PERTURB-PATH: BENCHMARKING H&E PATHOLOGY FOUNDATION MODELS WITH *in vivo* GENETIC PERTURBATIONS IN A LUNG CANCER MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology foundation models (PFMs) trained on datasets of histopathology images can learn robust patient representations that transfer to downstream disease modeling and drug response prediction tasks. While PFMs are trained on large amounts of unlabeled histopathology data, the scarcity of meaningful clinical labels forces these models to be evaluated on much smaller cohorts. Moreover, the labels in these evaluation cohorts are often noisy and corrupted by hard-to-resolve confounded factors. To mitigate these difficulties, we sought to create a PFM evaluation dataset with rich biological diversity, carefully-measured labels, and biological replicates. We leveraged Perturb-map to generate hundreds of genetically diverse tumors in a mouse model of lung cancer with paired hematoxylin & eosin (H&E) and multiplex immunofluorescence (mIF) imaging readouts. The resulting Perturb-Path dataset includes 84 knockout perturbations across 32 mice, resulting in 19,018 identified tumors with 116,383 H&E patches. We release a new PFM benchmark suite around this dataset and assessed eleven publicly-available PFMs and three ImageNet-based baselines. The benchmark suite evaluates the models on their ability to represent a range of phenotypic factors, including retrieval of genotypic perturbation labels and prediction of tumor microenvironment composition derived from the paired mIF readouts. This is, to our knowledge, the first publicly-available pathology benchmark dataset involving gene perturbations. We find that PFMs varied in their ability to retrieve specific perturbations, suggesting the presence of a genetically-associated phenotype that is captured by some models but not others. Further, we identify differences in ability to predict macrophage fraction versus T cell fraction across models, highlighting specific areas of improvement for model development. Overall, we found the PFMs outperform the ImageNet baselines, indicating that training on histopathology data of human origin translates to improvement on tasks in a mouse model.

1 BENCHMARKING PATHOLOGY FOUNDATION MODELS

Pathology foundation models (PFMs) are neural networks trained on large collections of histopathology images (Veremis et al., 2026; Ochi et al., 2025). Vector embeddings from PFMs can be extracted and aggregated at different visual scales to capture different scales of biology. Ideally, these embeddings encapsulate representations of patient biology that are useful for clinically important tasks such as classifying tumor mutations or predicting which patients will respond to a treatment. However, since useful and reliable labels in biomedicine are scarce or expensive to collect, PFMs are trained in a self-supervised manner (Balestriero et al., 2023; Oquab et al., 2024), allowing them to leverage vast quantities of unlabeled digital pathology data. Thus, benchmarks with high-quality labels are needed to assess PFM performance on relevant tasks, similar to how large language models (LLMs) are benchmarked.

The last several years have witnessed the birth of many PFMs and corresponding PFM benchmarks, which have focused mainly on oncology applications. PathBench (Ma et al., 2025b) evaluates 19 PFMs on over 15,000 histopathology whole-slide images (WSIs) across 5 different cancer types on histological subtyping, molecular subtyping, and survival prediction tasks. In contrast to using WSIs, THUNDER evaluated 23 models across 16 histopathology datasets at the tile level, which enabled

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.

evaluations that were independent of aggregation strategies (Marza et al., 2025). (Bareja et al., 2026) compares 31 PFMs in curated tasks from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Zhang et al., 2014; Vasaikar et al., 2019; Satpathy et al., 2025), finding surprisingly that vision-language PFMs often lag behind vision-only PFMs, and model size is not necessarily correlated with downstream performance. Meanwhile, PathoROB (Kömen et al., 2026) compares the robustness of 20 PFMs to spurious experimental confounders, such as scanner and stain intensity, across 34 different medical centers. Some PFM benchmarks even introduce additional modalities: HEST (Jaume et al., 2024) and SpaPath-Bench (Zhao et al., 2026) pairs histopathology images with spatial transcriptomics to evaluate the abilities of PFMs to predict spatial gene expression patterns and identify useful biomarkers. PFM benchmarks can also help us glean insights into why certain PFMs perform well and how to improve performance further: in benchmarking 19 PFMs on morphology and prognosis prediction tasks with samples from 6,818 patients across 4 cancer types, Neidlinger et al. (2026) discovered that ensembling PFMs could outperform individual models on a majority of their benchmark tasks.

All the benchmark tasks mentioned above only involve human data[1]. Most PFM applications revolve around treating human patients, so benchmarking on human data might appear entirely advantageous. However, human data in this context poses several statistical complications: the number of patients is often small due to the difficulty of acquiring labeled patient data; there are few (if any) biological replicates; the readouts to be predicted are often noisy or arbitrarily dichotomized, e.g. responder vs. non-responder; and there are many confounding factors such as tumors being resected at different timepoints in the course of treatment or variance in staining or imaging protocols across clinical sites (Goh et al., 2017). These factors make it hard to decouple model performance from data quality.

We address these limitations by leveraging an engineered biological dataset from an *in vivo* mouse model of lung cancer. This dataset was generated with Perturb-map (Dhainaut et al., 2022) and features thousands of unique tumors with nearly a hundred distinct genotypes. To the best of our knowledge, this is the *first pathology benchmark dataset* involving *genetic perturbations*. Then, we develop a suite of benchmark tasks leveraging this dataset, which allows us to quantify 1) whether PFMs can identify different perturbations from H&E, and 2) whether PFMs can map phenotypes visible in H&E to tumor microenvironment (TME) state, which plays a critical role in cancer progression and metastasis (Visser and Joyce, 2023).

2 THE PERTURB-PATH DATASET

To build the Perturb-Path dataset, we leveraged Perturb-map, an *in vivo* spatial functional genomics technology. In contrast to other approaches (Shi et al., 2023), Perturb-map identifies genetic perturbations directly in tissue sections, preserving the spatial architecture and unlocking multimodal spatial profiling of genetic perturbations at scale (Dhainaut et al., 2022). Perturb-map relies on Protein Barcodes (Pro-Codes), a set of protein reporters obtained from the combinatorial arrangement of triplets of short epitope tags onto a scaffold protein (Wroblewska et al., 2018). Each Pro-Code is associated with a specific CRISPR gRNA and serves as the molecular proxy to identify the genetic perturbation carried by each cell (Figure 1 (a)). Crucially, the Pro-Code identity of each cell can be determined with non-destructive spatial methods, such as multiplex immunofluorescence (mIF), which can be run prior to H&E staining on the same slide.

We applied Perturb-map to the KP mouse model of non-small cell lung cancer (NSCLC), which is characterized by an activating mutation in *Kras* and a functional knockout of *Tp53* (Wieckowski et al., 2015). KP cells were first engineered *in vitro* with Pro-Code/CRISPR libraries targeting genes that are (i) mutated in NSCLC, (ii) whose expression is dysregulated in NSCLC and/or (iii) part of major signaling pathways involved in tumor or immune biology. Each library contained a maximum of 56 unique KP cell populations, representing up to 56 gene knockout perturbations per experiment. The resulting cell pools, injected into the tail vein of immunocompetent mice, seed the lungs where they develop spatially distinct tumor lesions (Figure 1 (a)). Mice were further randomized between two treatment groups, each animal receiving either a control (condition 1) or a therapeutic treatment (condition 2). At the experiment endpoint, lung tissue was collected, dissected into the 5 distinct lobes, fixed and embedded in paraffin for sectioning.

[1]While the HEST-1k dataset includes murine samples, these data are not part of the HEST-Benchmark itself.

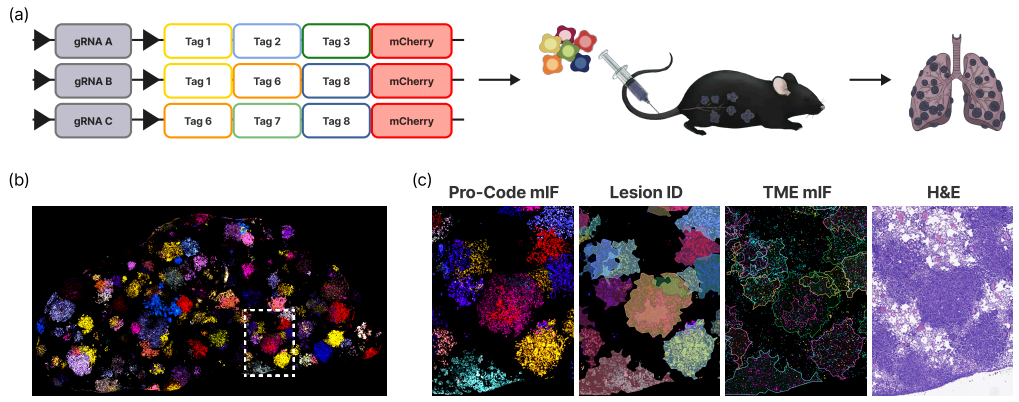


Figure 1: **High-throughput *in vivo* perturbation with Perturb-map is compatible with multimodal spatial phenotyping** (a) Protein barcodes are built by combinatorial arrangement of short epitope tags on a scaffold protein (mCherry). Each protein barcode is linked to a specific CRISPR perturbation (gRNA). Cells engineered with Pro-Code/gRNA pairs are injected into mice, where they will seed the lung and initiate distinct tumor lesions. (b) mIF of one lobe of mouse lung, seeded with Pro-Code labeled tumors. Pro-Codes (and their associated genetic perturbation) are visualized by mIF using antibodies specific to the short epitope tags. Each mouse samples consists of 5 lung lobes. (c) Magnification of the white rectangle in b. Perturb-map is compatible with multimodal phenotyping on the same slide. Pro-code mIF defines tumor lesion borders and genetic perturbation ID. Each slide is also stained with an antibody panel quantifying the tumor microenvironment (TME) composition and with H&E.

