DENSITY-BASED ONLINE AUTOCURATION FOR PATHOLOGY IMAGE MODEL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tile streams for pathology image model pre-training are curated at the slide level, or offline at the tile level with an oracle encoder. Morphologies are unevenly distributed within a slide, so random tile sampling carries each slide's imbalance into the training stream and common morphologies overshadow rare ones. This work curates the stream online, during DINOv2 pretraining, from a density estimate the model produces itself. A *density cache* keeps a fixed-size memory of M stored tile signatures and counts how often new tiles land on each; the counts give the typicality of an incoming tile, and tiles enter the batch with probability inversely proportional to it. One hyperparameter, the tilt, sets how far the batch leans toward low-typicality tiles. ViT-B and ViT-L encoders trained with DINOv2 on 287M tiles were evaluated on twelve biomarker mutation-prediction tasks from H&E. Fold-paired against the uncurated baseline at the same checkpoint, every curated run improved mutation prediction on average, by 0.3 to 1.0 percentage points of AUROC over the twelve tasks and by 2 to 4 on THCA/NRAS at ViT-B, and the two thyroid tasks improved in every run at both scales. At ViT-L the improvement is on the tasks with the most consistent folds, and two of the three runs are significant when each task is weighted by the precision of its measurement, with a tilt of 1.0 the best of the three. At ViT-B it is on the tasks with few positive slides per fold and is resolved in one run of seven. Across the eight 124k runs whose caches did not starve, the gain rises with the spread of hit-rates over the cache, its bandwidth (Spearman $+0.75$). A 500k-step pair at ViT-B, four times the schedule, is at parity on aggregate ($+0.3$ percentage points, 95% interval $[-0.4, +0.9]$), and the thyroid and colorectal tasks improve again under this third independent baseline. Mapped back onto a slide, the stored signatures separate morphologically coherent regions at tile resolution, so the cache also yields an unsupervised tissue segmentation.

1 INTRODUCTION

Pathology image models, when trained on vast institutional data corpora, are becoming clinically valuable in the detection of mutation status directly from tissue sections stained with hematoxylin and eosin (H&E), a task that otherwise requires the expensive and slow sequencing of each tissue sample. Such an application requires image models to have a diverse understanding of tissue morphology, since a mutation that admits morphological cues is usually spread across many tissue components and arrangements across the whole slide. An encoder that has learned mainly the morphologies it sees most often would be inadequate for the task.

Pathology image models are trained on tiles, small square crops of a few hundred pixels cut from whole-slide images (WSIs) at a fixed magnification, sampled from the slides in a stream. Within and across slides, morphologies occur with very different frequencies and can dominate the data stream by abundance alone. A model trained on such a stream spends most of its updates on morphologically imbalanced batches, and is therefore biased toward the more commonly occurring morphologies. In a hospital this imbalance is permanent, often exacerbated by institutional specialization, and cannot be accounted for by current data curation approaches. Slide-level curation decides which slides enter the corpus and is typically biased towards diagnostic occurrence and not morphology. Offline tile-level curation embeds and clusters the corpus before training, as in DINOv2 (Oquab et al., 2024), but it needs an external encoder, an oracle, whose representation already decides what

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.

is redundant or rare, and in pathology such encoders are themselves trained on equally imbalanced streams. Curation, thus, has to be lock-step with model pre-training. However, prior autocuration, in which the training model selects its own stream, scores tiles by loss or by gradient magnitude. In the autocuration approach presented in this work, a tile is scored by the density of its neighborhood in a morphology signature space, the bottleneck representation of the DINO head, estimated online by the model being trained, and admitted in inverse proportion to it.

Contributions.

- A Density-based autocuration approach of the training stream inside self-supervised pre-training (Section 3, Algorithm 1, Eqs. 1 and 2).
- Its application to histopathology, where curation improved a subset of the tasks fold-paired against the baseline at the same checkpoint (Figure 1, Section 4, Table 2), and that the precision-weighted improvement rises with the spread of hit-rates over the cache, i.e., the range of typicality the cache resolves (Table 1, Figure 4).
- An unsupervised segmentation approach at the slide-level utilizing the same cache, with no further training (Algorithm 2, Figure 3).

2 Related Work

Data curation and selection. Offline curation of pretraining corpora is well established. DINOv2 assembles LVD-142M by self-supervised retrieval and deduplication from an uncurated pool (Oquab et al., 2024), hierarchical k-means rebalances long-tailed concepts (Vo et al., 2024), SemDeDup removes semantic duplicates (Abbas et al., 2023), and data-pruning metrics discard examples to beat power-law scaling (Sorscher et al., 2022). Sampling approaches reweight examples during training, including online batch selection by loss rank (Loshchilov & Hutter, 2015), importance sampling by the magnitude of the per-example gradient (Katharopoulos & Fleuret, 2018), prioritized training on learnable points (Mindermann et al., 2022), and domain-mixture reweighting (Xie et al., 2023). Autocuration, in which the training model scores and selects its own stream, has been explored through joint example selection over super-batches (Evans et al., 2024) and active curation acting as implicit distillation (Udandarao et al., 2025). In all of these the selection signal comes from the training objective, a loss or its gradient, alone or against a reference model. Feature-space sampling has rebalanced long-tailed distributions during self-supervised training (Bai et al., 2023), typicality defined by density has selected examples for active learning (Hacohen et al., 2022), and Assran et al. (2023) treat the imbalance in the loss by relaxing the uniform-cluster prior that joint-embedding methods carry. The curator presented in this work selects by typicality estimated with a fixed-size memory whose structure is that of density-based stream clustering: micro-clusters with time-decayed weights and an outlier buffer with promotion in DenStream (Cao et al., 2006), fixed-radius assignment in DBSTREAM (Hahsler & Bolaños, 2016). New here are the readout, an unnormalized sum of hit-rates that cancels the density of the stored entries, its use as an admission signal inside self-supervised training, and its attachment to DINOv2, the established pretraining pipeline of histopathology image models.

Pathology image models. Pathology image models pretrain on tiles extracted from whole-slide images, with curation at the slide level, as in UNI (Chen et al., 2024), or unstated (Vorontsov et al., 2024; Zimmermann et al., 2024; Xu et al., 2024; Filiot et al., 2023). Chen et al. (2025) curate at the tile level with hierarchical clustering trees over 350 million pre-extracted embeddings, sampling datasets balanced across the embedding space; like slide-level filtering (Campanella et al., 2023), this is an offline step before training. No pathology pretraining effort known to the authors curates its own tile stream online.

3 Method

The curator rebalances the tile stream during training so that the composition of a batch no longer mirrors the frequency of each morphology in the corpus. Each candidate tile is reduced to a *signature* $s(x) \in \mathbb{R}^{256}$, and $p(x)$, the local density of signatures at $s(x)$, is the *typicality* of the tile. A *density cache*, a counted cover of signature space with least-frequently-used eviction, estimates it online on

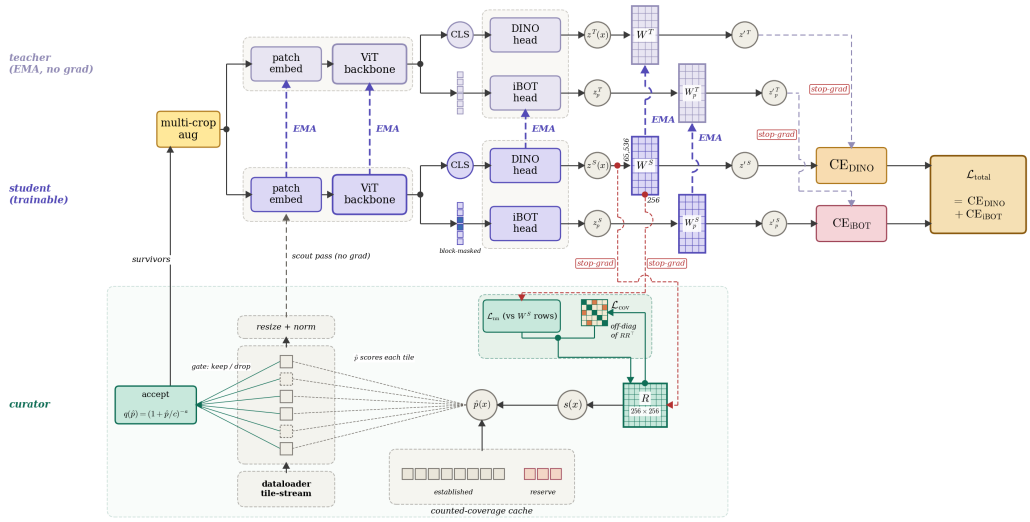

Figure 1: **Where the curator joins DINOv2.** DINOv2 (student, EMA teacher, DINO and iBOT heads; the KoLeo term is not drawn) trains on a batch the curator has thinned. A no-grad scout pass reduces each candidate tile to its signature $s(x) = R\,z(x)$, the density cache reads $\hat{p}(x)$, and the tile is kept with probability $q(\hat{p})$. R is trained online by $\mathcal{L}_{\mathrm{nn}}$ and $\mathcal{L}_{\mathrm{cov}}$. The loss is unweighted; the method acts only on which tiles enter the batch.

a fixed budget, and the estimate $\hat{p}(x)$ enters in one place: a candidate is admitted to the batch with a probability that decreases with $\hat{p}(x)$. The loss on admitted tiles is unchanged. Figure 1 shows where the curator attaches to DINOv2 and Algorithm 1 gives one step; Appendix A.1 collects the definitions used below.

From tile to signature. The signature is $s(x) = R\,z(x)$: $z(x)$ is the 256-dimensional bottleneck of the student DINO head on an un-augmented view of the tile, under stop-gradient, and R is a trained matrix with $K' = 256$ unit-norm rows, each a representative prototype drawn from the 65,536 prototypes of the DINO head, so that a coordinate of $s(x)$ is the projection of the bottleneck onto one prototype. With orthonormal rows R is a rotation of the bottleneck up to the residual of the decorrelation loss (in run 11 two row pairs are collinear, so R spans 254 directions); it preserves distances, and its role is to make the coordinates readable as prototype similarities. R is trained online with a nearest-prototype loss $\mathcal{L}_{\mathrm{nn}}$, which pulls each row toward the prototype it currently matches best, and a decorrelation loss $\mathcal{L}_{\mathrm{cov}}$, which keeps the rows orthogonal; the prototypes of the DINO head are heavily redundant (Govindarajan et al., 2024), and orthogonality caps the number of rows at the width of the bottleneck. Curation begins once the bottleneck's centered kernel alignment (CKA) with the final checkpoint of the finished baseline reaches 90%, at 50k steps for ViT-B and 62k for ViT-L; the class token is at 55% at 50k for ViT-B, which is why the bottleneck was preferred as the basis of the signature. The signature of a stream tile is called an *arrival*.

The density cache. The cache holds M *entries*, each a stored signature drawn from the arrivals with two counters: the hit count S_i, incremented when an arrival lands on the entry, and the step count E_i, incremented every step, both decayed at the same rate, so that their ratio compares recent hits to recent steps. The stored signatures form the *cover* of the region of signature space the stream has visited. A small *reserve* holds arrivals that lie far from every stored signature.

Seed-on-miss. An arrival within the hit radius r of its nearest stored signature is a *hit* and increments that entry's count. An arrival farther than r from every stored signature is a *miss* and is inserted as a new entry, directly while the cache fills and through the reserve once it is full. A new

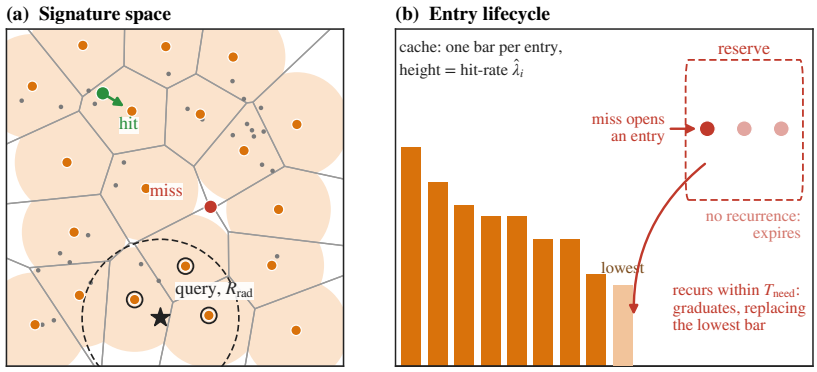

Figure 2: **The density cache.** (a) Stored signatures (orange) sit at least r apart; gray lines are their nearest-entry cells, and the shaded part of a cell is the hit zone within r of its entry. An arrival (gray) in a hit zone is a hit for that entry (green), one outside every hit zone a miss (red). A query (star) reads the hit-rates of the entries within $R_{\text{rad}} = 1.5\,r$ (outlined). (b) Each bar is one entry, height $\hat{\lambda}_i = S_i/E_i$. A miss opens a reserve entry; one that recurs within T_{need} graduates by replacing the lowest-rate entry, and one that does not expires.

entry can only appear where no entry lies within r, so no two stored signatures are closer than r and the cover spans the visited region at a spacing of about r.

The scale of the signature space changes as the encoder trains, so r is re-measured at regular intervals (Section 4): the seed-on-miss rule is replayed over a buffer of recent arrivals for a grid of candidate radii (Table 4), the largest radius at which all M slots still fill is selected, and r is updated by an exponential moving average to avoid jitter between re-measurements; $R_{\text{rad}} = 1.5\,r$ moves with it. Figure 6 shows r rising after the start of curation and contracting as the encoder converges.

Density from counts. The hit-rate $\hat{\lambda}_i = S_i/E_i$ of an entry estimates how often arrivals land on it. The density at a query tile x is the kernel-weighted sum of the hit-rates of all stored signatures within a readout radius $R_{\text{rad}} = 1.5\,r$ of $s(x)$:

$$\hat{p}(x) = \sum_{i:\ \|s(x)-b_i\| \le R_{\text{rad}}} \hat{\lambda}_i\, K\left(\frac{\|s(x)-b_i\|}{R_{\text{rad}}}\right), \qquad K(u) = (1-u^2)^3. \tag{1}$$

A query with no stored signature within R_{rad} reads $\hat{p} = 0$. The local density g of stored signatures is a seeding artifact, set by whichever tiles arrived while the cover formed, and the estimate must not depend on it. Under nearest-signature assignment an entry's hit-rate estimates p/g, and the unnormalized sum over a fixed radius restores p (Lemma 1 of Appendix A.1); a normalized readout does not (Remark 1). R_{rad} has to exceed the spacing r, so that the window holds several entries, and stay below the correlation length L of the density field, below which there is no structure to blur. On a probe of 384k signatures from the ViT-B baseline, L measured about 1.5 times the cover spacing (Appendix A.1), hence $R_{\text{rad}} = 1.5\,r$. Against a k-nearest-neighbour density on 50k signatures of the run-11 panel, $\hat{p}$ agrees in rank (Spearman 0.65) and compresses the range (Remark 4).

