Supplementary Notes:
Towards Robust Foundation Models for Digital Pathology

These supplementary notes provide additional data, experiments, and analyses extending those presented in the main paper. The aim is to further support the claims of the main paper regarding the (limited) robustness of pathology foundation models (FMs), the consequences, and robustification strategies.

Contents

1 PathoROB dataset details **3**
- 1.1 Camelyon . . . 3
- 1.2 TCGA . . . 3
- 1.3 Tolkach ESCA . . . 4

2 Robustness index computation and further results **6**
- 2.1 Data pre-processing: l2 normalization . . . 6
- 2.2 Selecting k values . . . 6
- 2.3 Efficient computation of the robustness index . . . 7
- 2.4 Trading off performance vs robustness . . . 8
- 2.5 Standard deviation estimation . . . 8
- 2.6 Results per dataset . . . 8
- 2.7 Robustness index: numerical results . . . 13
- 2.8 Robustness index per biological and confounding category . . . 16
- 2.9 Relation between biological representation robustness and out-of-distribution (OOD) generalization . . . 16

3 Robust foundation models prevent entangled features **18**
- 3.1 Spurious correlation in downstream datasets . . . 18
- 3.2 Spurious correlation can hamper generalizability . . . 18
- 3.3 Polysemy aggravates robustification . . . 19
- 3.4 Illustrative example: unrobust foundation models enable Clever Hans learning . . . 19
- 3.5 Connection to real-data experiments . . . 20
- 3.6 Summary . . . 20

4 Downstream model training **22**
- 4.1 General training details . . . 22
- 4.2 Training with spuriously correlated data . . . 22
- 4.3 Slide-level downstream model training . . . 23
- 4.4 Domain-adversarial training . . . 24

5 Clustering analysis **26**
- 5.1 Clustering score . . . 26
- 5.2 Implementation details . . . 27

6 Analysis of the FM representation space **27**
- 6.1 FM representation space visualizations . . . 27
- 6.2 Analysis of FM representation spaces with PCA . . . 30

7 Further downstream experiment results **34**
- 7.1 Full finetuning of foundation models on spuriously correlated data . . . 34
- 7.2 Low-rank adaptation of foundation models on spuriously correlated data . . . 35
- 7.3 Slide-level downstream model comparison . . . 36
- 7.4 Patch retrieval from heterogeneous databases . . . 36
- 7.5 Downstream model experiments on robustified FM representations . . . 37
- 7.6 Clustering performance independent of the silhouette score-based K selection . . . 40
- 7.7 Clustering experiments on robustified FM representations . . . 42

Supplementary Note 1. PathoROB dataset details

1.1 Camelyon

The CAMELYON16 dataset (`https://camelyon17.grand-challenge.org/Data`, License: Public Domain: CC0) [1, 2] comprises 400 sentinel lymph node slides from women with breast cancer, some of which contain precisely annotated regions of lymph node metastasis of different sizes. The dataset was issued by two hospitals in the Netherlands (RUMC, UMCU). The CAMELYON17 dataset [3] extends the CAMELYON16 dataset by 1000 slides from five medical centers (RUMC, UMCU, CWZ, RST, LPON), from which we only considered CZW, RST, and LPON as additional medical centers.

We extracted patches from the slides of 256×256 pixels without overlap at 20x magnification (0.5 microns per pixel). We identified and excluded background patches via Otsu's method [4] on slide thumbnails and applied a patch-level minimum standard deviation of 8. Each patch was labeled with the medical center from which its slide originated. We additionally labeled those patches fully inside the metastasis annotations as tumor and those fully outside as normal. For RUMC and UMCU, we sampled 300 normal patches each from 17 slides without metastases and 300 tumor patches each from 17 slides with metastases, making up our in-domain Camelyon dataset. For CWZ, RST, and LPON, where less data were available, we sampled 335 normal patches per slide from 5 slides without metastases and 327-335 tumor patches from 3-6 slides with metastases, respectively. These patches were used solely for out-of-domain evaluation. Supplementary Table 1 details the resulting dataset statistics.

	RUMC	UMCU	CWZ	RST	LPON	Total
Normal	5,100 (17)	5,100 (17)	335 (5)	335 (5)	335 (5)	11,205 (49)
Tumor	5,100 (17)	5,100 (17)	327 (5)	335 (6)	335 (3)	11,197 (48)
Total	10,200 (34)	10,200 (34)	662 (10)	670 (11)	670 (8)	22,402 (97)

Supplementary Table 1: **Dataset statistics for Camelyon.** We report the number of patches together with the number of slides / cases in parentheses.

1.2 TCGA

The TCGA-UT dataset (`https://zenodo.org/records/5889558`, License: Non-Commercial Use: CC-BY-NC-SA 4.0) [5, 6] provides tumor patches from 32 distinct cancer types within The Cancer Genome Atlas (TCGA). Each patch was extracted from an annotated tumor region, and each slide comes with multiple annotations created by trained pathologists. We used the highest available resolution level for the patches, corresponding to 0.5 μm/pixel or roughly 20x magnification. The patches have a size of 256x256 pixels.

To cover a large fraction of the dataset while maintaining class-center balancing, we extracted combinations of 2 cancer types and 2 medical centers (2x2 dataset). For each combination, we sampled 10 slides, from which we sampled 10 patches from 3 regions-of-interest, respectively. The resulting dataset statistics are shown in Supplementary Table 2. In total, we found 94 2x2 combinations for which sufficient data were available. The complete list is provided in Supplementary Table 5. It covers combinations of similar cancer types (e.g. lung adenocarcinoma and lung squamous cell carcinoma) as well as unrelated tumor entities (e.g. lung adenocarcinoma and glioblastoma). We further composed a single dataset comprising 4 classes and 4 tissue source sites (4x4 dataset). In this case, we sampled 12 slides with 10 patches from 3 regions-of-interest, each, and added data from 4 further medical centers for out-of-domain assessment, resulting in the statistics of Supplementary Table 3.

	MCT 1	MCT 2	Total
CLS 1	300 (10)	300 (10)	600 (20)
CLS 2	300 (10)	300 (10)	600 (20)
Total	600 (20)	600 (20)	1,200 (40)

Supplementary Table 2: **Dataset statistics for TCGA 2x2 per 2x2 combination.** We report the number of patches together with the number of slides / cases in parentheses. CLS = tumor class, MCT = medical center. The total dataset includes 94 double pairs, totaling 112,800 patches. Note that the same patch can be present in multiple 2x2 combinations.

	Asterand	Christiana	Roswell Park	Pittsburgh	...
BRCA	360 (12)	360 (12)	360 (12)	360 (12)	
COAD	360 (12)	360 (12)	360 (12)	360 (12)	
LUAD	360 (12)	360 (12)	360 (12)	360 (12)	
LUSC	360 (12)	360 (12)	360 (12)	360 (12)	
Total	1,440 (48)	1,440 (48)	1,440 (48)	1,440 (48)	...

...	Cureline	IGC	Poland	Johns Hopkins	Total
	300 (10)	300 (10)	0 (0)	0 (0)	2,040 (68)
	0 (0)	300 (10)	300 (10)	0 (0)	2,040 (68)
	0 (0)	0 (0)	300 (10)	300 (10)	2,040 (68)
	300 (10)	0 (0)	0 (0)	300 (10)	2,040 (68)
...	600 (20)	600 (20)	600 (20)	600 (20)	8,160 (272)

Supplementary Table 3: **Dataset statistics for TCGA 4x4.** We report the number of patches together with the number of slides / cases in parentheses. BRCA = breast invasive carcinoma; COAD = colon adenocarcinoma; LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma. Christiana = Christiana Healthcare; Pittsburgh = University of Pittsburgh; Poland = Greater Poland Cancer Center; IGC = International Genomics Consortium.

	WNS	CHA	UKK	TCGA	Total
TUMOR	900 (9)	900 (9)	500 (5)	500 (5)	2,800 (28)
REGR_TU	900 (9)	900 (9)	500 (5)	0	2,300 (23)
SH_OES	900 (9)	900 (9)	500 (5)	500 (5)	2,800 (28)
SH_MAG	900 (9)	900 (9)	500 (5)	500 (5)	2,800 (28)
MUSC_PROP	900 (9)	900 (9)	500 (5)	500 (5)	2,800 (28)
ADVENT	900 (9)	900 (9)	500 (5)	500 (5)	2,800 (28)
Total	5,400 (14)	5,400 (31)	3,000 (17)	2,500 (15)	16,300 (77)

Supplementary Table 4: **Dataset statistics for Tolkach ESCA.** We report the number of patches together with the number of slides / cases in parentheses. TUMOR = vital tumor tissue; REGR_TU = regression areas; SH_OES = oesophageal mucosa; SH_MAG = gastric mucosa; MUSC_PROP = muscularis propria; ADVENT = adventitial tissue. UKK = University Hospital Cologne; WNS = Landesklinikum Wiener Neustadt; CHA = University Hospital Berlin–Charité; TCGA = several participating institutions from the Cancer Genome Atlas, TCGA-ESCA project. Note that the same case may contribute patches to multiple biological classes from the same medical center.

1.3 Tolkach ESCA

The Tolkach ESCA dataset (`https://zenodo.org/records/7548828`, License: Custom, Non-Commercial Use) [7, 8] contains specimens from surgically resected oesophageal adenocarcinoma and adenocarcinoma of the oesophagogastric junction. The 320 slides from 126 patients were obtained from 4 medical centers, 3 of which (UKK, CHA, WNS) utilized the same tissue scanner.

	Disease Classes	Medical Centers
1	BLCA, BRCA	University of Pittsburgh, MD Anderson
2	BLCA, CESC	Asterand, Barretos Cancer Hospital
3	BLCA, COAD	Asterand, University of Pittsburgh
4	BLCA, ESCA	Asterand, Barretos Cancer Hospital
5	BLCA, HNSC	Asterand, Barretos Cancer Hospital
6	BLCA, LIHC	Asterand, ILSBio
7	BLCA, LUAD	Asterand, University of Pittsburgh
8	BLCA, LUSC	Asterand, University of Pittsburgh
9	BLCA, PRAD	University of Pittsburgh, University of California San Francisco
10	BLCA, SKCM	Asterand, University of Pittsburgh
11	BLCA, STAD	Barretos Cancer Hospital, ILSBio
12	BLCA, THCA	University of Pittsburgh, Memorial Sloan Kettering
13	BRCA, CESC	Asterand, International Genomics Consortium
14	BRCA, COAD	Asterand, Christiana Healthcare
15	BRCA, GBM	UCSF, Duke
16	BRCA, HNSC	Asterand, University of Miami
17	BRCA, KIRC	University of Pittsburgh, International Genomics Consortium
18	BRCA, KIRP	International Genomics Consortium, Roswell Park
19	BRCA, LGG	Duke, Cureline
20	BRCA, LIHC	Asterand, ILSBio
21	BRCA, LUAD	University of Pittsburgh, Christiana Healthcare
22	BRCA, LUSC	Cureline, Christiana Healthcare
23	BRCA, PRAD	International Genomics Consortium, Roswell Park
24	BRCA, READ	Greater Poland Cancer Center, Christiana Healthcare
25	BRCA, SKCM	Asterand, Cureline
26	BRCA, STAD	Asterand, International Genomics Consortium
27	BRCA, THCA	University of Pittsburgh, International Genomics Consortium
28	BRCA, UCEC	University of Pittsburgh, Duke
29	CESC, COAD	Asterand, International Genomics Consortium
30	CESC, ESCA	Asterand, Barretos Cancer Hospital
31	CESC, HNSC	Asterand, Barretos Cancer Hospital
32	CESC, LUAD	Asterand, International Genomics Consortium
33	CESC, LUSC	Asterand, International Genomics Consortium
34	CESC, STAD	Asterand, Barretos Cancer Hospital
35	CESC, THCA	Asterand, International Genomics Consortium
36	COAD, HNSC	University of Pittsburgh, International Genomics Consortium
37	COAD, KICH	MSKCC, Harvard
38	COAD, KIRC	University of Pittsburgh, MSKCC
39	COAD, KIRP	MSKCC, Roswell Park
40	COAD, LUAD	University of Pittsburgh, Roswell Park
41	COAD, LUSC	University of Pittsburgh, Roswell Park
42	COAD, PRAD	University of Pittsburgh, International Genomics Consortium
43	COAD, READ	Greater Poland Cancer Center, Christiana Healthcare
44	COAD, SKCM	Asterand, University of Pittsburgh
45	COAD, STAD	Greater Poland Cancer Center, International Genomics Consortium
46	COAD, THCA	Asterand, International Genomics Consortium
47	COAD, UCEC	University of Pittsburgh, MSKCC
48	ESCA, HNSC	Asterand, Barretos Cancer Hospital
49	ESCA, STAD	Asterand, Barretos Cancer Hospital
50	ESCA, THCA	Asterand, University Health Network
51	GBM, LGG	University of Florida, Duke
52	HNSC, KIRC	University of Pittsburgh, International Genomics Consortium
53	HNSC, KIRP	UNC, International Genomics Consortium
54	HNSC, LIHC	Asterand, UNC
55	HNSC, LUAD	UNC, International Genomics Consortium
56	HNSC, LUSC	Asterand, International Genomics Consortium
57	HNSC, PRAD	University of Pittsburgh, International Genomics Consortium
58	HNSC, SKCM	Asterand, University of Pittsburgh
59	HNSC, STAD	Asterand, Barretos Cancer Hospital
60	HNSC, THCA	Asterand, Johns Hopkins
61	KICH, KIRC	MSKCC, Harvard
62	KICH, UCEC	University of North Carolina, MSKCC
63	KIRC, KIRP	UNC, International Genomics Consortium
64	KIRC, LUAD	University of Pittsburgh, International Genomics Consortium
65	KIRC, LUSC	MSKCC, International Genomics Consortium
66	KIRC, PRAD	University of Pittsburgh, International Genomics Consortium
67	KIRC, THCA	University of Pittsburgh, International Genomics Consortium
68	KIRC, UCEC	University of Pittsburgh, MSKCC
69	KIRP, LIHC	Moffitt Cancer Center, UNC
70	KIRP, LUAD	UNC, International Genomics Consortium
71	KIRP, LUSC	International Genomics Consortium, Roswell Park
72	KIRP, PAAD	Moffitt Cancer Center, International Genomics Consortium
73	KIRP, PRAD	University of California San Francisco, International Genomics Consortium
74	LGG, LUSC	Mayo Clinic - Rochester, Cureline
75	LGG, UCEC	MSKCC, Duke
76	LIHC, LUAD	Asterand, UNC
77	LIHC, LUSC	Asterand, Mayo Clinic - Rochester
78	LIHC, PAAD	Moffitt Cancer Center, Alberta Health Services
79	LIHC, STAD	Asterand, ILSBio
80	LUAD, LUSC	Roswell Park, Johns Hopkins
81	LUAD, PRAD	International Genomics Consortium, Roswell Park
82	LUAD, SKCM	Asterand, University of Pittsburgh
83	LUAD, STAD	Asterand, International Genomics Consortium
84	LUAD, THCA	International Genomics Consortium, Johns Hopkins
85	LUSC, PRAD	International Genomics Consortium, Roswell Park
86	LUSC, SKCM	Asterand, University of Pittsburgh
87	LUSC, STAD	Asterand, International Genomics Consortium
88	LUSC, THCA	Asterand, University of Pittsburgh
89	LUSC, UCEC	University of Pittsburgh, MSKCC
90	PRAD, THCA	University of Pittsburgh, International Genomics Consortium
91	PRAD, UCEC	Washington University, University of Pittsburgh
92	SKCM, THCA	Asterand, University of Pittsburgh
93	STAD, THCA	Asterand, International Genomics Consortium
94	THCA, UCEC	University of North Carolina, University of Pittsburgh

Supplementary Table 5: **Disease class and medical center combinations in the TCGA 2x2 dataset.**

Further, they only contributed samples from patients who received neoadjuvant chemotherapy and had similar staging characteristics. In the remaining medical center (TCGA), different scanners were used, and no neoadjuvant treatment was given. Each slide was manually annotated and segmented into 11 tissue classes by board-certified pathologists, including tumor, regression tissue, and mucosa. Patches were extracted from these annotated regions at a resolution of approximately 0.75 microns per pixel and with a size of 256x256 pixels.

We selected the biological classes that have at least 900 (WNS, CHA) or 500 (UKK, TCGA) patches from 9 or 5 different cases in each medical center, and sampled 100 patches per case for each class-center combination. Notice that within one medical center, a case may contribute patches to multiple biological classes. The TCGA subset was only used for out-of-domain generalization considerations. The resulting dataset statistics are summarized in Supplementary Table 4.

Supplementary Note 2. Robustness index computation and further results

This section provides details regarding the implementation and calculation of the robustness index.

2.1 Data pre-processing: l2 normalization

Each representation vector was normalized using l2 normalization before computing the robustness index. By virtue of this choice, applying the Euclidean distance to the normalized embeddings, as used here, will result in the same neighbor ranking as using the cosine distance.

2.2 Selecting k values

The robustness index metric is defined for all values of k. To compare models, it's preferable to obtain a single scalar metric; below, we discuss how this selection is performed.

In this work, we choose to analyze the exact set of neighbors used by a KNN classifier, so that the robustness index value represents the robustness of the KNN classifier. We therefore base our choice of k on values for k that would be used in practice when applying KNN classification. We note that considerations for choosing k will depend on the context and application of a foundation model; for example, if the nearest neighbors are to be used for reference search rather than prediction, then the number of desired similar samples could guide the choice of k.

When a KNN model is applied for a prediction task, k will in current practice typically be chosen so as to facilitate high prediction performance (Our work suggests this should not be the only consideration; since the choice of k influences the robustness of the model, we argue it's advisable to select k such that a favorable tradeoff between prediction performance and robustness is attained). Therefore, per model, we first determine the value of k that maximizes the prediction performance of the biological class. Once all per-model optimal k values have been determined, the median of these is used as the common choice of k for a particular dataset.

