SAMPLES, SOURCES, SPACE: DECOMPOSING DATA SCALE IN SPATIALLY STRUCTURED REPRESENTATION LEARNING OF HUMAN BRAIN MICROARCHITECTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling studies typically represent training data by a single count of samples. For hierarchically and spatially structured data, however, the same number of samples can be drawn from few or many sources and distributed differently across the underlying domain. We therefore study data scaling as an allocation problem, separating unique sample count, source diversity, and spatial coverage. We study this decomposition in microscopic whole-brain histology, where a source is an individual brain, and a sample is an image patch at a specific spatial location. Across 93 controlled pretraining runs of a contrastive model that uses spatial proximity for supervision, we vary data allocation, compute, and model capacity over 11.6 million spatially anchored image patches from 21 human brains. Performance improves with more unique samples, broader spatial coverage, additional compute, and larger model capacity. At fixed sample count, distributing samples across one to 18 subjects produces no detectable improvement, even though representations generalize substantially better to subjects encountered during pretraining. Inter-subject variation therefore strongly affects generalization, but additional subjects provide no benefit when a fixed sample budget is distributed across more sources. These results establish sample count, source diversity, and spatial coverage as distinct axes of data scaling in spatially structured representation learning.

1 INTRODUCTION

Empirical scaling studies characterize how learning changes with model capacity, training compute, and dataset size, typically representing data scale by a single count of training examples N (Kaplan et al., 2020; Hoffmann et al., 2022; Henighan et al., 2020). In many scientific datasets, however, samples are *hierarchically* and *spatially* organized, so the same N can correspond to substantially different data compositions. Hierarchical organization determines how samples are distributed across sources: increasing N may mean sampling existing sources more densely or acquiring additional sources. Spatial organization determines how samples are distributed over the underlying domain: at fixed sample and source counts, samples may densely cover a restricted region or sparsely cover a broader one. Source diversity, sampling density, and domain coverage are therefore distinct properties of a dataset.

This raises a basic question: *What does it mean to add data?* We characterize data composition by the number of distinct samples N_{unique}, the number of contributing sources B, and domain coverage ρ. A *unique sample* is one distinct sampled position, counted once irrespective of how often it is presented to the optimizer. The resulting allocation problem is to determine what is gained by additional samples, additional sources, or broader domain coverage under controlled training compute and model capacity.

Variants of this problem arise across fields. In pathology, patch count competes with diversity across slides (Bosch et al., 2025), while ecological sampling allocates fixed effort between replication within locations and coverage of an environmental gradient (Schweiger et al., 2026). Analogous hierarchical and spatial structure also occurs in remote sensing (Wu et al., 2025) and spatial transcriptomics (Xu et al., 2023). Across these settings, the methodological question is whether different

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.

ways of increasing data are interchangeable or have distinct effects on learning. Answering it requires the corresponding factors to be varied independently.

In this study, we investigate this problem in microscopic whole-brain histology, a particularly informative setting for controlled variation of sample count, source diversity, and spatial coverage. In this setting, whole brains are sectioned, stained, digitized, and represented by image patches sampled from the resulting sections (Amunts et al., 2013; 2020; Schiffer et al., 2025). A source is an individual subject, a sample is an image patch, and the domain is anatomical space (fig. 1). Each brain provides millions of samples, whereas adding another source requires processing an entire postmortem brain. Its explicit spatial organization provides a learning signal for representation learning without manual labels (Schiffer et al., 2025). We consider 11.6 million sample locations from 21 human brains across 93 controlled pretraining runs spanning ViT-B to ViT-H, varying unique sample count, source count, spatial coverage, training compute, and model capacity. Representations are evaluated by linear probing for classification of cytoarchitectonic areas (Amunts et al., 2020).

Our experiments reveal a distinction between variation that affects generalization and variation that is valuable to acquire under a fixed sampling budget. Representations generalize substantially better to subjects encountered during pretraining, showing that capturing inter-subject variation is critical for the learned representation. However, redistributing a fixed number of samples across more subjects produces no detectable improvement across the tested model capacities. The value of additional sources therefore depends on how they trade off against sampling density, rather than following directly from the magnitude of source-level variation.

More broadly, increasing dataset size can correspond to qualitatively different changes in data composition. Additional samples, additional sources, and broader domain coverage need not contribute equally to representation quality. We introduce a controlled experimental decomposition that makes these contributions separately measurable under matched compute and model capacity. This turns data scaling in hierarchically and spatially structured datasets into an explicit allocation problem and provides a framework for identifying which form of additional data is most valuable.

2 Related work

Scaling laws and the data axis. Empirical scaling laws relate performance to model size, dataset size, and compute (Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022). Data scale is typically represented by example or token count, without distinguishing additional sources from within-source samples. Data-curation methods show that composition matters within a fixed count through deduplication, pruning, and sample selection (Lee et al., 2022; Sorscher et al., 2023; Tirumala et al., 2023), but do not decompose scale into source diversity, within-source density, and spatial coverage. Closest to our setting, Muennighoff et al. (2023) study uniqueness versus repetition when data are constrained, whereas we ask how a fixed budget of unique samples should be allocated across sources and space.

Hierarchical data and sampling allocation. Many scientific datasets contain observations nested within larger acquisition units, creating a trade-off between denser sampling within sources and broader sampling across sources. In computational pathology, tiles are nested within slides, specimens, and patients, and foundation models are pretrained on large numbers of both tiles and sources (Chen et al., 2024; Vorontsov et al., 2024; Lu et al., 2024). Cross-model evidence suggests that neither slide nor tile count alone explains downstream performance (Campanella et al., 2025), but these comparisons jointly vary architectures, objectives, datasets, and compute and cannot isolate source diversity from within-source density. Related allocation problems arise in ecology, where fixed sampling effort is divided between replication within locations and coverage across environmental gradients (Schweiger et al., 2026), and in geospatial learning, where observations are spatially clustered and acquisition must be distributed across locations (Betti et al., 2026). Sample dependence is also recognized through subject-level splitting, spatial generalization benchmarks, and avoidance of pseudoreplication (Yagis et al., 2021; Butsko et al., 2025; Hurlbert, 1984). These works establish that sampling structure matters, but not how to allocate a fixed pretraining budget across samples, sources, and spatial coverage.

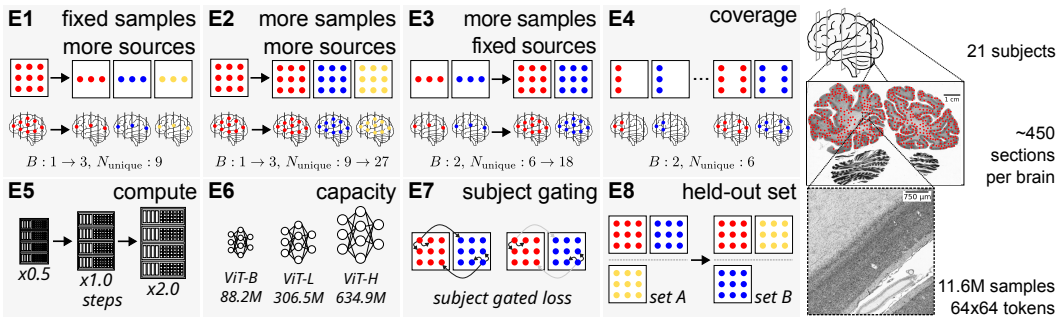

Figure 1: **Experimental decomposition of data scale and its instantiation in whole-brain histology. E1–E4** vary the composition of the training data by redistributing a fixed number of samples across more sources, increasing samples together with sources, increasing samples within fixed sources, or changing spatial coverage at fixed sample and source counts. The lower pictograms show the corresponding manipulations for brain histology, where sources are individual subjects and samples are spatially anchored image patches. **E5–E8** provide complementary controls over training compute, model capacity, cross-subject contributions to the SpatialNCE objective, and the choice of held-out evaluation subjects. See table 1 for details. The data hierarchy on the right maps the abstract units to the imaging pipeline from whole brains, to histological section, and image patches.

Spatially supervised representation learning. Tile2Vec (Jean et al., 2018), SGCL (Lin et al., 2023), and related histological approaches (Oberstrass et al., 2024) define positive relationships from spatial proximity. SpatialNCE, introduced with CytoNet for brain histology (Schiffer et al., 2025) and used here, instead weights pairwise relationships continuously by spatial distance. The spatial composition of the training set therefore determines the supervisory relationships available to the objective, motivating our treatment of spatial coverage as an explicit experimental factor.

3 Study Design

3.1 Data, Task, and Generalization Levels

The dataset comprises whole-brain histological images from 21 postmortem human subjects (Amunts et al., 2020), sectioned at 20μm, stained for neuronal cell bodies (Merker, 1983), and digitized at 1μm per pixel (see appendix A.1). Approximately every fifteenth physical section is used, yielding about 450 sections per brain across coronal and horizontal cutting planes, with the plane recorded per subject and used to stratify subject subsets. Candidate sample locations were generated independently within each section from gray- and white-matter masks (Schiffer et al., 2025), using greedy sampling based on repeated Euclidean distance transforms (see appendix A.2). After quality control, this yielded $11\,644\,201$ locations, approximately $554\,000$ per subject. Each location is mapped to the Colin27 reference space (Holmes et al., 1998), so Euclidean distances are defined between samples from different subjects (Schiffer et al., 2025), and centers a 2048×2048-pixel input patch read at 2μm per pixel (approximately 4×4mm) during pretraining.

Three subjects were excluded from every controlled pretraining run and from probe training, providing the primary evaluation subjects for separating pretraining from probe exposure. The remaining 18 subjects form the pretraining pool, of which 17 carry brain area annotations. The downstream task is classification of 191 cytoarchitectonically defined areas using expert annotations from the Julich Brain Atlas (Amunts et al., 2020). We train a linear classifier on frozen representations, compute macro-F1 over the classes present in each evaluation subject, and then average across subjects. The class set is fixed before evaluation, and probe optimization is identical for every encoder. Probe-training sets contain at most 1000 samples per class, selected without replacement and round-robin across subjects. No physical section is shared between probe training and evaluation. Full probe-training and evaluation details are provided in appendix A.3. Generalization level is defined per model–evaluation-subject pair, since a subject may be in the pretraining set of one model and excluded from that of another. An *unseen* subject is absent from both pretraining and probe training,

a *transfer* subject is present in pretraining and absent from probe training, and a *held-out-section* subject contributes other sections to both while the evaluated sections remain held out.