Cache churn. Figure 2 follows one arrival through the cache: a hit adds one to an entry's count, an entry that stops being hit decays until it holds the lowest rate, and a miss opens an entry in the reserve, which graduates by replacing the lowest-rate entry if the same morphology recurs within the residency T_{need} and otherwise expires. Stored signatures are frozen at seeding, so graduation is the only way the cover changes. A miss can mean a morphology new to the stream or a known one whose signature has moved beyond r of its entry; the recurrence test treats both alike, which is how the cache tracks the drifting encoder. As the learning rate decays and the representation settles, churn becomes dominated by new morphology: on a ViT-B run with M=8192, replacement falls in lockstep with signature motion while the reserve stays at capacity (Appendix A.1). The churn touches a fringe of low-rate slots; most of the cover persists for tens of thousands of steps (Appendix A.7).

Both counters are multiplied by $\eta = 2^{-1/H}$ each step, so a hit-rate averages over an effective window of $n_{\text{eff}} \approx 2.9H$ steps (Proposition 1 of Appendix A.1), 721 at H=250. H has to be long enough for eviction to rest on several times the $2 \ln M$ observations a reliable $\arg\min$ over M noisy hit-rates needs (Proposition 2), and short enough to stay current with signature drift. Traffic per entry thins as M grows, so for ViT-B H was 250, 500, and 1000 steps at $M = 8192$, 16384, and 32768, and for ViT-L, with three times the traffic, 110 steps at M=8192 (Table 1). Graduation requires two hits, three at M=32768.

Rebalancing by thinning. A batch of size N is filled from a pool of fN candidates drawn from the dataloader; $\hat{p}$ is read for each candidate and each is kept independently with probability

$$q(\hat{p}) = \left(1 + \frac{\hat{p}}{c}\right)^{-a}, \qquad c = c_{\text{frac}}\, p_{\text{ref}}. \tag{2}$$

p_{ref} is an exponential moving average, coefficient 0.999 per step, of the batch-median $\hat{p}$ (Algorithm 1), and $c_{\text{frac}} = 0.25$, so q depends only on $\hat{p}/c$. $\hat{p}$ carries no fixed units and its scale drifts as the encoder trains; the normalization makes every acceptance probability invariant to that drift (Proposition 3 of Appendix A.1). A tile with $\hat{p} \ll c$ is almost always kept, the median tile with probability $(1 + c_{\text{frac}}^{-1})^{-a} \approx 0.45$ at a=0.5, and common tiles are suppressed as $(\hat{p}/c)^{-a}$, so the committed batch has density proportional to $p\,(1 + p/c)^{-a}$, the rare end passed through and the common end flattened toward p^{1-a} (Proposition 4). The exponent a is the *tilt*. The first N survivors in pool order are committed; on a step with fewer than N survivors the shortfall is filled from the rejects with the highest q. The cache updates on the full pool, not on the thinned batch, since a cache that counted only admitted tiles would measure its own output. Each candidate is scored on an un-augmented view by a forward pass without gradient, the *scout pass*, so a rejected candidate costs one such pass. Augmentation of the scout view was not considered.

4 Experimental Setup

Corpus. The corpus is 287M tiles of 224 × 224 pixels at 20× (0.5 μm per pixel) from 69,790 H&E whole-slide images; TCGA contributes 18.1% of the tiles, CPTAC 6.8%, and an in-house cohort 75.1%, and each batch holds the three at these proportions. The loader keeps 16 tile archives open and draws at random among them, so consecutive tiles come from different slides. Table 5 gives the tissue distribution.

Pretraining. ViT-B/16 and ViT-L/16 encoders were trained with DINOv2 for 125k steps on 4 and 8 H100 GPUs, at a global batch of 1024 and 3072, which is 0.45 and 1.34 epochs of the corpus; Table 3 lists every constant of the recipe. The ViT-L schedule is UNI's (Chen et al., 2024), 125k steps at batch 3072. The corpora differ in shape: 287M tiles from 69,790 slides here against 100M tiles from 100,402 slides in Mass-100K, four times as many tiles per slide. At the same step count the ViT-L run covers 1.34 epochs of its tiles where UNI covered 3.8, and 5,500 draws per slide where UNI made 3,800; the ViT-B runs at batch 1024 cover 0.45 epochs and 1,800 draws per slide, the 500k pair 1.8 epochs and 7,300. Runs 1, 9, and 13 of Table 1 are the baselines and carry no curator. Every 125k-schedule encoder was evaluated at the checkpoint of iteration 124k, and the 500k pair at 500k.

Curated runs. Each curated run resumes from its baseline at T_{warm}, 50k steps for ViT-B and 62k for ViT-L (Section 3), with student, teacher, optimizer state, R, and data position retained, so it shares every step before curation with its baseline. Thinning runs per GPU with N the per-GPU batch, 256 for ViT-B and 384 for ViT-L: a pool of fN tiles is scouted, the signatures are all-gathered so the cache updates on all fB arrivals, and the survivors of Eq. 2 are committed and augmented. The oversample factor f was set by trial and error above the reciprocal of the mean acceptance, and the settled χ stayed below f in every run. The scout passes raise the step time by 1.65 and 1.68 times at f=6 and 2.40 at f=8 for ViT-B, 0.96 and 1.32 at f=3 and 6 for ViT-L, and 2.36 for run 14, measured from checkpoint intervals over the final 10k iterations; the ratio below one at ViT-L f=3 is as measured and has no explanation offered here. Table 1 lists the cache constants of every run; $c_{\text{frac}} = 0.25$ throughout. Traffic per entry is fB times the hit fraction, the share of arrivals within r of an entry, divided by M, and H was set so that an entry collects 240 to 630 hits within one window $n_{\text{eff}} \approx 2.9H$. r was re-measured every 500 steps with an EMA weight of 0.2 for ViT-B

		cache size		tilt and pool				renewal				
#	Enc.	M°	hit frac.$^\bullet$	a°	f°	$\chi^\bullet$	$\hat{\lambda}_{90}/\hat{\lambda}_{10}{}^\bullet$	$\lvert\mathcal{R}\rvert^\circ$	T°_{need}	k°	H°	$\tau/n_{\text{eff}}{}^\bullet$
1	ViT-B						*— baseline —*					
2	ViT-B	8192	0.90	0.5	3	2.0	3.0	550	300	2	250	3.7
3	ViT-B	8192	0.90	0.5	6	2.0	2.8	550	300	2	250	1.8
4	ViT-B	8192	0.87	1.0	6	3.3	3.1	550	300	2	250	2.0
5	ViT-B	16384	0.77	0.5	6	2.0	3.0	1100	600	2	500	3.0
6	ViT-B	16384	0.80	1.0	6	3.2	3.2	1100	600	2	500	2.8
7	ViT-B	32768	0.48	0.5	8	1.9	11.1	2200	1200	3	1000	29.6
8	ViT-B	32768	0.38	1.0	8	2.4	61.6	2200	1200	3	1000	10.9
9	ViT-L						*— baseline —*					
10	ViT-L	8192	0.87	0.5	3	2.0	3.1	280	300	2	110	4.6
11	ViT-L	8192	0.88	1.0	6	3.3	4.0	150	300	2	110	4.7
12	ViT-L	8192	0.86	1.5	6	4.7	3.5	150	300	2	110	3.0
13	ViT-B						*— baseline, 500k —*					
14	ViT-B	8192	0.88	1.5	6	5.6	3.6	150	300	2	110	14.7

Table 1: **Run list.** $^\circ$ set, $^\bullet$ measured as medians over the final 10k steps; for run 14 that is the end of a 500k schedule at zero learning rate, with a frozen encoder (at 114k to 124k its hit fraction, cost, spread, and turnover were 0.82, 4.7, 6.1, 5.6). Curation starts at 50k steps (ViT-B) and 62k (ViT-L); global batch 1024 and 3072. M sets the hit fraction, the share of arrivals within r of an entry; a sets the cost $\chi = 1/\mathbb{E}[q]$, candidates scouted per committed tile, which f must exceed, and the hit-rate spread $\hat{\lambda}_{90}/\hat{\lambda}_{10}$ over entries, called the cache bandwidth in the text, is the typicality spectrum the cache sees under that tilt; $|\mathcal{R}|$, T_{need}, k, and H are set so that the cache stays healthy, which shows in the turnover time τ/n_{eff}, M over graduations per step, in effective windows. Thresholds, set by inspection: hit fraction 0.5, spread 2.5, turnover 2 windows.

and every 150 steps with weight 0.6 for ViT-L. Runs 7 and 8 never settle (Table 1, Figure 6) and are reported with that noted. Runs 13 and 14 train ViT-B from scratch on a 500k-step schedule, a baseline and a curated run at a=1.5, f=6, with the cache constants of run 12, including its half-life of 110 steps, which at the ViT-B batch gives at most 238 hits per window and departs from the rule above; curation activated at 50k steps, and the pair is compared only with itself, at 500k.

Figure 6 follows the cache state of every run over training; run 3 sits below the turnover threshold, at 1.8 windows, and is kept with that noted.

Evaluation. Frozen encoders were evaluated on twelve mutation-prediction tasks from the in-house cohort (Table 6) with the GOLDMARK pipeline (Vanderbilt et al., 2026): a slide is the full bag of its 224×224 tissue tiles, each encoded to the post-norm class token of the student (768-dimensional for ViT-B, 1024 for ViT-L), and a gated-attention head is trained per task and split for 120 epochs with an evaluation every 20 (constants in the caption of Table 6). The five stratified shuffle splits hold one third of the slides out and are identical for every encoder. They are drawn at the slide level, so a patient with several slides can fall on both sides of a split, which holds for 8% of the held-out slides on average and 21% in BLCA/FGFR3, and the held-out sets of different splits overlap. The evaluation slides are not part of the pretraining corpus. There is no validation set; the reported AUROC of a split is the best of its six held-out evaluations, applied identically to every encoder. Table 7 reports the mean and standard deviation of this AUROC over the five splits per encoder and task. Four tasks hold fewer than 50 positive slides in a held-out fold (Table 6) and are read with that in mind. AUROC is reported in percent, and every difference between encoders is absolute, in percentage points (pp).

5 Results and Analysis

Mutation prediction. Table 2 gives, per curated run, the fold-paired AUROC difference in pp against the baseline on each task, the pooled mean with its interval, and the precision-weighted mean (D7 of Appendix A.1); a run counts as improved on a task when its mean difference is positive. At ViT-L every run has a positive weighted mean, +0.73, +0.86, and +0.47 for runs 10 to 12 (p=0.002, < 0.001, and 0.053), and run 11 has the largest pooled mean of the three, +0.91 with interval $[+0.4, +1.4]$; the improvement is on the tasks of highest precision, COAD/BRAF by 2.0

	ViT-B, 124k							**ViT-L**, 124k			**ViT-B**, 500k
run	2	3	4	5	6	7	8	10	11	12	14
M	8192	8192	8192	16384	16384	32768	32768	8192	8192	8192	8192
a	0.5	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0	1.5	1.5
f	3	6	6	6	6	8	8	3	6	6	6
UCEC/PTEN	−0.9(0)	−0.3(1)	−0.4(2)	−0.2(2)	−1.2(2)	0.0(3)	+0.3(3)	−0.1(3)	+0.4(3)	−0.2(2)	−0.4(2)
COAD/MSI	−0.3(2)	+0.5(3)	−0.9(1)	+0.4(4)	+0.8(3)	−0.7(2)	0.0(2)	+0.7(4)	+1.0(5)	+0.6(3)	+1.0(4)
COAD/BRAF	+1.4(2)	+1.0(3)	+2.2(3)	−0.7(2)	+3.0(3)	−0.4(3)	+2.4(3)	+3.0(5)	+2.0(4)	+2.2(4)	+0.7(4)
LUAD/EGFR	+2.4(3)	+0.6(2)	−0.3(2)	−0.2(3)	+0.2(3)	+2.7(3)	−0.9(2)	+0.4(3)	+1.5(3)	+0.7(4)	−1.4(0)
LGG/IDH1	+0.2(2)	−0.7(2)	−1.2(0)	−1.6(1)	+0.1(2)	−0.6(1)	+1.2(4)	+1.1(2)	+1.9(3)	+1.7(3)	+1.2(3)
BLCA/FGFR3	−0.7(3)	−0.1(1)	+1.0(3)	−1.7(1)	−0.6(1)	−0.5(2)	−0.6(1)	+0.7(3)	+0.8(3)	+0.8(3)	−0.5(2)
SKCM/BRAF	+0.4(2)	+1.8(4)	+1.1(4)	+1.9(4)	+2.2(4)	0.0(4)	+2.0(4)	0.0(3)	+0.1(3)	−2.1(1)	+0.4(3)
SKCM/NRAS	0.0(1)	−0.4(2)	+1.2(3)	+1.6(3)	+1.9(3)	+0.4(3)	+1.0(4)	+0.3(3)	+1.9(3)	+1.7(4)	−2.2(1)
STAD/PIK3CA	+2.6(4)	+2.0(3)	+0.6(3)	+4.1(4)	−0.2(3)	+1.3(3)	+2.8(2)	−1.3(2)	+1.3(3)	+0.2(3)	+1.9(4)
THCA/BRAF	+0.4(3)	+0.7(4)	+1.3(4)	+0.8(4)	+1.0(5)	+1.3(4)	+0.4(4)	+0.6(3)	+0.1(3)	+1.4(5)	+0.9(4)
THCA/NRAS	+2.0(4)	+4.4(5)	+3.0(4)	+2.6(4)	+3.2(4)	+2.6(3)	+4.1(4)	+0.7(3)	+1.1(3)	+0.6(3)	+1.4(5)
HNSC/PIK3CA	+2.0(3)	−0.6(3)	+1.3(3)	+0.5(3)	0.0(3)	+3.9(4)	−0.2(2)	−2.9(2)	−1.2(2)	−2.8(2)	+0.2(3)
All tasks	+0.80	+0.75	+0.74	+0.62	+0.87	+0.82	+1.04	+0.26	+0.91	+0.42	+0.28
folds won	29	33	32	35	36	35	35	36	38	37	35
s.e.	0.46	0.50	0.48	0.49	0.58	0.51	0.53	0.35	0.38	0.47	0.36
weighted	−0.25	+0.45	+0.06	−0.06	+0.86	+0.07	+0.50	+0.73	+0.86	+0.47	+0.00
wtd. s.e.	0.24	0.27	0.21	0.25	0.29	0.27	0.27	0.24	0.17	0.24	0.22
CI low	+0.15	+0.05	+0.05	−0.27	+0.12	+0.04	+0.24	−0.52	+0.38	−0.46	−0.41
CI high	+1.46	+1.60	+1.42	+1.58	+1.64	+1.70	+1.88	+0.98	+1.38	+1.14	+0.88

Table 2: **Fold-paired results per curated run at the checkpoint of iteration 124k, in percentage points of AUROC.** Each cell is the mean over the five held-out folds of the AUROC difference between the run and its baseline on the same fold, with folds won in parentheses; the last column is run 14 against run 13 at 500k. The pooled rows, over the 60 task-folds: mean, folds won, standard error, precision-weighted mean (wtd.), each task weighted by the inverse variance of its fold-paired differences (D7 of Appendix A.1), its standard error, and the low and high ends of a 95% interval of the pooled mean from a bootstrap over tasks. The interval and the weighted mean treat tasks as independent; the colorectal, thyroid, and melanoma task pairs share slides, so the twelve tasks are closer to ten independent ones. Runs 7 and 8 never settled. Mean AUROC per cell is in Table 7.