The following median optimal k values are obtained:

- TCGA Uniform: 61
- Camelyon16: 11
- Tolkach ESCA: 46

Leave one-case-out cross-validation for KNN classification To evaluate KNN prediction performance in the above procedure to identify the optimal value of k, we first determine all k nearest neighbors for each sample in the dataset once for a high number of k. Next, for each sample, we remove all neighbors that come from the same case, as using neighbors from the same case would mean that case is split across training and evaluation. This evaluation can be viewed as leave-one-case-out cross-validation; each case is used in evaluation once, using all remaining cases for training. It can equivalently be viewed as n-fold case-level cross-validation, where n is maximized to equal the total number of cases.

This yields the largest possible training set for each evaluation sample, and thus the most accurate, lowest variance result. It also implies that the procedure is deterministic, and has a single true outcome; i.e., there is no variation in training/evaluation splits, and thus no variance in the resulting outcome. Given that the dataset can be viewed as a sample from a larger, unknown full distribution, we estimate the variance in the resulting robustness index values using bootstrapping.

Efficient optimization of k for KNN classification As noted above, we first determine all k neighbors for a high value of k once. Next, we iteratively decrease k, compute the corresponding predictions, and track the resulting prediction accuracies, as measured using balanced accuracy. Since the sets of neighbors for decreasing numbers of k are strict subsets of the neighbors for larger values of k, we do not need to recompute the set of neighbors, but can simply take increasingly small subsets.

2.3 Efficient computation of the robustness index

To efficiently compute the robustness index for a range of k values, analogous to the above, a single KNN model with the highest k value of the range can be fit. After retrieving the set of nearest neighbors from the fitted model (e.g. using the sklearn kneighbors function) and retrieving binary matrices indicating, for each combination of sample $\times$ neighbor, whether the biological and confounding class are the same, the robustness index can be obtained by taking the cumulative sum of the column sums for SO and OS, and normalizing these vectors to sum to one, after which the normalized cumulative SO vector represents the robustness index.

Two dataset scenario The robustness index can also be computed for an evaluation set D_{eval} in a two-dataset setup where neighbors are selected from a different dataset D', provided that both datasets share the same set of biological and confounding classes.

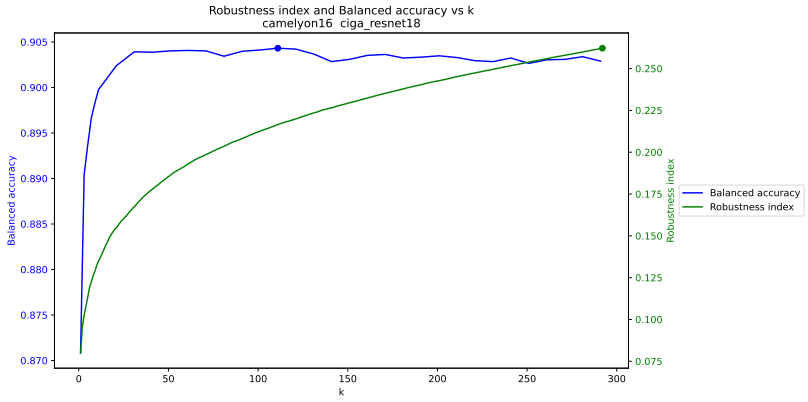
Dataset details The details of the datasets used for robustness index calculation are as follows:

- Camelyon: We used the patches from the RUMC and UMCU medical centers (see Supplementary Table 1), yielding 20,400 patches.
- TCGA 2x2: We used the full TCGA 2x2 dataset as specified in Supplementary Table 2, yielding a dataset of $4 \cdot 300 \cdot 94 = 112,800$ patches in total.
- Tolkach ESCA: We used the selection of the Tolkach ESCA dataset, described in Supplementary Table 4, with 5 randomly selected cases per combination of biological and confounding class, 100 patches per case, and excluding TCGA (as this contains multiple centers), yielding 9,000 patches in total (The case ID for this dataset was kindly provided by Yuri Tolkach upon request).

TCGA 2x2: paired version of the robustness index The robustness index, as defined in this article, can be applied to any dataset or pair of datasets for which both biological and confounding class information is available. To analyze whether biological or confounding information is more strongly represented, ideally, a balanced dataset should be used that has a similar number of biological and confounding classes, and where the number of samples for each combination of biological and confounding is similar (see Sup. Note 1). An example is the TCGA-2k dataset described in [9], which has five biological and confounding classes chosen such that for 20 of the 25 combinations of these, an equal number of samples can be selected.

The requirement of similar distributions of biological and confounding classes is a limiting factor. To overcome this, when many biological and confounding classes are available, as is the case in TCGA, a paired variant of the robustness index can be calculated as follows:

- Double pairs of two biological classes (TCGA projects) $\times$ two confounding classes (tissue source sites) are selected. The TCGA 2x2 subset in the PathoROB benchmark contains 94 such double pairs.
- The robustness index statistics (i.e. the SO and OS frequencies) are calculated separately per double pair, and then aggregated over all 94 double pairs.



Supplementary Figure 1: **Trading off performance vs robustness**. By choosing different values for k, different tradeoffs can be obtained. Source data are provided as a Source Data file.

2.4 Trading off performance vs robustness

We argue that prediction performance and robustness are both important objectives that must be included in the evaluation of medical foundation models. Both objectives are to be optimized.

In the context of KNN classification, selecting a value for k implies making a tradeoff between prediction performance and robustness. Our results show that for many models, robustness can easily be improved compared to the selected optimal k value by selecting a different value for k. We encourage practitioners to carefully consider what tradeoff between robustness and prediction performance is advisable, and note that for medical applications, robustness is of the utmost importance, given that the lives of patients may be at stake.

Supplementary Figure 1 shows how the choice of k affects the prediction performance and robustness. While maximal prediction performance is obtained for k values around 100, substantially higher robustness values can be achieved by choosing larger values of k, with relatively limited impact on prediction performance.

2.5 Standard deviation estimation

For a given dataset, the robustness index calculation is deterministic, and so there is no variance in the value. We can view the dataset as having been sampled from the hypothetical larger distribution of all possible datasets that could have been obtained in the same manner from the same population, and the robustness index values for such different instantiations do have a variance.

Bootstrapping [10] approximates this variance by treating the original dataset as a proxy for the population, generating multiple datasets via resampling, and computing the statistic's variability across these resamples. Given the knn neighbor lists for all samples in the dataset, we randomly sample a dataset of the same size with replacement. The resulting dataset will contain some samples multiple times and will not include all samples. We calculate the robustness index over the sampled set. This procedure is repeated $n = 1000$ times, yielding 1000 robustness index values. The estimated standard deviation of these robustness values is computed by taking the standard deviation over the computed robustness index values.

2.6 Results per dataset

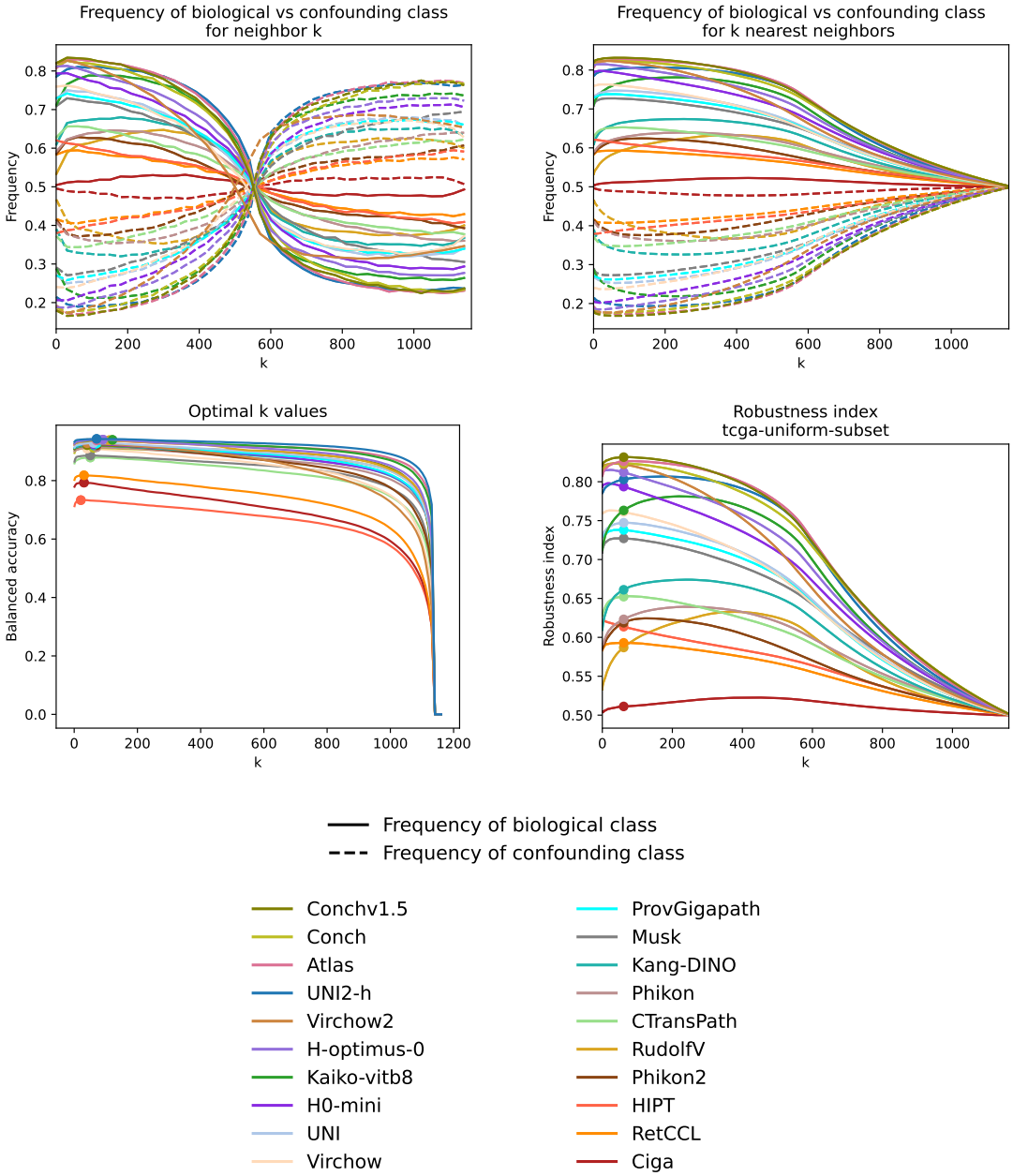
This section provides overview graphs with:

- Top left: the frequencies of SO (solid lines) and OS (dashed lines) for neighbor k

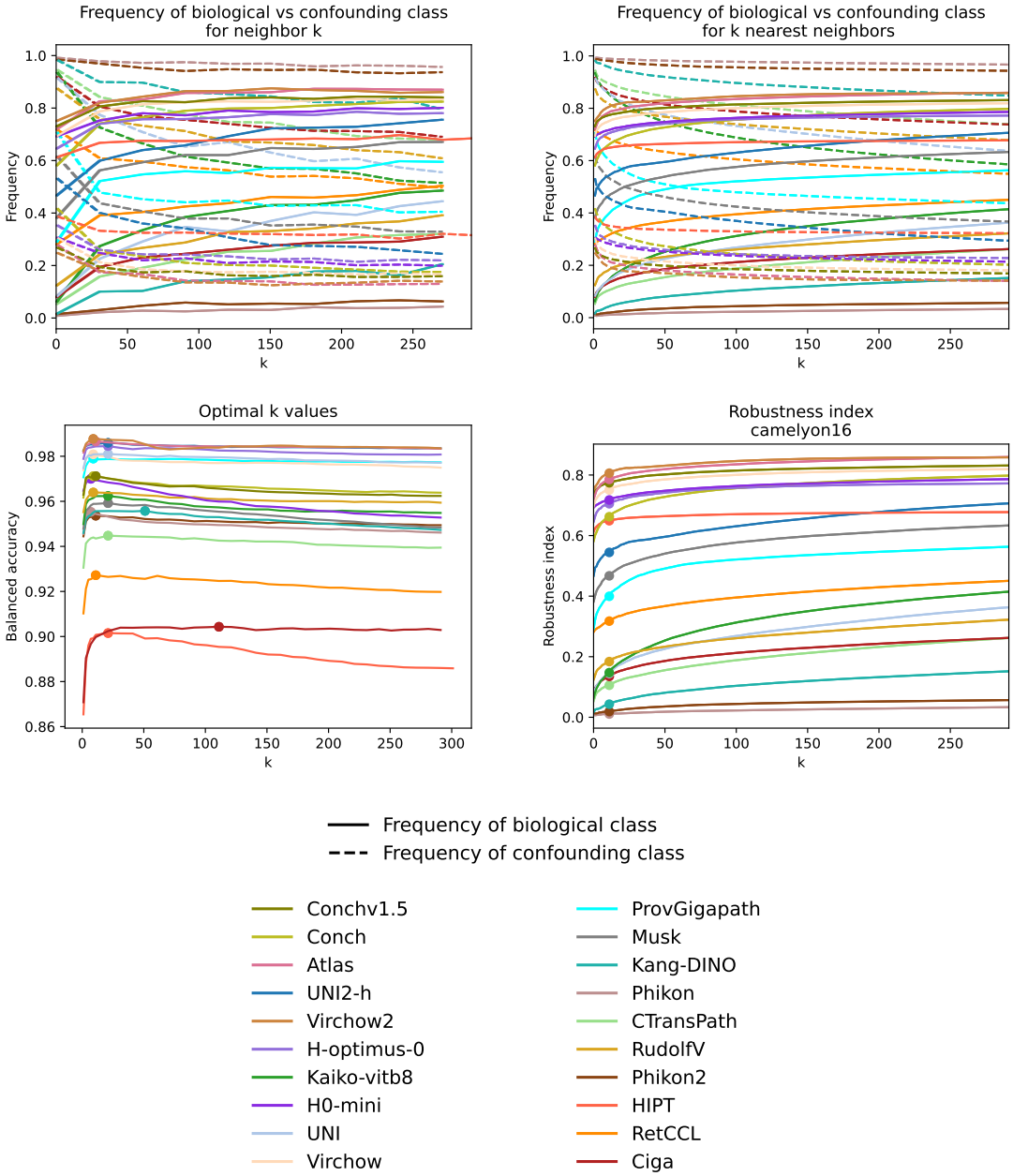
- Top right: the same information, aggregated over the k nearest neighbors
- Bottom left: the effect of k on prediction accuracy; the dot shows the optimal choice of k per model. The median of these optimal k values is chosen as the common choice of k.
- Bottom right: robustness index for all models; the value at the median k value is used to compare models.

Supplementary Figure 2 shows the robustness index results on TCGA 2x2 for all models. As the starting point (top left), the frequency with which the kth-nearest neighbor (e.g. k=3 represent the third nearest neighbor) has the Same biological class (e.g. cancer type) but the Other confounding class (e.g. medical center) is measured (combination SO; the solid lines show this for the $k = 1$ to 600 (the total number of SO + OS neighbors for one of the 94 TCGA project combinations). The complement of this, i.e. the fraction of neighbors at distance rank k that have the Other biological class but the Same confounding class (combination OS) is represented by the dashed lines. As the analysis is restricted to the SO and OS combinations, these two lines sum to one, and the total number of neighbors considered is restricted to an exclusive upper bound of 600 of the $4 \cdot 300 = 1200$ for each TCGA project combination. The top right show the same information in aggregated form, reflecting these frequencies of all nearest neighbors from distance rank 1 up to the specified value of k, as used in KNN classifiers. The solid lines at the top equal the robustness index. To select a value for k, the bottom left shows the performance for predicting the biological class as a function of k; dots mark the optimal value of k for each model. The median of these optimal k values, 61, is used in the bottom right zoomed-in robustness index plot to obtain the scalar robustness index value.

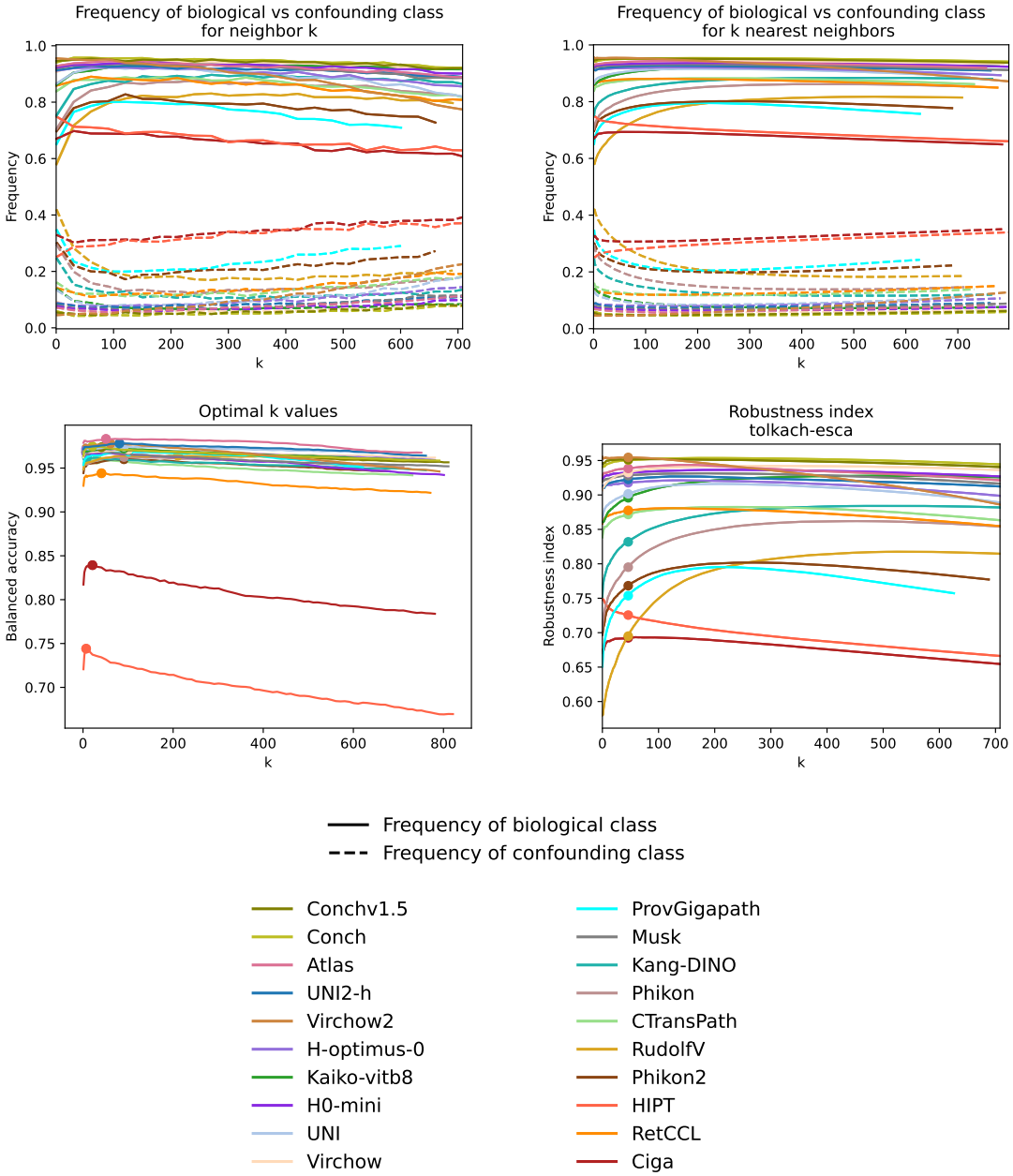
Supplementary Figures 3 and 4 on the following pages show the same graphs for the Camelyon and Tolkach ESCA datasets.