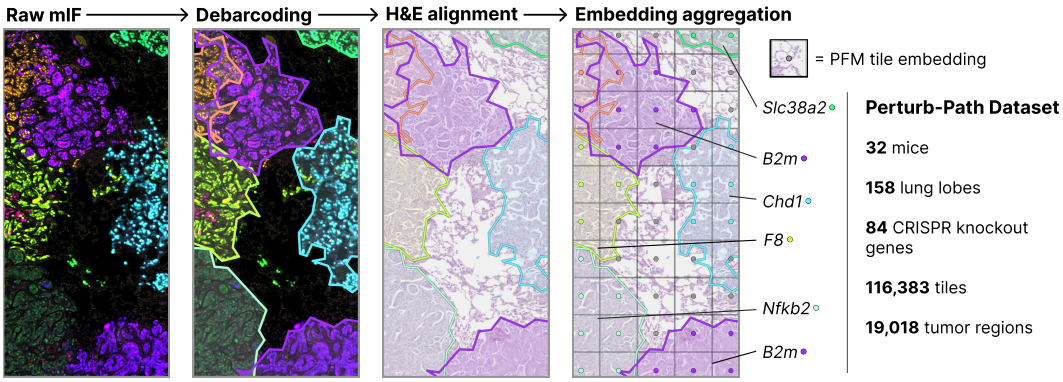


Figure 2: Overview of dataset construction for the Perturb-Path eval, starting from the Pro-Code readouts from mIF to the final aggregated tumor embeddings.

Lung tissue sections were profiled using multiplex immunofluorescence (mIF), and subsequently stained with H&E on the *same slide* (Figure 1 (b) and (c)). The mIF served two purposes: identifying the Pro-Code (and by extension, gene knockout) carried by each KP cell and profiling the tumor microenvironment, including the identification of immune cells. (See Section A.1.1 for more details on debarcoding and cell type labeling and quantification). Pro-Code staining revealed that each individual tumor lesion is clonal, i.e. originates from a single KP cell and presents a unique genetic perturbation (as shown on a representative lung lobe on Figure 1 (b)). This finding validates that each tumor lesion can be analyzed independently, and that from a single tissue slide, we access the H&E stain of hundreds of individual tumor lesions, each carrying a specific gene perturbation.

The resulting Perturb-Path dataset covers 84 perturbations distributed into 2 different libraries, each inoculated in a single cohort of animals. For each library and treatment condition, 8 mice with a consistently high tumor burden were selected, bringing the total dataset to 32 mice, and 19,018 independent tumor lesions.

Table 1: Description of models. All models except the ResNet50 take in $224\,\text{px} \times 224\,\text{px}$ tiles ($= 112\,\mu\text{m}$ at $0.5\,\mu\text{m/px}$).

		Patch tokenization			
Model	d	Tokens	Grid	Size (px)	License
Pathology foundation models					
GigaPath (Xu et al., 2024)	1536	196	14×14	16	Apache 2.0
GigaPath-Flash (Usuyama et al., 2026)	384	196	14×14	16	Apache 2.0
UNI (Chen et al., 2024)	1024	196	14×14	16	CC-BY-NC-ND 4.0
UNI2-h (Chen et al., 2024)	1536	256	16×16	14	CC-BY-NC-ND 4.0
Virchow (Vorontsov et al., 2024)	2560[a]	256	16×16	14	Apache 2.0
Virchow2 (Zimmermann et al., 2024)	1280 / 2560[a]	256	16×16	14	CC-BY-NC-ND 4.0
Hibou-L (Nechaev et al., 2024)	1024	256	16×16	14	Apache 2.0
H-Optimus-0 (Saillard et al., 2024)	1536	256	16×16	14	Apache 2.0
Midnight-12k (Karasikov et al., 2025)	1536	256	16×16	14	MIT
OpenMidnight (Kaplan et al., 2025)	1536	256	16×16	14	Apache 2.0
GPFM (Ma et al., 2025a)	1024	256	16×16	14	MIT
ImageNet baselines					
ViT-Base (Dosovitskiy et al., 2021)	768	196	14×14	16	Apache 2.0
ViT-Huge (Dosovitskiy et al., 2021)	1280	256	16×16	14	Apache 2.0
ResNet50[b] (He et al., 2015)	2048	49	7×7	–	Apache 2.0

[a] CLS token concatenated with the mean patch token.
[b] Convolutional backbone, so the 7×7 grid is the final feature map, not ViT patch tokens.

3 Models and Extracting Embeddings

We chose 11 state-of-the-art PFMs, as shown in Table 1, based on accessibility, compatibility with 224px tile inputs, and performance on existing PFM benchmarks (Ma et al., 2025b; Marza et al., 2025). All of the PFMs evaluated here are built on the same ViT (Dosovitskiy et al., 2021) architecture. Tile embeddings for these models were computed using TRIDENT (Zhang et al., 2025; Vaidya et al., 2025), a standardized toolkit for processing whole-slide images. To quantify the benefit of training on histopathology data, we also added three baselines trained on natural images from ImageNet (Deng et al., 2009) rather than histopathology slides: ViT-Base, ViT-Huge, and ResNet50.

As seen in Figure 2, the ViT operates on crops of H&E images called tiles (e.g., 224 px by 224 px). Each tile is converted into a grid of smaller patches (e.g., 16 px by 16 px). As part of its processing, the ViT model computes patch-level and tile-level embeddings (via a CLS token). To aggregate this into a tumor-level representation, we performed mean-pooling of the tile embeddings for all tiles within a given tumor regions using the extracted tumor masks. For Virchow and Virchow2 models, we performed mean-pooling of the concatenation of the CLS token with the mean patch token, as recommended in TRIDENT.

4 The Perturb-Path Evaluations

The Perturb-Path benchmark features four tasks: `robustness`, `mAP`, `mIF`, and `tme-effect`.

4.1 `robustness`: Measuring embedding robustness to nuisance variables

This first task captures the robustness of the embeddings to nuisance variables driven by batch effects. We adapt the robustness index as defined by PathoROB (Kömen et al., 2026), which measures how robust a model is to batch effects while calibrating to the biological signal it captures. Given a perturbation g in a batch b, we construct a profile $q = (g, b)$ by averaging over all tumor embeddings corresponding to perturbation g in the given batch b. For a profile q, we construct the sets $P_{SO}(q)$ consisting of all profiles with the same perturbation g but from a different batch b', and $P_{OS}(q)$ consisting of all tumors with a different perturbation g' but from the same batch b. Let Q_b denote the set of all profiles in batch b with both sets non-empty, and let $\text{Emb}(q)$ be the d-dimensional embedding associated with profile q. Then, the robustness index computes the probability that the embedding of a uniformly randomly sampled profile p from $P_{SO}(q)$ will lie closer to q than a

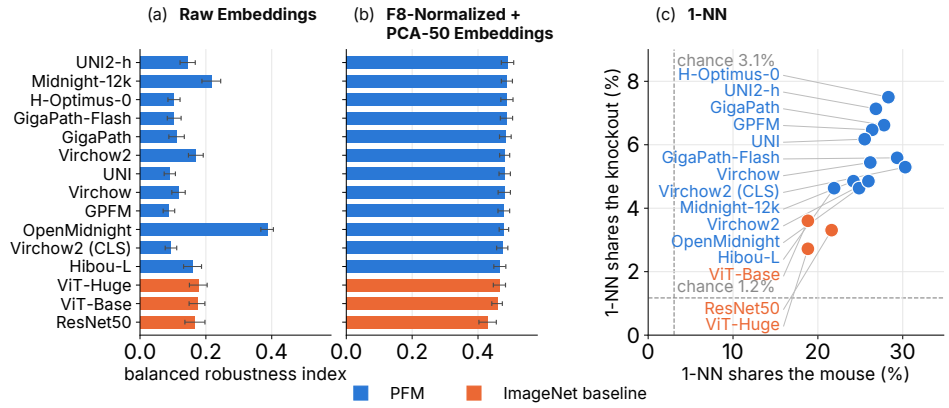


Figure 3: Robustness task performance across models. (a,b) Robustness index of each model with donor mouse as the batch variable using raw (a) vs *F8*-normalized + PCA-50 embeddings (b). (c) 1-NN rates of sharing a donor mouse vs 1-NN rates of sharing a perturbation.

uniformly randomly sampled profile p' from $P_{OS}(q)$, sampling q from Q_b uniformly and taking the mean over all b:

$$\text{RobustnessIndex} = \frac{1}{|B|}\sum_{b\in B}\frac{1}{|Q_b|}\sum_{q\in Q_b} P\left[\cos(\text{Emb}(q),\text{Emb}(p)) > \cos(\text{Emb}(q),\text{Emb}(p'))\right].$$

Our metric improves on the robustness measure introduced in (Kömen et al., 2026) because it is base-rate normalized; a random model would get a robustness index score of 0.5, no matter what $|P_{SO}(q)|$ and $|P_{OS}(q)|$ are for any q.