to 3.0 pp, and COAD/MSI, LUAD/EGFR, LGG/IDH1, and BLCA/FGFR3, in every ViT-L run. At ViT-B the pooled means of the seven runs are $+0.62$ to $+1.04$ pp, but the weighted means are within ± 0.5 of zero for five of them and resolved only for run 6, $+0.86 \pm 0.29$, with run 8 at $+0.50 \pm 0.27$: the pooled means rest on THCA/NRAS, $+2.0$ to $+4.4$ in every run, and STAD/PIK3CA, both under 50 positives per fold, while UCEC/PTEN and BLCA/FGFR3, the two tasks of highest precision, are at or below the baseline in most runs. Runs 7 and 8, whose caches never settled, sit inside that range, since a starved cache still admits every arrival farther than r from all stored entries. The two thyroid tasks improve in all ten runs; glioma improves in every ViT-L run and in runs 8 and 14 at ViT-B, and holds a class almost to itself in the run-11 cache (Figure 15, class XII); HNSC/PIK3CA, the rarest evaluated tissue, loses in every ViT-L run. The 500k pair asks whether the effect survives a baseline trained from scratch for four times the schedule. On aggregate it does not, $+0.28$ pp, $[-0.4, +0.9]$, p=0.44, a weighted mean of $+0.00 \pm 0.22$; per task it does: both thyroid tasks, both colorectal tasks, and STAD/PIK3CA improve under this third independent baseline as at ViT-L, THCA/NRAS by $+1.4$ on five of five folds, whereas LUAD/EGFR loses 1.4 pp on every fold and SKCM/NRAS 2.2. The thyroid and colorectal mutations are the ones with a distinctive minority morphology within the slide, papillary against follicular architecture, mucinous and lymphocyte-rich patterns; LUAD/EGFR, SKCM/NRAS, BLCA/FGFR3, and UCEC/PTEN, which improve in some runs and lose in others, are mutations without a settled morphological correlate at this magnification. The 500k baseline itself sits 2.8 pp above the 124k baseline and level with the ViT-L baseline at 124k. Across the eight 124k runs with a hit fraction above 0.5, runs 2 to 6 and 10 to 12, the weighted gain follows the hit-rate spread $\hat{\lambda}_{90}/\hat{\lambda}_{10}$ of the run's cache, its bandwidth, with Spearman $+0.75$, p=0.03 (Figure 4), $+0.36$ within the five ViT-B runs and $+0.5$ within the three ViT-L runs; runs 7 and 8, whose spreads are counter noise, and run 14, whose spread is that of a frozen encoder, lie off it, and over all eleven curated runs the correlation is $+0.42$. The bandwidth of the cache, how far apart the hit-rates of its common and rare entries are, is the best single predictor of downstream mutation prediction among the quantities measured in this study.

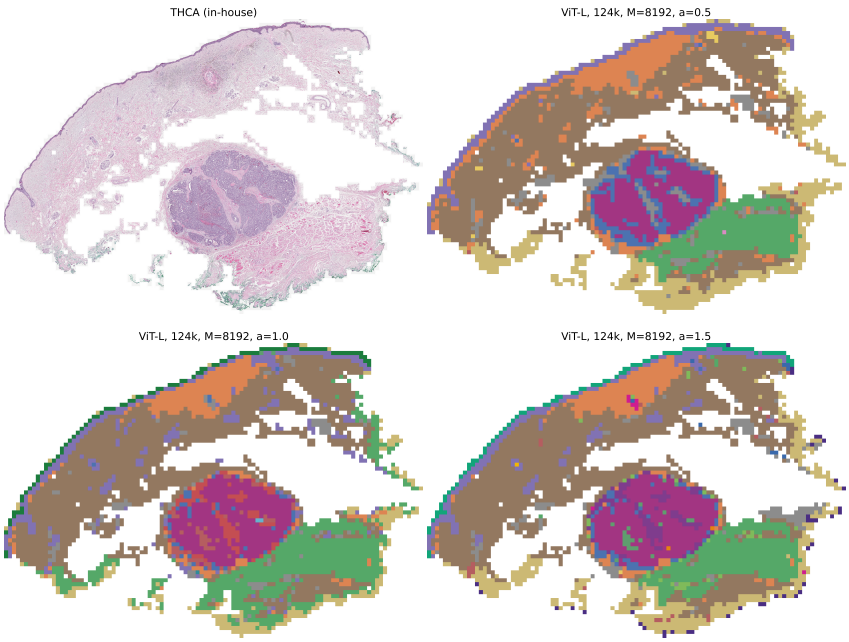

Figure 3: **Cache-derived segmentation of one slide under three tilts.** An in-house thyroid carcinoma slide, H&E top left, then the maps of runs 10, 11, and 12 (ViT-L, M=8192, 124k) at their operating points K=16, 18, and 28 (Algorithm 2). Colors are aligned to run 10; black tiles are beyond the support radius of every cache entry and white cells are tiles absent from the archives (conventions in Appendix A.9).

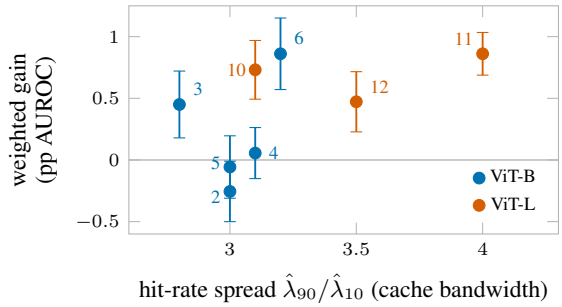

Figure 4: **Cache bandwidth against downstream gain.** Precision-weighted gain, in pp of AUROC, of runs 2 to 6 and 10 to 12 at 124k (Table 2) against the hit-rate spread $\hat{\lambda}_{90}/\hat{\lambda}_{10}$ of its cache, the bandwidth of Table 1, with standard errors; labels are run numbers. Spearman $+0.75$, p=0.03.

The signature space. Figure 5 shows the stored signatures of the three ViT-L caches, with exemplar tiles in Figures 14 to 16. The classes form separated islands in a UMAP supervised on the class labels (Appendix A.10). A pathologist's reading of the run-11 plate names them as adipose tissue, three kinds of stroma, necrosis, blood, lymph node, brain, blank glass, artifacts, and several growth patterns of carcinoma. Five classes hold non-diagnostic material. The exemplar tiles were chosen for distinct tumor types (Appendix A.10), so their spread across types says nothing about the classes. At a=1.0 the largest class, 29% of the scouted pool, is 3% of the committed batch, no class exceeds 10%, and the five non-diagnostic classes move from 25% to 28% (Appendix A.8). The classes are a vocabulary of the stream's morphology, and the tilt flattens the batch over that vocabulary without removing non-diagnostic material from it.

Tissue segmentation. Slides were segmented with Algorithm 2; Figure 3 shows an in-house thyroid slide under the three runs, the carcinoma separated from the surrounding tissue without supervi-

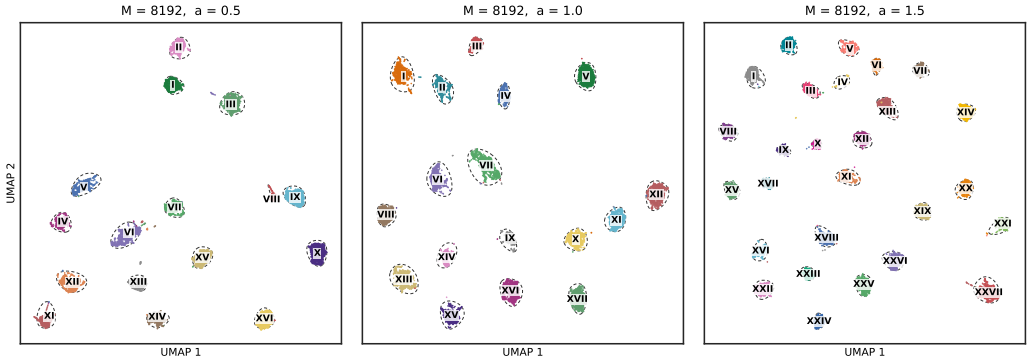

Figure 5: **The signature space of the three ViT-L caches.** UMAP of the 8192 stored signatures of runs 10, 11, and 12 at the checkpoint of iteration 124k, colored by their consensus classes at the operating points K=16, 18, and 28. Exemplar tiles per class are in Figures 14 to 16.

sion, and Figures 9 to 13 five other tumor types. Each cache is segmented at a number of classes read off its consensus-stability curve (Figure 8), and that number moves with the tilt: at ViT-L the stability is highest at K=12, 14, and 30 for a=0.5, 1.0, and 1.5, and the caches were segmented at K=16, 18, and 28; at ViT-B the order holds at M=8192 and reverses at M=16384 (Appendix A.9). The maps show that the stored signatures carry tissue structure at tile resolution; without an annotated reference their accuracy is not measured, and the number of classes is a choice within a band.

6 Conclusion and Limitations

Conclusion. The density cache estimates morphology density inside DINOv2 pretraining on a fixed memory budget, and thinning the batch by its estimate shifts training toward rare morphology. Fold-paired against the baseline at the same checkpoint, every curated run improved mutation prediction on average, by 0.3 to 1.0 percentage points of AUROC pooled over twelve tasks and by 2 to 4 on THCA/NRAS at ViT-B, with the two thyroid tasks improved in every run; at ViT-L the improvement is on the tasks of highest precision, COAD/BRAF by 2 to 3 pp in every run, and two of the three runs are significant under the precision weighting of Appendix A.1, the third at p=0.053, at ViT-B it is on the low-precision tasks and resolved in one run of seven. A stronger tilt resolves finer morphological structure in the cache, the best result at ViT-L was at a tilt of 1.0, and across the settled runs of both scales the weighted gain follows the hit-rate spread of the run's cache, a scale-free ratio that is the quantity to raise. The signature space separates morphology well enough that the stored entries alone segment tissue. Over a 500k-step schedule at ViT-B, curation at a=1.5 was at parity with its own baseline on aggregate, which had itself gained 2.8 pp over the 125k schedule, while the thyroid and colorectal tasks improved under this third independent baseline.

Limitations. A mechanistic understanding of the curation process is incomplete. The ten curated runs were unable to say which cache size or which tilt is preferable. A thorough treatment of the mechanism is the next step, and it is likely to give a more regular behavior across settings than the one reported here. The second is the churn mechanism, which is a baseline to improve on (Appendix A.7). The reserve is fast enough that a morphology enters the cache only if it recurs almost immediately, which is the opposite of what a stronger tilt needs. Therefore, tilts above 1.5 were not attempted under this mechanism, and the larger gains that a stronger tilt would test could not be shown. The corpus, 69,790 slides against the $\mathcal{O}(10^5)$ to $\mathcal{O}(10^6)$ behind published foundation models, under-represents some morphologies, and the same curator on a larger corpus could gain more; even so, the trained encoders come close to those models on several tasks (Table 8). Furthermore, the segmentations provided in this manuscript are qualitative, and finally, the evaluation is on one institution's cohort at slide-level splits.

REPRODUCIBILITY STATEMENT

The code is available at `https://anonymous.4open.science/r/Density-Autocurator-4528`. The repository holds the DINOv2 pretraining pipeline, the curator with its unit tests, the segmentation of Appendix A.9, and one launch script per run of Table 1, which fixes every constant of Tables 3 and 4. The pretraining corpus and the mutation labels of the evaluation cohort are not public as of this writing.

AI USE STATEMENT

A generative AI assistant (Claude Fable) was used to draft and edit prose, with considerable human supervision, to structure the manuscript, to assist with analysis and plotting code, and to formalize the definitions and statements of Appendix A.1 from ideas developed by the authors. The assistant also extracted the technique presented here from a larger unpublished codebase into a standalone repository and prepared its anonymized release. All analyses, numerical results, and figures were produced from the authors' own data and pipelines and were verified by the authors, who take responsibility for the final content, including all claims and artifacts.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Mahmoud Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *International Conference on Learning Representations*, 2023.

Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. *arXiv preprint arXiv:2306.04934*, 2023.

Gabriele Campanella, Ricky Kwan, Eugene Fluder, Jennifer Zeng, Aryeh Stock, Brandon Veremis, Alexandros D. Polydorides, Cyrus Hedvat, Adam Schoenfeld, Chad Vanderbilt, Patricia Kovatch, Carlos Cordon-Cardo, and Thomas J. Fuchs. Computational pathology at health system scale – self-supervised foundation models from three billion images. *arXiv preprint arXiv:2310.07033*, 2023.

Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou. Density-based clustering over an evolving data stream with noise. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pp. 328–339, 2006.

Boqi Chen, Cédric Vincent-Cuaz, Lydia A. Schoenpflug, Manuel Madeira, Lisa Fournier, Vaishnavi Subramanian, Sonali Andani, Samuel Ruiperez-Campillo, Julia E. Vogt, Raphaëlle Luisier, Dorina Thanou, Viktor H. Koelzer, Pascal Frossard, Gabriele Campanella, and Gunnar Rätsch. Revisiting automatic data curation for vision foundation models in digital pathology. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15965 of *LNCS*, pp. 554–564, 2025. arXiv:2503.18709.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.17711.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. On partial prototype collapse in the dino family of self-supervised methods. *arXiv preprint arXiv:2410.14060*, 2024.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *PMLR*, pp. 8175–8195, 2022.

Michael Hahsler and Matthew Bolaños. Clustering data streams based on shared density between micro-clusters. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1449–1461, 2016.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 2525–2534, 2018.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 15630–15649, 2022.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

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 Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 19523–19536, 2022.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J. Hénaff. Active data curation effectively distills large-scale multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.18674.

Chad Vanderbilt, Gabriele Campanella, Siddharth Singi, Swaraj Nanda, Jie-Fu Chen, Ali Kamali, Amir Momeni Boroujeni, David Kim, Mohamed Yakoub, Jamal Benhamida, Meera Hameed, Neeraj Kumar, and Gregory Goldgof. GOLDMARK: Governed outcome-linked diagnostic model assessment reference kit, 2026. URL https://arxiv.org/abs/2603.20848.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26 (4):359–372, 1964.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.10429.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Appendix

A.1 Definitions and results

Statements are given without proof; each is a short calculation from the definitions, and the numbering below is the one used in the main text.

Definitions.

D1 *Signature.* $s(x) = R\,z(x) \in \mathbb{R}^{K'}$, where $z(x)$ is the student DINO-head bottleneck of tile x on an un-augmented view and R has K' unit-norm rows. The signature of a stream tile is an *arrival*.

D2 *Entry.* A stored signature b_i with a hit count S_i and a step count E_i, both multiplied by $\eta = 2^{-1/H}$ at every step, E_i incremented by 1 every step and S_i by 1 at every hit. Its *hit-rate* is $\hat{\lambda}_i = S_i/E_i$.