Supplementary Figure 2: **Robustness index results on TCGA 2x2 for all models**. See text. Source data are provided as a Source Data file.



Supplementary Figure 3: **Robustness index results on Camelyon**. Source data are provided as a Source Data file.



Supplementary Figure 4: **Robustness index results on Tolkach ESCA**. Source data are provided as a Source Data file.

2.7 Robustness index: numerical results

This subsection summarizes the robustness index results for all processings. For the default processing (no robustification method applied), we include the results of bootstrapping, see above under Section 2.5; the second and third numerical columns report the mean and standard deviation of the resulting 1000 robustness index values.

Processing	Default	Default	Default	Reinhard	Combat	Reinhard Combat
Model	RI	RI mean	std dev	RI	RI	RI
Atlas	0.826	0.826	0.001	0.829	0.847	0.849
Ciga	0.511	0.511	0.001	0.565	0.627	0.635
CONCH	0.824	0.824	0.001	0.819	0.815	0.804
CONCHv1.5	0.832	0.832	0.001	0.840	0.822	0.818
H-optimus-0	0.812	0.812	0.001	0.823	0.849	0.847
H0-mini	0.794	0.794	0.001	0.805	0.836	0.834
HIPT	0.614	0.614	0.001	0.629	0.632	0.639
Kaiko ViT-B/8	0.763	0.763	0.002	0.776	0.813	0.816
Kang-DINO	0.661	0.661	0.002	0.670	0.804	0.801
Phikon	0.623	0.623	0.002	0.682	0.761	0.768
Phikon-v2	0.619	0.619	0.002	0.659	0.762	0.761
Prov-GigaPath	0.738	0.738	0.001	0.783	0.809	0.819
RetCCL	0.593	0.593	0.001	0.606	0.690	0.679
RudolfV	0.587	0.587	0.002	0.595	0.729	0.730
CTransPath	0.652	0.652	0.001	0.677	0.747	0.737
UNI	0.747	0.747	0.001	0.761	0.822	0.812
UNI2-h	0.803	0.803	0.001	0.831	0.839	0.850
MUSK	0.727	0.727	0.001	0.747	0.762	0.767
Virchow	0.761	0.761	0.001	0.714	0.782	0.711
Virchow2	0.822	0.822	0.001	0.829	0.841	0.838

Supplementary Table 6: **Robustness index results for different processings on TCGA 2x2**. The first column shows the robustness index computed over the full dataset with leave-one-case-out cross-validation for the default processing (no robustification method applied). The second and third columns show the mean and standard deviation of the robustness index computed using bootstrapping. The remaining columns show the robustness index for the different robustification methods compared.

Processing	Default	Default	Default	Reinhard	Combat	Reinhard Combat
Model	RI	RI mean	std dev	RI	RI	RI
Atlas	0.785	0.786	0.008	0.823	0.927	0.926
Ciga	0.135	0.135	0.004	0.569	0.458	0.654
CONCH	0.662	0.662	0.007	0.889	0.921	0.931
CONCHv1.5	0.774	0.774	0.006	0.892	0.925	0.930
H-optimus-0	0.705	0.705	0.009	0.890	0.907	0.931
H0-mini	0.718	0.717	0.007	0.840	0.916	0.918
HIPT	0.649	0.649	0.004	0.734	0.755	0.790
Kaiko ViT-B/8	0.147	0.147	0.007	0.742	0.631	0.835
Kang-DINO	0.043	0.043	0.004	0.290	0.758	0.776
Phikon	0.011	0.011	0.002	0.462	0.575	0.740
Phikon-v2	0.019	0.019	0.002	0.484	0.446	0.724
Prov-GigaPath	0.399	0.399	0.012	0.831	0.800	0.896
RetCCL	0.318	0.318	0.006	0.665	0.757	0.801
RudolfV	0.184	0.183	0.008	0.232	0.616	0.634
CTransPath	0.106	0.106	0.005	0.589	0.668	0.772
UNI	0.145	0.145	0.009	0.712	0.751	0.875
UNI2-h	0.544	0.545	0.013	0.872	0.868	0.931
MUSK	0.467	0.467	0.008	0.782	0.812	0.866
Virchow	0.751	0.750	0.008	0.891	0.896	0.918
Virchow2	0.806	0.806	0.007	0.923	0.941	0.953

Supplementary Table 7: **Robustness index results for different processings on Camelyon**. The first column shows the robustness index computed over the full dataset with leave-one-case-out cross-validation for the default processing (no mitigation method applied). The second and third columns show the mean and standard deviation of the robustness index computed using bootstrapping. The remaining columns show the robustness index for the different mitigation methods compared.

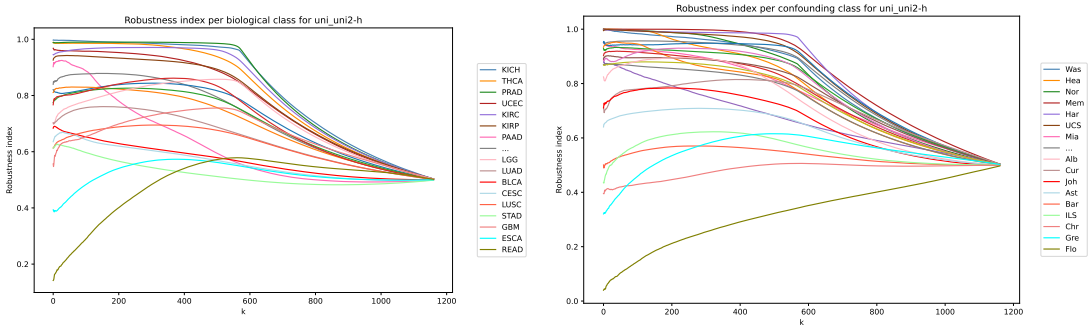
Processing	Default	Default	Default	Reinhard	Combat	Reinhard Combat
Model	RI	RI mean	std dev	RI	RI	RI
Atlas	0.938	0.939	0.003	0.940	0.972	0.971
Ciga	0.693	0.693	0.004	0.685	0.801	0.781
CONCH	0.951	0.951	0.002	0.945	0.972	0.965
CONCHv1.5	0.951	0.951	0.002	0.956	0.969	0.969
H-optimus-0	0.918	0.918	0.003	0.935	0.953	0.958
H0-mini	0.932	0.932	0.003	0.940	0.963	0.962
HIPT	0.726	0.725	0.004	0.728	0.765	0.774
Kaiko ViT-B/8	0.896	0.896	0.004	0.920	0.955	0.954
Kang-DINO	0.832	0.832	0.005	0.863	0.941	0.939
Phikon	0.795	0.795	0.006	0.835	0.934	0.922
Phikon-v2	0.768	0.768	0.006	0.820	0.923	0.910
Prov-GigaPath	0.754	0.754	0.007	0.887	0.915	0.939
RetCCL	0.878	0.877	0.003	0.853	0.943	0.925
RudolfV	0.695	0.695	0.007	0.794	0.883	0.899
CTransPath	0.872	0.872	0.004	0.872	0.947	0.937
UNI	0.902	0.901	0.004	0.907	0.957	0.951
UNI2-h	0.923	0.923	0.003	0.944	0.957	0.963
MUSK	0.928	0.928	0.002	0.931	0.958	0.954
Virchow	0.932	0.932	0.003	0.940	0.959	0.959
Virchow2	0.955	0.955	0.002	0.955	0.970	0.967

Supplementary Table 8: **Robustness index results for different processings on Tolkach ESCA**. The first column shows the robustness index computed over the full dataset with leave-one-case-out cross-validation for the default processing (no mitigation method applied). The second and third columns show the mean and standard deviation of the robustness index computed using bootstrapping. The remaining columns show the robustness index for the different mitigation methods compared.

2.8 Robustness index per biological and confounding category

The main results measure the degree to which biological features dominate confounding ones in the embedding space in total, computed over the full dataset. To see to what extent the robustness of a foundation model varies across different biological or confounding categories, we computed the robustness index per class for all biological and confounding categories.

Supplementary Figure 5 shows the results for UNI2-h. For both biological and confounding classes, robustness varies substantially across different categories.



Supplementary Figure 5: **Robustness index per biological class (left) and per confounding class (right) for UNI2-h**. This breakdown shows that robustness varies substantially across different categories, for both biological and confounding categories. Source data are provided as a Source Data file.

2.9 Relation between biological representation robustness and out-of-distribution (OOD) generalization

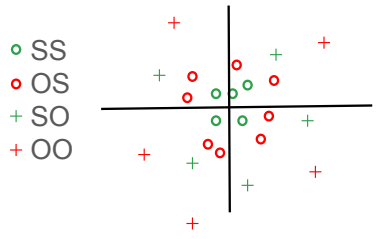
Biological representation robustness was defined here as the ability to capture biological features while ignoring confounding features. To clarify the importance of robustness for medical ML models, we describe an example. Imagine two patients with identical conditions treated at different hospitals, each using distinct tissue processing labs and scanners. A robust medical ML model should produce comparable diagnoses for both patients, regardless of these technical differences.

Achieving such comparable results requires Out-Of-Distribution (OOD) generalization—the ability to maintain high performance when a model trained on hospital A's data is applied to hospital B's data. While OOD generalization requires robustness, it does not imply robustness; a model can work well for different medical centers, yet produce robustness issues when data from different centers is combined. This can occur when the representation of biological classes differs across medical centers. A schematic example of such center-specific representations is described below.

Example: center-specific representations affect prediction performance

Supplementary Figure 6 shows a schematic example. Consider a patient whose embedding lies at the origin. Green circles represent correctly classified samples from the patient's hospital (Hospital A), while red circles show misclassified ones from the same hospital. Similarly, green and red crosses represent correct and incorrect samples from Hospital B.

A KNN model trained solely on Hospital A's data can achieve perfect in-distribution (ID) performance by using only green circles as neighbors. Likewise, a KNN model trained exclusively on Hospital B's data can achieve perfect OOD performance using only green crosses.



Supplementary Figure 6: Example illustrating how non-robust models can accurately predict within individual data subsets yet fail when these subsets are combined.

However, when data from both hospitals are combined, the model fails. Technical differences between hospitals (e.g., scanner characteristics) can cause Hospital A's negative samples to cluster closer to the patient than Hospital B's positive samples, leading to misclassification.

Robustness requires models to be invariant to hospital-specific characteristics. As the example demonstrates, this extends beyond achieving high accuracy on individual hospital datasets—models must maintain performance when data sources are mixed. Thus, robustness represents a more stringent requirement than OOD generalization alone. The following section discusses and compares different metrics of robustness.

Empirical comparison of robustness metrics

We showed that spurious correlations between biological and confounding features led to performance drops (Main Paper: Figure 3). We attribute these performance drops to non-robust models relying on confounding features rather than biological signals. A valid robustness index should be predictive of the magnitude of these performance drops. We therefore empirically validate this relationship and benchmark against alternative robustness metrics; see Main Paper: Methods for a definition of the main comparison metrics.

The following additional robustness metrics were compared:

- The Robustness metric only captures distance-based predictability of the confounding class. The Confounder logistic regression AUC generalizes this by measuring how well logistic regression can predict the confounding class from embeddings, with lower AUC indicating greater robustness.
- A robust model should generalize to unseen confounding classes (e.g. medical centers or scanners), with OOD performance matching ID performance. The Generalization Index $\mathcal{G}$ measures the ratio OOD/ID, which approaches 1 for robust models.

$$\mathcal{G} = \frac{\text{OOD performance}}{\text{ID performance}} = \frac{SO}{SO + OO} \cdot \frac{SS + OS}{SS} = \frac{SO \cdot (SS + OS)}{SS \cdot (SO + OO)} \tag{1}$$

Of the five robustness metrics compared (see Supplementary Table 9), the robustness index is most predictive of both the APD ID and the APD OOD reported in the previous section, across both correlation measures. In absolute terms, the robustness index strongly correlates with the APD, with Spearman rank correlations of $r_s = 0.80$ (APD OOD) and $r_s = 0.90$ (APD ID). This provides empirical validation of the utility of the robustness index.

Confounder Insensitivity correlates only weakly with the APD, and (especially for APD OOD) has high p-values, indicating a weak relation. This suggests that the common notion of robustness as insensitivity to confounders may require revision.

Supplementary Table 9: **Correlations between representation-level robustness metrics and downstream model robustness**. Pearson (r) and Spearman rank (r_s) correlations (n=20, two-sided permutation test) of robustness metrics with the in-domain (ID) and out-of-domain (OOD) average performance drop (APD). The robustness index is most predictive of the APD across both correlation measures and APD variants.

Robustness metric	APD ID				APD OOD			
	r	(p-value)	r_s	(p-value)	r	(p-value)	r_s	(p-value)
Robustness index	**0.92**	(0.00)	**0.90**	(0.00)	**0.68**	(0.00)	**0.80**	(0.00)
Bio-vs-confounding ratio	0.85	(0.00)	0.85	(0.00)	0.51	(0.02)	0.66	(0.00)
Generalization index	-0.43	(0.06)	-0.34	(0.15)	-0.13	(0.60)	-0.15	(0.52)
Confounder Insensitivity	0.39	(0.09)	0.49	(0.03)	-0.06	(0.82)	0.27	(0.26)
Confounder logistic regression AUC	-0.29	(0.22)	-0.33	(0.16)	0.15	(0.52)	-0.07	(0.76)

Supplementary Note 3. Robust foundation models prevent entangled features

3.1 Spurious correlation in downstream datasets

Multi-site histopathology datasets are typically imbalanced across contributing centers; for example, a disease class may be overrepresented at one medical site and underrepresented at others. This is particularly pronounced for rare diseases or tumor subtypes. Such imbalances, coupled with center-specific technical artifacts, induce spurious correlations between biological labels and the site signatures: a deceptive relationship. While these correlations reflect co-occurrence, they do not correspond to a medically meaningful coherence or causation [11, 12]. The risk of spurious correlation increases particularly in small datasets from a limited number of medical centers. Without deliberate data curation and balancing, it is unlikely that the distributions of biological classes, confounding factors, or their combinations will be uniform across sites.

3.2 Spurious correlation can hamper generalizability

Spurious correlation can significantly impair the generalizability of downstream models trained on non-robust foundation model representations. A foundation model with unrobust embeddings might encode both biologically meaningful and confounding information, thereby potentially propagating spurious correlation from the input to its representation space; its representations hold biological and confounding features, each correlated with the prediction target. A feature refers to a dimension or a one-dimensional direction in the representation space. Although the representations may contain sufficient information to learn the correct association between the biological features and the target label, the downstream model might just as well base its predictions on the confounding feature instead. Crucially, the model has no inherent way

of distinguishing or disentangling whether the correlation between a given feature and the target is medically meaningful or merely spurious. As a result, downstream models trained on spuriously correlated representations may inadvertently leverage these medical center-specific signals as Clever Hans (or shortcut) features. Although these Clever Hans features may enhance apparent model accuracy during development, they compromise generalization when such spurious correlations are absent at deployment [13–19].

In contrast to the abundance of unlabeled datasets used for pre-training foundation models, only limited labeled datasets are typically available for training on a downstream task, effectively increasing the risk of having spurious correlations in downstream inputs. A FM that encodes confounding variables within its representations then inherently snowballs this spurious correlation, possibly resulting in Clever Hans learning in downstream tasks. Conversely, representation spaces that are free of such confounding information effectively preclude the risk of Clever Hans learning even under strongly biased data — after all, downstream models cannot exploit signals that are not encoded.

3.3 Polysemy aggravates robustification

The problem of Clever Hans learning is further exacerbated by polysemy. Polysemantic features (cf. [20]) that encode a mixture of biologically meaningful and confounding information make it substantially more difficult for the downstream model to ascertain whether a given relationship is medically meaningful or the result of spurious correlation. Polysemy also presents a major challenge for robustification: if a feature encodes both types of information, removing it to eliminate confounding signals will also discard relevant biological information, thereby reducing the overall utility of the representation. This issue is also discussed in Sup. Note 6.2.

3.4 Illustrative example: unrobust foundation models enable Clever Hans learning

To illustrate this point, Supplementary Table 10 presents a synthetic example resembling downstream model training data with maximum correlation between medical center and biological target. A small dataset is derived from two centers, A and B. The downstream task is to classify tumor status. In the training set, however, the center of origin is perfectly (spuriously) correlated with the tumor label. All samples from center A are normal, and all from center B are tumorous. Thus, the model could learn to associate tumor status with center identity rather than true morphological features. In the test set, this spurious correlation is absent: samples from center B are normal, and those from center A are tumorous.