We instantiate the robustness index with the batch variable b as the individual mouse donor ID. Figure 3(a) shows that the robustness index for the raw embeddings is quite low due to experimental variability. Visually, a UMAP of the raw embeddings exhibits clustering by mouse donor (Figure 9 left panels). Leveraging the negative controls available in Perturb-map, we investigated a simple method to correct these batch effects for a given embedding Emb with dimension d. Given a mouse m, we compute the dimension-wise mean $\mu_{F8} \in \mathbb{R}^d$ and standard deviation $\sigma_{F8} \in \mathbb{R}^d$ over all tumor embeddings created by the knockout of the *F8* gene. *F8* is a blood coagulation factor which is not expressed in KP cancer cells. We normalize the tumor embedding of a given tumor q by z-scoring relative to *F8* tumors in the same mouse as q: $\text{Emb}_{F8}(q) \equiv (\text{Emb}(q) - \mu_{F8})/\sigma_{F8}$.

Figure 3 (b) demonstrates the improvements from *F8*-normalization combined with projection to the first 50 principal components (computed independently for each encoder) so that all encoders are compared at the same embedding dimension. After the normalization procedure, all models exhibit higher robustness with a clear improvement in PFMs relative to the ImageNet baselines. The resulting tumor embeddings show increased clustering by knockout and decreased clustering by mouse (Figure 9 right panels). However, all models perform worse than chance, indicating that the models still pick up on individual characteristics of the mice. Despite this, we find that the PFMs are still capable of identifying biologically relevant features, as described in the next tasks. Moreover, to maximize the signal of the perturbation relative to the batch effects, we utilize *F8*-normalization for all remaining tasks.

In Figure 3 (c), we look at two relevant quantities: 1) how often does the nearest neighbor (1-NN) of a profile share the same perturbation identity (i.e., in P_{SO}) and 2) how often does the nearest neighbor of a profile share the same batch variable (i.e., in P_{OS}). The robustness index normalizes by these rates, which makes it harder to see which features the models are picking up on. We see that the PFMs have both 1-NN rates significantly higher than the ImageNet baseline and chance, indicating that they are capturing both biological and batch signals.

4.2 mAP: Measuring retrieval accuracy of identical knockout tumors

Next, we evaluate how the models distinguish embeddings of each perturbation, which we formulate as a retrieval task. For the retrieval task, we compare the *F8*-normalized tumor embedding for each

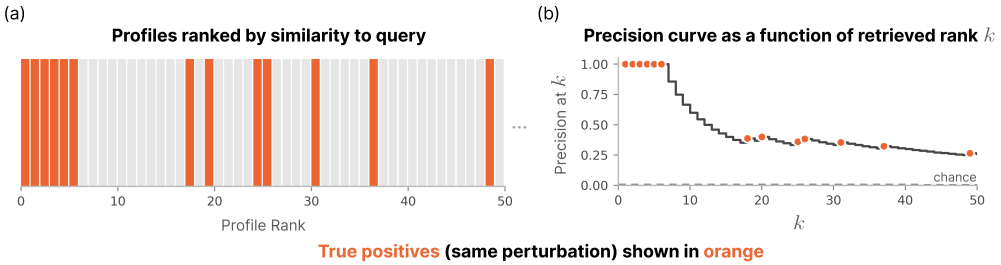


Figure 4: A visual description of the mAP task. For a given profile q, we sort all of the other profiles by cosine similarity to the embedding of q. Using the retrieved ranks of the true positives (shown in orange, indicating profiles with the same perturbation as q), we can obtain a precision curve as a function of rank k. The average precision (AP) of q is obtained by averaging the precision at the retrieved ranks of the true positives. The mAP is obtained by averaging the AP over all profiles q.

perturbation with that of other perturbations. Depending on the specific task, tumor embeddings can be aggregated within lobe (lobe-level), mouse (mouse-level), or not aggregated at all (tumor-level), to define an embedding profile. As shown in Figure 4, given a query embedding profile q, we rank all other embedding profiles p by cosine-similarity S and ask how many of the top-ranked profiles share the same perturbation. We measure retrieval performance using mean average precision (mAP) over all queries. If p_i denotes the profile at rank i after sorting in descending order by S, the precision P_k at rank k, the average precision AP and the mAP are defined as:

$$P_k(q) \equiv \frac{1}{k}\sum_{i=1}^{k} R(q,p_i) \quad \text{AP}(q) \equiv \frac{\sum_{k=1}^{|D|} P_k(q)\, R(q,d_k)}{\sum_{k=1}^{|D|} R(q,d_k)} \quad \text{mAP} \equiv \frac{1}{|Q|}\sum_{q\in Q} \text{AP}(q) \qquad (1)$$

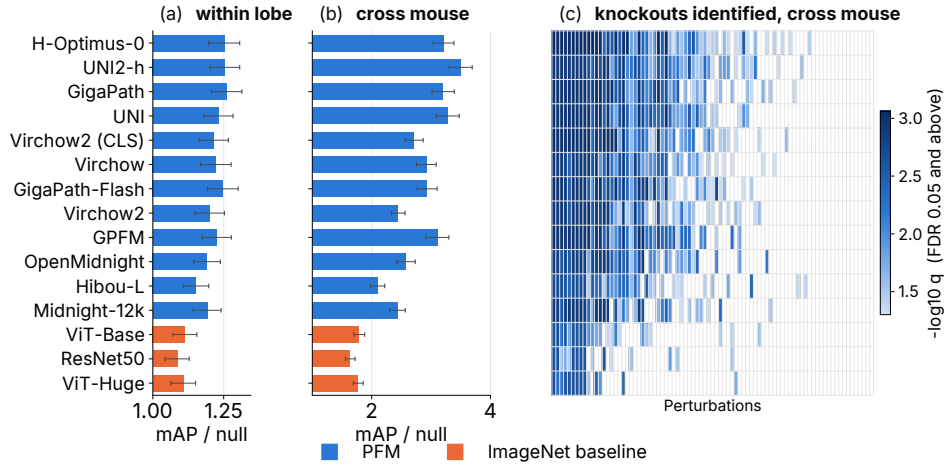


Figure 5: mAP task performance across models. (a) Within-lobe performance (b) Cross-mouse performance, (c) Heatmap across perturbations

First, we measured mAP within a lobe, using tumor-level embedding profiles queried against all other embedding profiles within the same lobe. This analysis removes the cross-lobe and cross-mouse nuisance variables at the cost of increased noise due to the lack of aggregation (Figure 5). Here we observed significant, but mild, enrichment across the models, with the pathology foundation models still performing better than the ImageNet baselines. We hypothesized that aggregation by mean-pooling across a lobe or mouse would further enhance the signal by reducing the inter-tumor variability. We calculated embedding profiles by mean-pooling the *F8*-normalized tumor embeddings within each mouse and use these as a cross-mouse mAP calculation (Figure 5(b)). We observed a higher mAP value in this setting, with several models achieving greater than 3-fold enrichment over the null, and observed a greater separation between the PFMs and the ImageNet baselines.

The retrieval task described here can also be formulated as a classification task using a linear probe. As seen in Section A.4, the model performance rankings using this approach remained generally consistent with the mAP model rankings.