D3 *Hit and miss.* An arrival is a hit for its nearest entry if it lies within the hit radius r of it, and a miss otherwise.

D4 *Cover.* The set of stored signatures. By seed-on-miss, no two are closer than r, so the cover spans the visited region of signature space at spacing about r; its local density is g.

D5 *Readout.* $\hat{p}(x) = \sum_{i:\|s(x)-b_i\|\le R_{\mathrm{rad}}} \hat{\lambda}_i\, K(\|s(x) - b_i\|/R_{\mathrm{rad}})$ with $K(u) = (1-u^2)^3$ and $R_{\mathrm{rad}} = 1.5\,r$; $\hat{p} = 0$ when no entry lies within R_{rad}.

D6 *Acceptance.* $q(\hat{p}) = (1 + \hat{p}/c)^{-a}$ with $c = c_{\mathrm{frac}}\, p_{\mathrm{ref}}$ and p_{ref} a slow moving average of the batch-median $\hat{p}$.

D7 *Precision-weighted mean over tasks.* For a curated run, let d_{tk} be the fold-paired AUROC difference against the baseline on task t and fold k, $k = 1, \ldots, 5$, and $\bar{d}_t$ and s_t^2 their mean and sample variance over the five folds. The precision of task t is $w_t = 5/s_t^2$, the inverse variance of $\bar{d}_t$, and the precision-weighted mean is $\bar{d}_w = \sum_t w_t \bar{d}_t / \sum_t w_t$ with standard error $(\sum_t w_t)^{-1/2}$. A task is well measured when its folds agree, that is when s_t is small; on this panel that is the tasks with many positive slides per fold (UCEC/PTEN, COAD/MSI, COAD/BRAF, BLCA/FGFR3, THCA/BRAF, Table 6), and the tasks with fewer than 50 positives per fold (THCA/NRAS, SKCM/NRAS, STAD/PIK3CA, HNSC/PIK3CA) carry the least weight. The unweighted pooled mean counts every task-fold once.

Lemma 1 (hit-rate and readout). Let p be the density of arrivals at a location of signature space and g the density of stored signatures there. Under nearest-entry assignment with the hit radius r, an entry collects the arrivals that fall in its hit zone, the part of its Voronoi cell within r of it. Where the cover is dense enough that the cell lies within r of its entry, the hit zone is the cell, of volume proportional to $1/g$, and the hit-rate estimates p/g; the unnormalized readout over a fixed radius then sums over a number of entries proportional to g, so its expectation is proportional to $g \cdot p/g = p$, up to the kernel constant, whenever R_{rad} exceeds the spacing r and stays below the correlation length L of p. Where the cover is sparser than r, the hit zone is the ball of radius r, the hit-rate estimates p times a constant, and the readout scales as $p \cdot g$; the cancellation holds in the dense part of signature space and not in the sparse part, and Table 1 gives the share of arrivals that fall in hit zones as the hit fraction, 0.77 to 0.90 in the settled runs.

Remark 1 (failure modes). A readout normalized by the number of entries within the radius, the Nadaraya-Watson form (Nadaraya, 1964; Watson, 1964), estimates p/g, which depends on the seeding density. Crediting a hit to every entry within the radius makes each hit-rate estimate p and the sum $p \cdot g$. An arrival-count denominator in place of E_i grows with S_i and drives the ratio to a constant. Cancelling the seeding density requires nearest-entry assignment together with the unnormalized sum.

Proposition 1 (effective window). A mean with weights η^k, $k \ge 0$, has effective sample size $(\sum_k \eta^k)^2 / \sum_k \eta^{2k} = (1+\eta)/(1-\eta)$. With $\eta = 2^{-1/H}$ this is $\approx 2.9H$ steps, independent of the length of the stream; the noise of every hit-rate is therefore set by H alone.

Proposition 2 (minimum half-life). For M hit-rates with independent Gaussian noise of equal variance, the expected maximum fluctuation is at most $\sigma\sqrt{2\ln M}$ (Gaussian maximal inequality). For the $\arg\min$ over entries to select on rate rather than on noise, each hit-rate must average over several times $2\ln M$ observations, which sets the minimum half-life at fixed traffic per entry and requires H to grow with M as traffic per entry thins.

Proposition 3 (scale invariance of the acceptance). q depends on $\hat{p}$ only through $\hat{p}/c$ with $c \propto p_{\mathrm{ref}}$, and p_{ref} scales with $\hat{p}$, so a rescaling of the signature space or of the readout leaves every acceptance probability unchanged. The median tile is kept with probability $(1 + c_{\mathrm{frac}}^{-1})^{-a}$.

Proposition 4 (committed density under the tilt). If candidates arrive with density p and are kept independently with probability $q(p) = (1 + p/c)^{-a}$, the committed batch has density proportional to $p\,(1 + p/c)^{-a}$, which equals p for $p \ll c$ and is proportional to $c^a p^{1-a}$ for $p \gg c$. The rare end of the stream is passed through unchanged and the common end is flattened by the exponent $1 - a$; at $a{=}1$ the common end is flat.

Remark 2 (correlation length). The upper bound on R_{rad} is the e^{-1} decorrelation length L of the two-point correlation of the hit-rate field over entry pairs binned by their separation, a construction borrowed from length-scale measurement in turbulent flows. On a fixed probe of 384k signatures from the ViT-B baseline run, the signature cloud had intrinsic dimension $d^* \approx 9$ (two-nearest-neighbour estimator) and $L \approx 1.5$ times the cover spacing, so $R_{\mathrm{rad}} = 1.5\,r$ sits at the limit of the lemma rather than below it; L was measured once, on this run, and used for every run.

Remark 3 (measured churn). On a ViT-B run with $M{=}8192$, across checkpoints from 54k to 124k steps, probe-signature motion fell from $0.34\,r$ to $0.12\,r$ per 2k steps, slot replacement fell in lockstep from 0.49 to 0.14, and the reserve stayed at capacity throughout. A replacement fraction of 0.14 per 2k steps agrees with the entry ages of Appendix A.7, where the median entry of the ViT-B caches at 124k is 12k to 14k steps old and the churn is confined to a fringe of slots.

Remark 4 (readout against a reference). On $50{,}000$ signatures drawn from the $102{,}244$ cached signatures of the eighteen panel slides encoded by run 11 at 124k, $\hat{p}$ from the run-11 cache was compared with a k-nearest-neighbour density estimate ($k{=}50$, intrinsic dimension 9) on the same points. Four entries created on the checkpoint step carry $E_i = 0$ and read hit-rates of 10^8; the 1.5% of points within their reach were excluded. On the rest, $\log\hat{p}$ and the log reference density have Spearman correlation 0.65 and Pearson correlation 0.66, with a log-log slope of 0.16: the readout orders density as the reference does and compresses its range, so it grades commonness in rank rather than in proportion.

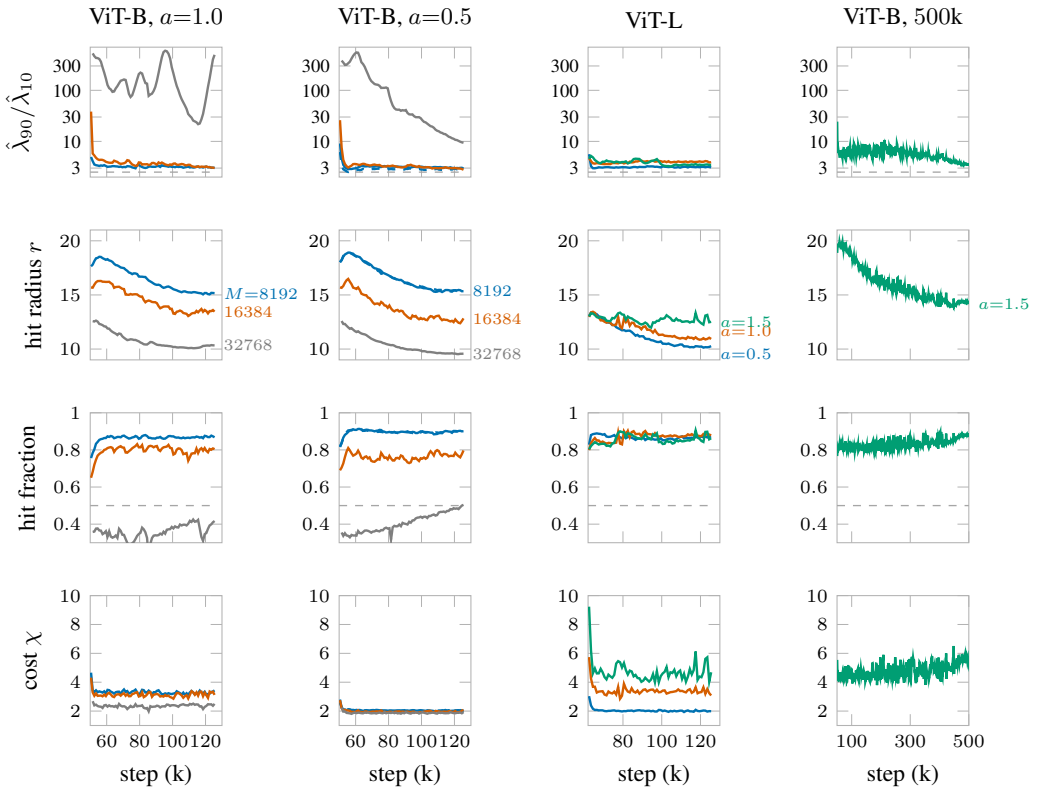

Figure 6: **Cache state during training.** Rows: hit-rate spread $\hat{\lambda}_{90}/\hat{\lambda}_{10}$ over the stored entries (log scale; the dashed line at 2.5 marks the value below which the cache cannot grade commonness), hit radius r in signature-space units, hit fraction (dashed line at 0.5, below which the counters starve), and realised cost χ. Medians over 1000-step bins from the start of curation to the end of training. Columns: the ViT-B runs at a=1.0 (runs 4, 6, 8) and at a=0.5 (runs 2, 3, 5, 7; dashed, f=6), the ViT-L runs 10 to 12, and the 500k-step ViT-B run 14. The level of r scales with M and with the scale of each run's bottleneck and is not comparable across runs; its shape is.

A.2 ALGORITHMS

Algorithm 1 gives one step of the density cache as run during training, and Algorithm 2 the tissue map of one slide.

Algorithm 1 Density cache: one step

Input: batch signatures $\{s(x)\}$; established set $\mathcal{E}$ (capacity M) and reserve $\mathcal{R}$; hit radius r; decay η; warmup T_{warm}; reserve residency T_{need}.
Output: per-tile density $\{\hat{p}(x)\}$ ($\perp$ when the cache declines to report).

if step $< T_{\text{warm}}$ **then return** $\perp$ for all x ▷ signatures not yet stable
end if
for each live entry i **do**
  $S_i \leftarrow \eta S_i$; $E_i \leftarrow \eta E_i + 1$; $\text{age}_i \leftarrow \text{age}_i + 1$ ▷ decay, age
  if $i \in \mathcal{R}$ **and** $\text{age}_i > T_{\text{need}}$ **then** evict i ▷ never-recurring reserve entry expires
  end if
end for
gather $\{s(x)\}$ across workers $\rightarrow X$
if $|\mathcal{E}| < M$ **or** $\text{median}_{i\in\mathcal{E}}\, E_i < \frac{1}{2}(1-\eta)^{-1}$ **then**
  $\hat{p}(x) \leftarrow \perp$ for all x ▷ filling or counters immature: no thinning
else
  for $x \in X$ **do**
    $$\hat{p}(x) \leftarrow \sum_{i\in\mathcal{E}:\, \|s(x)-b_i\|\le 1.5r} \frac{S_i}{E_i+\varepsilon}\, K\big(\|s(x)-b_i\|/1.5r\big)$$ ▷ ε: small constant
  end for
  $p_{\text{ref}} \leftarrow 0.999\, p_{\text{ref}} + 0.001\ \text{median}_{x\in X}\, \hat{p}(x)$ ▷ live reference scale
end if
for $x \in X$ **do** ▷ update runs every step, including fill
  $i^* \leftarrow$ nearest entry to $s(x)$ over $\mathcal{E} \cup \mathcal{R}$
  if $\|s(x) - b_{i^*}\| \le r$ **then** ▷ hit
    $S_{i^*} \leftarrow S_{i^*} + 1$
    if $i^* \in \mathcal{R}$ **and** its hits reach k (2, or 3 at M=32768) **then**
      if $|\mathcal{E}| = M$ **then** evict $\arg\min_{i\in\mathcal{E}} S_i/(E_i+\varepsilon)$; move i^* to $\mathcal{E}$ ▷ promote
    end if
  else ▷ miss (novel)
    if $|\mathcal{E}| < M$ **then** add $(s(x), 0, 0, 0)$ to $\mathcal{E}$ **else** evict oldest reserve entry if full, add to $\mathcal{R}$
  end if
end for
return $\{\hat{p}(x) : x \text{ local}\}$

A.3 PRETRAINING CONFIGURATIONS

Table 3 lists every constant of the DINOv2 recipe, read from the run configuration files, and Table 4 the constants of the curator that are not varied in Table 1.

Table 3: Pretraining recipe, from the run configuration files. A value spanning both columns is shared by the two scales.

Constant	ViT-B	ViT-L
Backbone	ViT-B/16	ViT-L/16
Blocks / heads / width	12 / 12 / 768	24 / 16 / 1024
Register tokens	4	
qk-normalization	yes	
LayerScale init	10^{-5}, uniform	
Stochastic depth	0.4, uniform	
Input size	224	
Global crops	2 at 224	
Global crop scale	0.32 to 1.0	
Local crops	8 at 96	
Local crop scale	0.05 to 0.32	
DINO prototypes	65,536	
iBOT head	separate, 65,536	
iBOT mask ratio	0.1 to 0.5	
iBOT loss weight	1.0	
KoLeo weight	0.1	
Optimizer	AdamW	
Peak learning rate	2.0×10^{-4}	3.5×10^{-4}
Learning-rate scaling	$\sqrt{B/1024}$	
Warmup steps	12,500	
Final learning rate	10^{-6}, cosine	
Patch-embedding lr multiplier	0.2	
Weight decay	0.04 to 0.4, cosine	
Teacher momentum	0.992 to 1, cosine	
Teacher temperature	0.04 to 0.07	
Temperature warmup	37,500 steps	
Last layer frozen	1,250 steps	
Gradient clipping	3.0	
Precision	fp16	
Gradient checkpointing	yes	
GPUs	4 H100	8 H100
Batch per GPU	256	384
Global batch B	1024	3072
Schedule length	125k steps	
Last checkpoint	124k	
Epochs of corpus	0.45	1.34
Step time	1.45 s	5.27 s
Seed	42	

Table 4: Curator constants common to all curated runs, from the run configuration files and the curator code, with the flag or the code location that sets each. Per-run values of M, a, f, $|\mathcal{R}|$, T_{need}, k, and H are in Table 1.