We compare the possible prediction outcomes of a downstream model trained on non-robust foundation model representations. An unrobust representation space might encode three features: Feature 1 captures biologically relevant information (e.g., tissue morphology) that accurately reflects tumor status. A downstream model relying on this feature learns the correct, generalizable association between input and label, and performs reliably on the test set. In contrast, feature 2 encodes a confounding feature which corresponds to a site-specific artifact (e.g., staining) and is spuriously correlated with the biological target in the training set. A downstream model that relies on this feature performs well during training but fails to generalize, leading to misclassification once the spurious correlation is absent. Feature 3 is polysemantic, encoding both morphological and staining information. On the test set, it partially activates in response to either the tumor status or the medical center artifact. As a result, a downstream model trained on this feature exhibits ambiguous behavior, potentially leading to misclassification and poor generalization.

Crucially, all three features yield identical prediction outcomes on the training set. Consequently, a downstream model has no inherent way of distinguishing which feature encodes biologically meaningful information and which relies on spurious correlations. It may therefore

base its predictions on any of the available features, including those that do not generalize. In contrast, a foundation model with purely biological representations (feature 1) ensures that any downstream model learns the true, medically meaningful relationship between input and target.

3.5 Connection to real-data experiments

In the main paper, we illustrated these issues using real clinical data rather than a synthetic example. In this experiment, the degree of spurious correlation between the biological target and the medical center of origin was systematically varied in the training data. This design allowed us to assess the consequences of spurious correlations and the corresponding risk of Clever Hans learning. Notably, in all splits, the downstream model had both the capacity (evident from its strong performance on the first split) and the necessary information (even for the skewed splits, the training data held information about *all* biological classes) to learn the true association between the representations and the biological target.

However, as the level of spurious correlation between the medical center and the biological target increased, the risk of Clever Hans learning became more pronounced. Although the FMs continued to encode all relevant biological information, the influence of confounding features became increasingly apparent. In these settings, downstream models were more likely to base their predictions on these confounding signals, leading to degraded performance across nearly all FMs. Only the FMs with high representation robustness maintained relatively high accuracy across all splits, which is also reflected in a high correlation between the robustness index and the average performance drop.

At first glance, these results may appear to contradict the high medical center classification accuracies shown in Figure 1d of the Main Paper: even the FMs with the highest robustness index results enabled a reliable prediction of the medical center from their representations. However, a feature space analysis (described in detail in Sup. Note 6.2) clarified this apparent discrepancy. While confounding information is encoded by all foundation models, some FMs tend to encode it more strongly and blatantly (linearly, in directions of largest variance). This makes such features more readily accessible to downstream models, which tend to rely on easily extractable signals [15, 21]. Furthermore, with an increasing level of spurious correlation between the medical center and the biological target, the apparent predictivity of the confounding feature is heightened, ultimately reaching the same alignment with the target as the biological feature. Thus, unrobust FM representations pose a severe risk of Clever Hans learning because they accessibly encode confounding features that have a fallacious predictive value in spuriously correlated datasets. While perfect insensitivity to non-biological features might not be achievable (yet), ensuring — possibly through robustification — that biological features dominate the representation space reduces the risk of Clever Hans learning in downstream models.

3.6 Summary

To summarize, the downstream model Clever Hans learning experiment described in the Main Paper highlighted a fundamental issue: foundation models with non-robust representations increase the likelihood of Clever Hans learning and poor generalization, especially in high-stakes clinical tasks. Reducing or eliminating confounding features from FM representation spaces is therefore essential for maximizing robustness and ensuring reliable performance in downstream applications.

Biological target	Medical center	Feature 1 morphology	Prediction based on f1	Feature 2 staining	Prediction based on f2	Feature 3 polysemantic	Prediction based on f3
Training set							
normal	A	0	✓	0	✓	0	✓
normal	A	0	✓	0	✓	0	✓
normal	A	0	✓	0	✓	0	✓
normal	A	0	✓	0	✓	0	✓
tumor	B	1	✓	1	✓	1	✓
tumor	B	1	✓	1	✓	1	✓
tumor	B	1	✓	1	✓	1	✓
tumor	B	1	✓	1	✓	1	✓
Test set							
normal	A	0	✓	0	✓	0	✓
normal	B	0	✓	1	×	0.5	?
tumor	A	1	✓	0	×	0.5	?
tumor	B	1	✓	1	✓	1	✓

Supplementary Table 10: **Effect of foundation model robustness on downstream model learning with spuriously correlated training data.** Non-robust foundation models (FMs) encode both biological (f1) and technical (confounding) features (f2, f3) in their representations, enabling downstream models to exploit spurious correlations between medical center signatures and prediction targets. FMs without confounding features in their representations enforce that downstream models to learn from biological signals alone.

Supplementary Note 4. Downstream model training

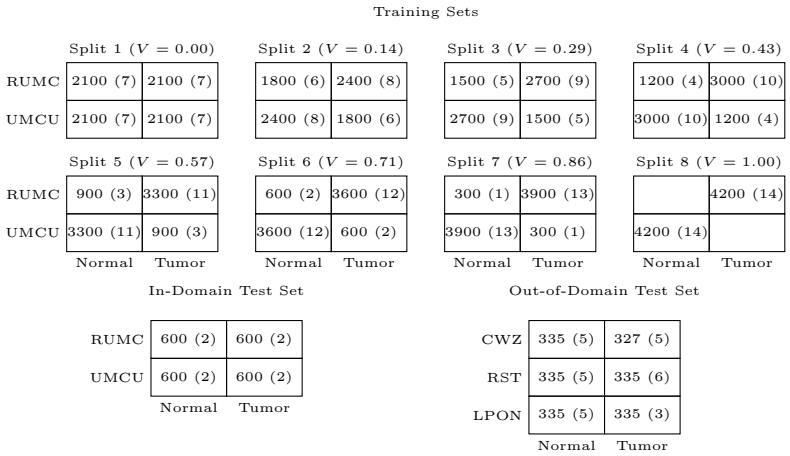
4.1 General training details

We trained various downstream models from FM representations of (small) labeled training data cohorts to evaluate FMs from different perspectives. This is a common way of assessing foundation models, and, in addition, an effective way of deriving special-purpose models for medical applications (see e.g. [22–24]).

As a downstream model, we used linear probing (aka. logistic regression), i.e., training a simple neural network head without hidden layers to predict the classes of interest, which is a common choice in the field [22–24].

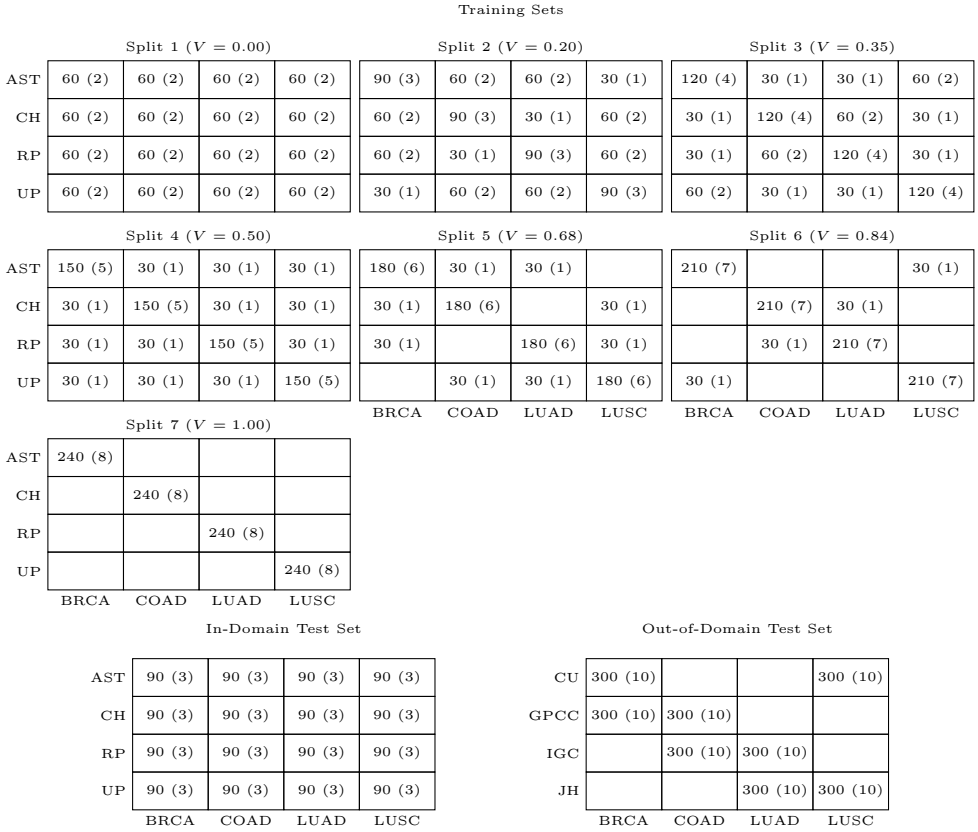
Using the sklearn implementation (`https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`), we selected the optimal regularization parameter C based on validation performance, choosing from 15 logarithmically spaced values in the range $[10^{-8}, 10^{4}]$. To account for variations due to splitting, we repeated all linear probing trainings 20 times by resampling the training, validation, and test data from the respective PathoROB dataset. Notice that the number and composition of samples in training, validation, and test sets always remained constant across repetitions.

4.2 Training with spuriously correlated data



Supplementary Figure 7: **Training and test data compositions, including the number of patches (slides), for the Camelyon downstream task.** For each training split, we report Cramer's V between the medical center and biological class as a measure of correlation between the two variables. Each split's validation set maintains the same correlation and balance as the training set. For each of the four training categories, the validation set is composed of patches coming from one single held-out validation slide, while, in total, rows and columns add up to 300 patches.

For assessing Clever Hans learning in supervised downstream models, we trained linear probing models on training data with increasing spurious correlations between the medical center and the biological class (see Main Paper: Figure 3). The detailed statistics of the training, validation, and test splits are depicted in Supplementary Figures 7, 8, 9. As a validation set, we took a small amount of data, maintaining the same correlation as the training set – further specified in the captions of Supplementary Figures 7, 8, 9.

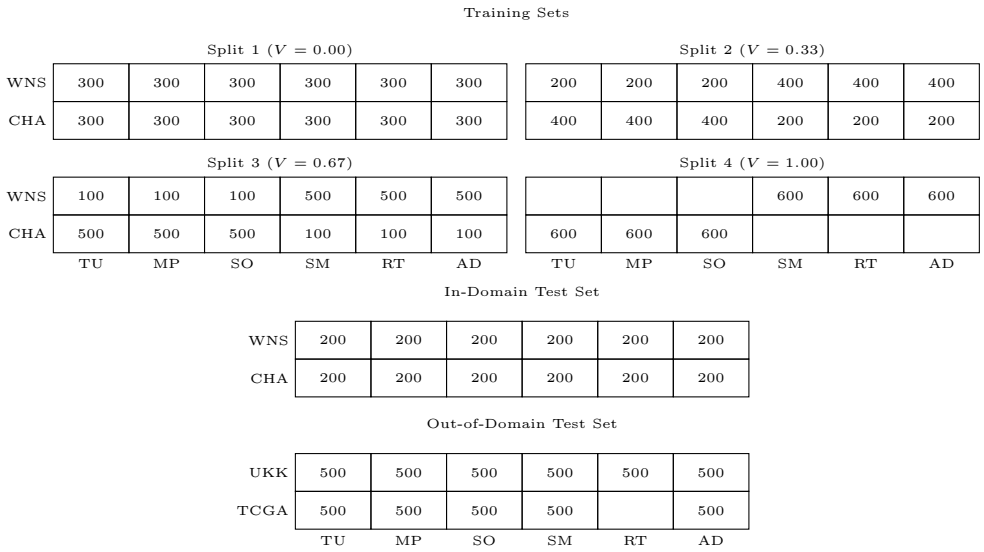


Supplementary Figure 8: **Training and test data compositions, including the number of patches (slides), for the TCGA downstream task based on the TCGA 4x4 dataset.** For each training split, we report Cramer's V between the medical center and biological class as a measure of correlation between the two variables. Each split's validation set maintains the same correlation and balance as the training set. For each of the 16 training categories, the validation set is composed of patches coming from one single held-out validation slide, while, in total, rows and columns add up to 30 patches.

To ensure a fair estimation of the generalization performance drop when increasing correlation, we always used the same data points across training splits (within a single repetition). That is, if a patch was present in one training split, it was also present in every other training split, with at least the same amount of data for the respective class-center combination. To estimate the stability of the experiment across repetitions, we repeated the split sampling procedure 20 times. That is, we randomly filled the numbers depicted in (Supplementary Figures 7, 8, 9) with different patches sampled from the PathoROB datasets.

4.3 Slide-level downstream model training

The split statistics for the slide-level downstream tasks are presented in Supplementary Figure 10. Note that for the OOD test sets in Camelyon, we only used the slides with pixel-level annotations, i.e., 20 slides per medical center. For the TCGA datasets, we only considered diagnostic slides from the primary tumor (with barcode suffix DX). We extracted patches of 256×256 pixels without overlap at 20x magnification (0.5 microns per pixel). We identified and excluded background patches via Otsu's method [4] on slide thumbnails and applied a patch-level minimum standard



Supplementary Figure 9: **Training and test data compositions, including the number of patches, for the Tolkach ESCA downstream task.** For each training split, we report Cramer's V between the medical center and biological class as a measure of correlation between the two variables. Each split's validation set maintains the same correlation and balance as the training set. For each of the 12 training categories, the validation set is composed of held-out validation patches, while, in total, the columns add up to 100 patches.

deviation of 8. Due to the high number of patches (e.g., over 12 million patches in TCGA-BRCA), we restricted the feature extraction to six foundation models (Virchow2, CONCHv1.5, H0-mini, UNI2-h, CTransPath, Phikon-v2) that represent different training paradigms and levels of robustness. Furthermore, we only extracted CLS token representations for the Transformer-based FMs in the TCGA datasets.

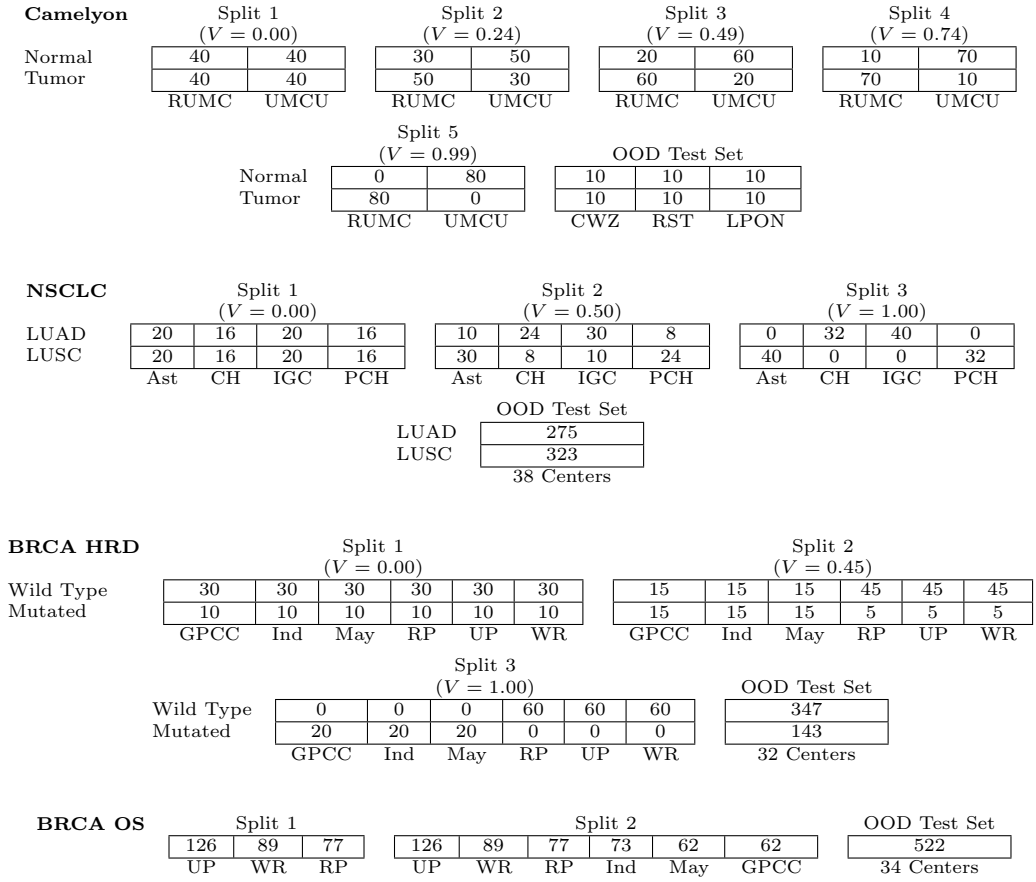
For training the Attention MIL models [25], we used a batch size of 32, a learning rate of 0.002, weight decay of 0.0 or 0.01, and trained for 200 epochs with stochastic gradient descent. The best checkpoint during training and the best weight decay values were chosen according to validation set performance.

4.4 Domain-adversarial training

We assessed domain-adversarial neural network (DANN) [26] training as a method for robustifying FM representation spaces. Recall from the main text that the DANN objective can be defined through two competing loss parts:

$$\mathcal{L}_{\mathsf{DANN}}(\hat{y}, \hat{c}; y, c) = \mathcal{L}_{\mathsf{CL}}(\hat{y}, y) + \lambda \cdot \mathcal{L}_{\mathsf{DA}}(\hat{c}, c)\,, \tag{2}$$

where the domain discriminator part $\mathcal{L}_{\mathsf{DA}}$ makes use of a gradient reversal layer [26]. While originally developed in the context of domain adaptation (cf. [26]), we revisit it here for post-hoc removing medical center signatures from FMs. The technical difference from the latter is that we do not consider unlabeled domains; instead, we assume complete access to the biological class and confounding information for all domains. This means our goal is not to propagate label information to an unsupervised domain to perform classification, but instead, we aim to learn a feature space that is free from non-biological features. To summarize, our approach consists of two steps: We first try to learn an unbiased feature representation of the original training data



Supplementary Figure 10: **Training and test data compositions for the slide-level downstream tasks.** For each training split, we report Cramer's V between the medical center and biological class as a measure of correlation between the two variables. Each split's validation set maintains the same correlation and balance as the training set. OS = overall survival; OOD = out-of-domain.

using DANN. Then, we trained a new linear probe (identical to the one used in the experiments without mitigation) on the training data embedded in the new feature space.