3.2 Objective and models

All encoders are pretrained with SpatialNCE (Schiffer et al., 2025), which encourages patches from nearby anatomical locations to have more similar representations than patches from distant ones. Let x_i denote an image patch, $p_i \in \mathbb{R}^3$ its 3D reference location, and z_i the normalized output of encoder and projection head. For a batch, the loss for sample i and the continuous target weight between two samples are given by

$$\mathcal{L}_i = -\frac{1}{\sum_{j\neq i}\omega_{ij}} \sum_{j\neq i} \omega_{ij} \log \frac{\exp\left(z_i^\top z_j/\tau\right)}{\sum_{k\neq i}\exp\left(z_i^\top z_k/\tau\right)}, \qquad \omega_{ij} = \exp\left(-\frac{\|p_i - p_j\|_2^2}{2\sigma^2}\right). \quad (1)$$

Following Schiffer et al. (2025) we use $\sigma = 10\,\mathrm{mm}$, $\tau = 0.1$, and compute the objective over all non-identical pairs in a global batch of 2048, gathered across all accelerators. Coordinates determine the supervisory relationship and are never given to the encoder, which must infer spatially consistent structure from image content alone. The kernel contains no subject indicator, so anatomically corresponding samples receive high weight regardless of subject identity. The normalizing term $\sum_{j\neq i}\omega_{ij}$ is the total positive kernel weight associated with anchor i in a batch, and we call its average across anchors and batches the *positive weight* M. M depends only on the data composition and the batching procedure, so it can be computed before training. Dataset composition therefore changes the effective objective: concentrating samples spatially increases M, while changing subject count alters the balance of within- and cross-subject pairs in the kernel. To isolate the role of cross-subject correspondences, one experiment (**E7**) sets all cross-subject weights to zero. We refer to this modification as *subject gating*.

We use encoders ViT-B, ViT-L, and ViT-H following the ViT-5 modernization of plain Vision Transformers (Dosovitskiy et al., 2020; Wang et al., 2026). Each input is tokenized into a 64×64 grid of 4096 tokens, together with a class token and four register tokens (Darcet et al., 2023). All models train with AdamW (Loshchilov & Hutter, 2018) at a constant learning rate 1×10^{-4} on 64 to 256 NVIDIA GH200 accelerators (see appendices B.3 and D.4 for details). The 93 runs consumed 316 000 accelerator-hours over 7.32 million optimization steps and 94 200 training exaFLOPs. Including the 124 linear probes, the study totals 349 000 accelerator-hours. Across comparisons, the architecture family, optimizer, batch size, and evaluation protocol are held fixed, so experiments isolate changes in data composition, compute, or model capacity. Architecture dimensions, augmentations, optimizer settings, and FLOP accounting are given in Appendices B.1 and B.2.

3.3 Experiments

We organize the experiments around five factors, describing representation quality as $S = f(P, C, N_{\mathrm{unique}}, B, \rho)$. S is macro-F1 on subjects excluded from both pretraining and probe training. P denotes model capacity, C compute measured in optimization steps, N_{unique} the number of distinct sample locations, B the number of contributing sources, and ρ the spatial coverage condition. Sources are individual subjects here, so B is a subject count throughout the results. The data axes are not fully independent: at fixed N_{unique}, increasing B necessarily reduces the mean number of locations per source N_{unique}/B. Table 1 summarizes the 93 runs. Unless compute is the manipulated factor, runs use a default budget of 73 200 steps, or 149 913 600 sample presentations. Presentations per unique location therefore vary inversely with N_{unique}, from roughly 3400 in the smallest single-subject condition to roughly 15 in the largest, and equal 333 throughout **E1**. The unique-count axis is therefore an axis of *uniqueness at a fixed presentation budget*, which differs from the data term in compute-optimal scaling laws (Hoffmann et al., 2022; Muennighoff et al., 2023). Subject subsets were chosen before any downstream result existed, from subject metadata with a fixed seed, stratified by cutting plane. For each level of B, we construct as many disjoint groups as the pool permits ($\lfloor 18/B \rfloor$), capped at two groups for the fixed-sample experiment **E1** and four for the more-sources-more-samples experiment **E2**. Each configuration is trained with a single optimization seed, so replication arises from the independently composed subject groups. Their spread therefore captures sensitivity to *which* subjects are sampled.

Table 1: **Overview of the eight experiments and 93 pretraining runs.** Run counts are incremental, with reused configurations counted only once. Unless stated otherwise, runs use ViT-B and the default presentation budget. In this study, a source corresponds to a subject, and a section is an ordered spatial subdivision of that source along the anatomical domain.

	Experiment	Manipulation	Levels	Runs
E1	Fixed samples, more sources	redistribute $N_{\text{unique}} = 450\,000$ across sources	$B \in \{1, 2, 4, 8, 18\}$	9
E2	More samples, more sources	retain selected sections as B increases	$B \in \{1, 2, 4, 8, 18\} \times$ {all, every 10th}	30
E3	More samples, fixed sources	vary retained-section fraction at fixed B	$B \in \{1, 9, 18\} \times$ {all, every 3rd, every 10th}	5
E4	Coverage	arrange fixed N_{unique} into k section blocks	$B \in \{1, 4, 13\}$, $k \in$ {every 10th, 16, 4, 1}	19
E5	Compute	vary optimization steps	$0.5\times$, $1\times$, $2\times$	7
E6	Capacity	vary encoder size	ViT-B, ViT-L, ViT-H	14
E7	Subject gating	zero cross-subject kernel weights	$B \in \{2, 4, 8, 18\}$	5
E8	Held-out sources	vary excluded source triple	three held-out assignments at $B = 18$	4

Experiments **E1**–**E4** define data compositions. **E5**–**E8** are crossed with a subset of them, so a change in compute, capacity, gating, or held-out assignment is always measured against an otherwise identical composition. The corresponding detection thresholds are defined in section 3.4. Appendix F lists every run. *Fixed samples, more sources* (**E1**) sets $N_{\text{unique}} = 450\,000$ and varies $B \in \{1, 2, 4, 8, 18\}$, selecting sections at equidistant intervals within each subject so that broad anatomical coverage is preserved as within-subject sampling density decreases. *Coverage* (**E4**) holds N_{unique} fixed and varies how those locations are arranged along each subject's section series. Every subject contributes k disjoint contiguous section blocks, grown outward from evenly spaced anchors until they reach the target sample budget, so larger k yields more and shorter blocks and therefore broader coverage. At $k = 1$, a single localized block is placed at an anterior, central, or posterior anchor, and these three positions are analyzed separately. The intermediate levels matter because anatomical coverage and positive weight M change at different rates as k decreases. 16 blocks preserve nearly the full anatomical extent and leave M almost unchanged, whereas four blocks introduce larger gaps and raise M by half. The primary contrast is 16 blocks against every-10th-section sampling at $B = 13$, which matches both N_{unique} and M and isolates the effect of introducing spatial gaps. *Subject gating* (**E7**) leaves within-subject distances unchanged and sets cross-subject weights to zero, so gated and standard objectives coincide at $B = 1$ and diverge as B grows. *Held-out sources* (**E8**) uses two additional disjoint triples of held-out subjects. Existing eligible models are re-evaluated on these triples. Four additional models are trained with the alternative triples excluded from pretraining.

3.4 Statistical analysis

Each of the three primary evaluation subjects contributes one score to every configuration. Let S_{cd} denote the macro-F1 score of configuration c on evaluation subject d. We fit a linear mixed-effects model to logit-transformed scores across the experimental grid:

$$\begin{aligned}\operatorname{logit}(S_{cd}) = \beta_0 + \beta_N \log_2 N_{\text{unique},c} + \beta_B \log_2 B_c + \beta_C \log_2 C_c + \beta_M \log_2 M_c \\ + \beta_G G_c + \gamma_P(P_c) + \gamma_\rho(\rho_c) + u_{\text{subset}(c)} + v_d + e_{cd}.\end{aligned} \tag{2}$$

Because macro-F1 is bounded to $[0, 1]$, we model its logit transformation, so coefficients are expressed on the logit scale. Continuous predictors enter as base-two logarithms, so their coefficients quantify the change in logit macro-F1 associated with a doubling of the corresponding quantity. Model capacity P_c and coverage condition ρ_c enter as categorical fixed effects through γ_P and γ_ρ, respectively, and G_c is an indicator for subject gating. The random intercept $u_{\text{subset}(c)}$ accounts for differences among pretraining-subject subsets, while v_d accounts for systematic differences among evaluation subjects. A second model is restricted to the fixed-sample experiment and adds a

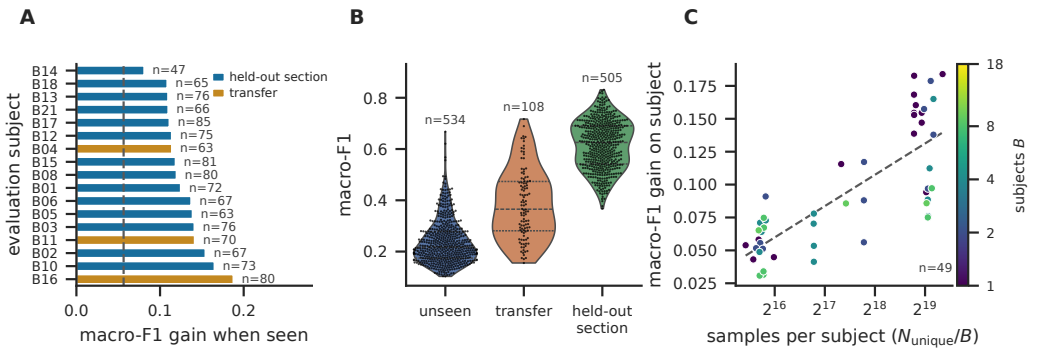

Figure 2: **Subject inclusion in pretraining dominates performance.** **A**: Macro-F1 score gain when a subject was included in the pretraining pool of a model vs. when it was not; averaged over n models (gray line: MDE 0.057), at identical evaluation samples, labels, and class denominator. **B**: Score distribution by generalization level: unseen subjects are absent from pretraining and probe training, transfer subjects only from probe training, and held-out-section subjects contribute other sections to both. **C**: Familiarity gain against N_{unique}/B, the mean number of locations a model saw per subject. Each point is one model from **E1**, **E2**, **E3**, **E6**, giving its subject-centered macro-F1 gain on pretraining subjects over other subjects.

capacity-by-subject-count interaction, $\gamma_{BP}(P_c)\log_2 B_c$, testing whether the effect of increasing B differs by model capacity. The full-grid fit uses 96 configurations and 288 subject-level scores, comprising 93 final pretraining runs and three retained intermediate checkpoints used as shorter-compute conditions in experiment **E5**. The fixed-sample fit uses 18 configurations and 54 scores. Both models are fitted by restricted maximum likelihood. We report 95% confidence intervals throughout, using profile-likelihood intervals for the scaling coefficients and capacity interaction and Wald intervals elsewhere.[1]

To distinguish unresolved differences from effects that the design was capable of detecting, we define minimum detectable effects (MDEs) from the fitted variance components at 80% power. The MDE is 0.226 logit for contrasts across pretraining-subject subsets and 0.166 logit for contrasts within the same subset. Differences below the applicable MDE are reported as *unresolved*. These thresholds apply to pairwise configuration contrasts and not to regression coefficients, which are evaluated using their confidence intervals. The MDE procedure and thresholds were fixed before interpreting the experimental contrasts, and primary and exploratory analyses are specified in appendix C.4. Appendix C.1 gives the full MDE derivation, and appendix C.2 reports collinearity diagnostics for the fixed-effect design.