Quantity	Value	Where used	Flag or code
Signature width K'	256	Section 3	`typicality_K_prime`
Learning rate of R	10^{-3}	Section 3	`typicality_repr_lr`
Readout radius R_{rad}	$1.5\,r$	Eq. 1	`typicality_radius_mult`
Kernel $K(u)$	$(1-u^2)^3$, triweight	Eq. 1	`counted_coverage_bank.py`
Counter decay η	$2^{-1/H}$	Section 3	from `typicality_halflife_steps`
Density fraction c_{frac}	0.25	Eq. 2	`typicality_c_frac`
p_{ref} update	EMA of the batch-median $\hat{p}$, coefficient 0.999	Eq. 2	`counted_coverage_bank.py`
Radius re-measurement	every 500 steps (ViT-B), 150 (ViT-L)	Section 3	`typicality_s_sweep_interval`
Radius EMA weight	0.2 (ViT-B), 0.6 (ViT-L)	Section 3	`typicality_s_ema_alpha`
Radius buffer	last $60{,}000$ arrivals (ViT-L)	Section 3	`typicality_s_buffer_size`
Candidate radii	9, log-spaced over 0.3 to $2.0\times$ the buffer scale	Section 3	`typicality_s_grid_points`, `typicality_s_grid_span`
Buffer-scale sample	2048 signatures	Section 3	`counted_coverage_bank.py`

A.4 Data Corpus

Table 5 gives the tissue distribution of the pretraining corpus, and Table 6 the evaluation cohort.

Table 5: Pretraining corpus tissue distribution (TCGA, CPTAC, and the in-house cohort combined). Shares are over the fully labelled corpus..

Tissue / organ system	Tiles	% corpus	Slides	% slides
Gynecologic	40,150,011	14.01	7,860	11.26
Lower GI / colorectal	32,204,729	11.24	7,945	11.38
Lung / thoracic	31,971,874	11.15	8,526	12.22
Breast	26,983,367	9.41	7,308	10.47
Soft tissue & bone	23,382,078	8.16	5,099	7.31
Prostate / testis	17,866,806	6.23	5,042	7.22
CNS (brain)	16,394,931	5.72	3,423	4.90
Bladder / urothelial	16,143,831	5.63	3,493	5.01
Pancreas	15,955,890	5.57	4,714	6.75
Skin / melanoma	12,695,971	4.43	3,139	4.50
Kidney	10,671,278	3.72	2,197	3.15
Upper GI (esoph. / gastric)	8,393,057	2.93	2,722	3.90
Hepatobiliary	7,953,190	2.77	2,127	3.05
Thyroid	7,277,341	2.54	1,492	2.14
Head & neck	7,164,972	2.50	1,789	2.56
Adrenal / paraganglia	5,186,080	1.81	946	1.36
Cancer of unknown primary	3,310,501	1.15	1,306	1.87
Hematolymphoid	1,517,784	0.53	383	0.55
Thymus	1,047,536	0.37	208	0.30
Other / rare	356,810	0.12	71	0.10
Total	286,628,037	100.00	69,790	100.00

Table 6: Evaluation cohort for the twelve biomarker tasks: slide counts in the in-house cohort. Prevalence is the positive fraction; Pos./fold is the number of positive slides in the held-out third of a split. Four tasks (†) hold fewer than 50 positives per held-out fold, and their AUROC carries the widest split-to-split spread in Table 7. The gated-attention head has 512 hidden units, a 384-dimensional attention, and a dropout of 0.25, trained with AdamW at a learning rate of 10^{-4} and weight decay 10^{-4}, one bag per step.

Tumor	Gene	Total	Pos.	Neg.	Prev.%	Pos./fold
BLCA	FGFR3	2031	393	1638	19.3	131
LGG	IDH1	449	216	233	48.1	72
LUAD	EGFR	923	273	650	29.6	91
UCEC	PTEN	2062	972	1090	47.1	324
COAD	MSI	2959	350	2609	11.8	117
COAD	BRAF	2959	318	2641	10.7	106
THCA	BRAF	920	399	521	43.4	133
THCA	NRAS†	920	130	790	14.1	43
SKCM	BRAF	886	213	673	24.0	71
SKCM	NRAS†	886	143	743	16.1	48
STAD	PIK3CA†	742	69	673	9.3	23
HNSC	PIK3CA†	485	84	401	17.3	28

A.5 FULL EVALUATION SWEEP

Table 7 reports mean AUROC and its spread over the five splits for every encoder and task, the 500k pair included.

Table 7: Full evaluation sweep across the twelve biomarker tasks, part 1 of 2. Runs 1 to 12 are at the checkpoint of iteration 124k and runs 13 and 14 at 500k. Each cell reports mean AUROC in percent with the standard deviation over the five splits as a subscript and the difference Δ, in percentage points, between that mean and the mean of the same-scale baseline at 124k; runs 13 and 14 are compared with each other only, so run 13, their baseline, carries a zero difference and run 14 is measured against it. Since the difference of fold means equals the mean of the fold-paired differences, Δ agrees with Table 2 by construction.

					BLCA/FGFR3		LGG/IDH1		LUAD/EGFR	
run	enc.	M	a	f	AUC	Δ	AUC	Δ	AUC	Δ
1	B	–	–	–	$81.0_{\pm 2.6}$	–	$85.7_{\pm 2.8}$	–	$64.1_{\pm 3.2}$	–
2	B	8192	0.5	3	$80.4_{\pm 1.8}$	-0.7	$85.9_{\pm 1.7}$	$+0.2$	$66.5_{\pm 4.2}$	$+2.4$
3	B	8192	0.5	6	$81.0_{\pm 1.5}$	-0.1	$85.0_{\pm 2.3}$	-0.7	$64.8_{\pm 4.6}$	$+0.6$
4	B	8192	1	6	$82.0_{\pm 0.9}$	$+1.0$	$84.5_{\pm 2.5}$	-1.2	$63.9_{\pm 2.1}$	-0.3
5	B	16384	0.5	6	$79.4_{\pm 1.0}$	-1.7	$84.1_{\pm 3.0}$	-1.6	$63.9_{\pm 3.6}$	-0.2
6	B	16384	1	6	$80.5_{\pm 1.6}$	-0.6	$85.9_{\pm 1.7}$	$+0.1$	$64.4_{\pm 4.3}$	$+0.2$
7	B	32768	0.5	8	$80.6_{\pm 1.1}$	-0.5	$85.2_{\pm 2.9}$	-0.6	$66.9_{\pm 2.1}$	$+2.7$
8	B	32768	1	8	$80.5_{\pm 1.9}$	-0.6	$87.0_{\pm 2.7}$	$+1.2$	$63.2_{\pm 2.4}$	-0.9
9	L	–	–	–	$82.4_{\pm 1.9}$	–	$85.8_{\pm 3.2}$	–	$68.8_{\pm 3.0}$	–
10	L	8192	0.5	3	$83.1_{\pm 1.5}$	$+0.7$	$86.9_{\pm 1.8}$	$+1.1$	$69.2_{\pm 4.0}$	$+0.4$
11	L	8192	1	6	$83.2_{\pm 2.5}$	$+0.8$	$87.7_{\pm 1.7}$	$+1.9$	$70.3_{\pm 2.5}$	$+1.5$
12	L	8192	1.5	6	$83.2_{\pm 2.2}$	$+0.8$	$87.6_{\pm 1.7}$	$+1.7$	$69.5_{\pm 5.3}$	$+0.7$
13	B	–	–	–	$83.3_{\pm 2.0}$	–	$86.1_{\pm 3.8}$	–	$69.4_{\pm 2.3}$	–
14	B	8192	1.5	6	$82.8_{\pm 1.4}$	-0.5	$87.3_{\pm 3.4}$	$+1.2$	$67.9_{\pm 2.6}$	-1.4

					UCEC/PTEN		COAD/MSI		COAD/BRAF	
run	enc.	M	a	f	AUC	Δ	AUC	Δ	AUC	Δ
1	B	–	–	–	$82.9_{\pm 0.9}$	–	$87.8_{\pm 1.5}$	–	$65.4_{\pm 4.3}$	–
2	B	8192	0.5	3	$82.0_{\pm 0.4}$	-0.9	$87.5_{\pm 2.3}$	-0.3	$66.8_{\pm 1.7}$	$+1.4$
3	B	8192	0.5	6	$82.6_{\pm 1.3}$	-0.3	$88.3_{\pm 0.7}$	$+0.5$	$66.4_{\pm 2.0}$	$+1.0$
4	B	8192	1	6	$82.5_{\pm 1.0}$	-0.4	$86.9_{\pm 1.5}$	-0.9	$67.6_{\pm 3.2}$	$+2.2$
5	B	16384	0.5	6	$82.6_{\pm 1.0}$	-0.2	$88.2_{\pm 1.0}$	$+0.4$	$64.7_{\pm 2.2}$	-0.7
6	B	16384	1	6	$81.7_{\pm 1.8}$	-1.2	$88.6_{\pm 0.9}$	$+0.8$	$68.4_{\pm 2.6}$	$+3.0$
7	B	32768	0.5	8	$82.9_{\pm 0.7}$	0.0	$87.1_{\pm 0.6}$	-0.7	$65.0_{\pm 2.6}$	-0.4
8	B	32768	1	8	$83.1_{\pm 0.8}$	$+0.3$	$87.8_{\pm 0.6}$	0.0	$67.8_{\pm 4.0}$	$+2.4$
9	L	–	–	–	$85.2_{\pm 0.5}$	–	$90.8_{\pm 1.4}$	–	$67.4_{\pm 3.5}$	–
10	L	8192	0.5	3	$85.1_{\pm 1.6}$	-0.1	$91.5_{\pm 1.2}$	$+0.7$	$70.4_{\pm 4.2}$	$+3.0$
11	L	8192	1	6	$85.6_{\pm 1.1}$	$+0.4$	$91.8_{\pm 1.1}$	$+1.0$	$69.4_{\pm 1.9}$	$+2.0$
12	L	8192	1.5	6	$85.0_{\pm 0.7}$	-0.2	$91.3_{\pm 1.6}$	$+0.6$	$69.6_{\pm 3.2}$	$+2.2$
13	B	–	–	–	$84.7_{\pm 1.0}$	–	$90.2_{\pm 0.7}$	–	$68.1_{\pm 1.1}$	–
14	B	8192	1.5	6	$84.3_{\pm 0.8}$	-0.4	$91.2_{\pm 0.6}$	$+1.0$	$68.8_{\pm 1.6}$	$+0.7$

Table 7: Full evaluation sweep, part 2 of 2; conventions as in part 1.

					THCA/BRAF		THCA/NRAS		SKCM/BRAF	
run	enc.	M	a	f	AUC	Δ	AUC	Δ	AUC	Δ
1	B	–	–	–	$83.3_{\pm 1.6}$	–	$70.7_{\pm 2.5}$	–	$63.4_{\pm 2.1}$	–
2	B	8192	0.5	3	$83.8_{\pm 1.2}$	$+0.4$	$72.7_{\pm 5.6}$	$+2.0$	$63.8_{\pm 2.0}$	$+0.4$
3	B	8192	0.5	6	$84.1_{\pm 1.4}$	$+0.7$	$75.1_{\pm 1.2}$	$+4.4$	$65.2_{\pm 3.4}$	$+1.8$
4	B	8192	1	6	$84.7_{\pm 1.2}$	$+1.3$	$73.7_{\pm 3.7}$	$+3.0$	$64.4_{\pm 2.9}$	$+1.1$
5	B	16384	0.5	6	$84.1_{\pm 1.9}$	$+0.8$	$73.3_{\pm 4.9}$	$+2.6$	$65.3_{\pm 2.8}$	$+1.9$
6	B	16384	1	6	$84.3_{\pm 2.0}$	$+1.0$	$73.9_{\pm 3.6}$	$+3.2$	$65.6_{\pm 1.7}$	$+2.2$
7	B	32768	0.5	8	$84.6_{\pm 1.6}$	$+1.3$	$73.3_{\pm 5.1}$	$+2.6$	$63.3_{\pm 1.8}$	0.0
8	B	32768	1	8	$83.7_{\pm 0.9}$	$+0.4$	$74.8_{\pm 3.5}$	$+4.1$	$65.3_{\pm 3.1}$	$+2.0$
9	L	–	–	–	$87.8_{\pm 1.4}$	–	$80.0_{\pm 3.5}$	–	$68.3_{\pm 2.0}$	–
10	L	8192	0.5	3	$88.5_{\pm 2.4}$	$+0.6$	$80.8_{\pm 1.9}$	$+0.7$	$68.2_{\pm 2.3}$	0.0
11	L	8192	1	6	$88.0_{\pm 2.7}$	$+0.1$	$81.1_{\pm 1.0}$	$+1.1$	$68.3_{\pm 1.7}$	$+0.1$
12	L	8192	1.5	6	$89.3_{\pm 1.3}$	$+1.4$	$80.7_{\pm 3.1}$	$+0.6$	$66.2_{\pm 2.0}$	-2.1
13	B	–	–	–	$87.0_{\pm 1.6}$	–	$78.2_{\pm 3.0}$	–	$66.4_{\pm 3.0}$	–
14	B	8192	1.5	6	$87.9_{\pm 2.2}$	$+0.9$	$79.6_{\pm 3.6}$	$+1.4$	$66.8_{\pm 1.2}$	$+0.4$

					SKCM/NRAS		HNSC/PIK3CA		STAD/PIK3CA	
run	enc.	M	a	f	AUC	Δ	AUC	Δ	AUC	Δ
1	B	–	–	–	$55.8_{\pm 3.9}$	–	$58.5_{\pm 3.9}$	–	$55.4_{\pm 5.9}$	–
2	B	8192	0.5	3	$55.8_{\pm 4.6}$	0.0	$60.5_{\pm 2.3}$	$+2.0$	$58.0_{\pm 5.3}$	$+2.6$
3	B	8192	0.5	6	$55.4_{\pm 2.3}$	-0.4	$57.9_{\pm 4.6}$	-0.6	$57.4_{\pm 6.7}$	$+2.0$
4	B	8192	1	6	$56.9_{\pm 6.4}$	$+1.2$	$59.7_{\pm 3.5}$	$+1.3$	$56.1_{\pm 5.5}$	$+0.6$
5	B	16384	0.5	6	$57.4_{\pm 5.5}$	$+1.6$	$59.0_{\pm 3.0}$	$+0.5$	$59.5_{\pm 5.6}$	$+4.1$
6	B	16384	1	6	$57.7_{\pm 4.2}$	$+1.9$	$58.4_{\pm 5.2}$	0.0	$55.2_{\pm 5.3}$	-0.2
7	B	32768	0.5	8	$56.1_{\pm 4.0}$	$+0.4$	$62.4_{\pm 4.7}$	$+3.9$	$56.8_{\pm 2.3}$	$+1.3$
8	B	32768	1	8	$56.8_{\pm 3.7}$	$+1.0$	$58.3_{\pm 3.4}$	-0.2	$58.3_{\pm 5.9}$	$+2.8$
9	L	–	–	–	$54.1_{\pm 4.3}$	–	$62.4_{\pm 2.5}$	–	$57.5_{\pm 1.6}$	–
10	L	8192	0.5	3	$54.4_{\pm 2.9}$	$+0.3$	$59.5_{\pm 6.1}$	-2.9	$56.2_{\pm 3.8}$	-1.3
11	L	8192	1	6	$56.0_{\pm 3.5}$	$+1.9$	$61.2_{\pm 6.9}$	-1.2	$58.8_{\pm 4.9}$	$+1.3$
12	L	8192	1.5	6	$55.8_{\pm 3.8}$	$+1.7$	$59.6_{\pm 6.5}$	-2.8	$57.8_{\pm 7.3}$	$+0.2$
13	B	–	–	–	$55.7_{\pm 4.6}$	–	$62.8_{\pm 5.5}$	–	$55.7_{\pm 5.2}$	–
14	B	8192	1.5	6	$53.5_{\pm 4.2}$	-2.2	$63.0_{\pm 5.4}$	$+0.2$	$57.7_{\pm 2.7}$	$+1.9$

A.6 PUBLISHED ENCODERS ON THE SAME TASKS

Table 8 gives, for the twelve tasks, the AUROC of five published pathology encoders under the same GOLDMARK evaluation, taken from the GOLDMARK benchmark (Vanderbilt et al., 2026), beside run 11 and run 14, with the number of slides in each encoder's pretraining corpus.