For the DANN architecture, we found that a one-layer feature extractor with ReLU, which reduces the original feature space by half, combined with a linear classification head and a shallow non-linear domain discriminator, results in the best overall model performances. The idea of halving the original feature space is to compress the downstream accessible information, hoping to discard confounding signals in favor of biological signals. We trained DANN for 20 epochs and chose the loss-balancing weight λ to increase linearly from zero to approach one over the epochs (i.e. in each epoch we slightly increase the impact of the domain discriminator loss while the weight for the classification loss term remains the same).

Supplementary Note 5. Clustering analysis

5.1 Clustering score

The clustering score evaluates the robustness of the global organization within the representation space. Unlike the robustness index, which assesses the composition of each embedding's nearest neighbors and, thus, provides a local perspective, the clustering score assesses the composition of entire groups of embeddings, offering a global view. Similar to the robustness index, it quantifies the extent to which the spatial arrangement of the representations is influenced by biologically meaningful factors as opposed to confounding variables.

To examine the global structure, we performed unsupervised K-means clustering using cosine distances. The number of clusters K was not predefined but selected by maximizing the silhouette score [27]. The silhouette score quantifies cluster cohesion (within cluster distances) compared to cluster separation (distances to closest, distinct cluster), thereby serving as a proxy for the cluster quality in the absence of ground-truth labels. Let $\hat{\mathcal{C}} = \{\hat{C}_1, \ldots, \hat{C}_K\}$ denote the predicted partition (clustering) of a data set $\mathcal{D}$ such that $\bigcup_k \hat{C}_k = \mathcal{D}$, and $d(i,j)$ is the distance between samples $i \in \hat{C}_I$ and $j \in \hat{C}_J$, then

$$\text{silhouette score}(\hat{\mathcal{C}}) = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} s(i), \quad \overbrace{s(i) = \frac{a(i)-b(i)}{\max\{a(i),b(i)\}}}^{\text{silhouette of sample } i}, \tag{3}$$

$$\underbrace{a(i) = \frac{1}{|\hat{C}_I|-1}\sum_{i\neq i'\in\hat{C}_I} d(i,i')}_{\text{average inter-cluster distances}}, \quad \underbrace{b(i) = \min_{\hat{C}_I\neq\hat{C}_J\subseteq\hat{\mathcal{C}}} \frac{1}{|\hat{C}_J|}\sum_{j\in\hat{C}_J} d(i,j)}_{\text{average intra-cluster distances}}. \tag{4}$$

This fully unsupervised setup deliberately reflects realistic conditions in clinical knowledge discovery and novelty detection, where prior cluster information is typically unavailable. See Sup. Note 7.6 for an ablation study that eliminates the effect of the silhouette score-based selection of K on the clustering performance.

The quality of the clustering is evaluated by comparing the predicted cluster assignments $\hat{\mathcal{C}}$ to the known biological $\mathcal{C}_{\text{bio}}$ and confounding labels $\mathcal{C}_{\text{mc}}$ using the adjusted Rand index (ARI) [28]. The ARI measures the agreement between two clusterings mainly by the number of pairs of samples correctly clustered together or apart, accounting for chance. It is symmetric, invariant to label permutations, and bounded in $[-0.5, 1]$, with 1 indicating perfect agreement, 0 random labeling, and negative values suggesting poor, independent alignment:

$$\text{ARI}(\hat{\mathcal{C}}, \mathcal{C}) = \frac{\text{RI} - \mathbb{E}(\text{RI})}{\max(\text{RI}) - \mathbb{E}(\text{RI})}, \quad \underbrace{\text{RI} = \frac{n+m}{|\mathcal{D}|(|\mathcal{D}|-1)/2}}_{\text{Rand index}} \tag{5}$$

n : number of pairs (i,j) that are in the same cluster in $\hat{\mathcal{C}}$ and in the same cluster in $\mathcal{C}$

m : number of pairs (i,j) that are in different clusters in $\hat{\mathcal{C}}$ and in different clusters in $\mathcal{C}$

$|\mathcal{D}|(|\mathcal{D}|-1)/2$: total number of pairs (i,j)

$\mathbb{E}(\text{RI})$: expectation under random cluster assignments.

To evaluate the influence of confounding factors, we compute the ARI twice: once comparing the predicted clusters to the biological class labels (e.g., diagnosis, tissue type), and once to

the confounding labels (e.g., medical center origin). The clustering score is then defined as the difference between these two ARI values

$$\text{clustering score} = \overbrace{\text{ARI}(\hat{\mathcal{C}}, \mathcal{C}_{\text{bio}})}^{\text{higher is better}} - \overbrace{\text{ARI}(\hat{\mathcal{C}}, \mathcal{C}_{\text{mc}})}^{\text{lower is better}} \tag{6}$$

The clustering score inherits ARI's desirable properties — chance adjustment, symmetry, and boundedness (in $[-1.5, 1.5]$). It ranges from approximately $[-1, 1]$, where values near zero suggest the clustering is influenced by both biological and confounding information (or neither), positive values indicate clustering aligned with medically relevant features, and negative values reflect confounding-driven structure.

5.2 Implementation details

To ensure comparability between $\text{ARI}(\hat{\mathcal{C}}, \mathcal{C}_{\text{bio}})$ and $\text{ARI}(\hat{\mathcal{C}}, \mathcal{C}_{\text{mc}})$, an equal number of biological and confounding classes was selected, with a balanced number of representations per class. Specifically, the clustering experiments were conducted on the paired settings Camelyon 2x2, Tolkach 2x2, and all 2x2 class pairings from TCGA 4x4 (for reducing the complexity). Prior to clustering, all embeddings were normalized to unit ℓ_2-norm, so that clustering with Euclidean distances effectively corresponded to clustering based on cosine distances. We chose the optimal number of clusters K among $[2, 30]$ by maximizing the silhouette score and finally quantifying the quality of the resulting K-means clustering by the proposed clustering score.

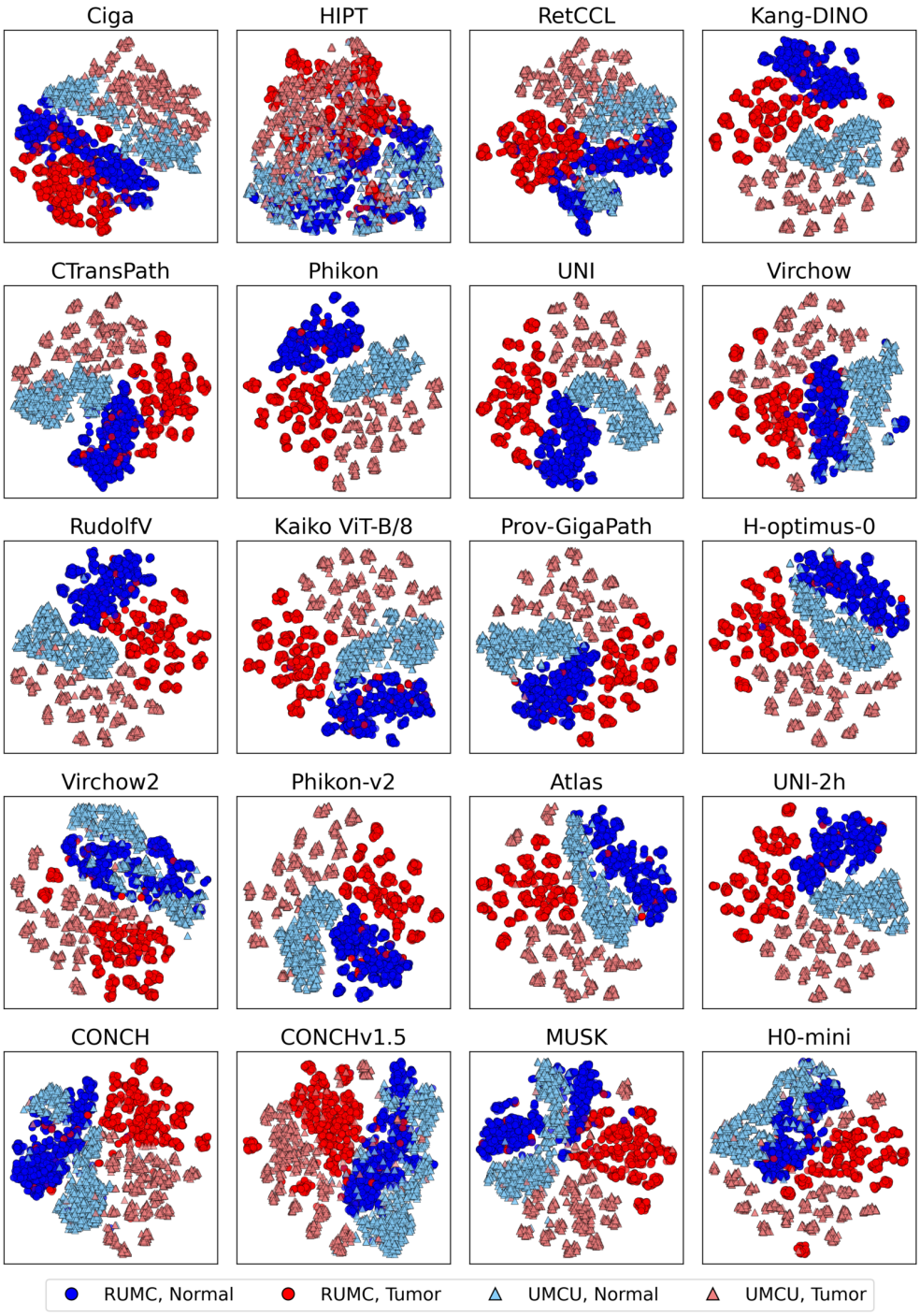
Because K-means clustering is sensitive to the random cluster initialization, the best solution among consecutive random runs is generally selected based on the lowest inertia. For finding K, the clustering is repeated for 20 and the final clustering only for 5 random initializations. Additionally, the estimation of an average clustering score and its standard deviation was enabled by repeating the final clustering for 50 trials (with each 5 initializations) with different random seeds. Thereby, the dependency of the K-means clustering on its random cluster initialization is diminished, which is reflected by very low standard deviations of the clustering score.

Supplementary Note 6. Analysis of the FM representation space

6.1 FM representation space visualizations

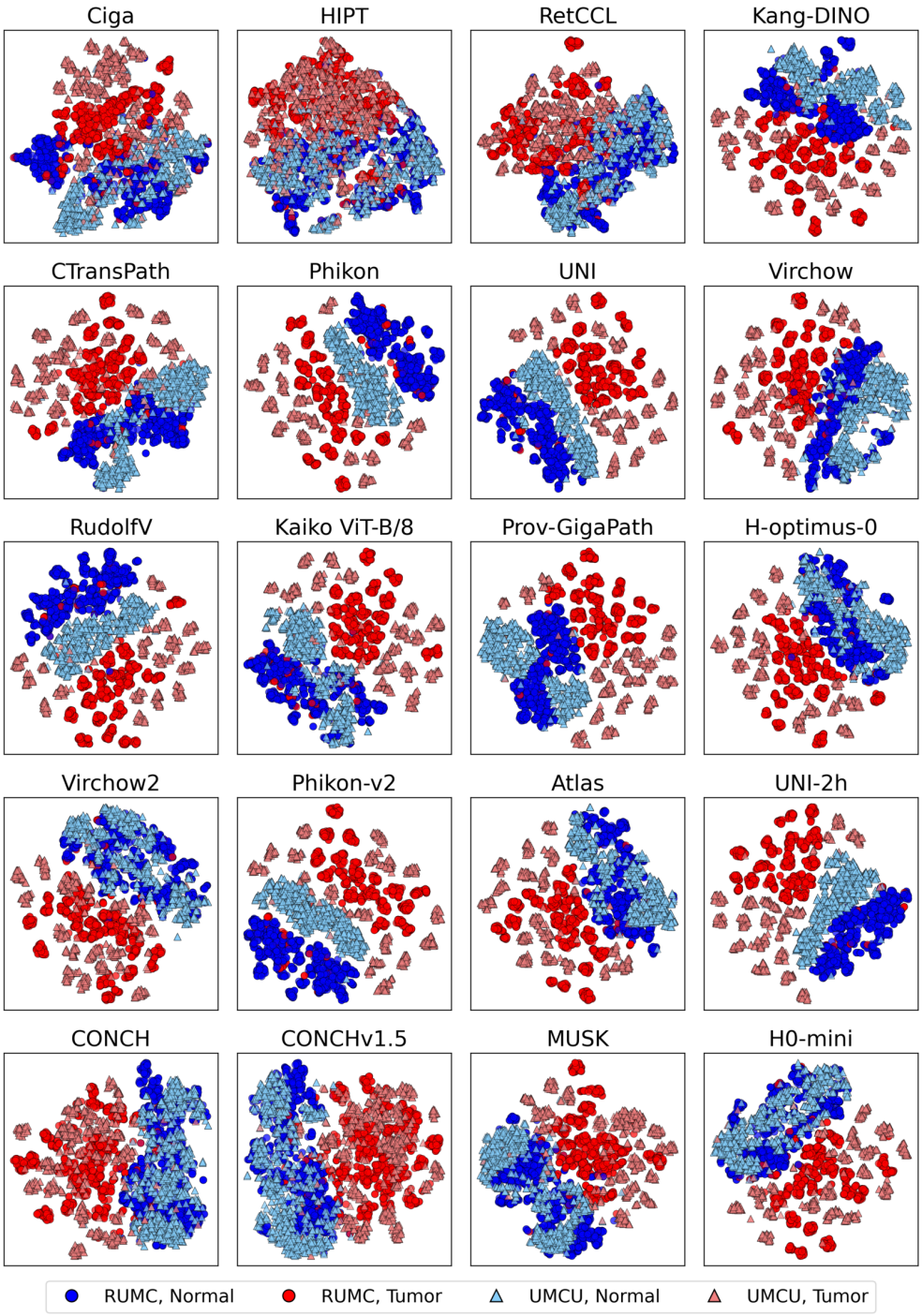
Supplementary Figure 11 shows additional t-SNE plots computed on Camelyon representation vectors of all foundation models. The underlying patches were sampled from the RUMC and UMCU medical centers. For the computation of the t-SNE, we chose the perplexity as the optimal k value derived during the robustness index calculation (see Sup. Note2). Notably, in older and less robust models (e.g. Ciga, RetCCL), tumor and normal patches tend to cluster more closely within individual medical centers. In contrast, newer and more robust models like H0-mini and CONCHv1.5 show more overlap of centers within the same biological subclass. However, most models exhibit a recognizable clustering pattern corresponding to the four cancer/normal $\times$ RUMC/UMCU subclasses, suggesting that the learned feature spaces are at least partially structured by medical center information at some hierarchical level.

In comparison, Supplementary Figure 12 visualizes the embedding spaces after robustification using ComBat batch correction (RR) [29, 30]. It appears that robustification successfully leads to stronger overlap between embeddings of the same morphology, especially if there was already correct organization based on the tumor status prior to robustification, e.g., see Virchow2, CONCH variants, H0-mini. In some cases, robustification altered the organization such that it now reflects the biology rather than the medical center origin, e.g., RetCCL, UNI variants, Ciga,



Supplementary Figure 11: **t-SNE visualization of the representation spaces per foundation model on Camelyon.** Groups of embeddings are separated based on their medical center origin (RUMC: dark circles / UMCU: bright triangles) and tumor status (blue / red).

CTransPath, Kaiko ViT-B/8, and so forth. However, ComBat also caused stronger overlaps between groups of embeddings of different medical status, e.g., the CONCH variants, RudolfV,



Supplementary Figure 12: **t-SNE visualization of the representation spaces after robustification by ComBat (RR) per foundation model on Camelyon.** Groups of embeddings are mostly separated based on their tumor status (blue / red) while overlapping by their medical center origin (RUMC: dark circles / UMCU: bright triangles).

Kaiko ViT-B/8, RetCCL.

While t-SNE captures and visualizes local neighborhood structures, longer distances are not properly reflected. Due to the small choice of the perplexity, an embedding is only represented in terms of its closest neighbors, most likely corresponding to patches of the same slide. Relations between sub-groups, e.g., different WSIs, are not properly represented. Explaining the calculated robustness indices or clustering scores in terms of the t-SNE representations is limited. Yet, some trends can be observed. Overlaps between normal and tumorous embeddings have a negative impact on the robustness index and the clustering score. Non-isotropic clusters, e.g., a banana-shaped group of tumors around the normal embeddings for UNI, violate the assumption of spherical-shaped, isotropic clusters made by the K-means clustering.

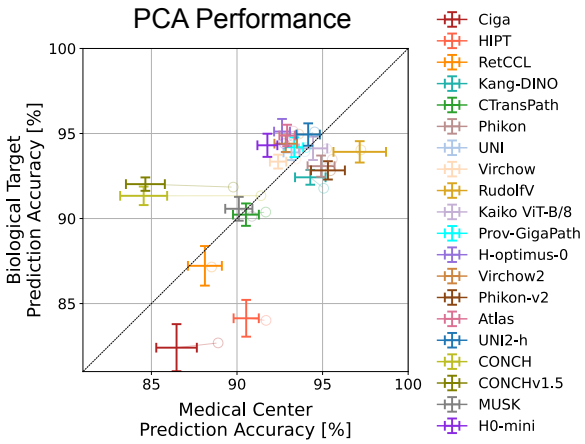
6.2 Analysis of FM representation spaces with PCA

To better understand how medical center and biologically relevant information are encoded in the representation space, we performed multiple analyses based on Principal Component Analysis (PCA). PCA reduces the embeddings to linear directions of greatest variance (principal components, PCs), i.e., to those directions in which the embeddings vary the most.