We then asked whether certain perturbations were easier to retrieve than other perturbations, and whether this differs across models. We evaluated the average precision for each query perturbation, and evaluated the significance by a permutation test and multiple hypothesis correction with false discovery rate (FDR) (Figure 5(c), Section A.3). Across all models, a subset of genes were consistently significantly retrieved, and better performing models showed an increasing number of perturbations with significant retrieval. Interestingly, we did not observe that better performing models simply retrieved a strict superset of perturbations, suggesting that there are some model-dependent strengths related to specific perturbations. Finally, there were several genes that did not show significant retrieval across all models.

4.3 mIF: Predicting immune cell fractions in tumor regions

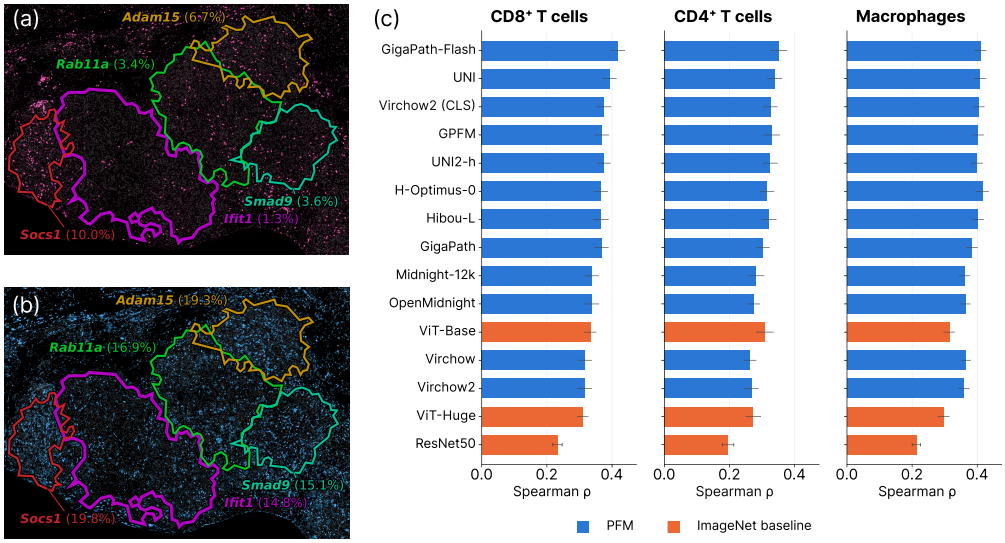


Figure 6: (a) Variance in T cell infiltration across tumors, with combined CD3 and CD8 IF signal shown in magenta. (b) Macrophage infiltration across the same tumors, with F4/80 signal shown in light blue. (c) Performance on mIF tasks: immune cell composition prediction on held-out mice.

The mIF task was designed to evaluate PFMs on their ability to predict the proportions of CD8⁺ T cells, CD4⁺ T cells, and macrophages using the H&E of the tumor regions. These proportions are derived from the mIF tumor microenvironment channels (Section A.1). Importantly, we find that neighboring tumors can have a drastically different immune composition, suggesting that the perturbation itself drives specific phenotypes and that each tumor region can be analyzed independently (Figure 6). Embeddings and cell fraction labels are both normalized by the mouse's *F8*-control tumor, with the *F8*-normalizations for the embeddings described above. For each model, we fit a ridge regression (with L_2 regularization factor $\alpha = 1$ and features standardized within each training fold) on five-fold cross-validation grouped by mouse, so that the predictions are performed in unseen mice. Performance is measured by Spearman correlation compared to ground-truth normalized cell type fractions in the out-of-fold predictions. We report the Spearman correlation on the pooled out-of-fold predictions, averaged over 5 repeats of 5-fold grouped CV. Standard error was measured using a leave-one-mouse-out jackknife procedure, recomputed on the retained mice from the same out-of-fold predictions.

All models exhibited significant performance on this task, and the top models exhibited a Spearman rho above 0.3 for all cell types (Figure 6(c)). Again, the PFMs generally performed better than the ImageNet baselines in this task. However, the degree of separation was more mixed, with some models performing worse at CD8⁺ and CD4⁺ T cell prediction. Further, we observed better overall predictions for macrophage, with greater separation between the PFMs and ImageNet baselines,

suggesting that pre-training on histopathology data enables improved prediction of macrophage populations from the H&E.

4.4 `tme-effect`: Measuring whether embeddings capture known TME biology

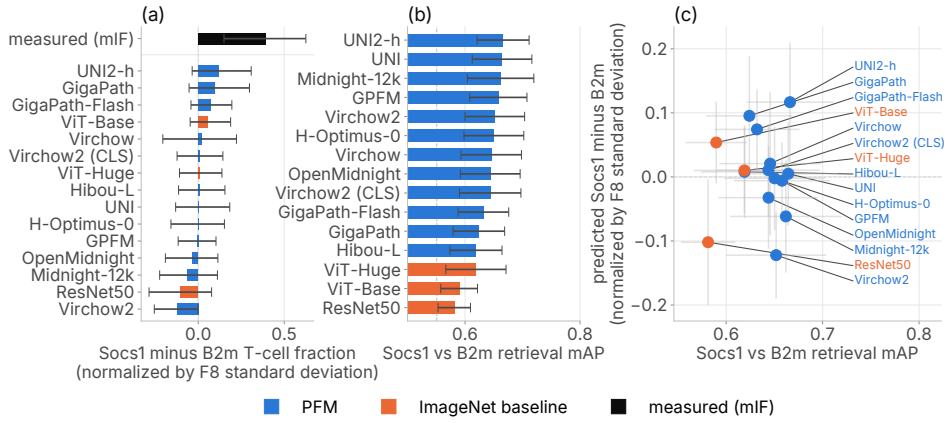


Figure 7: Performance on `tme-effect` tasks: (a) Predicted T cell fraction difference between *B2m* and *Socs1* across the models. The black bar is the measured T cell fraction difference based on the mIF. Units are *F8*-normalized cell fractions. (b) mAP score across models in a restricted set of *B2m* and *Socs1* only. (c) Comparison of results from the T cell fraction task and mAP task. Top-right points indicate better performance.

The evaluation in `mIF` focuses on many perturbations whose underlying genes are not fully understood in the context of TME regulation and immune response. Next, we focus on two well-characterized perturbations, *B2m* and *Socs1*, to evaluate whether the PFM is able to recapitulate a *known tumor-microenvironment relationship* from the literature. The *B2m* gene encodes a key component of the antigen presentation complex, and loss of function of *B2m* in tumor cells has been linked to both a decrease in local T cell infiltration and a lack of response to immune checkpoint inhibitors (Dhainaut et al., 2022; Zaretsky et al., 2016). *Socs1* is a negative regulator of inflammatory cytokine signaling, and *Socs1* knockout (KO) KP tumors have previously been shown to present an increase T cell infiltration (Dhainaut et al., 2022). Therefore, we reasoned that PFMs should be able to predict that a *B2m* KO tumor has fewer T cell infiltrates than a *Socs1* KO tumor. To perform this evaluation, model embeddings and mIF cellular composition values are used to fit a ridge regression on the T cell fraction, with *B2m* and *Socs1* held-out. Then, the ridge regression model is applied to the embeddings of the *Socs1* and *B2m* tumors to predict the *Socs1-B2m* difference in T cell fraction. As expected, we observed in the measured mIF data that *Socs1* perturbations exhibit higher T cell fraction compared to *B2m*. As seen in Figure 7, 9 out of the 15 models predicted this directionality in effect, but no model predicted a significant difference. The low performance across the models, despite the significant effect observed in the ground truth, suggests that this task may be particularly challenging for PFMs. To additionally test whether the models can differentiate between *B2m* and *Socs1*, we also evaluated the cross-mouse mAP using only the two perturbations. We then compared the scores for these two sub-tasks, observing minimal correlation in performance (Figure 7 (c)).

4.5 Overall performance

Comparing the performance of the models across our benchmark suite is tricky due to overlapping error bars; this is a phenomenon often seen in other PFM benchmarks (Ma et al., 2025b; Neidlinger et al., 2026) and can be attributed to similar model architectural choices (e.g., using a ViT (Dosovitskiy et al., 2021)), training strategies (e.g., DINOv2 (Oquab et al., 2024)), overlap between models on publicly available training data sets such as the TCGA (Weinstein et al., 2013), and noise in our benchmark dataset. Regardless, we attempted to provide an unbiased ranking, by summarizing the mean performance across all tasks, focusing on those tasks with higher discriminability and removing tasks with high rank correlation to the others. We encourage users of this benchmark suite to select the tasks that are most relevant to their model evaluation goals.

Our leaderboard is shown in Figure 8, observing that UNI2-h has the highest performance. We then compared the overall performance ranking in Perturb-Path to the overall ranking in previous benchmarks THUNDER (Marza et al., 2025) and PathBench (Ma et al., 2025b)(Figure 8 (b) and (c)). We observe here that there is significant correlation with those benchmarks, yet still significant differences; for example, Virchow2 performs relatively worse on Perturb-Path compared to the other two benchmark suites. Some of these differences can be explained by the fact that the H&E in Perturb-Path is mouse H&E, which can serve as a good out-of-distribution test for the models evaluated here trained on human H&E.