Table 8: AUROC in percent, mean and standard deviation over the five splits, of five published encoders, from the GOLDMARK benchmark (Vanderbilt et al., 2026), and of runs 11 and 14 on the twelve tasks under the same evaluation. The second header row gives the number of whole-slide images in each encoder's pretraining corpus; this work's corpus holds 69,790. Bold marks a value of run 11 or run 14 within one point of AUROC of at least one published encoder.

	UNI	Virchow	Virchow2	Prov-GigaPath	H-Optimus-0	run 11	run 14
slides	100,402	1.5M	3.1M	171,189	500k	69,790	69,790
UCEC/PTEN	$87.9_{\pm 0.8}$	$87.1_{\pm 1.2}$	$88.8_{\pm 0.8}$	$86.8_{\pm 1.0}$	$89.2_{\pm 0.6}$	$85.6_{\pm 1.1}$	$84.3_{\pm 0.8}$
COAD/MSI	$95.5_{\pm 1.4}$	$94.8_{\pm 1.1}$	$96.7_{\pm 1.0}$	$96.0_{\pm 1.3}$	$96.7_{\pm 1.2}$	$91.8_{\pm 1.1}$	$91.2_{\pm 0.6}$
COAD/BRAF	$78.6_{\pm 1.6}$	$76.5_{\pm 2.7}$	$80.9_{\pm 0.8}$	$79.6_{\pm 1.9}$	$80.5_{\pm 3.2}$	$69.4_{\pm 1.9}$	$68.8_{\pm 1.6}$
LUAD/EGFR	$74.9_{\pm 2.6}$	$74.0_{\pm 1.1}$	$76.6_{\pm 2.9}$	$78.1_{\pm 2.5}$	$77.6_{\pm 2.1}$	$70.3_{\pm 2.5}$	$67.9_{\pm 2.6}$
LGG/IDH1	$90.5_{\pm 2.2}$	$88.4_{\pm 2.8}$	$89.7_{\pm 2.1}$	$92.0_{\pm 2.1}$	$91.9_{\pm 1.8}$	$\mathbf{87.7}_{\pm 1.7}$	$87.3_{\pm 3.4}$
BLCA/FGFR3	$89.4_{\pm 1.6}$	$89.2_{\pm 0.7}$	$89.0_{\pm 1.9}$	$89.4_{\pm 1.4}$	$89.6_{\pm 0.6}$	$83.2_{\pm 2.5}$	$82.8_{\pm 1.4}$
SKCM/BRAF	$71.9_{\pm 3.7}$	$67.8_{\pm 2.9}$	$72.4_{\pm 2.7}$	$72.3_{\pm 3.3}$	$74.9_{\pm 2.3}$	$\mathbf{68.3}_{\pm 1.7}$	$66.8_{\pm 1.2}$
SKCM/NRAS	$62.9_{\pm 0.4}$	$60.5_{\pm 2.9}$	$65.0_{\pm 1.8}$	$63.3_{\pm 2.3}$	$62.6_{\pm 2.2}$	$56.0_{\pm 3.5}$	$53.5_{\pm 4.2}$
STAD/PIK3CA	$62.0_{\pm 2.5}$	$61.5_{\pm 5.3}$	$62.1_{\pm 6.0}$	$61.1_{\pm 2.7}$	$64.0_{\pm 5.6}$	$58.8_{\pm 4.9}$	$57.7_{\pm 2.7}$
THCA/BRAF	$92.7_{\pm 1.5}$	$90.5_{\pm 2.2}$	$92.8_{\pm 1.5}$	$92.0_{\pm 1.0}$	$93.4_{\pm 1.4}$	$88.0_{\pm 2.7}$	$87.9_{\pm 2.2}$
THCA/NRAS	$87.4_{\pm 2.4}$	$86.3_{\pm 2.7}$	$89.0_{\pm 3.0}$	$87.4_{\pm 2.6}$	$89.9_{\pm 1.6}$	$81.1_{\pm 1.0}$	$79.6_{\pm 3.6}$
HNSC/PIK3CA	$65.6_{\pm 5.3}$	$58.1_{\pm 4.5}$	$58.4_{\pm 3.1}$	$58.9_{\pm 2.3}$	$62.3_{\pm 5.1}$	$\mathbf{61.2}_{\pm 6.9}$	$\mathbf{63.0}_{\pm 5.4}$

A.7 CACHE CHURN

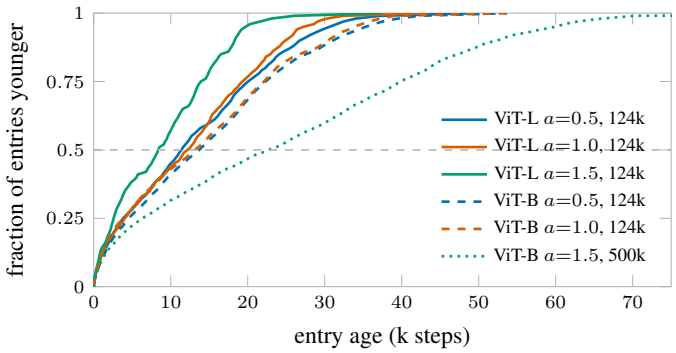

Figure 7: Cumulative distribution of entry age in the density cache, the number of steps since each stored entry was created: the three ViT-L caches and the ViT-B caches of runs 3 and 4 at the checkpoint of iteration 124k (62k and 74k curated steps), and the 500k-step ViT-B cache of run 14 at iteration 500k.

Figure 7 gives the age of every stored entry at the checkpoint. The cache is a long memory with a churning fringe. At 124k the median entry is 11,400 and 12,200 steps old in the ViT-L caches at a=0.5 and 1.0 and 12,400 to 13,700 in the ViT-B caches of runs 3 to 6; more than half of the entries are older than 10k steps and 9 to 12% are younger than 1k. At ViT-L a=1.5 the median falls to 8,400 and the share older than 10k to 43%. Graduation replaces a small set of low-rate slots repeatedly, 5.5 times per step at ViT-L a=1.0 and 8.4 at a=1.5, while most of the cover persists; the ratio M over graduations per step, about 1,500 and 990 steps for those two runs, is the turnover time of that fringe and not the lifetime of a typical entry. In every run the reserve empties by flushing rather than by expiry, since new misses push out older ones within a step at ViT-L and within a few steps at ViT-B, so a morphology enters the cache only if it recurs almost immediately and T_{need} never binds. Renewal in this version is driven by the encoder, not by the stream. An entry is replaced only when misses recur, and misses come from signatures moving out of the hit zones of their entries, so the churn rate follows the learning rate: in run 14 the graduations per step fall from 4.8 to 2.0 as the learning rate decays to zero over the last 150k steps, the hit fraction rises from 0.82 to 0.88, the hit-rate spread compresses from 6.1 to 3.6, and the median entry age rises from 5,500 steps at 440k to 22,700 at 500k and was still rising. The cache did not reach a steady turnover; the cover froze with the encoder, its entries seeded under earlier encoder states, and the density signal it reads flattened with it. A steady renewal set by the stream rather than by drift, eviction by age or by coverage in place of the lowest hit-rate, is the change for the next version.

A.8 BATCH COMPOSITION

Table 9 gives, for the run-11 cache at the checkpoint of iteration 124k, the share of each class in the scouted pool and in the committed batch, from 2,000 pools of $fN = 2{,}304$ tiles drawn from the training loader in its normal order and thinned with the run's own a, c_{frac}, and stored p_{ref}; each tile is assigned to the class of its nearest entry, or to none if it lies beyond that entry's support radius. The realised cost over these pools was $\chi = 3.47$. The largest class of the stream, a mixed carcinoma class that holds 29% of the pool, is 3% of the committed batch, and no class exceeds one tenth of the batch; blood, artifacts, and lymph node, the next three largest, are reduced, and every other class is raised, since at $a{=}1$ the acceptance flattens toward equal shares and lifts every class below the flat share, non-diagnostic classes included. The five non-diagnostic classes, adipose tissue, blank glass, artifacts, blood, and necrosis, rank ninth, tenth, third, second, and sixteenth of nineteen by pool share, and together they rise from 25.1% of the pool to 27.8% of the batch, while the carcinoma classes fall from 46.1% to 31.5%. Tiles with $\hat{p} = 0$, which the curator always admits, are 0.36% of the pool and 1.25% of the batch; by nearest entry, 26% of them lie beyond every support radius and the rest are spread over the tissue classes, with 2% nearest to the blank class, 2.6% to blood, and fewer than 0.1% to artifacts, so the admission of unrecognized tiles does not select for artifacts.

Table 9: Class shares of the scouted pool and of the committed batch for run 11 at 124k, and their ratio. Classes are the $K{=}18$ consensus classes of the cache with the pathologist's labels of Figure 15.

Class	pool %	batch %	ratio
Cancer, mixed	29.4	3.0	0.10
Blood	7.1	6.2	0.87
Artifacts	6.6	4.7	0.72
Lymph node	6.2	5.5	0.89
Cancer, squamoid	5.7	7.9	1.39
Cancer, NOS	5.2	7.7	1.47
Cancer, clusters	4.9	7.9	1.61
Brain, some necrosis	4.9	5.4	1.11
Adipose tissue	4.7	8.3	1.75
Blank	4.4	5.6	1.28
Cancer, mostly adenocarcinoma	4.4	8.2	1.89
Stroma, cell-poor	3.6	7.7	2.13
Stroma	3.0	4.7	1.58
Brain and necrosis	2.9	5.0	1.72
Stroma, cell-rich	2.4	3.9	1.66
Necrosis	2.3	3.0	1.29
Cancer, solid	2.3	4.8	2.10
beyond every support radius	0.1	0.4	3.42
12 entries, not on the plate	0.1	0.3	2.32

A.9 CACHE-DERIVED SEGMENTATION

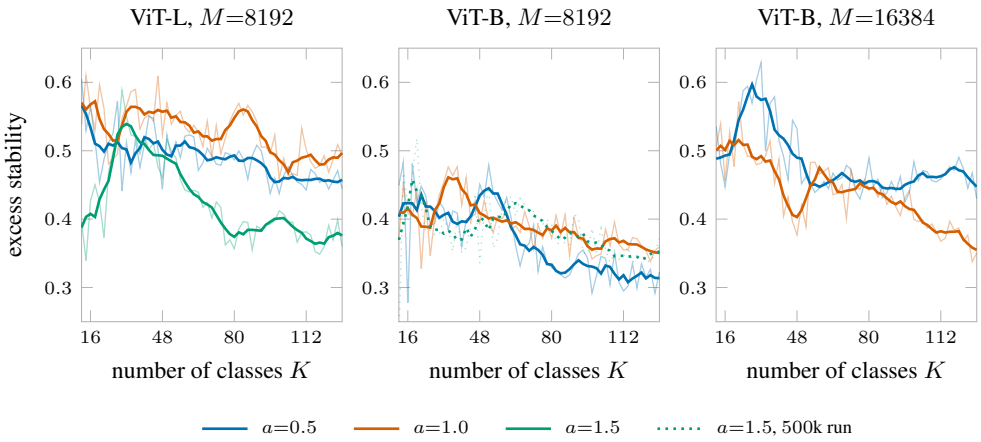

Figure 8: Calibrated stability against the number of classes K for every cache at the checkpoint of iteration 124k, and for run 14 at 500k. Excess of the consensus ARI between partitions from disjoint seed sets over the same protocol applied to a covariance-matched Gaussian sample; thin lines are the values at every second K, thick lines 5-point running means. Left, the ViT-L caches of runs 10 to 12; middle, the ViT-B caches of runs 3 and 4 at M=8192 with run 14 at 500k dotted; right, the ViT-B caches of runs 5 and 6 at M=16384.

Classes and the choice of K. The entries of each cache were partitioned by consensus clustering: twenty-five seeded k-means runs yield a co-association matrix whose elements are the fraction of runs in which two entries share a cluster, and the average-linkage agglomeration of this matrix was cut at K classes. Plain k-means was not reproducible, with an adjusted Rand index (ARI) between seeds of 0.46 to 0.64 at usable K. Since raw stability favors coarse partitions (a single-class partition trivially attains an ARI of 1), each K was scored by the excess of its stability (ARI between consensus partitions from disjoint seed sets, and against a 90% subsample of entries) over the same protocol applied to a multivariate Gaussian sample with the mean and covariance of the entries. Over $K = 12$ to 128 in steps of 2 (Figure 8), the excess is highest at K=12 (0.61) for run 10, at the low edge of the sweep, at K=14 (0.61) for run 11, within 0.02 of that value out to K=46, and at K=30 (0.59) for run 12; the 5-point running means peak at K=12, 18, and 32. The caches were segmented at K=16, 18, and 28, read off the curves by eye near those maxima. The same sweep on the ViT-B caches gave, at M=8192, a plateau within 0.02 of its maximum from K=14 to 52 for run 3 (a=0.5) and a peak at K=34 to 42 for run 4 (a=1.0); at M=16384, peaks at K=32 for run 5 (a=0.5) and K=22 for run 6 (a=1.0); and for run 14 (a=1.5) at 500k a raw peak at K=20 with a smoothed peak at K=18, each inside the other's 0.02 band. Each cache was swept once; the peaks of the ViT-B curves are shallow, the a=0.5 ViT-L peak lies at the low edge of the sweep, and every operating point was read off its curve by eye, a choice within a band rather than a computed optimum. Entries created on the save step of the checkpoint itself (E_i=0, namely 3, 5, and 12 entries in the caches of runs 10, 11, and 12) were excluded from clustering and assigned to their nearest clustered entry.

Algorithm 2 Cache-derived tissue map of one slide

Input: encoder and cache of one run (entries $\{b_i\}$, hit radius r, projection R); class table at that run's K; the tiles of the slide on their grid.
Output: one class per tile, or none.