First, we predicted the medical center and biological class from the reduced representations. We used the same linear probing model training setup as described in Sup. Note 4 but on the PCA-reduced embeddings derived using the sklearn implementation (`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`). We considered representations projected to their first 10% of principal components. The results are shown in Supplementary Figure 13. The biological target and the medical center classification performance using only the top 10% of the principal components remained comparable to that using the full-dimensional embeddings, implying that medically relevant and irrelevant medical center information are both strongly encoded in the linear directions of the representation space, and are readily available and accessible by downstream models. Given that downstream models tend to exploit easily accessible linear features (as a Clever Hans feature or shortcut), even if less predictive [15, 21], this explains both the strong supervised downstream model performance on balanced data and their vulnerability to spurious correlations (cf. Main Paper: Figure 3, see also Sup. Note 3 for a thorough discussion).

Second, we analyzed the class-separability of the biological class or the medical center within the individual PCs by measuring AUROC scores for the biological target and medical center classification. This is to assess how these signals are encoded in the embedding space, how prevalent they are in the most important directions, and whether they are polysemantic.

Specifically, we first projected the embeddings on each PC individually, which gave one-dimensional representations. Then, we analyzed how well these one-dimensional representations can be linearly separated with respect to either their medical center of origin or their biological class. For this, we measured their separability via a multi-class one-vs-one area under the receiver operating characteristic curve (AUROC) using the one-dimensional projections as scores and the medical center or biological classes as labels. To ensure that biological and medical center AUROCs are comparable, we used the paired datasets Camelyon 2x2, TCGA 4x4, and a reduced Tolkach 2x2, with the medical centers WNS and CHA and the biological classes SH_OES and SH_MAG. The AUROCs range from 0.5, indicating no separation is possible, to 1, with perfect separability (PC fully explains the class separability and target). They naturally handle class imbalances, guarantee comparability across models as they are always in $[0.5, 1]$, and are nonparametric since it does not rely on a particular choice of the classification algorithm. PCs for which both signals have an AUROC exceeding 0.6, which indicates that both signals are linearly pronounced, are considered polysemantic (cf. [20]). Polysemantic features encode both biological and confounding information. The depth of the encoding, i.e., the number of PCs for which a significant AUROC > 0.6 is observed, reflects the complexity of the data representation in the embedding space (loosely comparable to the dimensionality of the subspace that encodes



Supplementary Figure 13: **Classification accuracy for predicting the medical center and biological class before and after reducing the dimensionality of the representations using principal component analysis (PCA).** The representation spaces were reduced to their first 10% of principal components with PCA applied individually per dataset and foundation model. We report mean prediction accuracies with 95% confidence intervals over 20 repetitions and three held-out test sets (Camelyon, TCGA 4x4, Tolkach ESCA) (n=60), corrected to remove common variance due to dataset (Masson & Loftus [31]). Only minor differences in accuracy were observed, indicating that medically relevant and medical center information is strongly encoded in the representation space. Source data are provided as a Source Data file.

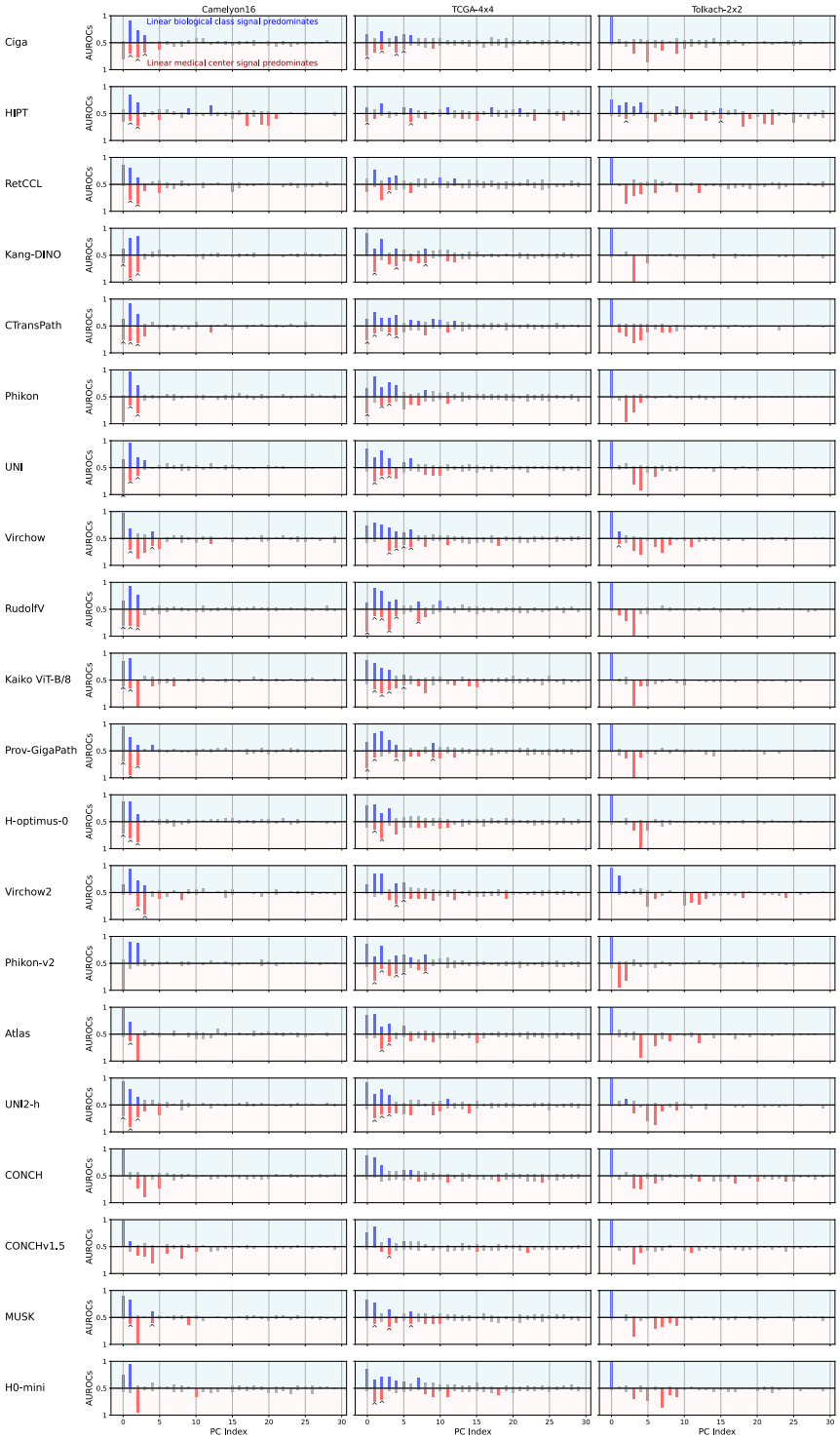
relevant information).

The results are shown in Supplementary Figure 14. All models encoded both biological and medical center information in their early PCs, with the highest separability often occurring within the first 5 to 10 PCs — sometimes even approaching perfect separation. Thus, the largest variations in the embeddings align with variations in the biological or medical center label. However, the degree of prominence and polysemy of these signals varied across models and datasets. For Tolkach ESCA, there is a strong encoding of the biological signal in the very first PC and almost no polysemy. This indicates that distances between patches in the embedding space mostly reflect biologically relevant similarities. This matches the good robustness indices and high clustering scores (cf. Figure 6b in Main Paper). In contrast, for TCGA, both signals are encoded deeply, and principal directions are polysemantic. This provides an explanation for the low robustness indices and clustering scores, as the data has complex embedding representations that reflect medically relevant and confounding information equally. For Camelyon, we observe a strong encoding in the first PCs and, depending on the FM, some polysemy. In general, the representation spaces of older models and RudolfV include polysemantic directions, in contrast to CONCH, CONCHv1.5, and H0-mini whose principal directions mostly encode either biological or center information. Due to the polysemy, selectively isolating biological signals without retaining confounding information might become difficult. Accordingly, robustification of FMs with polysemantic directions is precarious; removing these features to eliminate confounding signals will also discard relevant biological information, which reduces the utility of the representation.

Similarly, as the directions of greatest variance hold both relevant and irrelevant information, pairwise distances are affected by both and reflect differences in biology and artifacts. This explains the clustering performance and robustness indices observed in Main Paper: Figures 2 and 6, which verified that pairwise distances in the representation space reflect not only morphological differences of the patches, but also variations in their medical center origin.

While the previously described approaches already illuminated the structure of representation spaces, advanced Explainable AI (XAI) techniques might deliver deeper insights. Concept-

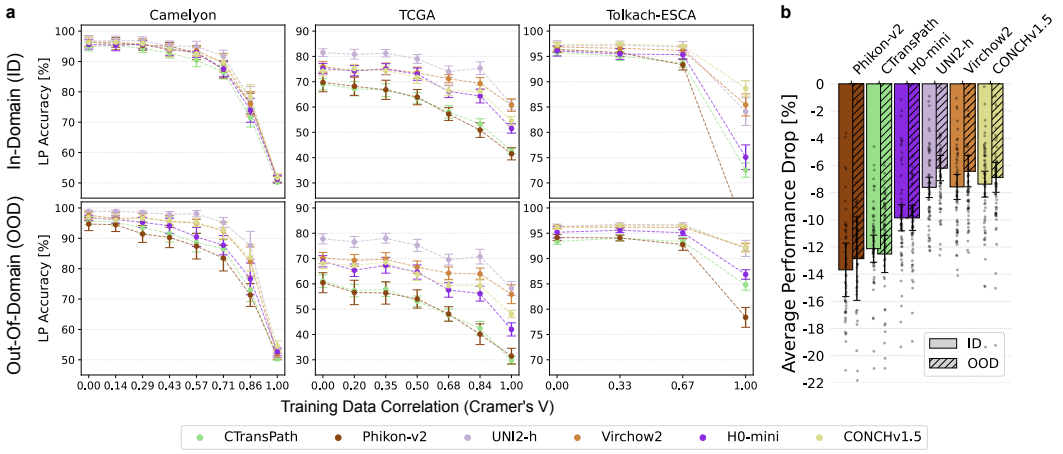
based XAI (e.g. [32, 33]), for instance, may reveal human-understandable concepts encoded by the latent features within these spaces. When applied to data shift analysis (e.g. [34]), such methods can uncover directions along which distributional differences are most pronounced, enabling assessment of whether they reflect meaningful variation or spurious signals—such as class separation driven by confounders like medical center bias. These techniques thus support robustness analysis by revealing whether observed class differences and learned representations are shaped by genuine biological signals or underlying biases.



Supplementary Figure 14: **Biological and medical center separability per principal component (PC)**. Separability is measured by pairwise AUROCs, with 1 indicating strong separability. For many foundation models (FMs), both signals were highly separable and polysemantic (corresponding PCs with both AUROCs > 0.6 are marked ($\hat{}$)). Source data are provided as a Source Data file.

Supplementary Note 7. Further downstream experiment results

7.1 Full finetuning of foundation models on spuriously correlated data



Supplementary Figure 15: **Finetuning model performance under spuriously correlated training data**. **a** In-/out-of-domain generalization performance on data from seen/unseen hospitals. Mean finetuning accuracies (center dots) are reported with 95% confidence intervals (error bars) over n=20 repetitions, each with randomly resampled but overlapping training data. **b** In-/out-of-domain average generalization performance drops per model averaged over all datasets and repetitions (n=60) with 95% confidence intervals corrected to remove common variance due to dataset (Masson & Loftus [31]). Source data are provided as a Source Data file.

To more thoroughly assess the foundation models' generalization capabilities under spuriously correlated training data, we repeated our patch-level downstream experiments (see Main Paper: Figure 3) with foundation model finetuning instead of using the simple linear-probe (LP/LR) heads trained on static feature embeddings.

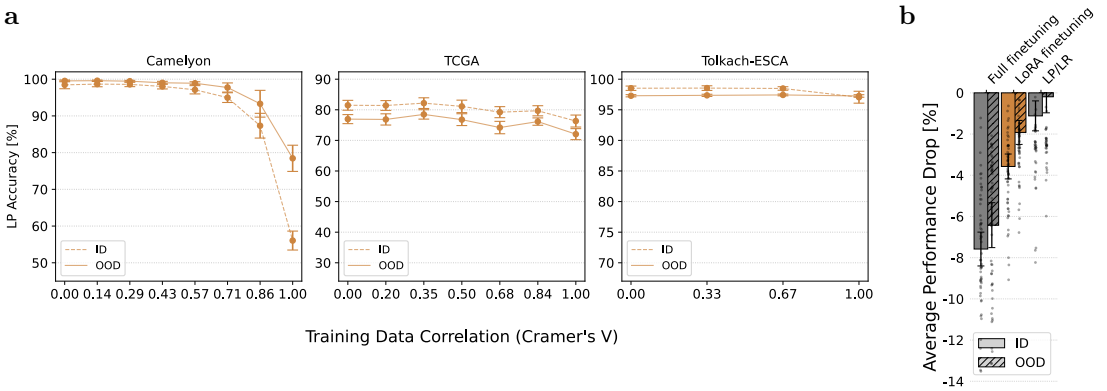
For finetuning, i.e., the supervised alignment of pretrained foundation model weights to the prediction task, we added a simple one-layer head to the models and unfroze all the model weights for training. To avoid large gradients at the start of the training, we zero-initialized the head. We further used a linear warm-up of the base learning rate $5 \cdot 10^{-5}$ for the first 10% of total training steps followed by a cosine decay. We trained for 10 epochs with a batch size of 32, using Cross-Entropy Loss and the AdamW optimizer with a weight decay of 10^{-4}, and selected the checkpoint with the highest balanced accuracy on the validation set. Finetuning a foundation model for a given dataset–iteration–split combination required up to three hours on a single A100 80GB GPU. We did not consider additional regularization, more sophisticated training strategies, or evaluations for more FMs for computational reasons. We also restricted our analysis to six FMs representing different training paradigms and robustness levels.

The generalization performance drops observed for spuriously correlated training data (Supplementary Figure 15a) were much more pronounced than in the LR experiments (see Main Paper Figure 3b). Especially for the Camelyon task, the accuracies collapsed to 50% for both ID- and OOD-settings across all FMs for fully correlated training data (Cramér's V of 1). Therefore, the resulting APDs (Supplementary Figure 15b) also exceed the LR APDs (see Main Paper Figure 3c) across FMs. In addition, FMs with better robustness index results (CONCHv1.5, Virchow2, and UNI2-h) also achieved better APD scores than the unrobust models, such as

Phikon-v2 and CTransPath. Yet, FM comparisons are more difficult in this setting due to computational costs, overall poorer performance, and less informative APD differences (especially among the models with the best APD results – CONCHv1.5, Virchow2, and UNI2-h).

We note that the overall generalization performance levels for finetuning were also slightly lower than with linear probing when using balanced training data (Cramér's V = 0), indicating that finetuned models may overfit more to other features, possibly not captured by medical center differences. Yet, the finetuned models still performed competitively in all three datasets, suggesting that such overfitting was not the key driver of the strong performance drops observed for imbalanced data (Cramér's V > 0), but Clever Hans learning of non-biological features.

7.2 Low-rank adaptation of foundation models on spuriously correlated data



Supplementary Figure 16: **LoRA-finetuned Virchow2 performance under spuriously correlated training data**. **a** In-/out-of-domain generalization performance on data from seen/unseen hospitals. Mean LoRA finetuning accuracies (center dots) are reported with 95% confidence intervals (error bars) over n=20 repetitions, each with randomly resampled but overlapping training data. **b** In-/out-of-domain average generalization performance drops per training paradigm over all datasets and repetitions averaged over all datasets and repetitions (n=60) with 95% confidence intervals corrected to remove common variance due to dataset (Masson & Loftus [31]).

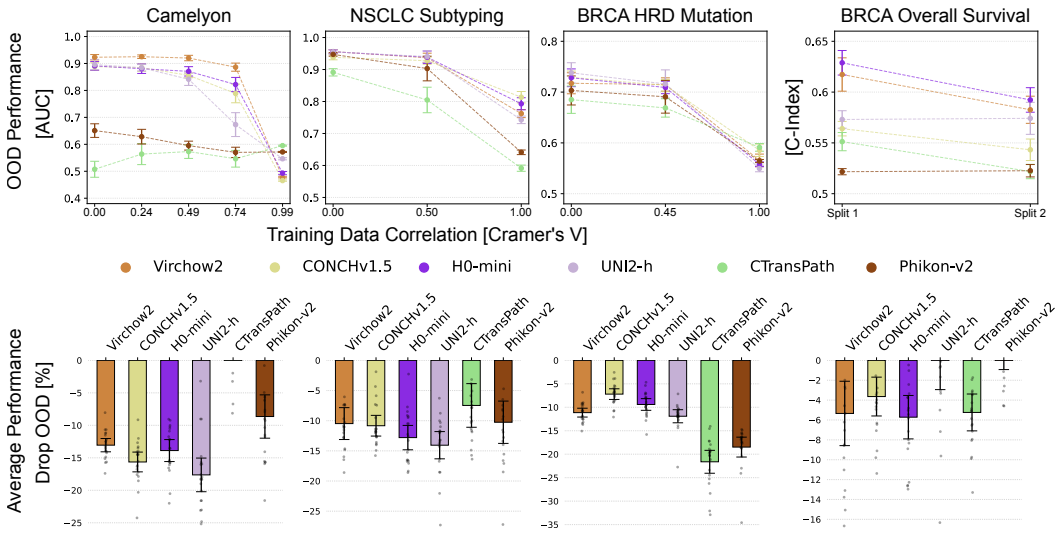
As a middle ground between full finetuning and linear probing, we also considered parameter-efficient finetuning using low-rank adaptation (LoRA) [35].

For LoRA finetuning, we added a simple one-layer head to the Virchow2 model and froze all other model weights for training. We then applied LoRA to QKV and projection modules, within each attention block. We chose $r = 8$ with LoRA-$\alpha = 16$ and a LoRA-dropout of 0.1. We normal-initialized the head and used a linear warm-up of the base learning rate $5 \cdot 10^{-4}$ for the first 10% of total training steps followed by a cosine decay. We trained for 10 epochs with a batch size of 32, using Cross-Entropy Loss and the AdamW optimizer with a weight decay of 10^{-4} for the head, and selected the checkpoint with the highest balanced accuracy on the validation set.