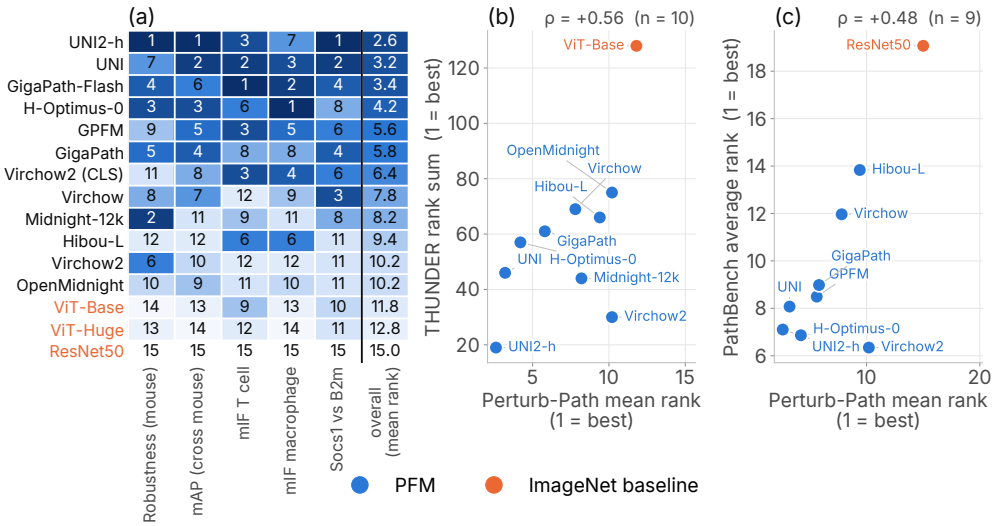


Figure 8: Summary of overall performance across models. (a) Ranking of performance, lower is better, across selected tasks. (b,c) Correlation of overall mean rank in Perturb-Path versus THUNDER (b) and PathBench (c) across models evaluated in both.

5 Conclusion

In this work, we introduced Perturb-Path, the first PFM benchmark suite built around an *in vivo* genetic perturbation dataset. This dataset contains 84 distinct perturbations across 19,018 tumor regions, with accompanying cell type composition metrics derived from paired mIF readout. This experimental design enabled a wide range of technical and biological benchmark tasks.

We found that ranking of PFM performance varies throughout the Perturb-Path tasks, although UNI2-h performed consistently near or at the top of the ranking. Models frequently performed similarly well, as exhibited by the overlapping error bars across the models. Such similarities in model performance have been noted in other PFM benchmarks (Ma et al., 2025b; Neidlinger et al., 2026) and can be attributed to similar architectural choices (e.g., ViT (Dosovitskiy et al., 2021)), training strategies (e.g., DINOv2 (Oquab et al., 2024)), overlap on publicly available training data sets such as the TCGA (Weinstein et al., 2013), and noise in this benchmark dataset. However, PFMs clearly outperformed the ImageNet baselines across most tasks, and were able to capture several biological aspects of the tumor microenvironment, yet are still far from saturating the performance on our tasks.

Our benchmark is unique in the large number of classes and presence of biological replicates, but also presents some limitations. First, the KO efficiency for each gene perturbation is variable and has not been quantified at scale. A given KO can also have a limited phenotypic impact on the tumor. Batch effects due to variations in staining and treatment protocols also still exist, although our data architecture allows us to mitigate them via internal normalization for each sample. Finally, more complex aggregation strategies for computing the tumor embeddings could be explored in the future. Nevertheless, these caveats are outweighed by the unique application of Perturb-Path for evaluating PFMs; Perturb-Path is the first histopathology dataset that can relate genetic perturbation in cancer cells not only to the cells' phenotype, but also to how they impact their surrounding TME.

AI USE STATEMENT

In this work, we used agentic AI tools to generate code for data processing pipelines, various plots and figures, and benchmarks tasks. All code outputs were thoroughly evaluated for correctness, oftentimes by multiple members of the team. None of the text that appears in this work was generated by AI, and although AI was sometimes used for initial ideation and literature search, we manually checked all references to ensure that none were hallucinated. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All animal protocols used in the present studies were approved by the relevant Institutional Animal Care and Use Committee.

REPRODUCIBILITY STATEMENT

To facilitate our reviewers' understanding of the dataset and codebase, we have uploaded the benchmark code and the H&E tiles, tumor regions, and accompanying immune cell proportions for an example lobe image as supplementary material.

Upon acceptance, both the entire Perturb-Path dataset and benchmark code will be publicly released. The dataset will be available in a processed format. For each tumor region the following will be released:

- H&E images cropped down to the tumor region, including the tumor region polygon.
- Cell type proportions calculated from mIF, as described in A.1.2.
- Associated gene perturbation and mouse metadata.

This dataset format alongside the benchmark code will facilitate easy reproducibility of all results in this paper, in addition to enabling potential downstream research applications mentioned in the Conclusion.

REFERENCES

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A Cookbook of Self-Supervised Learning, June 2023. URL http://arxiv.org/abs/2304.12210. arXiv:2304.12210 [cs.LG].

Rohan Bareja, Francisco Carrillo-Perez, Yuanning Zheng, Marija Pizurica, Tarak Nath Nandi, Lu Tian, Jeanne Shen, Ravi Madduri, and Olivier Gevaert. A benchmark study of vision and pathology foundation models for computational pathology. *Nature Communications*, 17(1):9012, July 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-76004-6. URL https://www.nature.com/articles/s41467-026-76004-6.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, and others. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Maxime Dhainaut, Samuel A. Rose, Guray Akturk, Aleksandra Wroblewska, Sebastian R. Nielsen, Eun Sook Park, Mark Buckup, Vladimir Roudko, Luisanna Pia, Robert Sweeney, Jessica

Le Berichel, C. Matthias Wilk, Anela Bektesevic, Brian H. Lee, Nina Bhardwaj, Adeeb H. Rahman, Alessia Baccarini, Sacha Gnjatic, Dana Pe'er, Miriam Merad, and Brian D. Brown. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, 185(7): 1223–1239.e20, March 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.02.015. URL https://www.sciencedirect.com/science/article/pii/S0092867422001957.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs.CV].

Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, 35(6):498–507, June 2017. ISSN 0167-7799. doi: 10.1016/j.tibtech.2017.02.012. URL https://www.sciencedirect.com/science/article/pii/S0167779917300367.

Noah F. Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J. McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, Cole Pavelchek, Sunny Cui, Isabella Camplisson, Omer Bar-Tal, Jaiveer Singh, Mara Fong, Gautam Chaudhry, Zion Abraham, Jackson Moseley, Shiri Warshawsky, Erin Soon, Shirley Greenbaum, Tyler Risom, Travis Hollmann, Sean C. Bendall, Leeat Keren, William Graf, Michael Angelo, and David Van Valen. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4): 555–565, April 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01094-0. URL https://doi.org/10.1038/s41587-021-01094-0.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs.CV].

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis, November 2024. URL http://arxiv.org/abs/2406.16192. arXiv:2406.16192 [cs.CV].

Daniel Kaplan, Ratna Sagari Grandhi, Connor Lane, Benjamin Warner, Tanishq Mathew Abraham, and Paul S. Scotti. How to Train a State-of-the-Art Pathology Foundation Model with $1.6k, 2025. URL https://sophont.med/blog/openmidnight.

Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. *arXiv preprint arXiv:2504.05186*, 2025. URL https://arxiv.org/abs/2504.05186.

Jonah Kömen, Edwin D. de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, Frederick Klauschen, and Klaus-Robert Müller. Towards robust foundation models for digital pathology. *Nature Communications*, 17(1):5218, 2026. doi: 10.1038/s41467-026-73923-2. URL https://doi.org/10.1038/s41467-026-73923-2.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, pages 1–20, 2025a.

Jiabo Ma, Yingxue Xu, Fengtao Zhou, Yihui Wang, Cheng Jin, Zhengrui Guo, Jianfeng Wu, On Ki Tang, Huajun Zhou, Xi Wang, Luyang Luo, Zhengyu Zhang, Du Cai, Zizhao Gao, Wei Wang, Yueping Liu, Jiankun He, Jing Cui, Zhenhui Li, Jing Zhang, Feng Gao, Xiuming Zhang, Li Liang, Ronald Cheong Kin Chan, Zhe Wang, and Hao Chen. PathBench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology, May 2025b. URL http://arxiv.org/abs/2505.20202. arXiv:2505.20202 [cs.CV].

