Supplementary Information

CervixFM: A General Foundation Model for Cervical Cytology Image Analysis

Hua Ye[1,2,#], Shijie Liu[3,#], Lanjie Pei[4,#], Jiaxin Bai[3], Qiqi Lu[1,2], Junbo Hu[5], Li Chen[6], Jing Cai[7], Xi Feng[8], Yuwei Xiao[9], Shaoqun Zeng[3], and Xiuli Liu[3,*], Shenghua Cheng[1,2,*]

[1]School of Biomedical Engineering, Southern Medical University, Guangzhou, China.

[2]Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China.

[3]Britton Chance Center and MoE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, Wuhan, China.

[4]Hubei Provincial Center for Disease Control and Prevention, NHC Specialty Laboratory of Food Safety Risk Assessment and Standard Development, Wuhan 430075, China

[5]Department of Pathology, Maternal and Child Hospital of Hubei Province, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[6]Department of Clinical Laboratory, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[7]Department of Obstetrics and Gynecology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

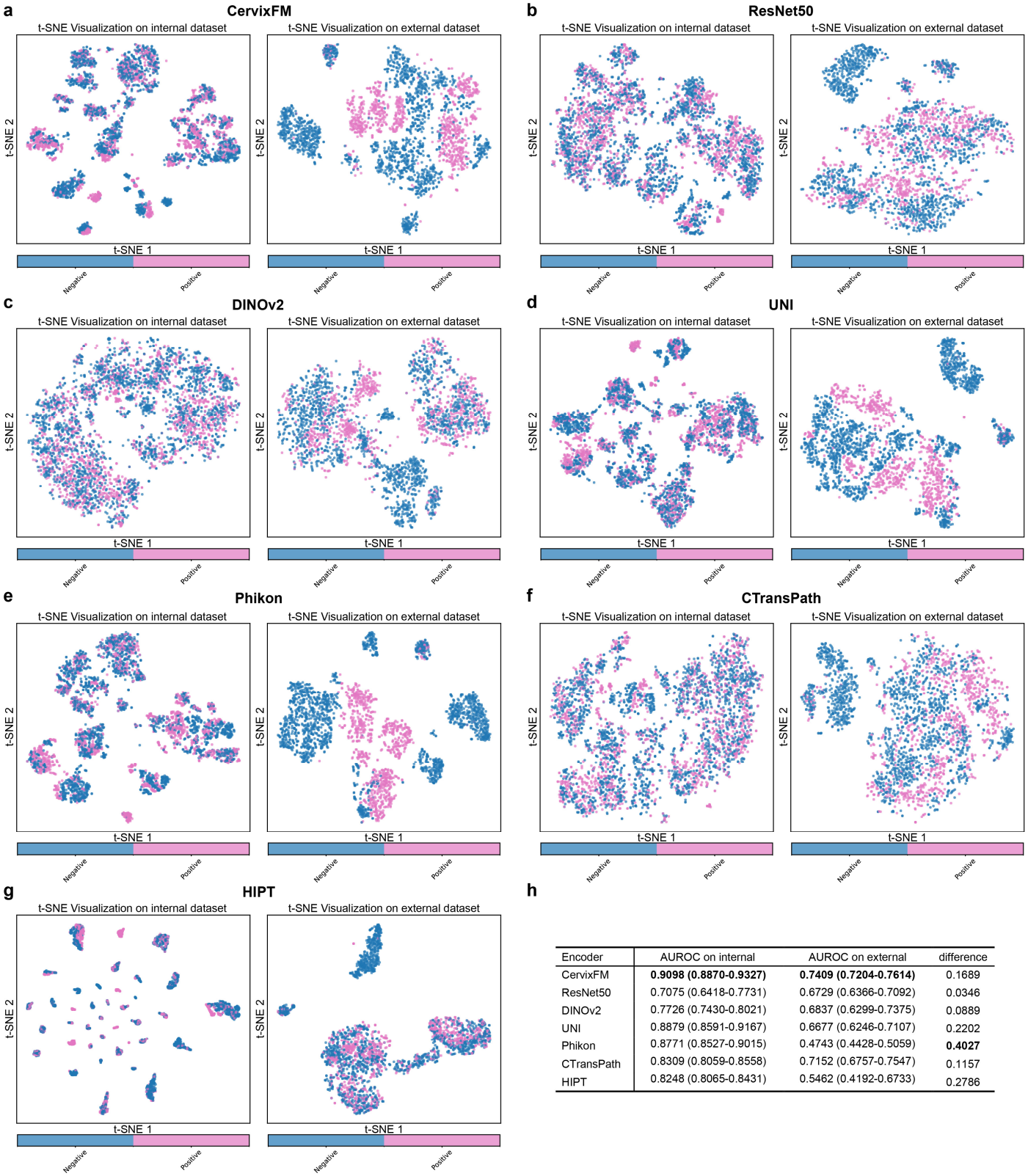
[8]Department of Pathology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[9]Medical Laboratory and Pathology Department, Wuhan Economic and Technological Development District (Hannan District) People's Hospital, Wuhan, Hubei, China

