derived perturbation signal tracks the orthogonal Lacoste mislocalization annotation. CellProfiler shows the smallest separation, consistent with its lower specificity at the strict 95th-percentile hit threshold (Table 1).

A.4.6 CLINVAR AGREEMENT

The main-text ClinVar analysis (§4.2) computes per-variant imaging phenotype scores using each representation's full feature space, and the unsupervised CellProfiler scores in particular show a markedly weaker pathogenic-vs-benign separation than the supervised XGBoost view. The motivation for the top-K analysis is to ask whether the variant-relevant signal is genuinely absent from CellProfiler or whether it is present but hidden among too many uninformative feature dimensions, with the bulk of the feature space contributing nuisance variance that washes out the localization signal in a global distance metric. We therefore ranked CellProfiler features by their importance in the per-variant XGBoost classifiers, retained the top K, and recomputed the unsupervised phenotype score and the corresponding ClinVar Pathogenic-vs-Benign comparison as a function of K.

Figure 9 shows the resulting pathogenic-vs-benign F_1 as K varies. Both CellProfiler channel settings show the same non-monotonic behavior: F_1 rises sharply as the feature set is reduced from the full panel down to a small top-K subset, peaking at $F_1 \approx 0.58$ for CellProfiler/GFP and $F_1 \approx 0.46$ for CellProfiler/ALL at $K \approx 10$–25, and decays back to the full-feature baseline ($F_1 \approx 0.10$) once K exceeds 100. This pattern indicates that the ClinVar-relevant signal is present in the CellProfiler embedding but is diluted by the surrounding uninformative dimensions: a global PA retrieval metric averages over them and underestimates the phenotype, whereas conditioning on a small set of variant-discriminative features recovers most of the signal that the supervised XGBoost classifier already exploits. The top-K view therefore reinforces the interpretation in §4.2: the supervised-vs-unsupervised gap for CellProfiler reflects feature-space dilution rather than missing biological information.

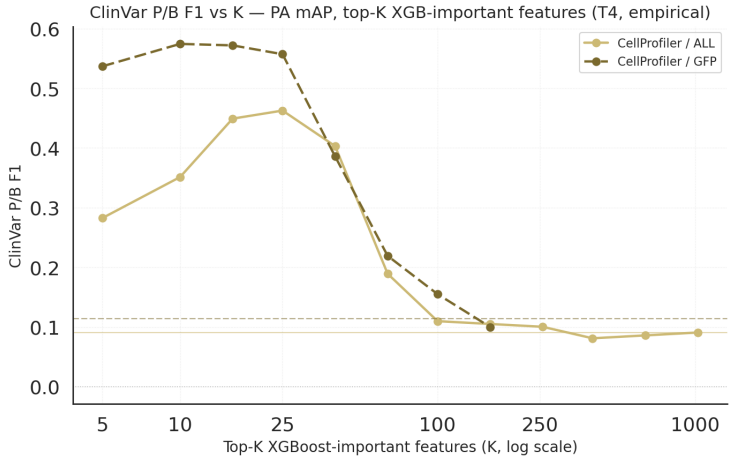

Figure 9: **ClinVar pathogenic-vs-benign F_1 as a function of the number of top-K XGBoost-important CellProfiler features used to compute the unsupervised phenotype score.** For each K, the CellProfiler feature set is restricted to the top-K features ranked by XGBoost importance in the per-variant variant-vs-reference classifiers; the unsupervised phenotypic-activity (PA) `mAP_norm` is recomputed on this restricted feature space and converted to a binary hit indicator using the same control-null calibration as in the main text. The hit indicator is then treated as a pathogenic-vs-benign classifier and we report its F_1. Solid line: CellProfiler/ALL; dashed line: CellProfiler/GFP. Horizontal reference lines mark the full-feature-set F_1 for each channel setting. Both curves peak at $K \approx 10$–25 ($F_1 \approx 0.58$ for GFP and ≈ 0.46 for ALL) and decay back toward the full-feature baseline ($F_1 \approx 0.10$) by $K \gtrsim 100$, indicating that the ClinVar-relevant signal is present in CellProfiler.

A.4.7 SEQUENCE-BASED VARIANT-EFFECT PREDICTOR BENCHMARK

To test whether imaging-derived phenotype scores align with computational variant-effect predictions beyond ClinVar, we repeated the pathogenic-versus-benign comparison using four variant-effect predictors. Each predictor was binarized using its published pathogenicity thresholds, and the same enrichment and score-separation tests used for ClinVar were applied independently for each predictor. This analysis is not intended to treat sequence-based predictors as ground truth; instead, it asks

Table 5: **Pathogenic/benign hit-call enrichment across four sequence variant-effect predictors under stringent cross-batch aggregation.** For each biological-replicate batch pair, a variant is counted as a hit only if its held-out imaging phenotype score exceeds the 95th-percentile control-null threshold in **both** batches. This stricter rule emphasizes variant–reference separations that are reproducible across independently screened batches. As in main Table 2, each row is one (representation, channel) combination. Columns are split into supervised XGBoost (XGB, left) and unsupervised phenotypic activity (PA, right) — each using labels derived from four predictors. Per cell we report the F_1 score of the binary hit indicator treated as a pathogenic-vs-benign classifier, and the BH-corrected one-sided Fisher p-value for hit enrichment among pathogenic alleles.