Pierre Marza, Leo Fillioux, Sofiène Boutaj, Kunal Mahatha, Christian Desrosiers, Pablo Piantanida, Jose Dolz, Stergios Christodoulidis, and Maria Vakalopoulou. THUNDER: Tile-level Histopathology image UNDERstanding benchmark. *Neural Information Processing Systems (NeurIPS) D&B Track*, 2025.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A Family of Foundational Vision Transformers for Pathology, 2024. URL https://arxiv.org/abs/2406.05074.

Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10(6):1113–1123, June 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01516-3. URL https://www.nature.com/articles/s41551-025-01516-3.

Mieko Ochi, Daisuke Komura, and Shumpei Ishikawa. Pathology Foundation Models. *JMA Journal*, 8(1):121–130, January 2025. ISSN 2433-328X. doi: 10.31662/jmaj.2024-0206. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11799676/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs.CV].

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Shankha Satpathy, Natalie M. Clark, Yi-Ju Chen, Noshad Hosseini, Ya-Hsuan Chang, Yi Hsiao, Jonathan T. Lei, Francesca Petralia, Jin-Shing Chen, Yifat Geffen, David I. Heiman, Indranil Paul, Hanbyul Cho, Michelle Hollenberg, Giacomo B. Marino, Kuen-Tyng Lin, Rahul Mannan, C. Jackson White, Joe Allen, Shayan C. Avanessian, M. Harry Kane, Ashley Wolfe, Miloni Kinarivala, Wenke Liu, Shankara Anand, Mong-Wei Lin, Moe Haines, Erik J. Bergstrom, Grant Hussey, Ginny Xiaohe Li, Deepak C. Mani, Hao Fang, Eric J. Jaehnig, Hasmik Keshishian, Brecca Miller, Kang-Yi Su, Yi-Jing Hsiao, Hsao-Hsun Hsu, Min-Shu Hsieh, Kuo-Hsuan Hsu, Alexi Monovoukas, Simone Gohsman, John R. Thorup, Yamei Deng, Yo Akiyama, Eden Deng, Eric Sheng-Wen Chen, Azra Krek, Rodrigo Espinoza, Weiping Ma, Daniel Charytonowicz, Robert Sebra, Jyun-Hong Lin, Yan-Si Chen, Yin-Chen Hsu, Ze-Shiang Lin, Kun-Chieh Chen, Chang-Wei Yeh, Yu-Tai Wang, Alexander J. Lazar, Mehdi Mesri, Eunkyung An, Xu Zhang, Karl R. Clauser, David Fenyö, Arul M. Chinnaiyan, Bing Zhang, Li Ding, Kelly Ruggles, Chelsea Newton, Hui Zhang, Pei Wang, Galen Hostetter, Gilbert S. Omenn, Chandan Kumar-Sinha, Mathangi Thiagarajan, Ramaswamy Govindan, Paul Paik, Abhijit Parolia, Qing K. Li, Avi Ma'ayan, Gad A. Getz, Saravana M. Dhanasekaran, Ana I. Robles, Gee-Chen Chang, Pan-Chyr Yang, Sung-Liang Yu, Hsuan-Yu Chen, Alexey I. Nesvizhskii, Steven A. Carr, D. R. Mani, Marcin P. Cieslik, Yu-Ju Chen, Michael A. Gillette, Chao-Wen Lu, Cheng-Hsiang Chu, Chi-Ya Shen, Chia-Li Han, Chien-Chia Lin, Chien-Yu Lin, Ching-Wen Chen, Chung-Hsien Lin, Hsiang-En Hsu, Hsing-Jui Tsai, Jia-Jun Wu, Jing-Wei Lin, Juani Waniwan, Ki-Hok Liao, Pei-Hsing Chen, Pei-Rong Huang, Sin-Ming Huang, Tai-Ching Lin, Wan-Chun Lai, Wei-Tzu Chiu, Xu-Heng Chiang, Ya-Ling Chang, Yan-Ming Chen, Yi-Ling Chen, Yi-Wen Wang, Yi-Wei Lin, Yu-Cheng Chang, Yu-Ting Huang, Yuju Lien, Zhe-Rong Zheng, Abdul Samad Hashimi, Adrij Mohan, Akhilesh Pandey, Alexander Pilozzi, Alex Webster, Amanda G. Paulovich, Aniket Dagar, Andrew K. Godwin, Barbara Pruetz, Bart O. Williams, Brian J. Druker, Daniel C. Rohrer, Daniel W. Chan, Danail Petrov, David Chesla, Diwakar Davaar, Elizabeth Duffy, George D. Wilson, Grace Zhao, Iga Kołodziejczak, Jan Lubinski, Jasmine Huang, Jason Hafron, Jeffrey W. Tyner, John Koomen, Kakhaber Zaalishvili, Karen A. Ketchum, Maciej Wiznerowicz, Marcin J. Domagalski, Meenakshi Anurag, Melissa Borucki, Nathan J. Edwards, Negin Vatanian, Pamela Grady, Paul D. Piehowski, Popovici Bogdan, Qin Li, Rafael Fonseca, Rashna Madan, Ratna R. Thangudu, Reese Crispen, Ronald Matteotti, Ross

Bremner, Sandra Cerda, Sandra L. Cottingham, Shirley Tsang, Shuang Cai, Tao Liu, Thomas Bauer, William W. Maggio, Xiaojun Jing, Yuping Zhang, Yvonne Shutack, and Zoran Andric. Integrative analysis of lung adenocarcinoma across diverse ethnicities and exposures. *Cancer Cell*, 43(9):1731–1757.e17, September 2025. ISSN 1535-6108, 1878-3686. doi: 10.1016/j.ccell.2025.07.011. URL https://www.cell.com/cancer-cell/abstract/S1535-6108(25)00317-4.

Hao Shi, John G Doench, and Hongbo Chi. CRISPR screens for functional interrogation of immunity. *Nature Reviews Immunology*, 23(6):363–380, 2023.

Naoto Usuyama, Jeya Maria Jose Valanarasu, Sicong Yao, Hanwen Xu, Jaspreet Bagga, Guanghui Qin, Robert E. Kramer, Cliff Wong, Soohee Lee, Hao Qiu, Theodore Zhengde Zhao, Racheli Ben Shimol, Angela Crabtree, Kevin Matlock, Eduardo Alejandro Lozano Garcia, Naiteek Sangani, Alberto Santamaria-Pang, Maximilian Rokuss, Yashna Hasija, Naisargi Manishkumar Patel, Jason Entenmann, Alexandra Q. Bartlett, Bill J. Wright, Bernard A. Fox, Brian Piening, Sheng Zhang, Sheng Wang, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. GigaPath-Flash and GigaTIME-Flash: Efficient Pathology Foundation Models for Whole-Slide and Tumor Microenvironment Analysis. *arXiv preprint arXiv:2607.18218*, 2026. doi: 10.48550/arXiv.2607.18218. URL https://arxiv.org/abs/2607.18218.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina ElHarouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. Molecular-driven Foundation Model for Oncologic Pathology, 2025. URL https://arxiv.org/abs/2501.16652.

Suhas Vasaikar, Chen Huang, Xiaojing Wang, Vladislav A. Petyuk, Sara R. Savage, Bo Wen, Yongchao Dou, Yun Zhang, Zhiao Shi, Osama A. Arshad, Marina A. Gritsenko, Lisa J. Zimmerman, Jason E. McDermott, Therese R. Clauss, Ronald J. Moore, Rui Zhao, Matthew E. Monroe, Yi-Ting Wang, Matthew C. Chambers, Robbert J. C. Slebos, Ken S. Lau, Qianxing Mo, Li Ding, Matthew Ellis, Mathangi Thiagarajan, Christopher R. Kinsinger, Henry Rodriguez, Richard D. Smith, Karin D. Rodland, Daniel C. Liebler, Tao Liu, Bing Zhang, Akhilesh Pandey, Amanda Paulovich, Andrew Hoofnagle, D. R. Mani, Daniel W. Chan, David F. Ransohoff, David Fenyo, David L. Tabb, Douglas A. Levine, Emily S. Boja, Eric Kuhn, Forest M. White, Gordon A. Whiteley, Heng Zhu, Hui Zhang, Ie-Ming Shih, Jasmin Bavarva, Jeffrey Whiteaker, Karen A. Ketchum, Karl R. Clauser, Kelly Ruggles, Kimberly Elburn, Linda Hannick, Mark Watson, Mauricio Oberti, Mehdi Mesri, Melinda E. Sanders, Melissa Borucki, Michael A. Gillette, Michael Snyder, Nathan J. Edwards, Negin Vatanian, Paul A. Rudnick, Peter B. McGarvey, Philip Mertins, R. Reid Townsend, Ratna R. Thangudu, Robert C. Rivers, Samuel H. Payne, Sherri R. Davies, Shuang Cai, Stephen E. Stein, Steven A. Carr, Steven J. Skates, Subha Madhavan, Tara Hiltke, Xian Chen, Yingming Zhao, Yue Wang, and Zhen Zhang. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell*, 177(4):1035–1049.e19, May 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.03.030. URL https://www.cell.com/cell/abstract/S0092-8674(19)30292-2.