4 Results

4.1 Source inclusion in pretraining dominates performance

Whether an evaluation subject was encountered during pretraining had the largest observed effect on performance among the factors examined here (fig. 2). We compare configurations pretrained with a given evaluation subject against configurations pretrained without that subject, while holding evaluation samples, labels, probe protocol, and class denominator fixed. Across the 14 held-out-section subjects, macro-F1 was higher by 0.124 on average when the subject had been seen during pretraining, with subject-level differences from 0.081 to 0.165. Across the three transfer subjects, the mean advantage was 0.148, with differences from 0.114 to 0.188. Configurations with and without a given subject necessarily use different pretraining-subject subsets, so the across-subset MDE of 0.226 logit applies. All 17 subject-level contrasts were positive and exceeded the applicable MDE on the logit scale, and 14 of 17 exceeded twice that threshold.

[1]Estimates are reported as coefficient = estimate [lower, upper], on the logit scale per doubling unless stated otherwise.

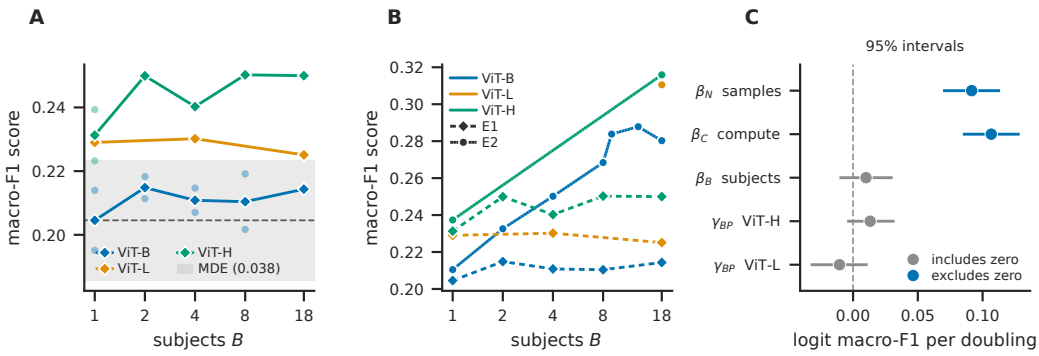

Figure 3: **Scaling subject count B without scaling N_{unique} does not improve representations. A**: Macro-F1 against subject count B at fixed $N_{\text{unique}} \approx 450\,000$ (**E1** in table 1), with solid lines giving the mean per capacity P (color) and the MDE band anchored at the $B{=}1$ mean (dashed line). The curve is flat at every capacity. **B**: The same axis overlaid with **E2**, where N_{unique} grows linearly with B. **C**: 95% intervals of coefficients per doubling on the logit scale. Intervals for β_B and the capacity interaction γ_{BP} include zero, while performance improves with sample count β_N and compute β_C.

The number of familiar evaluation subjects increases almost deterministically with B, so a cross-model regression cannot identify whether the gap itself changes with subject count. We therefore examine it against N_{unique}/B, the mean number of unique locations seen per pretraining subject. Across 49 models the familiarity advantage grows by 0.024 macro-F1 per doubling of N_{unique}/B, and in a joint model that coefficient remains $0.022\,[0.017, 0.027]$ while the coefficient for $\log_2 B$ is $-0.012\,[-0.019, -0.005]$. The advantage is therefore associated more strongly with per-subject sampling density than with subject count, which sharpens the question for the next experiment: whether more subjects improve generalization when the total sample budget is held fixed.

4.2 More sources do not help at a fixed sample budget

Distributing a fixed sample budget across more subjects produced no detectable improvement in representation quality (fig. 3). For ViT-B in the fixed-sample analysis, the estimated subject-count effect was $\beta_B = 0.0014\,[-0.0253, 0.0282]$ logit macro-F1 per doubling of B. Across the full range from $B = 1$ to $B = 18$ ($\log_2 18 = 4.17$ doublings), this interval bounds the fitted change to $[-0.106, 0.118]$ logit, or approximately $[-0.017, 0.020]$ macro-F1 at the reference score. The observed macro-F1 differences between $B = 1$ and $B = 18$ were 0.015 for ViT-B, -0.002 for ViT-L, and 0.017 for ViT-H, all below the across-subset MDE that applies to these contrasts. The subject-count effect also did not increase detectably with model capacity: relative to ViT-B, the capacity-by-subject-count interaction was $\gamma_{BP} = 0.0132\,[-0.0046, 0.0321]$ for ViT-H and $-0.0105\,[-0.0328, 0.0112]$ for ViT-L. The full-grid model gives the same conclusion, with $\beta_B = 0.0100\,[-0.0105, 0.0306]$. The sensitivity of **E1** to capacity-by-subject-count interactions is quantified by simulation in appendix C.3.

Two observations constrain the interpretation of this null result. First, the detection threshold was fixed in advance, and the upper confidence limit for β_B is only about one quarter of the estimated effect of doubling N_{unique} reported below. Second, performance was still responsive to additional unique samples within the same subject pool: holding the same 18 subjects fixed while increasing N_{unique} from $450\,000$ to 9.8 million raised macro-F1 from 0.214 to 0.300, well above the applicable MDE. The conclusion is therefore limited to the tested regime of $B = 1$–18, approximately $450\,000$ unique locations, SpatialNCE pretraining, and cytoarchitectonic targets. The transferable result is the distinction itself: source count B and coverage ρ are separate experimental quantities, and neither follows from N_{unique}. Together with section 4.1, inter-subject variability therefore strongly affects generalization, yet increasing subject diversity by reducing the number of samples per subject does not compensate for that variability. The main conclusions were unchanged under alternative held-out-subject assignments (appendix D.5).

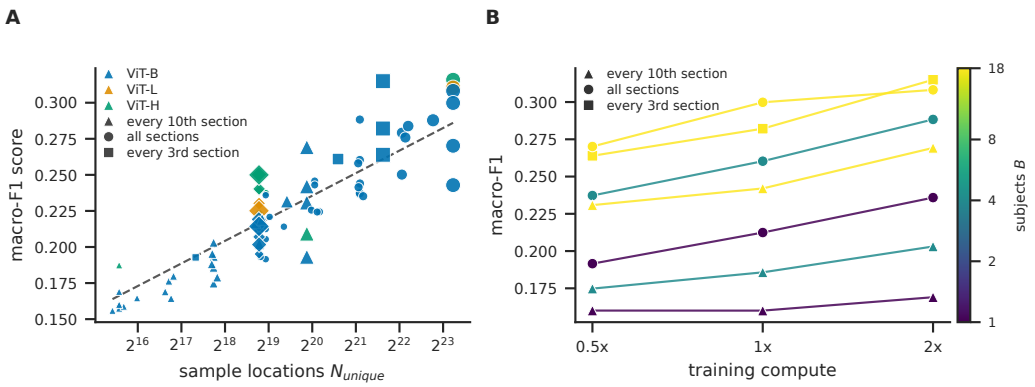

Figure 4: **Scaling samples N_{unique}, compute C, and capacity P improve representations.** **A**: Macro-F1 against N_{unique} across the grid ($\beta_N = 0.09$ / doubling, dashed line), with marker color, style, and size giving capacity P, coverage ρ, and subject count B. **B**: Macro-F1 against compute C, subject count B (color), and coverage ρ (marker style) across the 7 settings of **E5**.

4.3 More samples, compute, and capacity improve representations

Increasing unique sample count, compute, and model capacity each improved representation quality (fig. 4). Doubling the number of unique sample locations increased logit macro-F1 by $\beta_N = 0.0916\,[0.0695, 0.1135]$, while doubling optimization steps increased it by $\beta_C = 0.1067\,[0.0847, 0.1286]$. Because the presentation budget is fixed, β_N captures the joint effect of doubling uniqueness while halving repetition, whereas the fixed-sample estimate of β_B varies subject count at constant repetition. All seven configurations evaluated at both the default and doubled compute budgets improved, by 0.008 to 0.034 macro-F1, and the second doubling gave roughly 43% of the gain of the first.

Capacity point estimates were monotone. At $B = 18$, macro-F1 increased from 0.300 for ViT-B to 0.311 for ViT-L and 0.316 for ViT-H, none of which reaches the within-subset MDE that applies to these matched-subset contrasts. Evidence for a capacity effect comes instead from the categorical capacity terms pooled across the experimental grid, with $\gamma_P = 0.127\,[0.094, 0.159]$ for ViT-H and $0.081\,[0.031, 0.130]$ for ViT-L relative to ViT-B.

In practice, adding a subject increases both source diversity and the number of available samples. Under full sampling, ViT-B macro-F1 increased from 0.210 across the four single-subject subsets to 0.300 at $B = 18$, and under every-10th-section sampling from 0.160 to 0.242. Adding subjects is therefore beneficial when they also increase the number of available samples, while section 4.2 shows that subject count alone does not explain the gain. For context, the strongest published reference encoder under the same protocol reached 0.251 macro-F1 (Virchow2; appendix D.6).

4.4 Broad spatial coverage improves representations

Restricting a fixed sample budget to narrower anatomical coverage reduced representation quality at high subject count (fig. 5). All $B = 13$ coverage conditions use the same thirteen coronal subjects, so the within-subset MDE of approximately 0.028 macro-F1 applies. Relative to every-10th-section sampling at $B = 13$, macro-F1 decreased by 0.038 for four blocks, by 0.034, 0.028, and 0.021 for the three localized single-block conditions, and by 0.017 for 16 blocks. The four-block condition and two of the three localized conditions meet or exceed the MDE, while the 16-block condition and the remaining localized condition do not. Spatial concentration alone does not explain this ordering: the four-block condition spans 78% of the full section range, yet shows the largest deficit, so the arrangement of covered and uncovered regions also contributes. The primary contrast between 16 blocks and every-10th-section sampling matches both N_{unique} and positive weight M and is unresolved, so introducing many small gaps into otherwise broad coverage has no detectable cost under this design. The deficit also appears only at high subject count, since no coverage condition differs detectably from every-10th-section sampling at $B = 1$ or $B = 4$.