The generalization performance drops observed for spuriously correlated training data (Supplementary Figure 16a) were less pronounced than in the full finetuning setting, yet more substantial than in the linear probing experiments. For both the TCGA and Tolkach-ESCA tasks, the per-split mean accuracies were comparable to those obtained with linear probing, typically lying within confidence intervals across nearly all splits or falling only marginally below them. In contrast, for the Camelyon task, the accuracies collapsed to 56% (ID) and 78% (OOD) for

fully correlated training data (Cramér's V of 1), remaining superior to full finetuning but clearly inferior to linear probing (≈ 87% for ID and ≈ 98% for OOD). Consequently, the resulting APDs (Supplementary Figure 16b) fall between those of linear probing and full finetuning, indicating intermediate robustness to Clever Hans learning.

7.3 Slide-level downstream model comparison



Supplementary Figure 17: **Slide-level downstream model performances across clinically relevant prediction tasks.** Out-of-domain average performance drops per foundation model and dataset for comparison, aggregated over all repetitions (n=20) with 95% confidence intervals. Source data are provided as a Source Data file.

Comparing the average performance drops per FM showed that the differences between the best-performing models (Virchow2, CONCHv1.5, H0-mini, UNI2-h) were small on the slide-level tasks (Supplementary Figure 17). The models with a lower robustness index, i.e., CTransPath and Phikon-v2, escaped the comparison in some cases, as low initial performance on balanced data (Cramér's V of 0) limited the magnitude of the performance drop. Generally, FM comparisons were more difficult in the slide-level than in the patch-level tasks, due to higher error margins (in absolute numbers) driven by the small training and validation dataset sizes and more degrees of freedom in the Attention-based MIL architecture. This suggests that for benchmarking purposes, patch-level tasks with lower variance and higher data availability may be more suitable.

7.4 Patch retrieval from heterogeneous databases

We further assessed the efficacy of unsupervised patch retrieval in unrobust feature spaces: given a patch of interest, retrieve a biologically similar patch from a database using the foundation model representations. This clinically relevant setting has recently gained attention [36–38]. Although retrieval is more commonly performed on the slide level, algorithms are regularly based on patch-level distances.

Identical to the supervised downstream model experiments (see Main Paper: Figure 3), we sampled patch databases with increasing correlation between the medical center and biological class, and evaluated whether patches retrieved from the database had the same biological class as the query patches. For three PathoROB datasets (Camelyon, TCGA 4x4, Tolkach ESCA), we considered splits with increasing levels of correlation between medical centers and biological

targets (see Supplementary Figures 7, 8, 9). As a retrieval database, we used the FM embeddings of all patches of the respective training split. For each query patch from the test set (both in-domain and out-of-domain), we retrieved the most similar patch from the database based on the cosine distance between the two embedding vectors. We then assessed whether the retrieved patch belonged to the same biological class as the query patch, resulting in an accuracy score for the test set (= the set of query patches). From these accuracies, we derived the average performance drop measure and correlated it with the robustness index (see Main Paper: Methods). Notice that in practice, the patch retrieval experiments correspond to exchanging the linear probing downstream model for a k-nearest neighbors model with $k = 1$.

We observed retrieval performance drops when retrieval databases were heterogeneous across medical centers (Supplementary Figure 18), meaning that retrieval is prone to errors if the retrieval database cannot be guaranteed to have balanced data contributions from all medical centers. For Camelyon, for example, the in-domain retrieval performance dropped from 85%-98% to 55%-92%. For TCGA and Tolkach ESCA, the decreases were less steep, but still present for all foundation models. On out-of-domain data, the best foundation models (Virchow, H-optimus-0, Virchow2, CONCH, CONCHv1.5, MUSK) enabled stable performances on imbalanced retrieval databases, whereas other models also suffered from diminishing retrieval accuracy. At the same time, the magnitude of the average performance drop was strongly correlated with the robustness of the foundation model ($\rho = 0.925$ for ID; $\rho = 0.791$ for OOD), and more robust FM representation spaces led to more stable retrieval results.

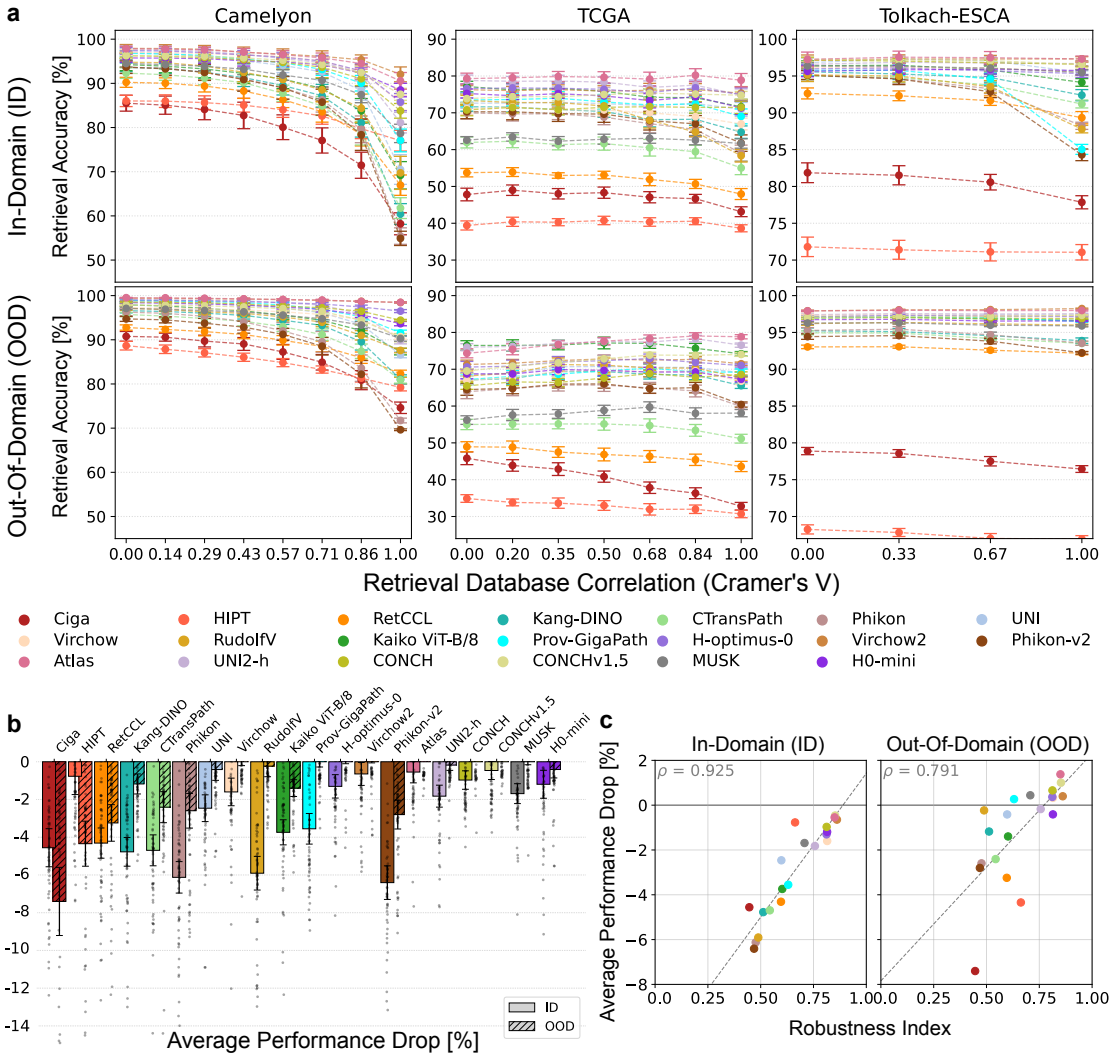
7.5 Downstream model experiments on robustified FM representations

We reran the experiments to assess Clever Hans learning of downstream models under heterogeneous training data (Main Paper: Figure 7c) on the robustified embedding spaces. Supplementary Figure 19 provides further insights on how robustification methods impact downstream models and Clever Hans learning [13–18].

ComBat [29, 39] showed in-domain generalization performance improvements on training data with low or moderate spurious correlations. It even mostly outperformed Reinhard and DANN up until Split 6 ($V = 0.71$) out of 8 (Camelyon) and Split 4 ($V = 0.5$) out of 7 (TCGA). This is in line with previously reported results [39]. Nevertheless, the method exacerbated generalization performance drops for fully correlated data in two datasets, specifically when the medical centers did not contribute any data for at least one biological class. In Camelyon, for example, all FMs drop below 60% performance, which is considerably worse than without any robustification (Base setting). This suggests that, in this case, ComBat also removes relevant biological information. Consequently, ComBat may be a promising robustification option if training data imbalances are guaranteed to be limited, and test data are known to come from the same medical centers as the training data.

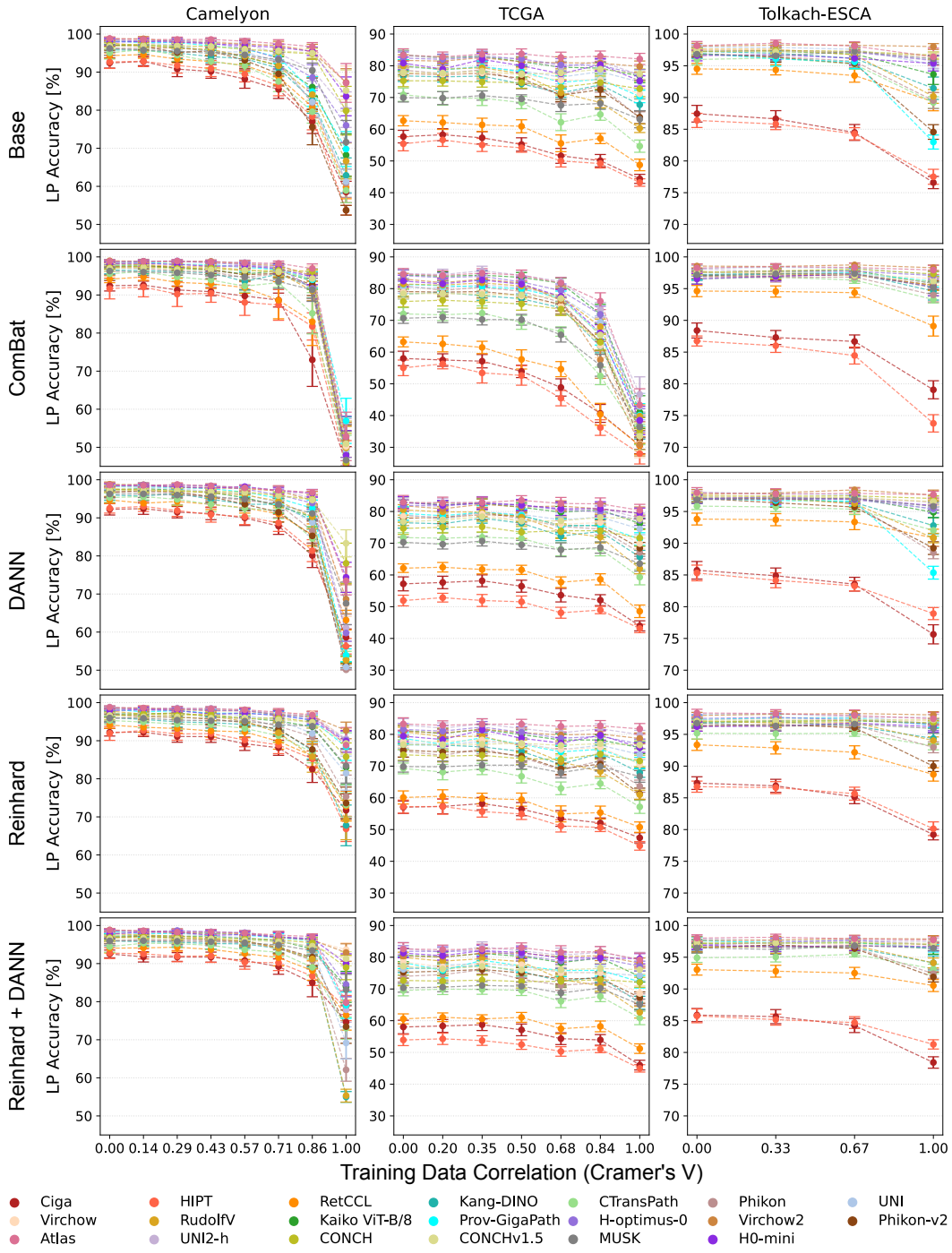
The DANN method [26] also led to better generalization performance for training data with low or moderate spurious correlations. On fully correlated training data ($V = 1$), though, it did not yield improvements over the baseline setting without embedding space robustification. This is not surprising, as the medical center-based domain loss does not provide complementary information to the biological class prediction loss in this case. Therefore, using DANN alone is arguably not advisable when some medical centers did not contribute any data for some biological classes.

In contrast, Reinhard stain normalization [40] strongly improved the accuracy under fully correlated training data ($V = 1$) in many cases (although not all cases). For Camelyon, in particular, the linear probing accuracies in the last split were between 67%-93% — substantially higher than for the baseline setting with 53%-87% average accuracy. At the same time, the generalization performances on splits with low to moderate correlations did not always improve compared to Base, indicating that stain normalization may also remove some biologically relevant information. Therefore, it is arguably most effective in extremely unbalanced scenarios.

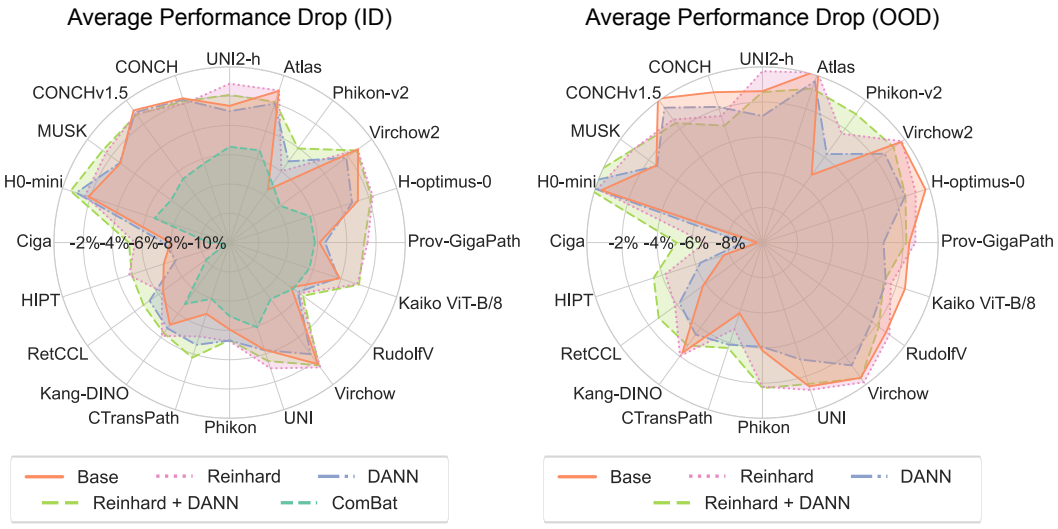


Supplementary Figure 18: **Retrieval performances from retrieval databases with heterogeneous data contributions across medical centers**. **a** In-/out-of-domain retrieval performance on data from seen/unseen hospitals. Mean retrieval accuracies (center dots) are reported with 95% confidence intervals (error bars) over n=20 repetitions, each with randomly resampled but overlapping retrieval databases. **b** In-/out-of-domain average retrieval performance drops per model averaged over all datasets and repetitions (n=60) with 95% confidence intervals corrected to remove common variance due to dataset (Masson & Loftus [31]). **c** Spearman rank correlation between the robustness index averages and the in-/out-of-domain average retrieval performance drops of the models (Spearman rank correlation (n=20); two-sided permutation test; ID: $\rho = 0.925$, $p = 0.00004$, 95% CI $[0.816, 0.970]$; OOD: $\rho = 0.791$, $p = 0.00012$, 95% CI $[0.536, 0.914]$). Source data are provided as a Source Data file.

Using Reinhard and DANN sequentially combined the advantages of both methods, i.e., it led to improved performances across foundation models, datasets, and splits. Therefore, it is arguably the best overall choice when the composition of the training data is unknown or variable. However, the experiments also show that only using one of the two may yield better results in specific situations. For example, only applying Reinhard stain normalization performed better



Supplementary Figure 19: **Downstream model generalization performance on robustified feature spaces**. In-domain generalization performance on data from seen hospitals for different post-hoc robustification methods. Mean linear probing accuracies (center dots) are reported with 95% confidence intervals (error bars) over n=20 repetitions, each with randomly resampled but overlapping training data. Source data are provided as a Source Data file.



Supplementary Figure 20: **Downstream model generalization performance on robustified feature spaces**. The effect of post-hoc robustness improvement methods on downstream performance drops across foundation models, in-domain (ID) and out-of-domain (OOD). For in-domain performance, ComBat is included. For out-of-domain performance, however, ComBat led to non-competitive results far below the Base setting. Source data are provided as a Source Data file.

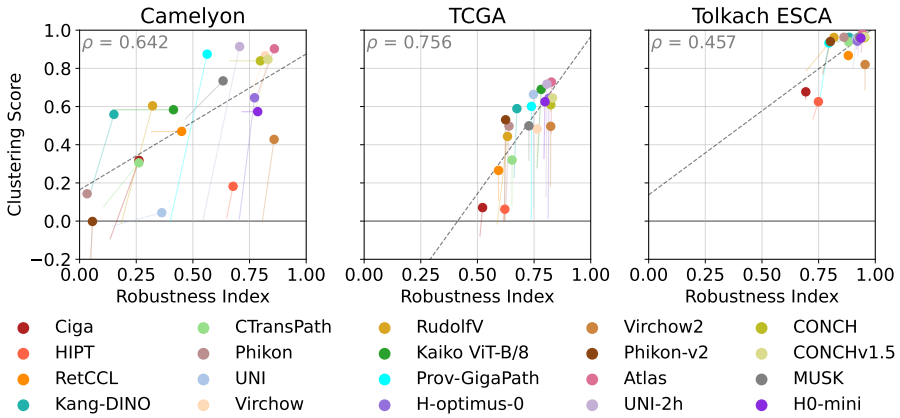
on Split 8 ($V = 1$) in Camelyon.