Brandon Veremis, Shengjia Chen, and Gabriele Campanella. An Introduction to Pathology Foundation Models. *Head and Neck Pathology*, 20(1):9, January 2026. ISSN 1936-0568. doi: 10.1007/s12105-025-01878-9. URL https://doi.org/10.1007/s12105-025-01878-9.

Karin E. de Visser and Johanna A. Joyce. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*, 41(3):374–403, 2023. ISSN 1535-6108. doi: https://doi.org/10.1016/j.ccell.2023.02.016. URL https://www.sciencedirect.com/science/article/pii/S1535610823000442.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 2024.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. ISSN 1546-1718. doi: 10.1038/ng.2764. URL https://www.nature.com/articles/ng.2764.

Sébastien Wieckowski, Teresa Hemmerle, Spasenja Savic Prince, Béatrice Dolder Schlienger, Sven Hillinger, Dario Neri, and Alfred Zippelius. Therapeutic efficacy of the F8-IL2 immunocytokine in a metastatic mouse model of lung adenocarcinoma. *Lung Cancer*, 88(1):9–15, April 2015. ISSN 1872-8332. doi: 10.1016/j.lungcan.2015.01.019.

Aleksandra Wroblewska, Maxime Dhainaut, Benjamin Ben-Zvi, Samuel A. Rose, Eun Sook Park, El-Ad David Amir, Anela Bektesevic, Alessia Baccarini, Miriam Merad, Adeeb H. Rahman, and Brian D. Brown. Protein Barcodes Enable High-Dimensional Single-Cell CRISPR Screens. *Cell*, 175(4):1141–1155.e16, November 2018. ISSN 1097-4172. doi: 10.1016/j.cell.2018.09.022.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

Jesse M Zaretsky, Angel Garcia-Diaz, Daniel S Shin, Helena Escuin-Ordinas, Willy Hugo, Siwen Hu-Lieskovan, Davis Y Torrejon, Gabriel Abril-Rodriguez, Salemiz Sandoval, Lucas Barthly, et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *New England Journal of Medicine*, 375(9):819–829, 2016.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating Data Processing and Benchmarking of AI Models for Pathology, 2025. URL https://arxiv.org/abs/2502.06750.

Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, Lisa J. Zimmerman, Kent F. Shaddox, Sangtae Kim, Sherri R. Davies, Sean Wang, Pei Wang, Christopher R. Kinsinger, Robert C. Rivers, Henry Rodriguez, R. Reid Townsend, Matthew J.C. Ellis, Steven A. Carr, David L. Tabb, Robert J. Coffey, Robbert J.C. Slebos, and Daniel C. Liebler. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, September 2014. ISSN 0028-0836. doi: 10.1038/nature13438. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4249766/.

Bokai Zhao, Yiyang Zhang, Yuanchi Zhu, Hanqing Chao, Long Bai, Tai Ma, Minfeng Xu, Ming Song, and Tianzi Jiang. Benchmarking Pathology Foundation Models for Spatial Domain Understanding, May 2026. URL http://arxiv.org/abs/2605.25764. arXiv:2605.25764 [cs.CV].

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Appendix

A.1 Details of Perturb-Map Data Processing

As mentioned in Section 2, the Perturb-Map readout includes two types of mIF channels: Pro-Code channels and tumor microenvironment channels. The former is used for debarcoding tumors, as described in Section A.1.1, while both types of channels are used for cell typing, as described in Section A.1.2.

A.1.1 Debarcoding

Recall from Section 2 that each Pro-Code is composed of a unique triplet of tags. There are 8 tags in total, each with its channel in the mIF readout. Debarcoding is the process by which tag mIF channels are used to identify clonal tumor populations and their corresponding gene perturbations.

Our debarcoding implementation has two steps: (1) label the tags present in each nucleus and (2) aggregate nearby cells labeled with the same Pro-Code, i.e. triplet tag combination, into tumor regions. The labeling step requires cell segmentations. To obtain these, we took advantage of the fact that all lung lobes were stained with DAPI or Hoechst (both fluorescent dyes that bind to DNA and mark the nucleus) and that the mIF readout included channels for CD45 and NaK ATPase (membrane proteins used to delimit the cell boundaries). We fed the DAPI/Hoechst readout, along with CD45 and NaK ATPase mIF channels, into Mesmer (Greenwald et al., 2022) to obtain the desired cell segmentations. 24,694 cells across 37 lung lobes were then manually assigned Pro-Code labels using a custom-built annotation tool.

Separately, a masked autoencoder (MAE) was pretrained in a self-supervised setting to reconstruct partially masked tag channel readouts from hundreds of separate Perturb-map experiments. This enabled the training of our tag labeling model, which features a hybrid two-arm architecture with a vision transformer (ViT) as the first arm and a simple MLP as the second. The ViT is fed MAE embeddings generated from the surrounding tag mIF signal along with a mask indicating the cell to label. The MLP is fed 32 precomputed 'summary statistics' that compare the Pro-Code mIF signal inside the target nucleus to the background signal across the lobe. For example, one set of summary statistics contains the average difference in average IF log-intensities between the target cell and the entire lobe for each of the 8 tags. Another MLP at the head of the model takes as input the CLS token from the ViT arm alongside the outputs from this summary statistic MLP arm. This classification head MLP produces 8 logits, one for each tag. Only cells with exactly three tags with probabilities all > 0.125 were labeled with a proper triplet-tag Pro-Code. These triplet-tag Pro-Codes can then be unambiguously mapped to a perturbation.

Finally, tumor cell are assembled into tumor regions, one Pro-Code at a time. For each triplet tag combination, we select every cell labeled with that combination and dilate the cells $25\,\mu\mathrm{m}$; each set of overlapping dilated cells is merged into its own tumor region. The region boundaries are then shrunk back down by $25\,\mu\mathrm{m}$ and all the interior holes in the regions are filled. Finally, the region contours are simplified by the Douglas-Peucker algorithm with $10\,\mu\mathrm{m}$ tolerance. Disjoint polygons are considered separate tumors, and tumors less than $12\,500\,\mu\mathrm{m}^2$ area are filtered out.

A.1.2 Cell Typing

Labeling of immune cell types happens through a process parallel to debarcoding. We begin with the same Mesmer cell masks described in A.1.1. The first step is to identify cells with increased intensity of CD45, a membrane protein found on the surface of almost all leukocytes, including macrophages, T cells, and B-cells. We accomplish this by finding a negative set: the top 10% of cells ranked by mean intensity for each Pro-Code tag. The mean CD45 intensity across this negative set becomes the cutoff for a cell to be considered CD45-positive. (Although this cutoff is replaced by the 5th percentile of CD45 expression if this is a stricter thresholder.)

At the same time, as part of this cell typing process parallel to debarcoding, cells are labeled as positive for each Pro-Code tag if the tag IF intensity is greater than the average intensity among the top 10% of cells ranked by mean CD45 intensity. (Or the 5th percentile of mean CD45 intensity if this is a stricter threshold.) A cell is labeled 'tumor' if it positive by this heuristic for at least two Pro-Code tags.

Next, intensities for all channels are normalized so that 0 and 1 because the 5th and 95th percentile. Percentiles are calculated at the lobe level. There is also a separate 'tumor intensity' that is calculated by taking the average across all 8 Pro-Code tag intensities. At this point, we have cells that are CD45-positive, and cells that are labeled 'tumor' by this system. But there can also be cells labeled both CD45-positive and tumor—these cells are given whichever label has the greater intensity, with ties resolving to immune.

Finally, each immune cell will be assigned a cell type based on their expression of pre-defined combinations of immune markers measured by mIF. These markers are treated as one-hot vectors. For instance, the marker vector for $CD8^+$ has 1's for CD45, CD3, and CD8; and 0's everywhere else. Dot products are computed between these cell type marker vectors and all non-tumor nuclei normalized intensity profiles. A cell is labeled as a particular cell type if dot product > 0.6. If there are multiple positive labels, the greatest one wins. If there are no positive labels, the cell is marked 'unassigned'. Cell type proportions for $CD8^+$ T cells, $CD4^+$ T cells, and macrophages are computed over the tumor regions constructed in A.1.1.