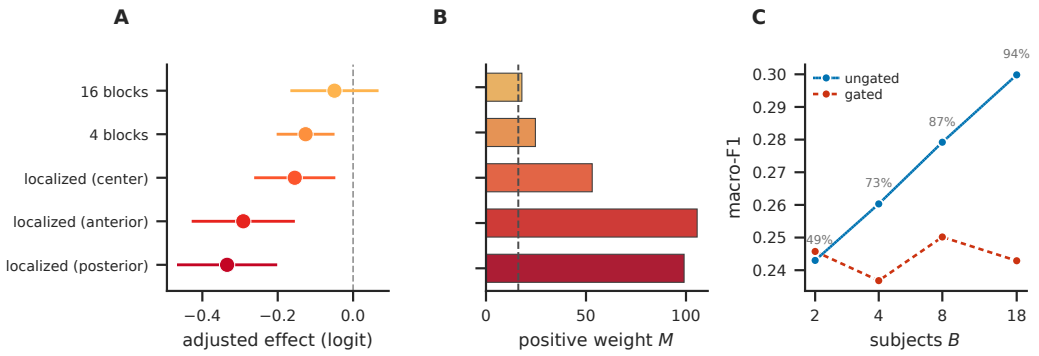

Figure 5: **Coverage geometry, positive weight, and subject gating at fixed sample count. A**: Adjusted logit macro-F1 differences among the coverage conditions in **E4** relative to every-10th-section sampling, with 95% confidence intervals, ordered by spatial concentration. **B**: Positive weight M for the corresponding **E4** conditions, with every-10th-section sampling as the reference. **C**: Effect of subject gating in **E7**, which sets all cross-subject SpatialNCE weights to zero. Each gated configuration is paired with an ungated configuration using the same data composition. Annotations indicate the fraction of M contributed by cross-subject pairs before gating. At $B = 2$, where cross-subject pairs contribute 49% of M, the gating contrast is unresolved. At $B = 18$, where cross-subject pairs contribute 94% of M, gating reduces macro-F1 by 0.057.

Most coverage conditions also change the positive weight M and thus do not isolate anatomical coverage as a single factor. At $B = 13$, M is 16.15 for every-10th-section sampling and 17.94 for 16 blocks, and ranges from 53.11 to 105.58 across the three localized conditions. The localized conditions therefore combine reduced coverage with a substantial change in objective composition, so this experiment does not separate the two contributions. Subject gating introduces an even larger change in positive weight. Cross-subject pairs supply 94.2% of the positive weight at $B = 18$, so removing them reduces M by roughly a factor of 17, and performance under gating was lower from $B = 4$ onward. At $B = 18$, macro-F1 decreased from 0.300 with the standard objective to 0.243 with gating, a difference of 0.057 that exceeds the within-subset MDE. Across the full-grid model, doubling positive weight is associated with an increase of $\beta_M = 0.0979\,[0.0597, 0.1357]$ in logit macro-F1, consistent with much of that deficit. Cross-subject anatomical correspondences therefore supply most of the positive mass available to SpatialNCE at high subject count and are important for maintaining performance in this regime. The composition of M across **E1**, **E4**, and **E7**, and the gating indicator after adjustment for M, are detailed in appendices D.1 and D.2.

5 Discussion

A central implication of our results is that the magnitude of source-level variation and the value of acquiring additional sources are distinct quantities. Inter-source variation strongly affects generalization, yet broader source sampling does not improve representations when it redistributes a fixed sample budget. For hierarchical datasets, observing substantial variation between sources therefore provides insufficient evidence for how additional acquisition effort should be allocated. The relevant quantity is the marginal value of samples from new versus already represented sources.

This distinction extends the usual notion of data scale. A sample count alone does not specify whether dataset growth adds observations, sources, or coverage of the underlying domain, and our experiments show that these forms of scale can have different value within the same learning problem. Making this decomposition explicit provides a way to reason about data acquisition in settings with repeated or spatially structured observations, where collecting another sample and collecting another source can have very different costs. Our results establish this distinction in spatially supervised whole-brain histology and motivate treating data composition as an explicit dimension of empirical scaling studies.

ETHICS STATEMENT

This study required no additional ethical approval beyond the existing approval covering use of the postmortem material. The corresponding ethics committee, approval number, and participating institutions are anonymized for double-blind review and will be restored in the camera-ready version. Postmortem brains were obtained through institutional body-donation programs in accordance with applicable legal and ethical requirements. All donors had provided written consent for the use of their bodies for research and teaching.

REPRODUCIBILITY STATEMENT

The analysis plan was pre-specified before the first pretraining run and fixed the unit of replication, mixed-effects specification, interval procedure, and minimum detectable effect. Contrasts outside this plan are identified as exploratory. Section 3.3 and table 1 define all 93 pretraining runs, and appendix F lists them individually. Code, configurations, and derived data were version-controlled with DataLad, providing provenance from pretraining manifests through the tables used to generate the reported results and figures. Analysis code, run configurations, pretraining manifests, and derived evaluation tables will be released with the final version; the underlying whole-brain histological images cannot be redistributed.

AI USE STATEMENT

Generative AI tools were used during the research workflow and manuscript preparation. They assisted with discussion of experimental design and statistical analysis, refinement of the conceptual framing, manuscript drafting and editing, and generation and revision of analysis and figure code. All AI-assisted text, code, analyses, and interpretations were reviewed by the authors. All reported numerical results were verified against their generating data or analysis outputs, and AI-generated code was inspected and its outputs validated. Model pretraining, evaluation, and the statistical fits reported in this work were produced by the authors' own computational pipelines. The authors take responsibility for all content, analyses, figures, and conclusions in the final manuscript.

REFERENCES

Katrin Amunts, Claude Lepage, Louis Borgeat, Hartmut Mohlberg, Timo Dickscheid, Marc-Étienne Rousseau, Sebastian Bludau, Pierre-Louis Bazin, Lindsay B. Lewis, Ana-Maria Oros-Peusquens, Nadim J. Shah, Thomas Lippert, Karl Zilles, and Alan C. Evans. BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. *Science*, 340(6139):1472–1475, 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1235381.

Katrin Amunts, Hartmut Mohlberg, Sebastian Bludau, and Karl Zilles. Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science*, 369(6506):988, 2020. doi: 10.1126/science.abb4588.

Livia Betti, Farooq Sanni, Gnouyaro Z. Sogoyou, Togbe Agbagla, Cullen Molitor, Tamma Carleton, and Esther Rolf. Mapping on a budget: Optimizing spatial data collection for ML. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38233–38241, 2026.

Christoph Bosch, John K. L. Wong, Martin Paulikat, Myroslav Zapukhlyak, Bharti Arora, Manasi Aichmüller-Ratnaparkhe, Jens Baumann, Shivani Karn, Rutuja Kamble, Swapnil Karnik, Bhushan Khedkar, Serey Vathana Chhut, Witali Aswolinskiy, and Christian Aichmüller. Diversity Over Scale: Whole-Slide Image Variety Enables H&E Foundation Model Training with Fewer Patches, November 2025.

Christina Butsko, Kristof Van Tricht, Gabriel Tseng, Giorgia Milli, David Rolnick, Ruben Cartuyvels, Inbal Becker Reshef, Zoltan Szantoi, and Hannah Kerner. Deploying Geospatial Foundation Models in the Real World: Lessons from WorldCereal, July 2025.

Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Ivan Shujski, Noora Neittaanmäki, Kuan-lin Huang, Ricky Kwan, Jane Houldsworth, Adam J. Schoenfeld, and Chad

Vanderbilt. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16(1):3640, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58796-1.

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, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*, October 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR 2020)*, 2020.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling Laws for Autoregressive Generative Modeling, November 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022.

Colin J. Holmes, Rick Hoge, Louis Collins, Roger Woods, Arthur W. Toga, and Alan C. Evans. Enhancement of MR Images Using Registration for Signal Averaging. *Journal of Computer Assisted Tomography*, 22(2):324–333, March 1998. ISSN 0363-8715.

Stuart H. Hurlbert. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, 54(2):187–211, 1984. ISSN 1557-7015. doi: 10.2307/1942661.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2Vec: Unsupervised representation learning for spatially distributed data, May 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577.

Tiancheng Lin, Zhimiao Yu, Zengchao Xu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. SGCL: Spatial guided contrastive learning on whole-slide pathological images. *Medical Image Analysis*, 89:102845, October 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102845.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, September 2018.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4.

Björn Merker. Silver staining of cell bodies by means of physical development. *Journal of Neuroscience Methods*, 9(3):235–241, 1983. ISSN 0165-0270. doi: 10.1016/0165-0270(83)90086-9.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling Data-Constrained Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376. Curran Associates, Inc., 2023. doi: 10.52202/075280-2191.

Alexander Oberstrass, Sascha E.A. Muenzing, Meiqi Niu, Nicola Palomero-Gallagher, Christian Schiffer, Markus Axer, Katrin Amunts, and Timo Dickscheid. Self-supervised representation learning for nerve fiber distribution patterns in 3D-PLI. *Imaging Neuroscience*, 2:imag–2–00351, November 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00351.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

Christian Schiffer, Zeynep Boztoprak, Jan-Oliver Kropp, Julia Thönnißen, Katia Berr, Hannah Spitzer, Katrin Amunts, and Timo Dickscheid. CytoNet: A Foundation Model for the Human Cerebral Cortex at Cellular Resolution, October 2025.

Andreas H. Schweiger, Aron Garthen, Michael Bahn, David Chalcraft, Nicolas Schtickzelle, Klaus Steenberg Larsen, and Jürgen Kreyling. How to optimally allocate sampling effort in experimental ecology. *Scientific Reports*, 16(1):6503, February 2026. ISSN 2045-2322. doi: 10.1038/s41598-026-38541-4.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning, April 2023.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53983–53995. Curran Associates, Inc., 2023. doi: 10.52202/075280-2348.

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*, 30(10):2924–2935, October 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03141-0.

Feng Wang, Sucheng Ren, Tiezheng Zhang, Predrag Neskovic, Anand Bhattad, Cihang Xie, and Alan Yuille. ViT-5: Vision Transformers for The Mid-2020s, February 2026.

Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, Yongjun Zhang, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. *Nature Machine Intelligence*, 7(8):1235–1249, August 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01078-8.