Supplementary Figure 20 gives an overview of the overall generalization performances of all robustification methods in the downstream model evaluation. Notably, ComBat exacerbated performance drops on both in-domain and out-of-domain data due to its poor performance on strongly correlated training data. On unseen out-of-distribution test data – where the base performance drops were lower than on in-distribution data – we find that DANN only led to better performance for a few FMs. Reinhard stain normalization, in contrast, could increase performance for 14 out of 20 FMs. Stain normalization was most effective for lower-performing models, as the prediction ability of the higher-performing models was already saturated. Notably, H0-mini in combination with domain-adversarial training enabled stable downstream models on out-of-distribution data.

7.6 Clustering performance independent of the silhouette score-based K selection

The initial experimental setup was designed to emulate a realistic scenario where the number of clusters K is determined automatically by maximizing the silhouette score. Although the silhouette score is not the final optimization objective, it serves as a useful heuristic for selecting K, particularly in scenarios such as subtyping, where the ground-truth number of clusters is typically unavailable. However, like any heuristic, the silhouette score may fail in certain cases.

To investigate inconsistencies observed between the robustness index and the clustering score — specifically on the Camelyon dataset, where some FMs exhibited high robustness indices but low clustering scores — we conducted an ablation study to remove the effect of a potentially suboptimal K selection. In this analysis, we performed K-means clustering and evaluated the clustering score across a range of $K \in [2, 100]$. The highest observed clustering score within this range serves as an upper-bound on the clustering score attainable with K-means clustering that is independent of K. Similarly, the number of neighbors k for evaluating the robustness index



Supplementary Figure 21: **Maximum clustering scores obtainable with K-means and maximum robustness index per dataset.** The number of clusters K was selected by maximizing the clustering score, rather than by automatic selection. Similarly, the number of neighbors k for the robustness index was chosen to maximize the robustness index. Selecting K using the silhouette score seemed to have a strong impact on the clustering performances for Camelyon and TCGA, while the robustness index showed only minor differences to its default setting. Yet, both datasets remained challenging to cluster. Strong correlations between the two maximal scores were observed (Spearman rank correlation (n=20); two-sided permutation test; Camelyon: $\rho = 0.642$, $p = 0.00284$, 95% CI $[0.279, 0.845]$; TCGA: $\rho = 0.756$, $p = 0.00012$, 95% CI $[0.472, 0.898]$; Tolkach ESCA: $\rho = 0.457$, $p = 0.04328$, 95% CI $[0.018, 0.748]$). Source data are provided as a Source Data file.

was not set to the median of the optimal values across FMs (for details see Sup. Note 2.6), but was instead selected individually for each FM by maximizing the robustness index. This yields an upper-bound on the robustness index that is independent of k.

As shown in Supplementary Figure 21, this upper-bound analysis revealed further insights. On the Tolkach ESCA dataset, already high clustering scores improved further, with many FMs achieving near-perfect clustering scores close to 1. The TCGA dataset, despite some improvement, remained the most challenging, reflecting an inherent difficulty in distinguishing certain tumor types. In particular, in most FM representation spaces, the LUAD-LUSC combination, especially from Asterand and Christiana Healthcare, remained difficult to cluster sensibly, while BRCA and COAD consistently formed well-separated clusters.

In contrast, stark discrepancies between the clustering scores, obtained with the silhouette score selection of K, and their upper-bounds were observed in the Camelyon dataset. For several FMs, such as RudolfV, Prov-GigaPath, UNI2-h, H-optimus-0, and Virchow2, clustering performance improved dramatically, with scores increasing from near 0 (reflecting a clustering based on biological information and the medical center origin jointly) to over 0.8 (indicating a compliance of the clustering with the tumor status). For these FMs, the silhouette score determined highly suboptimal values (e.g., $K > 26$) and exhibited instability, with low and fluctuating scores across the initially tested range of K.

Exceptions to this trend were noted for Phikon-v2 and UNI, where clustering performance remained poor regardless of K. For these FMs, even an optimal K choice yielded clustering scores close to 0. Their learned representation spaces appear to be dominated by medical center information rather than biological signals.

One potential reason for the silhouette score's failure in selecting appropriate K values is its sensitivity to its assumptions on the cluster structure. The silhouette score favors clusterings with high intra-cluster cohesion and well-separated inter-cluster distances. In cases where clusters

have heterogeneous densities or varying degrees of compactness, the silhouette score tends to penalize even semantically meaningful partitions, leading to poor selections of K. The inability of the silhouette score to identify a suitable K for some FMs (RudolfV, Prov-GigaPath, UNI2-h, H-optimus-0, and Virchow2), despite the datasets of the 2×2 pairings being completely balanced in terms of sample size, suggests the presence of heterogeneous subgroups of embeddings within the learned representation spaces. These subgroups appear not to correspond to either tumor status or the medical center origin, indicating that these FMs' representation spaces may be structured by an additional source of variation.

7.7 Clustering experiments on robustified FM representations

The clustering experiments described in Sup. Note 5 were reevaluated on the robustified embedding spaces. To ensure comparability with the baseline results (i.e., without embedding space robustification), the experimental setup was kept identical. The outcomes are summarized in Supplementary Figure 22. Overall, the impact of robustification on clustering performance varied considerably depending on both the dataset and the foundation model under evaluation.

Dataset-wise results For the Tolkach ESCA dataset, most FMs already exhibited relatively high clustering scores prior to robustification. Consequently, only minor improvements were observed, with the best-performing strategy — most frequently ComBat — yielding an average relative increase of 7.11%. In contrast, models with initially low clustering scores (e.g., Ciga, HIPT, Prov-GigaPath, and Virchow2 with < 0.7) benefited more substantially, achieving an average improvement of 17.51%. The largest relative gain (42.61%) was observed for Prov-GigaPath with the combined application of ComBat and Reinhard. These findings suggest that biologically relevant patterns are inherently well-represented in the Tolkach ESCA dataset and are effectively captured by most FMs.

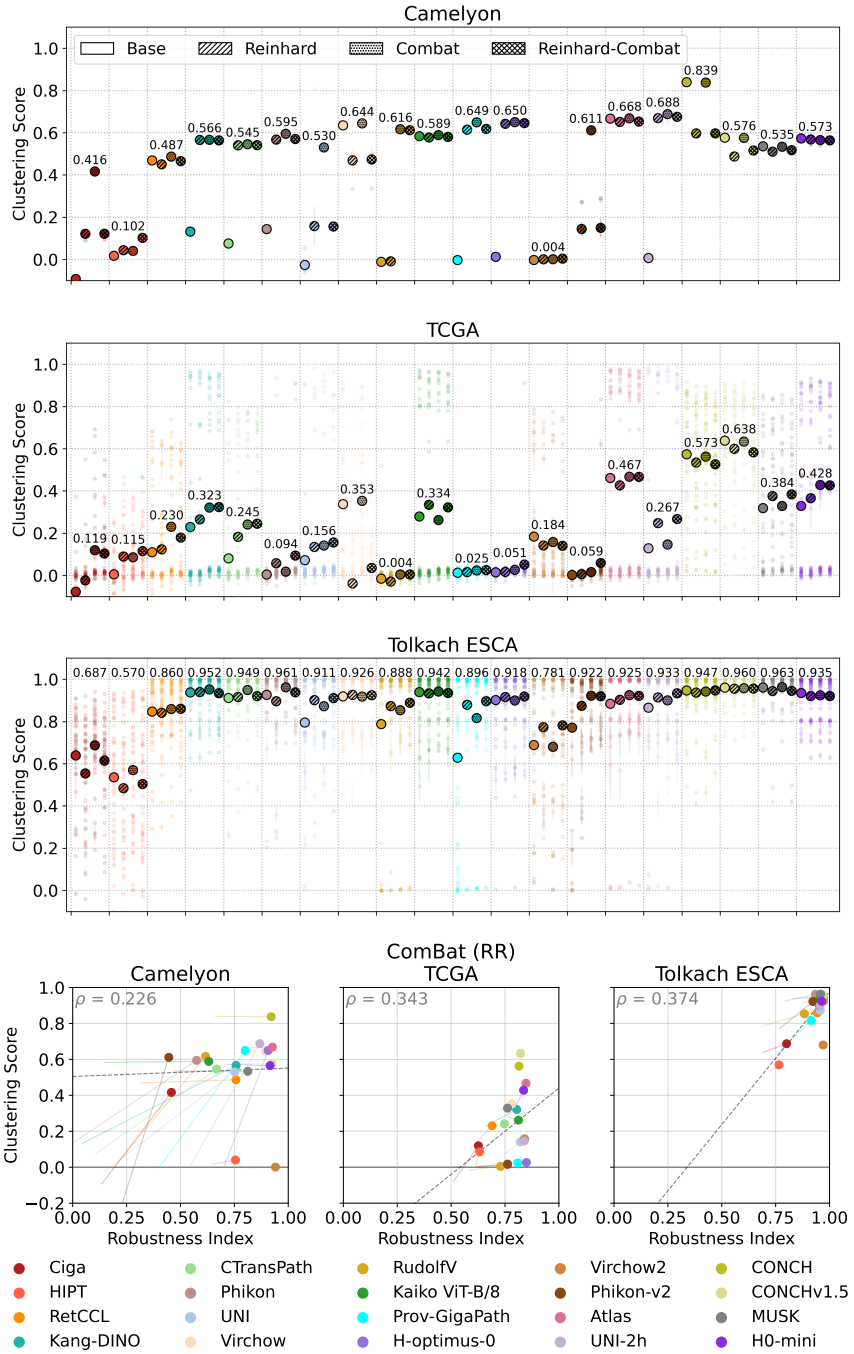
The TCGA dataset presented a greater challenge. It had the lowest baseline clustering scores across all datasets, and even after robustification, the average clustering score remained low (0.25). This suggests that medical center-specific information is strongly embedded and not easily removed by current robustification strategies. The feature space analysis (Sup. Note 6.2) supports this, revealing that TCGA embeddings span a more entangled subspace, where the medical center origin is tightly coupled with biologically relevant features. Distinguishing pairs of the biological classes varies in difficulty: LUSC–LUAD is particularly hard, whereas BRCA–COAD is easy. This difference explains the high variance in clustering scores for the TCGA data set. Nevertheless, the combined ComBat and Reinhard strategy yielded a fourfold increase in clustering scores, particularly benefiting FMs with negative or near-zero baseline scores. However, for models such as Virchow2, CONCH, and CONCHv1.5, robustification sometimes disrupted biologically meaningful groupings, reducing clustering scores by an average of -5.57%.

In the Camelyon dataset, robustification led to a marked unification and improvement in clustering outcomes across FMs, with an average 22-fold increase. It had a strong impact on both the clustering score and the robustness index, making the scores more aligned. This suggests that while the medical center origin was encoded, it is relatively straightforward to remove. For recent FMs (e.g., Atlas, CONCH, CONCHv1.5, MUSK, and H0-mini) that already exhibited biologically structured embeddings, robustification had a negligible impact. In contrast, models whose embeddings were initially influenced by confounding variables (e.g., HIPT, UNI, RudolfV, Phikon-v2) saw substantial improvements. ComBat was generally the most effective method on this dataset.

Robustification method performance ComBat, and in some cases its combination with Reinhard, proved most effective, particularly for embedding spaces heavily influenced by medical center information. When performance diverged between ComBat and Reinhard, their combination typically yielded results aligned with those from Reinhard alone, with a few notable

exceptions (e.g., RudolfV on Camelyon, Ciga, Kang-DINO, and CTransPath on TCGA). Robustification was particularly beneficial for FMs with low or negative baseline clustering scores, enabling a shift toward biologically meaningful groupings. In contrast, FMs with embeddings already structured according to biological differences showed limited or even adverse responses to robustification.

Foundation model-wise results Among all models, CONCH, CONCHv1.5, and Atlas consistently achieved the best clustering performance, reflecting a strong alignment of their embedding spaces with biological similarity; both with and without robustification. H0-mini, MUSK, and Kaiko ViT-B/8 also ranked highly. Conversely, older models such as Ciga and HIPT performed poorly, with embeddings strongly reflecting medical center origin. While most FMs benefited from robustification, RudolfV, Phikon variants, Virchow2, Prov-GigaPath, and H-optimus-0 exhibited inconsistent behavior across datasets and robustification strategies.



Supplementary Figure 22: **Effect of robustication on the clustering performance per dataset.** Robustifying the representation spaces led to strong improvements in clustering scores across foundation models. (top) Dots show mean clustering scores for each pairwise setting, with standard deviation as vertical lines. The large marker gives the overall mean. (bottom) The Spearman rank correlation between the robustness index and the clustering score was no longer significant on the ComBat (RR) robustified representation spaces (Spearman rank correlation; n=20; two-sided permutation test; Camelyon: $\rho = 0.226$, $p = 0.33803$, 95% CI $[-0.241, 0.607]$; TCGA: $\rho = 0.343$, $p = 0.13824$, 95% CI $[-0.117, 0.682]$; Tolkach ESCA: $\rho = 0.374$, $p = 0.10684$, 95% CI $[-0.081, 0.701]$), since ComBat is most effective on the less robust models. Source data are provided as a Source Data file.

References

1. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318,** 2199–2210 (2017).
2. Litjens, G. *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* **7,** giy065 (May 2018).
3. Bándi, P. *et al.* From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging* **38,** 550–560 (2019).
4. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics (SMC)* **9,** 62–66 (1979).
5. Komura, D. *et al.* Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports* **38,** 110424 (2022).
6. Komura, D. & Ishikawa, S. *Histology images from uniform tumor regions in TCGA Whole Slide Images (TCGA-UT)* version 1.0 (Zenodo, Jan. 2021). `https://doi.org/10.5281/zenodo.5889558`.
7. Tolkach, Y. *et al.* Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health* **5,** e265–e275 (2023).
8. Tolkach, Y. *Datasets for Publication "Artificial intelligence for tumor detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study"* Jan. 2023. `https://doi.org/10.5281/zenodo.7548828`.
9. De Jong, E. D., Marcus, E. & Teuwen, J. Current Pathology Foundation Models are unrobust to Medical Center Differences. *arXiv preprint arXiv:2501.18055* (2025).
10. Efron, B. in *Breakthroughs in statistics: Methodology and distribution* 569–593 (Springer, 1992).
11. De Angeli, K. *et al.* Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types. *Journal of Biomedical Informatics* **125,** 103957 (2022).
12. Guerrero, R. E. D., Carvalho, L., Bocklitz, T., Popp, J. & Oliveira, J. L. A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images. *Journal of Imaging Informatics in Medicine* **37,** 1767–1782 (2024).
13. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. & Müller, K.-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* **10,** 1096 (2019).
14. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2,** 665–673 (2020).
15. Hermann, K., Mobahi, H., Fel, T. & Mozer, M. C. *On the Foundations of Shortcut Learning* in. Presented at: International Conference on Learning Representations (ICLR). (2024).
16. Kauffmann, J., Dippel, J., Ruff, L., Samek, W., Müller, K.-R. & Montavon, G. Explainable AI reveals Clever Hans effects in unsupervised learning models. *Nature Machine Intelligence* **7,** 412–422 (2025).
17. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications* **12,** 4423 (2021).
18. Vaidya, A. *et al.* Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine* **30,** 1174–1190 (2024).

19. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A. & Schrouff, J. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications* **14,** 4314 (2023).

20. Elhage, N. *et al.* Toy Models of Superposition. *Transformer Circuits Thread* (2022).

21. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. & Brendel, W. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness* in. Presented at: International conference on learning representations (ICLR). (2019).

22. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A Simple Framework for Contrastive Learning of Visual Representations* in. Presented at: International Conference on Machine Learning (ICML). (2020).

23. Vorontsov, E. *et al.* Virchow: A Million-Slide Digital Pathology Foundation Model. *arXiv preprint arXiv:2309.07778* (2023).

24. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30,** 850–862 (2024).

25. Ilse, M., Tomczak, J. & Welling, M. *Attention-based deep multiple instance learning* in. Presented at: International Conference on Machine Learning (ICML). (2018).

26. Ganin, Y. *et al.* Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **17,** 59:1–59:35 (2016).

27. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20,** 53–65 (1987).

28. Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2,** 193–218 (1985).

29. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

30. Behdenna, A. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24,** 459 (2023).

31. Masson, M. E. J. & Loftus, G. R. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale* **57,** 203–220 (2003).

32. Achtibat, R. *et al.* From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence* **5,** 1006–1019 (2023).

33. Chormai, P., Herrmann, J., Müller, K.-R. & Montavon, G. Disentangled explanations of neural network predictions by finding relevant subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46,** 7283–7299 (2024).

34. Naumann, P., Kauffmann, J. & Montavon, G. Wasserstein Distances Made Explainable: Insights into Dataset Shifts and Transport Phenomena. *arXiv preprint arXiv:2505.06123* (2025).

35. Hu, E. J. *et al. LoRA: Low-Rank Adaptation of Large Language Models* in *International Conference on Learning Representations* (2022).

36. Dippel, J. *et al.* RudolfV: A Foundation Model by Pathologists for Pathologists. *arXiv preprint arXiv:2401.04079* (2024).

37. Tizhoosh, H. & Pantanowitz, L. On image search in histopathology. *Journal of Pathology Informatics* **15,** 100375 (2024).

38. Alfasly, S., Alabtah, G., Hemati, S., Kalari, K. R., Garcia, J. J. & Tizhoosh, H. Validation of histopathology foundation models through whole slide image retrieval. *Scientific Reports* **15,** 3990 (2025).

39. Murchan, P., Broin, P. Ó., Baird, A.-M., Sheils, O. & Finn, S. P. Deep feature batch correction using ComBat for machine learning applications in computational pathology. *Journal of Pathology Informatics* **15,** 100396 (2024).

40. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Computer Graphics and Applications* **21,** 34–41 (2001).