		XGB								PA							
		AlphaMissense		ESM-1b		REVEL		MutPred2		AlphaMissense		ESM-1b		REVEL		MutPred2	
Representation	Channel	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}	F_1	p_{BH}
SC ft/ViT	EMBED	0.390	1e-8	0.365	1e-6	0.332	1e-4	0.407	1e-5	**0.457**	1e-8	**0.388**	1e-4	**0.364**	1e-6	**0.457**	1e-7
SC ft/MAE	EMBED	0.393	1e-7	0.361	1e-4	0.335	1e-3	0.408	1e-4	0.422	1e-8	0.381	1e-4	0.338	1e-4	0.444	1e-6
SC rbg/MAE	EMBED	0.310	1e-5	0.286	1e-4	0.225	0.011	0.316	1e-3	0.365	1e-7	0.320	0.001	0.271	1e-3	0.368	1e-5
SC rbg/ViT	EMBED	0.318	1e-7	0.293	1e-6	0.227	0.002	0.317	1e-4	0.341	1e-6	0.280	0.010	0.233	0.004	0.334	1e-3
MorphEM	ALL	0.247	1e-4	0.248	1e-4	0.158	0.072	0.273	1e-4	0.209	1e-3	0.177	0.010	0.121	0.012	0.206	1e-3
MorphEM	GFP	**0.535**	1e-8	**0.491**	1e-8	**0.457**	1e-8	**0.499**	1e-5	0.369	1e-5	0.329	0.012	0.295	0.002	0.394	1e-5
Cytoself	combined	0.509	1e-8	0.468	1e-8	0.433	1e-6	0.464	1e-4	0.118	0.028	0.111	0.090	0.089	0.073	0.130	0.009
CellProfiler	ALL	0.232	1e-4	0.243	1e-3	0.180	0.018	0.263	0.003	0.011	0.824	0.012	0.637	0.021	0.213	0.005	0.791
CellProfiler	GFP	0.151	0.001	0.158	1e-4	0.137	0.050	0.138	0.113	0.038	0.756	0.031	0.637	0.050	0.213	0.028	0.791

whether microscopy-derived localization phenotypes recover a signal that is broadly concordant with established sequence-based functional priors, and whether any representation-specific differences observed with ClinVar persist when the annotation source changes. Table 5 reports the same enrichment summary as main Table 2 but under the stricter cross-batch `both-batch` aggregation rule, so that variants count as hits only when the imaging phenotype reproduces in both batches of a biological-replicate pair.

As expected from §A.4.1, F_1 scores drop and most of the BH-corrected Fisher p-values weaken or stay the same relative to the `any-batch` version, but the qualitative ordering of representations and predictors is mostly preserved. Under XGBoost, most deep representations remain significantly enriched for pathogenic alleles across all four predictors, with AlphaMissense the most concordant predictor with imaging phenotype scores and REVEL the least. Under unsupervised phenotypic activity, the SubCell fine-tuned variants and MorphEM-GFP retain the strongest signal, while CellProfiler reproduces the supervised-vs-unsupervised asymmetry described in §4.2: significant under XGBoost across all four predictors but collapsing under PA.

A.4.8 Additional reference-localization structure analyses

Table 6: **Image representation similarity for reference proteins annotated in various locations.** Per-representation location-level `mAP_norm` measuring how well each representation places profiles of reference proteins annotated to the same Human Protein Atlas (HPA) location close together in feature space. Rows are representation / channel pairs (GFP/ALL channel(s) for CellProfiler and MorphEM); columns are aggregated results across HPA locations, restricted to locations with at least three proteins (counts above each column).

Representation / Channel	n_{org}	$\overline{\text{mAP}_{\text{norm}}}$	$\overline{\text{mAP}^{\text{sig}}_{\text{norm}}}$	n_{sig}	Sig. %
SC ft/ViT / EMBED	21	0.079	0.171	8	38.1%
SC ft/MAE / EMBED	21	0.078	0.171	8	38.1%
SC rbg/MAE / EMBED	21	0.063	0.168	5	23.8%
SC rbg/ViT / EMBED	21	0.063	0.141	8	38.1%
MorphEM / ALL	21	0.051	0.120	4	19.0%
MorphEM / GFP	21	0.053	0.156	5	23.8%
Cytoself / combined	21	0.034	0.111	2	9.5%
CellProfiler / ALL	21	0.042	0.119	1	4.8%
CellProfiler / GFP	21	0.061	0.137	8	38.1%

Table 6 reports the per-representation summary of the HPA reference-localization analysis underlying Fig. 3. For each (representation, channel), we list the mean multi-label `mAP_norm` across the 21 HPA locations annotated to at least three reference proteins, the mean restricted to BH-significant locations, and the count and fraction of significant locations. The SubCell fine-tuned heads achieve the highest

aggregate retrieval (mean mAP_norm of 0.079 for SC ft/ViT and 0.078 for SC ft/MAE), and their significant-only means of 0.171 indicate that on the locations they do resolve, they cluster reference proteins very tightly. Two further patterns are worth noting. CellProfiler/GFP is competitive with the SubCell portable variants on aggregate mAP_norm (0.061 vs 0.063) and is tied with SC rbg/ViT for the largest number of BH-significant locations ($8/21$, 38.1%), whereas CellProfiler/ALL performs badly at 0.042 mean and a single significant location, indicating that the all-channel CellProfiler features actively dilute the GFP-channel localization signal rather than augmenting it. Cytoself, by contrast, returns the lowest mean mAP_norm (0.034) and only 2 significant locations, an inversion relative to its strong performance on variant discrimination in §4.1.