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*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07441-w.

Hao Xu, Shuyan Wang, Minghao Fang, Songwen Luo, Chunpeng Chen, Siyuan Wan, Rirui Wang, Meifang Tang, Tian Xue, Bin Li, Jun Lin, and Kun Qu. SPACEL: Deep learning-based characterization of spatial transcriptome architectures. *Nature Communications*, 14(1):7603, November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43220-3.

Ekin Yagis, Selamawet Workalemahu Atnafu, Alba García Seco de Herrera, Chiara Marzi, Riccardo Scheda, Marco Giannelli, Carlo Tessa, Luca Citi, and Stefano Diciotti. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11(1): 22544, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-01681-w.

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

Table 2: **Dataset overview.** For each subject, we report the number of sampling locations, retained sections, annotated sections, cutting plane, and role in the study. Subjects are identified by numeric IDs.

subject	locations	sections	annotated	plane	role
B01	644 224	468	332	coronal	pretraining
B02	522 551	462	253	coronal	pretraining
B03	610 975	506	298	coronal	pretraining
B04	570 373	475	372	coronal	transfer
B05	461 204	451	355	coronal	pretraining
B06	571 496	469	385	coronal	pretraining
B07	597 666	487	378	coronal	held out
B08	514 651	451	343	coronal	pretraining
B09	655 532	476	353	coronal	held out
B10	533 815	478	346	coronal	pretraining
B11	557 567	499	329	coronal	transfer
B12	496 407	465	221	coronal	pretraining
B13	499 678	431	264	coronal	pretraining
B14	500 528	460	259	coronal	pretraining
B15	607 156	387	105	horizontal	pretraining
B16	566 503	300	83	horizontal	transfer
B17	619 245	322	158	horizontal	pretraining
B18	532 668	339	138	horizontal	pretraining
B20	556 888	453	451	coronal	held out
B21	668 055	492	216	coronal	pretraining
B23	357 019	322	0	horizontal	pretrain only

A Data and task

A.1 Subjects, tissue processing, and consent

The dataset comprises 21 postmortem human brains. Subject characteristics and the tissue-processing and digitization procedures are described in Amunts et al. (2020) and Schiffer et al. (2025). Ethics approval and the body-donor consent procedure are stated in the ethics statement of the main text. Brains were chemically fixed, paraffin-embedded, sectioned at $20\,\mu\mathrm{m}$, stained for neuronal cell bodies with a modified silver protocol (Merker, 1983), and digitized at $1\,\mu\mathrm{m}$ per pixel. Approximately every fifteenth physical section was retained, yielding 300–506 sections per brain and $11\,644\,201$ sampling locations in total. Sixteen brains were cut coronally and five horizontally. Cutting plane was recorded for each subject and used to stratify subject subsets. **E4** is restricted to the 13 coronal subjects in the pretraining pool because its block-based coverage manipulation requires a common sectioning axis across subjects. Table 2 gives the per-subject inventory.

A.2 Sampling-location generation

Candidate locations were generated independently within each section from gray- and white-matter masks using repeated Euclidean distance transforms. At each iteration, the procedure selected the point with the largest distance to the nearest previously selected point or tissue boundary, excluded a $2 \times 2\,\mathrm{mm}$ square around it, and repeated until the maximum remaining distance fell below $300\,\mu\mathrm{m}$ in gray matter or $500\,\mu\mathrm{m}$ in white matter. Gray- and white-matter masks were generated as described in Schiffer et al. (2025) using thresholding and intensity transformations. Because sampling was performed independently within each section, the procedure did not require surface reconstruction or section-to-section alignment and could sample cortical and subcortical gray matter as well as white matter. The procedure yielded $11\,644\,201$ locations, approximately $554\,000$ per subject. Each location was mapped to the Colin27 reference space (Holmes et al., 1998) following Schiffer et al. (2025). The resulting common coordinate system defines Euclidean distances between samples from different subjects.

Table 3: **Encoder dimensions.** Depth, width, and head count follow the standard ViT variants, while token grid and register count are identical across capacities.

Variant	Width	Heads	Layers	Accelerators
ViT-B	768	12	12	64
ViT-L	1024	16	24	128
ViT-H	1280	16	32	256

A.3 Downstream task and evaluation semantics

The probe predicts one of 191 cytoarchitectonic areas from a class vocabulary fixed before evaluation and derived from the Julich Brain atlas (Amunts et al., 2020). Probe-training samples are capped at 1000 per class without replacement; capped classes are filled round-robin across subjects, with a fixed-seed random order within each subject. The probe is a single affine layer mapping the frozen encoder representation to the 191 classes and is trained with categorical cross-entropy. Encoder features are passed to the classifier without normalization or other transformation. The probe is optimized with AdamW at a constant learning rate of 10^{-3} and weight decay 0.05, using a global batch size of 2048 across 64 GH200 accelerators for 4710 steps. The fixed step budget keeps probe optimization effort independent of training-pool size. No physical section contributes to both probe training and evaluation. Macro-F1 is computed separately for each evaluation subject over the classes present in that subject—137, 171, and 172 of the 191 classes for the three primary unseen subjects, respectively—and then averaged across subjects. Three subjects are excluded from all controlled pretraining runs and from probe training and form the primary unseen evaluation set. Three additional subjects are excluded from probe training but remain eligible for pretraining and define the transfer subjects. The remaining annotated subjects contribute to probe training, with separate sections held out for evaluation. Generalization level is therefore defined for each model–subject pair.

B Models, training, and compute

B.1 Architecture and optimization

Encoders use the modernized ViT-5 architecture of Wang et al. (2026), with SwiGLU feed-forward blocks of hidden size $8/3$ times the embedding dimension, bias-free QKV projections, no stochastic depth, and four register tokens (Darcet et al., 2023). Each 2048×2048 input is divided into a 64×64 grid of 32×32-pixel patches, yielding 4096 patch tokens and a sequence length of 4101 after adding the class token and four register tokens; the class token provides the encoder representation.

All models are trained with AdamW (Loshchilov & Hutter, 2018) at a constant learning rate of 10^{-4}, without warmup or decay, using weight decay 0.05 for weights and 0 for normalization parameters and biases. Pretraining and linear-probe training use automatic mixed precision with bfloat16 autocast. The global batch size is fixed at 2048 for all capacities. Larger encoders use more accelerators while retaining the same global batch size and step budget, separating model capacity from training duration. Representations are gathered across all accelerators before computing SpatialNCE, so the objective is evaluated over the full global batch. The projection head maps the class-token representation to z_i using a two-layer MLP consisting of a linear layer to the encoder width, RMS normalization, GELU, and a bias-free linear layer to 256 dimensions. The resulting vector is L_2-normalized before entering SpatialNCE. The projection head is discarded after pretraining, and downstream evaluation uses the encoder class-token representation.

Augmentation follows the CytoNet protocol of Schiffer et al. (2025). Patches are rotated by $\theta \sim U[-\pi, +\pi]$, mirrored with probability 0.5, and translated in a random direction by up to $0.2\,\mathrm{mm}$. Pixel intensities $x \in [0, 1]$ are transformed using unbiased gamma augmentation (Pohlen et al., 2017) as $\alpha x^{\gamma} + \beta$, with $\alpha \sim U[0.9, 1.1]$, $\beta \sim U[-0.1, +0.1]$,

$$\gamma = \frac{\log\left(0.5 + 2^{-0.5} Z\right)}{\log\left(0.5 - 2^{-0.5} Z\right)},$$

and $Z \sim U[-0.05, +0.05]$. With probability 0.25 each, a patch is blurred by an isotropic Gaussian with $\sigma \sim U[0.125, 1.0]$ or sharpened as $x + \delta(x - G_{\sigma_u}(x))$, with $\sigma_u \sim U[0.125, 2.0]$ and $\delta \sim U[0.5, 1.5]$. Patches are standardized to $[-1, 1]$ after augmentation.

B.2 Compute effort

Pretraining consumed 315 973 accelerator-hours over 7 320 000 optimization steps and $94\,200 \times 10^{18}$ FLOPs (94 200 exaFLOPs). The 124 linear probes consumed an additional 32 751 accelerator-hours, giving 348 724 accelerator-hours in total. FLOPs are computed analytically from dense matrix operations at 4101 tokens, counting two operations per multiply–add and excluding normalization, softmax, activation functions, and optimizer updates. Superseded training attempts, data loading, and prediction jobs are excluded from these totals.

B.3 Software and execution environment

All runs were executed on the same supercomputer, whose identity is withheld for double-blind review and will be restored in the camera-ready version. Its accelerated partition uses NVIDIA GH200 Grace–Hopper superchips with four accelerators per node. Pretraining used 16 nodes for ViT-B, 32 nodes for ViT-L, and 64 nodes for ViT-H, corresponding to 64, 128, and 256 accelerators, respectively. The global pretraining batch size was fixed at 2048 across all capacities. Linear probes used 16 nodes with 64 accelerators. All pretraining and probe jobs ran under SLURM in the same Apptainer image, providing a common software environment across runs. Training used PyTorch (Paszke et al., 2019) and distributed data parallel (DDP) training. Statistical analysis and figure generation were performed in a separate environment using NumPy, pandas, SciPy, and statsmodels.

C Statistical analysis

C.1 Derivation of the two detection thresholds

We derive separate minimum detectable effects (MDEs) for contrasts between configurations trained on different pretraining-subject subsets and contrasts between configurations trained on the same subset. For configuration c evaluated on subject d, we write the fitted mixed-effects model as

$$Y_{cd} = \mu_c + u_{s(c)} + v_d + e_{cd},$$

where μ_c is the fixed-effect contribution for configuration c, $u_{s(c)}$ is the random intercept for its pretraining-subject subset $s(c)$, v_d is the evaluation-subject random intercept, and e_{cd} is the residual. The fitted variance components relevant to these contrasts are $\hat{\sigma}_u^2 = 0.00150$ and $\hat{\sigma}_e^2 = 0.00525$.

Every configuration is evaluated on the same held-out subjects, so configuration contrasts are paired by evaluation subject. For two configurations c_1 and c_2 evaluated on the same subject d,

$$Y_{c_1 d} - Y_{c_2 d} = (\mu_{c_1} - \mu_{c_2}) + (u_{s(c_1)} - u_{s(c_2)}) + (e_{c_1 d} - e_{c_2 d}).$$

The evaluation-subject effect v_d cancels in this paired contrast.