1: **Signature.** For each tile, average the backbone class token over its four 90° rotations and pass it through the DINO head and R: $s(x)$.
2: **Support.** Radius of entry i: its distance to its 10th-nearest entry, rescaled so that the median radius is $1.5\,r$. A tile farther than the radius of every entry is *unsupported*: it receives no class, is drawn black, and is excluded from steps 4 and 5.
3: **Class model.** For each class c, a diagonal Gaussian $\mathcal{N}(\mu_c, \Sigma_c)$ fitted to its entries, with prior π_c proportional to the number of entries.
4: **Slide adaptation.** Repeat four times: (E) for every supported tile i, compute its responsibility $q_i(c)$ for each class c from the current Gaussians and priors; (M) re-estimate each μ_c, Σ_c, and π_c from the tiles of the slide weighted by $q_i(c)$, blended with the cache's own estimate for that class as if the cache contributed 50 tiles. A class with few tiles on the slide therefore stays where the cache put it, and a class with many tiles moves toward them; no class can vanish or merge.
5: **Smoothing.** Ten sweeps of $q_i \leftarrow 0.5\,q_i^{\text{cache}} + 0.5\sum_{j\in\mathcal{N}_8(i)} \hat{w}_{ij}\,q_j$, with $w_{ij} = \exp\big((\cos(\text{CLS}_i, \text{CLS}_j) - 1)/0.10\big)$ from the class tokens of the same encoder and $\hat{w}$ row-normalized.
6: **return** $\arg\max_c q_i(c)$ for every supported tile.

Figure conventions. For each figure slide, the colors of runs 11 and 12 were aligned to those of run 10 by a rectangular Hungarian assignment on the tile-label confusion matrix of that slide; since the three caches have different K, unmatched classes retain their own colors, and the correspondence across runs is best-effort rather than exact. The figure slides were screened for a single tissue section, with at least 95% of the tiles in one connected component of the tile grid, one in-house slide per tumor type shown; the thyroid slide is in Figure 3 and the others in Figures 9 to 13.

A.10 SIGNATURE SPACE OF THE VIT-L CACHES

Figures 14 to 16 show, for each ViT-L cache, the UMAP of Figure 5 with six exemplar tiles per class, labeled by tumor type. The UMAP is supervised on the class labels (UMAP with 30 neighbors, minimum distance 0.25, Manhattan metric, target weight 0.35, seed 2026) on the 8192×256 stored signatures. Exemplars are drawn from a pool of tiles of the nine evaluation cohorts, scored against each class with the triweight kernel at R_{rad}, and the six with the highest score times class purity are taken under three passes that prefer distinct tumor types, then distinct slides, then distinct tiles; the spread of tumor types on a plate is therefore by construction. The classes of run 11 carry a pathologist's reading. Classes with fewer than 50 entries are not shown, one class of 12 entries in the run-11 cache, so seventeen and twenty-seven classes are shown for $K{=}18$ and 28.

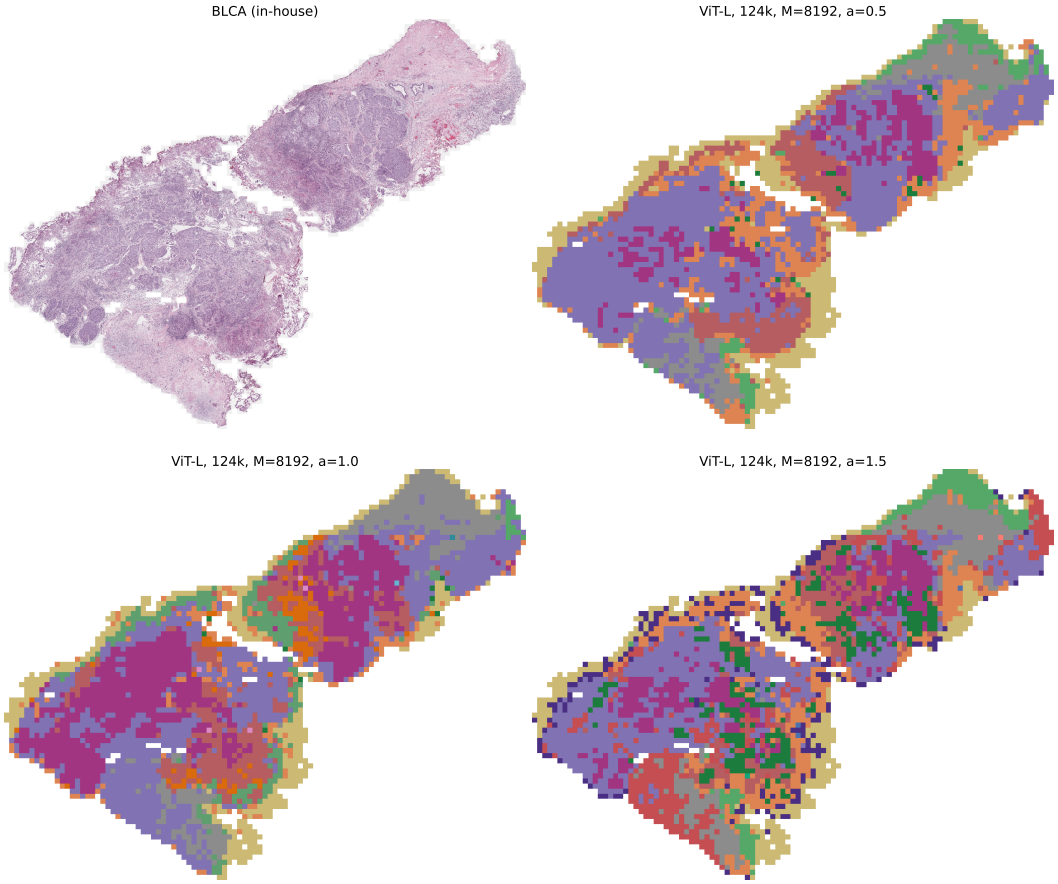

Figure 9: Cache-derived segmentation of an in-house bladder carcinoma slide (BLCA) under the three ViT-L caches: H&E top left, then runs 10, 11, and 12 at their operating points $K=16$, 18, and 28. Colors are aligned to run 10 by a rectangular Hungarian assignment on the tile-label confusion of this slide, with unmatched classes in their own colors; black marks tiles beyond the support radius of every entry, white marks tiles absent from the corpus archives.

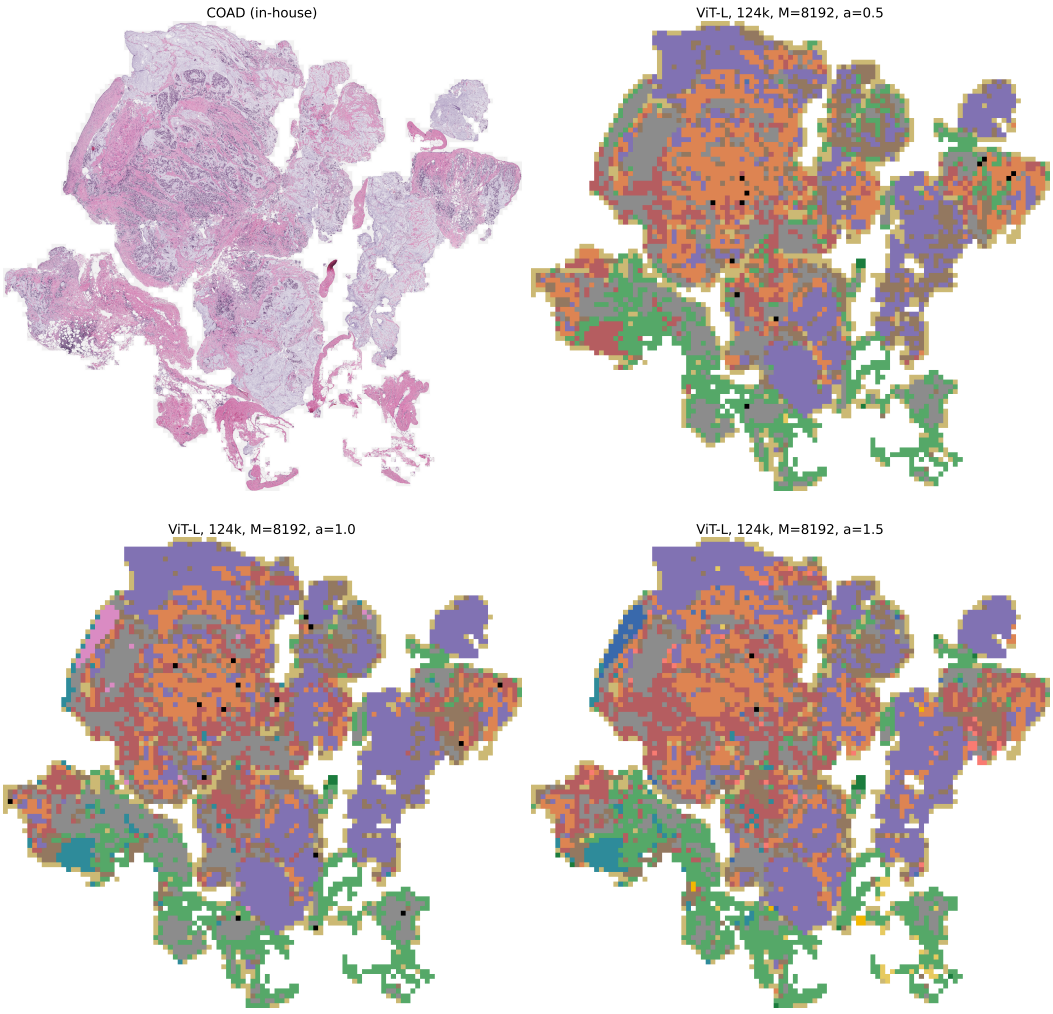

Figure 10: Cache-derived segmentation of an in-house colorectal adenocarcinoma slide (COAD) under the three ViT-L caches: H&E top left, then runs 10, 11, and 12 at their operating points K=16, 18, and 28. Colors are aligned to run 10 by a rectangular Hungarian assignment on the tile-label confusion of this slide, with unmatched classes in their own colors; black marks tiles beyond the support radius of every entry, white marks tiles absent from the corpus archives.

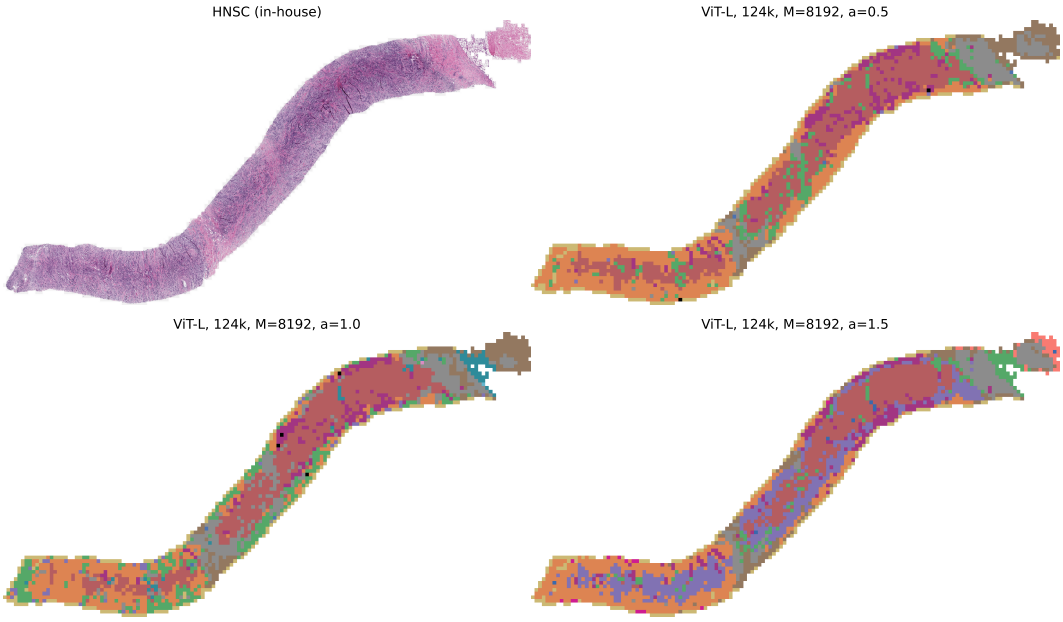

Figure 11: Cache-derived segmentation of an in-house head and neck carcinoma slide (HNSC) under the three ViT-L caches: H&E top left, then runs 10, 11, and 12 at their operating points K=16, 18, and 28. Colors are aligned to run 10 by a rectangular Hungarian assignment on the tile-label confusion of this slide, with unmatched classes in their own colors; black marks tiles beyond the support radius of every entry, white marks tiles absent from the corpus archives.

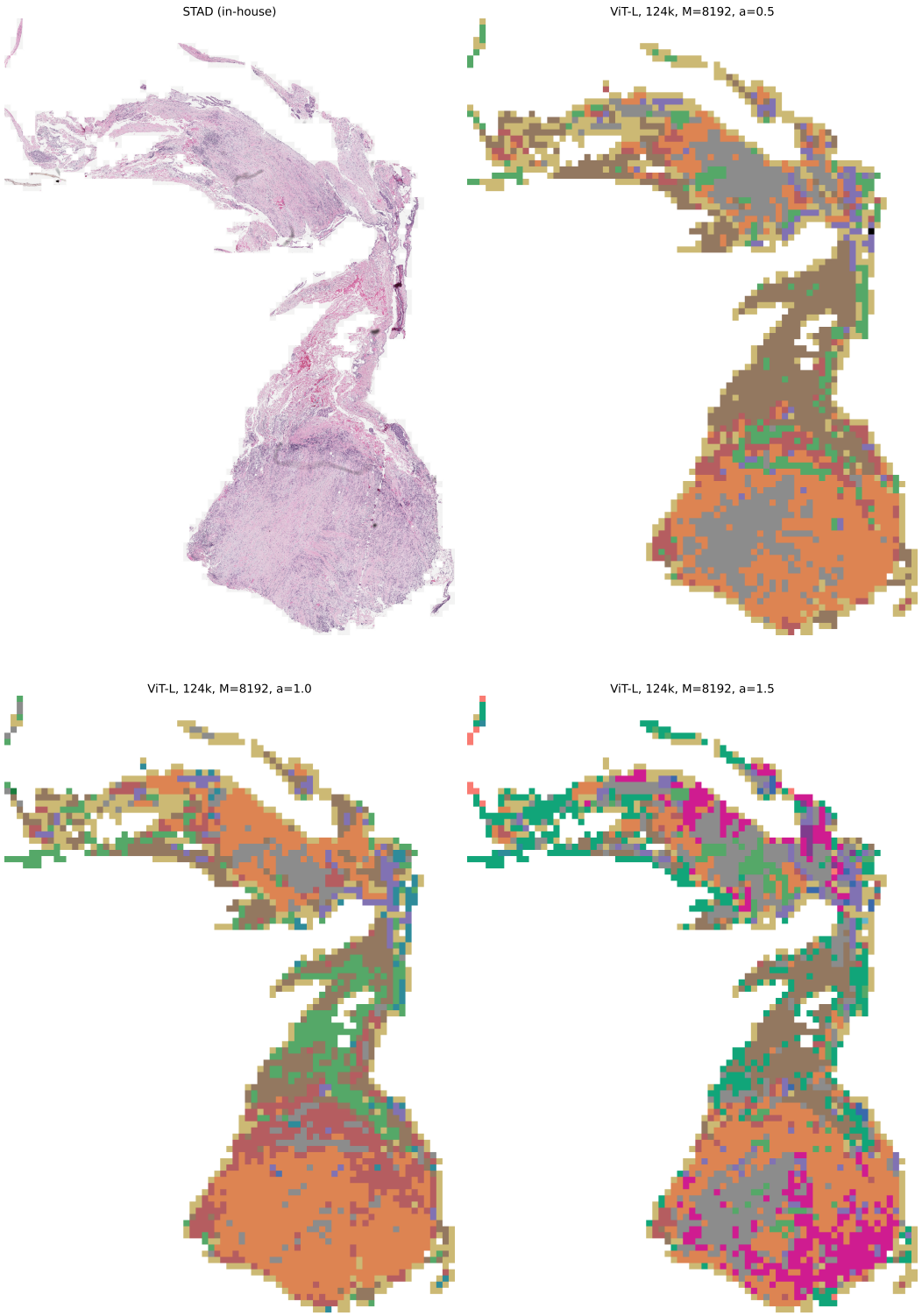

Figure 12: Cache-derived segmentation of an in-house gastric adenocarcinoma slide (STAD) under the three ViT-L caches: H&E top left, then runs 10, 11, and 12 at their operating points $K{=}16$, 18, and 28. Colors are aligned to run 10 by a rectangular Hungarian assignment on the tile-label confusion of this slide, with unmatched classes in their own colors; black marks tiles beyond the support radius of every entry, white marks tiles absent from the corpus archives.