A.2 Effect of *F8*-normalization

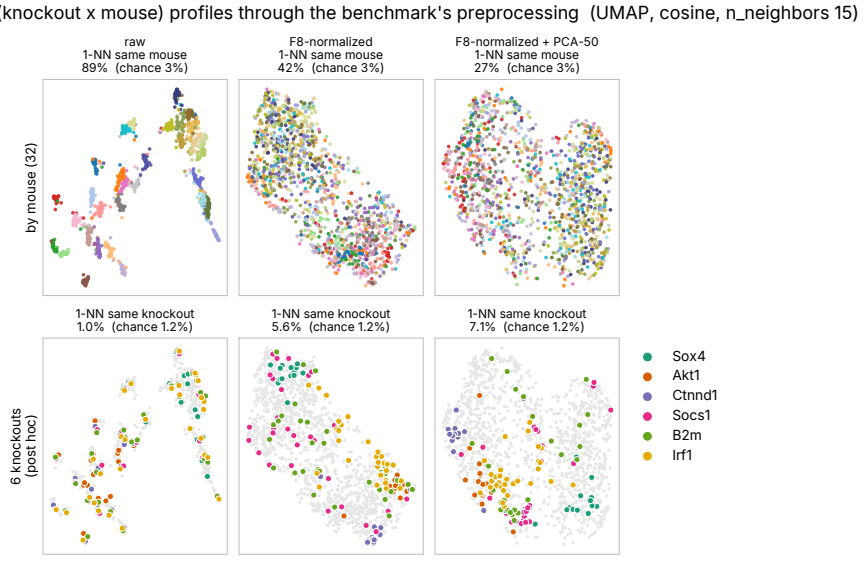


Figure 9: UMAP of mean-pooled embeddings for each perturbation, mouse pair across raw embeddings, *F8*-normalized embeddings, and *F8*-normalized with z-score+PCA normalization. Top row: UMAP is colored by mouse. Bottom row: UMAP is colored by 6 perturbations to depict the increased clustering with additional normalization. Embeddings here are generated by the UNI2-H model.

A.3 Per-perturbation mAP significance calculation

As part of the cross-mouse mAP task, we also calculate the significance that a model can retrieve a specific perturbation profile above chance. This significance is calculated by a permutation test with a multiple hypothesis correction. Specifically, we shuffle the perturbation labels over all the profiles 5000 times and re-score the per-perturbation mean AP after each shuffle. The one-sided p-value for perturbation k is thus

$$p_k = \frac{1 + \#\{\text{shuffles with null mean AP} \geq \text{observed}\}}{5001}. \tag{2}$$

To account for multiple hypothesis testing across all perturbations, we then apply the Benjamini–Hochberg (Benjamini and Hochberg, 1995) correction to these p-values and call a perturbation k significant if $q_k < 0.05$.

A.4 KNOCKOUT CLASSIFICATION

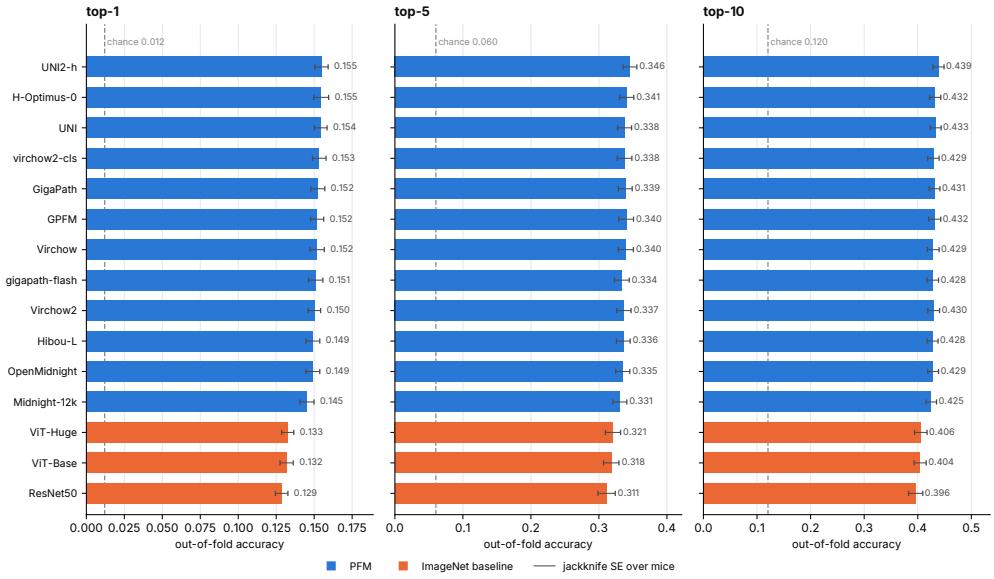


Figure 10: Predicting the gene knockout corresponding to each tumor from the corresponding embeddings.

The mAP retrieval task described above can also be formulated as a classification task. We evaluate whether the gene knockout itself is predictable from the tumor embeddings, which we set up as a $|G|$-way classification problem, where G is the set of gene knockouts performed across all CRISPR libraries in Perturb-Path. Whereas the mAP task relies on cosine similarity between tumor region embeddings, this task examines predictability based on a linear probe. Specifically, each tumor region's embedding is normalized and fit using a multinomial logistic regression with 5-fold cross validation grouped by mouse. As shown in Figure 10, the top-1, top-5 and top-10 accuracies of the PFMs are 1) significantly above chance, 2) are higher than the baseline ViT and ResNet50 models, and 3) generally consistent in ranking with the mAP results.

A.5 mAP TASK: ADDITIONAL RESULTS

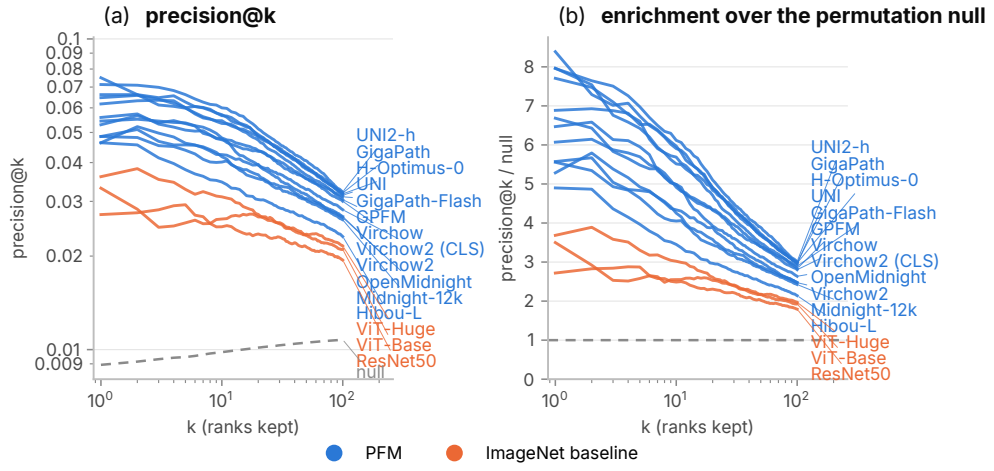


Figure 11: Precision as a function of rank k for a given seed across models, averaged across all queries.

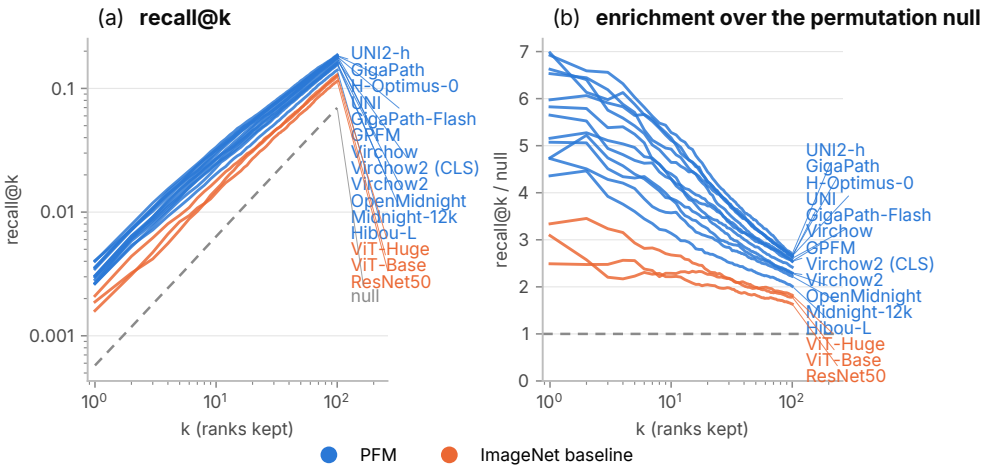


Figure 12: Recall as a function of rank k for a given seed across models, averaged across all queries.