For an across-subset contrast, the fitted model treats $u_{s(c_1)}$ and $u_{s(c_2)}$ as independent, giving

$$\mathrm{Var}\big(u_{s(c_1)} - u_{s(c_2)}\big) = 2\sigma_u^2.$$

This subset-level contribution is shared across evaluation subjects and is unchanged when their contrasts are averaged. The residuals are independent across evaluation subjects under the fitted model, so averaging over n_d subjects reduces the residual contribution to $2\sigma_e^2/n_d$. Writing $\bar{\Delta}$ for the configuration contrast averaged over the n_d evaluation subjects gives

$$\mathrm{Var}(\bar{\Delta}_{\mathrm{across}}) = 2\sigma_u^2 + \frac{2\sigma_e^2}{n_d}.$$

For a within-subset contrast, the pretraining-subject subset effect also cancels. The variance of the averaged contrast is then

$$\mathrm{Var}(\bar{\Delta}_{\mathrm{within}}) = \frac{2\sigma_e^2}{n_d}.$$

Table 4: **Variance inflation factors for the continuous predictors in the full-grid model.** All VIFs are below 2.

fit	term	VIF
full grid	β_N	1.92
full grid	β_B	1.73
full grid	β_C	1.00
full grid	β_M	1.14
fixed sample	β_B	1.37
fixed sample	β_M	1.37

With $n_d = 3$ evaluation subjects, the corresponding standard errors are

$$\mathrm{SE}_{\mathrm{across}} = \sqrt{2\hat{\sigma}_u^2 + \frac{2\hat{\sigma}_e^2}{3}}, \qquad \mathrm{SE}_{\mathrm{within}} = \sqrt{\frac{2\hat{\sigma}_e^2}{3}}.$$

Using a normal approximation for a two-sided test with $\alpha = 0.05$ and 80% power, we define

$$\mathrm{MDE} = \left(z_{1-\alpha/2} + z_{0.80}\right) \mathrm{SE} = \left(z_{0.975} + z_{0.80}\right) \mathrm{SE},$$

where $z_{0.975} + z_{0.80} \approx 2.80$. This gives an across-subset MDE of 0.226 logit units and a within-subset MDE of 0.166 logit units. Because the logit transformation is nonlinear, the corresponding macro-F1 difference depends on the baseline score S. At $S = 0.15$, the across- and within-subset MDEs correspond to approximately 0.029 and 0.021 macro-F1, respectively. At $S = 0.21$, they correspond to approximately 0.038 and 0.028 macro-F1, respectively. At $S = 0.30$, they correspond to approximately 0.047 and 0.035 macro-F1, respectively. These MDEs apply to pairwise contrasts between two configurations.

For the aggregated familiarity analysis in section 4.1, the across-subset MDE is used only as a reference effect-size scale. All 17 subject-level familiarity contrasts exceed this value.

C.2 Collinearity among the fixed effects

The full-grid model fits a common additive specification across experiments with different manipulations. Several fixed effects are correlated by design, particularly N_{unique}, the coverage indicators ρ, and the positive weight M. We quantify linear collinearity for each continuous predictor using variance inflation factors (VIFs) computed from the encoded fixed-effect design matrix.

For continuous predictor j,

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is obtained by regressing predictor j on all remaining columns of the fixed-effect design matrix.

The largest VIF is 1.92 for $\log_2 N_{\mathrm{unique}}$, indicating limited linear dependence among the encoded fixed effects (table 4). In **E5**, compute is varied through optimization steps while data composition is held fixed, providing direct variation in C independent of the data-allocation factors. The VIFs quantify linear dependence within the encoded design matrix but do not test whether a common additive model is an adequate specification across the heterogeneous experiments.

C.3 Sensitivity of the capacity interaction

To quantify sensitivity to a capacity-by-subject-count interaction, we simulated outcomes under the realized design of **E1** (fig. 6). The design matrix was fixed, while the imposed interaction coefficient γ_{BP} and residual variance were varied. All remaining model parameters were fixed to their estimates from the observed data. For each parameter setting, we generated repeated synthetic datasets and refit the fixed-sample mixed-effects model. Power was estimated as the fraction of fits whose 95% confidence interval for γ_{BP} excluded zero.

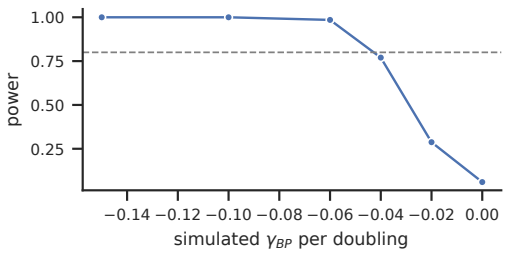

Figure 6: **Simulated power for detecting a capacity-by-subject-count interaction γ_{BP} under E1.** Power is shown as a function of the imposed interaction magnitude and residual variance. At the fitted residual variance, approximately 80% power is reached for an interaction of 0.05 logit units per doubling of subject count.

Table 5: **Geometry of the coverage conditions ρ in E4 at $B = 13$, averaged across the 13 coronal subjects.** All conditions use the same sample budget but differ in anatomical extent and spatial fragmentation.

condition ρ	sections	A–P span (mm)	largest gap (mm)
all	467	180	1.0
every 10th	47	180	5.2
16 blocks	78	175	11.9
4 blocks	52	142	42.5
loc. posterior	97	37	0.5
loc. center	31	11	0.5
loc. anterior	88	34	0.8

At the fitted residual variance, interactions smaller than approximately 0.05 logit units per doubling of subject count are not reliably detectable under **E1**. The absence of a detected capacity-by-subject-count interaction therefore excludes only effects above this approximate sensitivity threshold.

C.4 Pre-specification and multiplicity

The capacity-by-subject-count interaction in **E1** was designated as a primary analysis before inspection of outcomes. The contrast between the 16-block and every-10th-section conditions in **E4**, which approximately matches both N_{unique} and M, was also designated as primary. Individual localized-coverage contrasts in **E4**, subject-gating contrasts in **E7**, familiarity comparisons, and reference-encoder comparisons were treated as exploratory. Exploratory contrasts are reported with unadjusted 95% confidence intervals, without multiplicity correction. No primary conclusion relies solely on whether an exploratory confidence interval excludes zero.

D Additional results

D.1 Coverage geometry and subject gating

Figure 5 summarizes the performance and positive-weight changes induced by the coverage manipulation in **E4** and subject gating in **E7**. Table 5 details the section count, anatomical extent, and fragmentation of the **E4** coverage conditions at $B = 13$. The three localized conditions in **E4** match in sample count but differ in the number and anatomical distribution of sampled sections. Central sections contain substantially more tissue area, so 31 such sections provide approximately as many sampling locations as 97 posterior or 88 anterior sections. The localized-center condition therefore has the smallest anterior–posterior extent among the three localized conditions. Its positive weight M is not the largest, because M depends on the full three-dimensional distribution of pairwise distances.

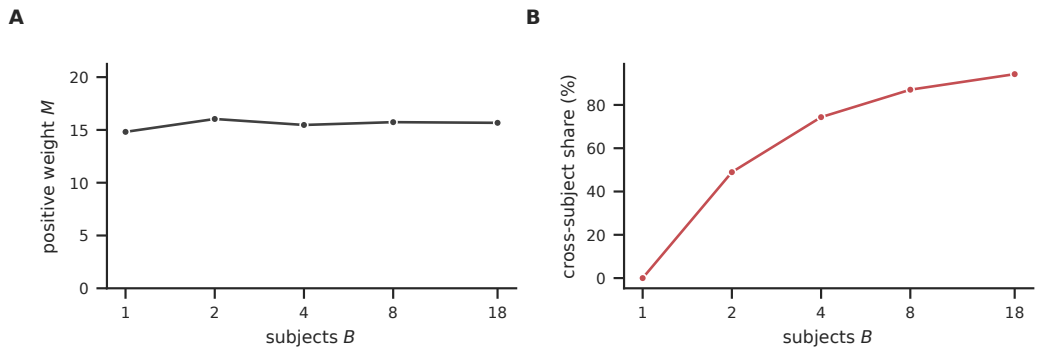

Figure 7: **Positive weight across E1. A**: Total positive weight M remains approximately constant as subject count increases from one to 18. **B**: The fraction of M contributed by cross-subject pairs increases from zero to 94.2% over the same range. At fixed N_{unique}, **E1** therefore preserves the overall positive weight while changing the composition of contributing pairs.

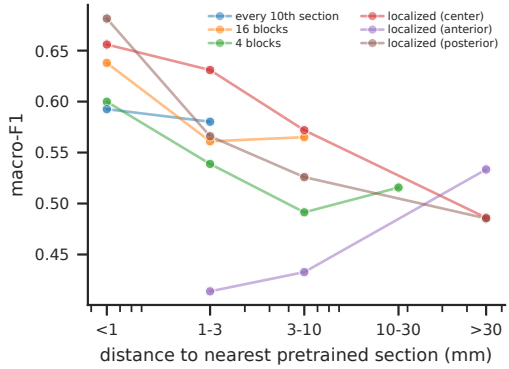

Figure 8: **Macro-F1 on held-out sections from familiar subjects as a function of distance to the nearest section seen during pretraining in the $B = 13$ runs of E4.** Each line corresponds to one coverage condition ρ. Macro-F1 decreases with distance from the nearest pretraining section in five of the six coverage conditions.

D.2 Composition of the SpatialNCE objective

Figure 7 shows the total positive weight M and its cross-subject contribution across **E1**, while table 6 reports both quantities for each pretraining configuration. Realized location counts in **E1** range from $448\,861$ to $450\,825$, a spread below 0.5%, and the number of sample presentations is identical across conditions. Across ungated configurations outside **E4**, M ranges only from 14.5 to 17.6 despite a 224-fold range in N_{unique} and variation across all tested subject counts. **E4** increases M as samples become more spatially concentrated, reaching 17.9 and 28.1 for the fragmented conditions and 53–115 for the localized conditions. **E7** decreases M by removing cross-subject contributions, leaving $M = 8.34$ at $B = 2$ and $M = 0.92$ at $B = 18$, approximately 17-fold below the corresponding ungated value at $B = 18$.

D.3 Spatial localization of the familiarity benefit

Figure 8 resolves the familiarity effect spatially across the $B = 13$ conditions of **E4**. Macro-F1 decreases with distance to the nearest section seen during pretraining in five of the six coverage conditions. This pattern is consistent with a familiarity benefit that is strongest near tissue encountered during pretraining. The analysis is confounded by spatial variation in the difficulty and class composition of cytoarchitectonic areas. For example, performance in the anterior-localized condi-

Table 6: **Positive weight M and its cross-subject contribution for every pretraining configuration, averaged across subject subsets at each level.** M is the covariate M_c used in the mixed-effects model and is computed from the configuration before training.