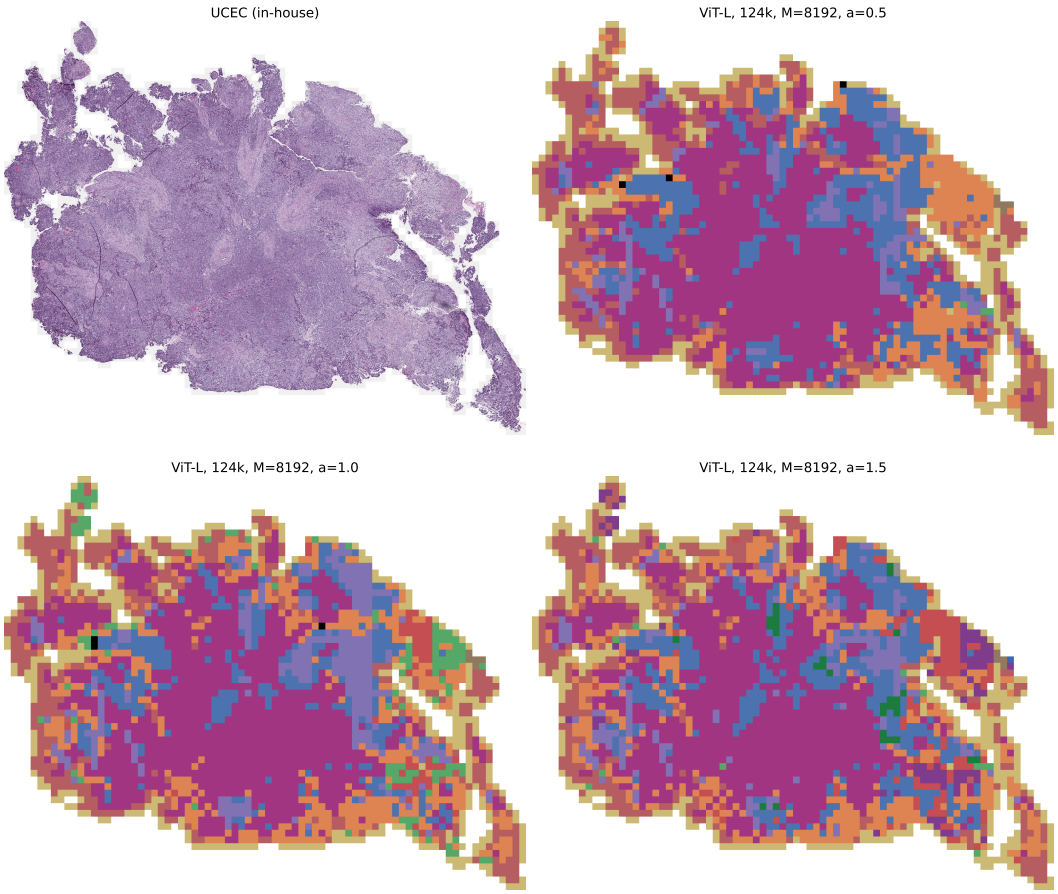

Figure 13: Cache-derived segmentation of an in-house endometrial carcinoma slide (UCEC) under the three ViT-L caches: H&E top left, then runs 10, 11, and 12 at their operating points $K{=}16$, 18, and 28. Colors are aligned to run 10 by a rectangular Hungarian assignment on the tile-label confusion of this slide, with unmatched classes in their own colors; black marks tiles beyond the support radius of every entry, white marks tiles absent from the corpus archives.

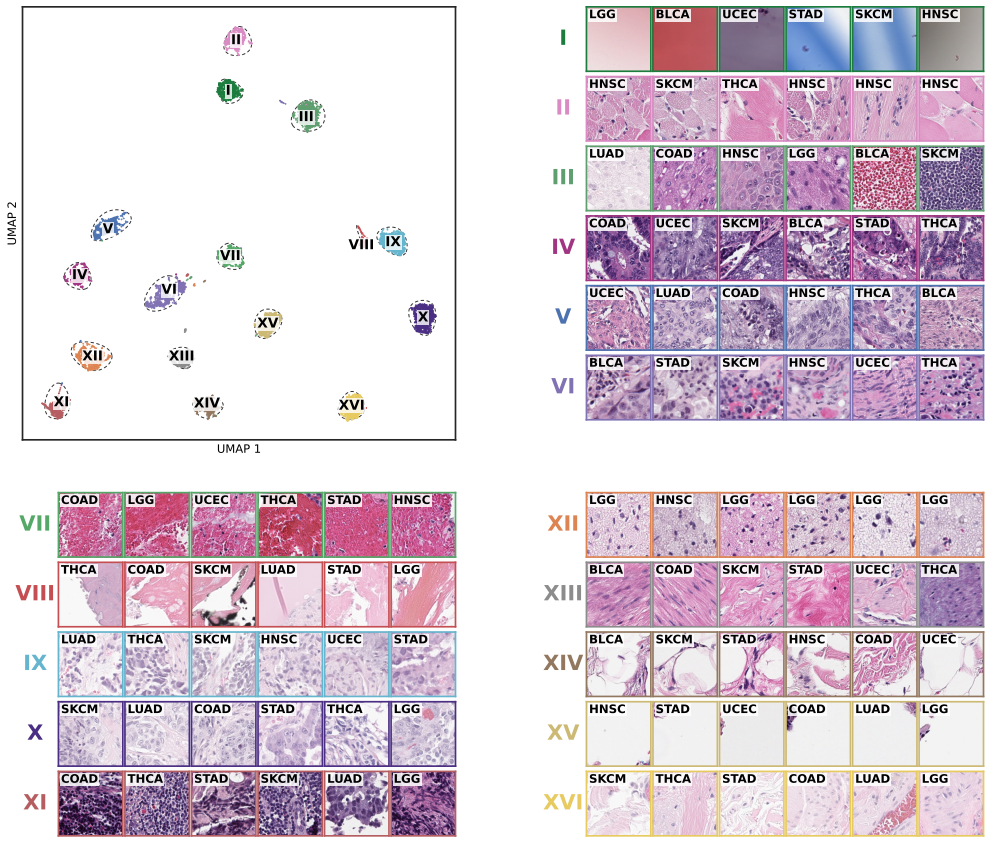

Figure 14: Run 10 (ViT-L, a=0.5), K=16: UMAP of the stored signatures with exemplar tiles per class.

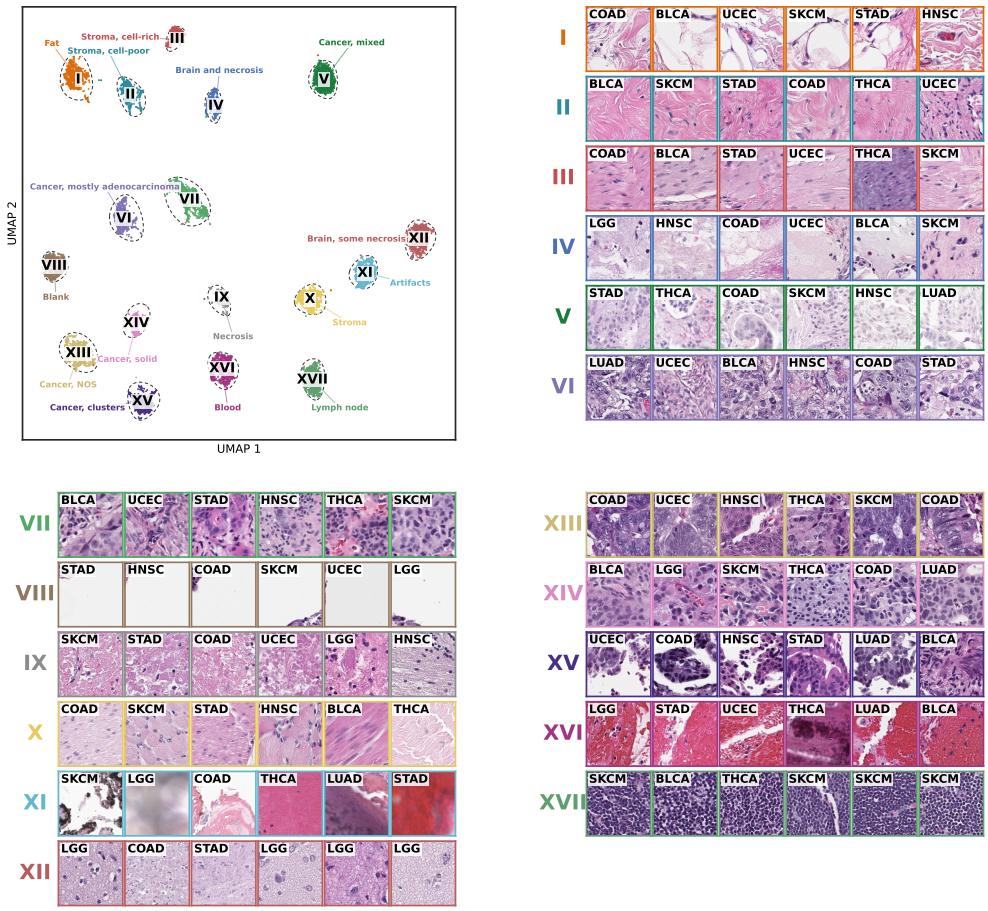

Figure 15: Run 11 (ViT-L, a=1.0), K=18: UMAP of the stored signatures with exemplar tiles per class and a pathologist's label for each class.

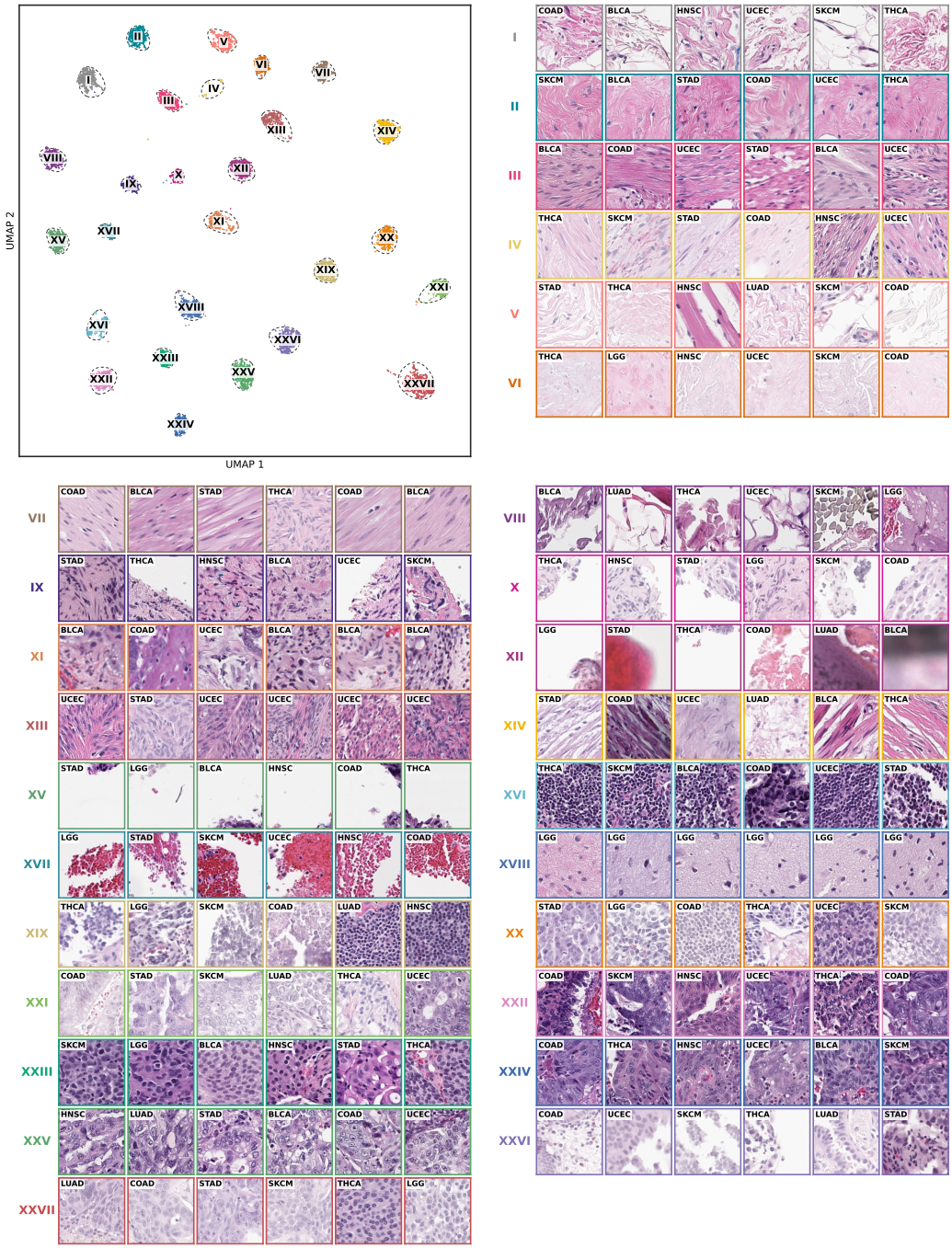

Figure 16: Run 12 (ViT-L, a=1.5), K=28: UMAP of the stored signatures with exemplar tiles per class.