B	sampling	gated	subsets	N_{unique}	M	cross-subject (%)
1	every 10th	✗	5	51 709	15.23	0.0
2	every 10th	✗	4	108 881	16.31	49.0
4	every 10th	✗	5	219 266	15.95	74.0
8	every 10th	✗	2	438 674	15.85	86.9
9	every 10th	✗	1	472 504	15.66	88.4
13	every 10th	✗	1	701 132	16.15	92.0
18	every 10th	✗	1	962 026	15.73	94.1
1	every 3rd	✗	1	164 378	14.65	0.0
9	every 3rd	✗	1	1 579 631	15.59	88.4
18	every 3rd	✗	1	3 225 383	15.74	94.2
1	4 blocks	✗	1	49 036	28.08	0.0
4	4 blocks	✗	1	223 273	27.11	70.4
13	4 blocks	✗	1	706 946	24.70	90.1
13	16 blocks	✗	1	711 367	17.94	91.7
1	all	✗	5	532 596	15.23	0.0
2	all	✗	4	1 108 908	16.33	49.1
4	all	✗	5	2 238 444	15.94	74.0
8	all	✗	2	4 481 222	15.77	86.9
9	all	✗	1	4 816 705	15.59	88.4
13	all	✗	1	7 151 524	16.12	92.0
18	all	✗	3	9 942 251	15.68	94.1
1	fixed budget	✗	2	450 358	14.82	0.0
2	fixed budget	✗	2	450 354	16.04	48.9
4	fixed budget	✗	2	449 824	15.47	74.4
8	fixed budget	✗	2	449 761	15.74	87.0
18	fixed budget	✗	3	449 192	15.67	94.2
1	loc. anterior	✗	1	48 872	115.43	0.0
4	loc. anterior	✗	1	221 468	107.81	72.9
13	loc. anterior	✗	1	708 859	105.58	91.5
1	loc. center	✗	1	50 094	64.96	0.0
4	loc. center	✗	1	223 022	63.40	69.4
13	loc. center	✗	1	712 904	53.11	89.5
1	loc. posterior	✗	1	49 110	98.57	0.0
4	loc. posterior	✗	1	221 797	106.65	73.0
13	loc. posterior	✗	1	707 964	99.05	91.4
18	every 10th	✓	1	962 026	0.92	0.0
2	all	✓	1	1 090 235	8.34	0.0
4	all	✓	1	2 233 423	3.97	0.0
8	all	✓	1	4 356 763	2.07	0.0
18	all	✓	1	9 834 115	0.92	0.0

tion increases with distance from the pretraining sections, which may reflect higher classification performance for areas located farther posteriorly.

D.4 Pretraining behavior

Figure 9 shows standard and doubled-compute ViT-B, capacity ViT-L/H, and subject-gated ViT-B trajectories. Loss trajectories were smoothed with a Savitzky–Golay filter using a window of 1000 steps before computing their gradients. SpatialNCE loss continues to decrease at the doubled budget of 146 400 steps, indicating that training has not reached a loss plateau within the tested compute

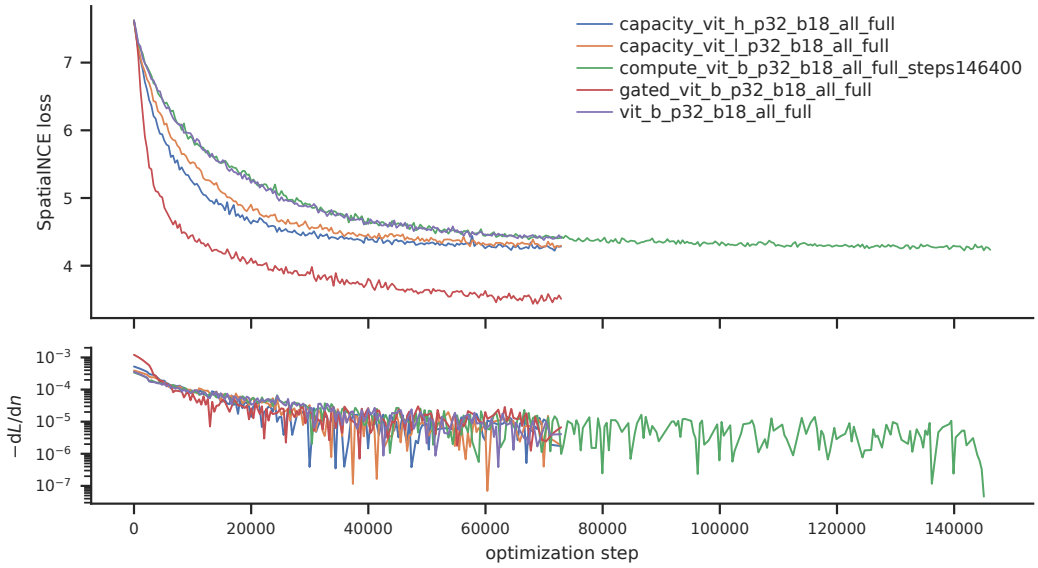

Figure 9: **Training trajectories for five $B = 18$ full-coverage configurations from E5–E7 and their shared baseline.** The top panel shows SpatialNCE loss, and the bottom panel shows the gradient of the smoothed loss trajectory. The loss continues to decrease at the doubled budget of 146 400 steps. See table 8 for the corresponding configurations.

range. Higher-capacity models reach lower SpatialNCE loss values over the same optimization budget.

D.5 Alternative held-out-subject assignments

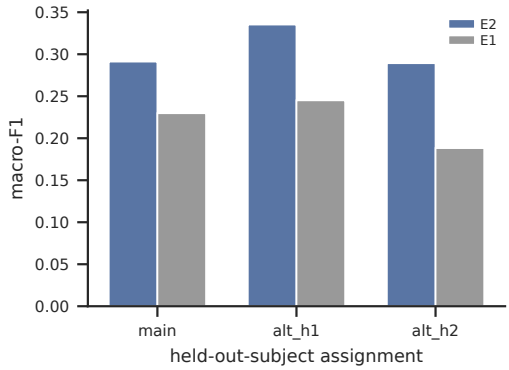

Figure 10: **Macro-F1 for the $B = 18$ configurations of E1 and E2 under two alternative unseen-subject assignments.** Absolute performance varies across assignments, while the direction and approximate magnitude of the difference between **E1** and **E2** are preserved.

Figure 10 compares the $B = 18$ configurations of **E1** and **E2** under the primary unseen-subject assignment and two alternative assignments. The two alternative subject triples are mutually disjoint, disjoint from the primary triple, and restricted to subjects with a coronal cutting plane. Scores are computed over the 188 classes represented in all three assignments to make the results directly comparable across assignments. This common class set causes the values in fig. 10 to differ slightly from the corresponding scores reported in the main text. Re-evaluating all eligible **E1** configurations

Table 7: **Reference encoders evaluated as frozen feature extractors using the same probe and unseen-subject evaluation protocol.** Our rows correspond to the $B = 18$ configurations of **E6** at each model capacity. The comparison provides an absolute performance reference across models with different architectures, pretraining data, and computational budgets. The chance row states where the macro-F1 scale begins under the same per-subject class denominator.

Encoder	Reference	Macro-F1
Uniform random guess	—	0.004
DINOv2 (ViT-B/14)	Oquab et al. (2024)	0.168
GigaPath (ViT-G/16)	Xu et al. (2024)	0.220
UNI (ViT-L/16)	Chen et al. (2024)	0.225
Virchow2 (ViT-H/14)	Zimmermann et al. (2024)	0.251
ViT-B/32	ours	0.300
ViT-L/32	ours	0.311
ViT-H/32	ours	0.316

on an alternative unseen-subject triple also left the subject-count effect unresolved, with a ViT-B slope of $\beta_B = 0.0131\,[-0.0164, 0.0429]$ logit macro-F1 per doubling of B.

D.6 Reference encoders

Table 7 compares published vision and pathology encoders with the $B = 18$ configurations of **E6** under the same frozen-feature probe and unseen-subject evaluation protocol. Each 2048×2048-pixel input was split into sixteen 512×512 tiles, converted from grayscale to RGB, and processed natively (224 pixels for UNI/Virchow2, 518 for DINOv2, and resize to 256 then center-crop to 224 for GigaPath); embeddings were averaged. Differences in architecture, pretraining, physical resolution, and modified-silver versus H&E staining make this an absolute-performance reference only. Among the evaluated reference encoders, Virchow2 achieves the highest macro-F1 at 0.251, compared with 0.300, 0.311, and 0.316 for our ViT-B/32, ViT-L/32, and ViT-H/32 models, respectively. A uniform predictor over the 191-class vocabulary yields 0.004 macro-F1 averaged across the three unseen subjects.

E Remaining limitations

The study has several limitations. The subject pool is finite, and subject subsets at different values of B overlap, so comparisons across subject-count conditions are not based on fully independent sets of subjects. The primary evaluation uses three unseen subjects, although the main comparisons retain their direction under two additional held-out-subject assignments (appendix D.5). Each configuration was pretrained with a single optimization seed because of the computational cost of the study. Replication across independently composed subject subsets captures sensitivity to subject selection but does not separately quantify optimization-seed variability. The coverage manipulation in **E4** changes both sampled anatomy and the spatial distribution of pairwise relationships entering SpatialNCE. Subject gating in **E7** directly changes the composition and total positive weight of the SpatialNCE objective. The corresponding results therefore do not isolate all contributing mechanisms individually, as discussed in section 4.4. **E1** evaluates fixed-sample subject scaling at a single budget of approximately $450\,000$ unique locations and over $B = 1$–18. Subject diversity may have a different effect at substantially smaller sample budgets or when each subject is sampled more sparsely. All downstream results use linear probes on frozen representations and cytoarchitectonic area labels. More expressive readouts may recover representation structure that is not linearly accessible, although the familiarity effect in section 4.1 shows that the linear probe is sensitive to substantial inter-subject differences.

F COMPLETE RUN INVENTORY

Table 8 lists all 93 pretraining runs conducted in this study. A pretraining run denotes one optimization trajectory, whereas a configuration in the statistical analysis denotes one retained checkpoint evaluated on the common primary unseen subjects. Four runs from **E8** use alternative held-out-subject assignments and are excluded from the common-subject full-grid fit. Each of the seven doubled-budget runs in **E5** contributes an additional retained half-budget checkpoint without requiring a separate optimization trajectory. The full-grid fit therefore contains $93 - 4 + 7 = 96$ configurations and $96 \times 3 = 288$ subject-level scores. The fixed-sample fit for **E1** contains 18 configurations: nine ViT-B, six ViT-H, and three ViT-L, yielding 54 subject-level scores.

Table 8: **Inventory of all 93 pretraining runs, grouped by experiment.** *Steps* gives the completed optimization budget. The seven runs at 146 400 steps also contribute a retained checkpoint at half that budget. *Gated* marks the runs whose objective admits no cross-subject positives. *Fit* marks inclusion in the full-grid model. The four runs on alternative held-out triples are analyzed only in appendix D.5.

exp.	run	arch	B	sampling	N_{unique}	steps	gated	fit
E1	vit_b_p32_b01_subset03_iso_unique	B	1	fixed budget	449 924	73 200	✗	✓
E1	vit_b_p32_b01_subset04_iso_unique	B	1	fixed budget	450 792	73 200	✗	✓
E1	vit_b_p32_b02_subset01_iso_unique	B	2	fixed budget	449 883	73 200	✗	✓
E1	vit_b_p32_b02_subset02_iso_unique	B	2	fixed budget	450 825	73 200	✗	✓
E1	vit_b_p32_b04_subset01_iso_unique	B	4	fixed budget	450 404	73 200	✗	✓
E1	vit_b_p32_b04_subset02_iso_unique	B	4	fixed budget	449 245	73 200	✗	✓
E1	vit_b_p32_b08_subset01_iso_unique	B	8	fixed budget	448 861	73 200	✗	✓
E1	vit_b_p32_b08_subset02_iso_unique	B	8	fixed budget	450 661	73 200	✗	✓
E1	vit_b_p32_b18_all_iso_unique	B	18	fixed budget	449 794	73 200	✗	✓
E2	vit_b_p32_b01_subset01_alt10	B	1	every 10th	52 461	73 200	✗	✓
E2	vit_b_p32_b01_subset02_alt10	B	1	every 10th	64 795	73 200	✗	✓
E2	vit_b_p32_b01_subset03_alt10	B	1	every 10th	48 732	73 200	✗	✓
E2	vit_b_p32_b01_subset04_alt10	B	1	every 10th	43 829	73 200	✗	✓
E2	vit_b_p32_b01_subset01_full	B	1	all	532 668	73 200	✗	✓
E2	vit_b_p32_b01_subset02_full	B	1	all	668 055	73 200	✗	✓
E2	vit_b_p32_b01_subset03_full	B	1	all	500 528	73 200	✗	✓
E2	vit_b_p32_b01_subset04_full	B	1	all	461 204	73 200	✗	✓
E2	vit_b_p32_b02_subset01_alt10	B	2	every 10th	107 329	73 200	✗	✓
E2	vit_b_p32_b02_subset02_alt10	B	2	every 10th	115 616	73 200	✗	✓
E2	vit_b_p32_b02_subset03_alt10	B	2	every 10th	101 308	73 200	✗	✓
E2	vit_b_p32_b02_subset04_alt10	B	2	every 10th	111 274	73 200	✗	✓
E2	vit_b_p32_b02_subset01_full	B	2	all	1 090 235	73 200	✗	✓
E2	vit_b_p32_b02_subset02_full	B	2	all	1 177 529	73 200	✗	✓
E2	vit_b_p32_b02_subset03_full	B	2	all	1 034 343	73 200	✗	✓
E2	vit_b_p32_b02_subset04_full	B	2	all	1 133 526	73 200	✗	✓
E2	vit_b_p32_b04_subset01_alt10	B	4	every 10th	218 659	73 200	✗	✓
E2	vit_b_p32_b04_subset02_alt10	B	4	every 10th	214 133	73 200	✗	✓
E2	vit_b_p32_b04_subset03_alt10	B	4	every 10th	212 405	73 200	✗	✓
E2	vit_b_p32_b04_subset04_alt10	B	4	every 10th	231 730	73 200	✗	✓
E2	vit_b_p32_b04_subset01_full	B	4	all	2 233 423	73 200	✗	✓
E2	vit_b_p32_b04_subset02_full	B	4	all	2 188 587	73 200	✗	✓
E2	vit_b_p32_b04_subset03_full	B	4	all	2 177 018	73 200	✗	✓
E2	vit_b_p32_b04_subset04_full	B	4	all	2 355 517	73 200	✗	✓
E2	vit_b_p32_b08_subset01_alt10	B	8	every 10th	425 687	73 200	✗	✓
E2	vit_b_p32_b08_subset02_alt10	B	8	every 10th	451 662	73 200	✗	✓
E2	vit_b_p32_b08_subset01_full	B	8	all	4 356 763	73 200	✗	✓
E2	vit_b_p32_b08_subset02_full	B	8	all	4 605 682	73 200	✗	✓
E2	vit_b_p32_b18_all_alt10	B	18	every 10th	962 026	73 200	✗	✓
E2	vit_b_p32_b18_all_full	B	18	all	9 834 115	73 200	✗	✓
E3	vit_b_p32_b01_subset03_alt33	B	1	every 3rd	164 378	73 200	✗	✓
E3	vit_b_p32_b09_subset01_alt10	B	9	every 10th	472 504	73 200	✗	✓
E3	vit_b_p32_b09_subset01_alt33	B	9	every 3rd	1 579 631	73 200	✗	✓
E3	vit_b_p32_b09_subset01_full	B	9	all	4 816 705	73 200	✗	✓
E3	vit_b_p32_b18_all_alt33	B	18	every 3rd	3 225 383	73 200	✗	✓

continued on next page

exp.	run	arch	B	sampling	N_{unique}	steps	gated	fit
E4	vit_b_p32_b01_coronal_subset01_alt10	B	1	every 10th	48 732	73 200	✗	✓
E4	vit_b_p32_b01_coronal_subset01_fragmented04	B	1	4 blocks	49 036	73 200	✗	✓
E4	vit_b_p32_b01_coronal_subset01_full	B	1	all	500 528	73 200	✗	✓
E4	vit_b_p32_b01_coronal_subset01_localized_matched_end	B	1	loc. anterior	48 872	73 200	✗	✓
E4	vit_b_p32_b01_coronal_subset01_localized_matched_middle	B	1	loc. center	50 094	73 200	✗	✓
E4	vit_b_p32_b01_coronal_subset01_localized_matched_start	B	1	loc. posterior	49 110	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_alt10	B	4	every 10th	219 403	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_fragmented04	B	4	4 blocks	223 273	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_full	B	4	all	2 237 676	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_localized_matched_end	B	4	loc. anterior	221 468	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_localized_matched_middle	B	4	loc. center	223 022	73 200	✗	✓
E4	vit_b_p32_b04_coronal_subset01_localized_matched_start	B	4	loc. posterior	221 797	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_alt10	B	13	every 10th	701 132	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_fragmented04	B	13	4 blocks	706 946	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_fragmented16	B	13	16 blocks	711 367	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_full	B	13	all	7 151 524	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_localized_matched_end	B	13	loc. anterior	708 859	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_localized_matched_middle	B	13	loc. center	712 904	73 200	✗	✓
E4	vit_b_p32_b13_coronal_all_localized_matched_start	B	13	loc. posterior	707 964	73 200	✗	✓
E5	compute_vit_b_p32_b01_subset03_alt10_steps146400	B	1	every 10th	48 732	146 400	✗	✓
E5	compute_vit_b_p32_b01_subset03_full_steps146400	B	1	all	500 528	146 400	✗	✓
E5	compute_vit_b_p32_b04_subset01_alt10_steps146400	B	4	every 10th	218 659	146 400	✗	✓
E5	compute_vit_b_p32_b04_subset01_full_steps146400	B	4	all	2 233 423	146 400	✗	✓
E5	compute_vit_b_p32_b18_all_alt10_steps146400	B	18	every 10th	962 026	146 400	✗	✓
E5	compute_vit_b_p32_b18_all_alt33_steps146400	B	18	every 3rd	3 225 383	146 400	✗	✓
E5	compute_vit_b_p32_b18_all_full_steps146400	B	18	all	9 834 115	146 400	✗	✓
E6	capacity_vit_h_p32_b01_subset03_alt10	H	1	every 10th	48 732	73 200	✗	✓
E6	capacity_vit_h_p32_b01_subset03_full	H	1	all	500 528	73 200	✗	✓
E6	capacity_vit_h_p32_b01_subset03_iso_unique	H	1	fixed budget	449 924	73 200	✗	✓
E6	capacity_vit_h_p32_b01_subset04_iso_unique	H	1	fixed budget	450 792	73 200	✗	✓
E6	capacity_vit_l_p32_b01_subset03_iso_unique	L	1	fixed budget	449 924	73 200	✗	✓
E6	capacity_vit_h_p32_b02_subset01_iso_unique	H	2	fixed budget	449 883	73 200	✗	✓
E6	capacity_vit_h_p32_b04_subset01_iso_unique	H	4	fixed budget	450 404	73 200	✗	✓
E6	capacity_vit_l_p32_b04_subset01_iso_unique	L	4	fixed budget	450 404	73 200	✗	✓
E6	capacity_vit_h_p32_b08_subset01_iso_unique	H	8	fixed budget	448 861	73 200	✗	✓
E6	capacity_vit_h_p32_b18_all_alt10	H	18	every 10th	962 026	73 200	✗	✓
E6	capacity_vit_h_p32_b18_all_full	H	18	all	9 834 115	73 200	✗	✓
E6	capacity_vit_l_p32_b18_all_full	L	18	all	9 834 115	73 200	✗	✓
E6	capacity_vit_h_p32_b18_all_iso_unique	H	18	fixed budget	449 794	73 200	✗	✓
E6	capacity_vit_l_p32_b18_all_iso_unique	L	18	fixed budget	449 794	73 200	✗	✓
E7	gated_vit_b_p32_b02_subset01_full	B	2	all	1 090 235	73 200	✓	✓
E7	gated_vit_b_p32_b04_subset01_full	B	4	all	2 233 423	73 200	✓	✓
E7	gated_vit_b_p32_b08_subset01_full	B	8	all	4 356 763	73 200	✓	✓
E7	gated_vit_b_p32_b18_all_alt10	B	18	every 10th	962 026	73 200	✓	✓
E7	gated_vit_b_p32_b18_all_full	B	18	all	9 834 115	73 200	✓	✓
E8	vit_b_p32_b18_alth1_all_full	B	18	all	10 041 128	73 200	✗	✗
E8	vit_b_p32_b18_alth2_all_full	B	18	all	9 951 511	73 200	✗	✗
E8	vit_b_p32_b18_alth1_all_iso_unique	B	18	fixed budget	448 344	73 200	✗	✗
E8	vit_b_p32_b18_alth2_all_iso_unique	B	18	fixed budget	449 438	73 200	✗	✗