[#]These authors contributed equally: Hua Ye, Shijie Liu, Lanjie Pei

*Correspondence: chengsh2023@smu.edu.cn (S.C.), xlliu@mail.hust.edu.cn (X.L.)

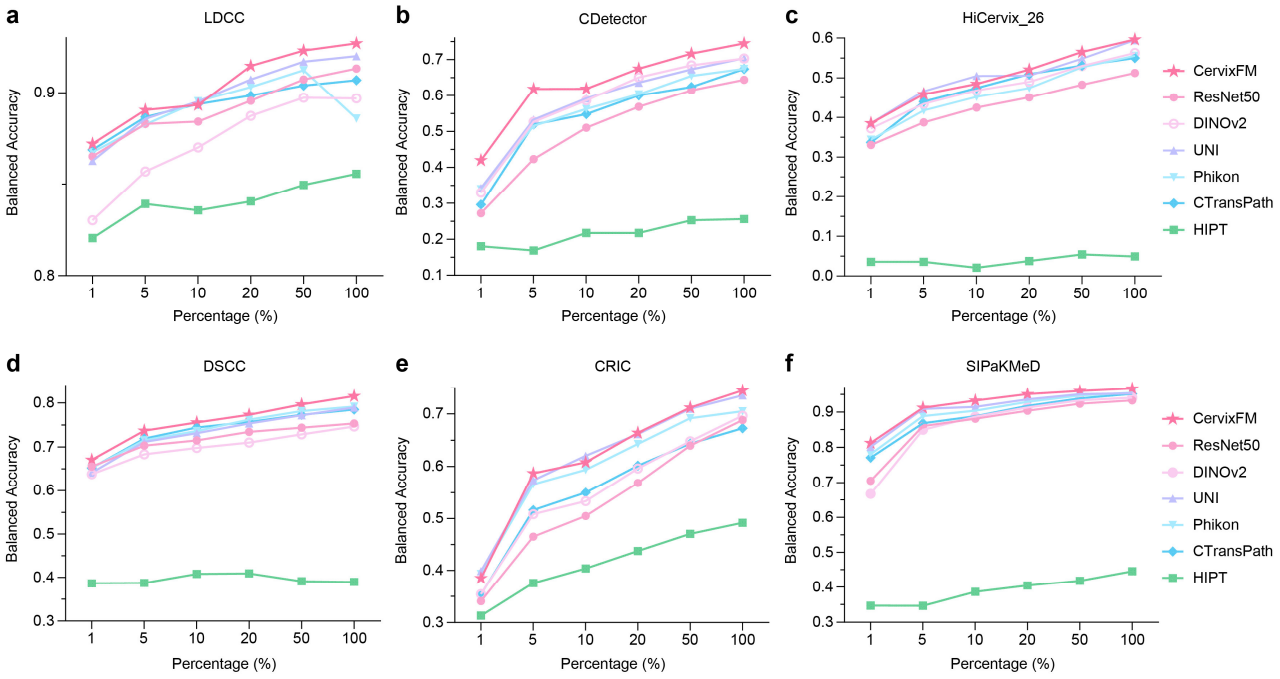
Supplementary Information Contents



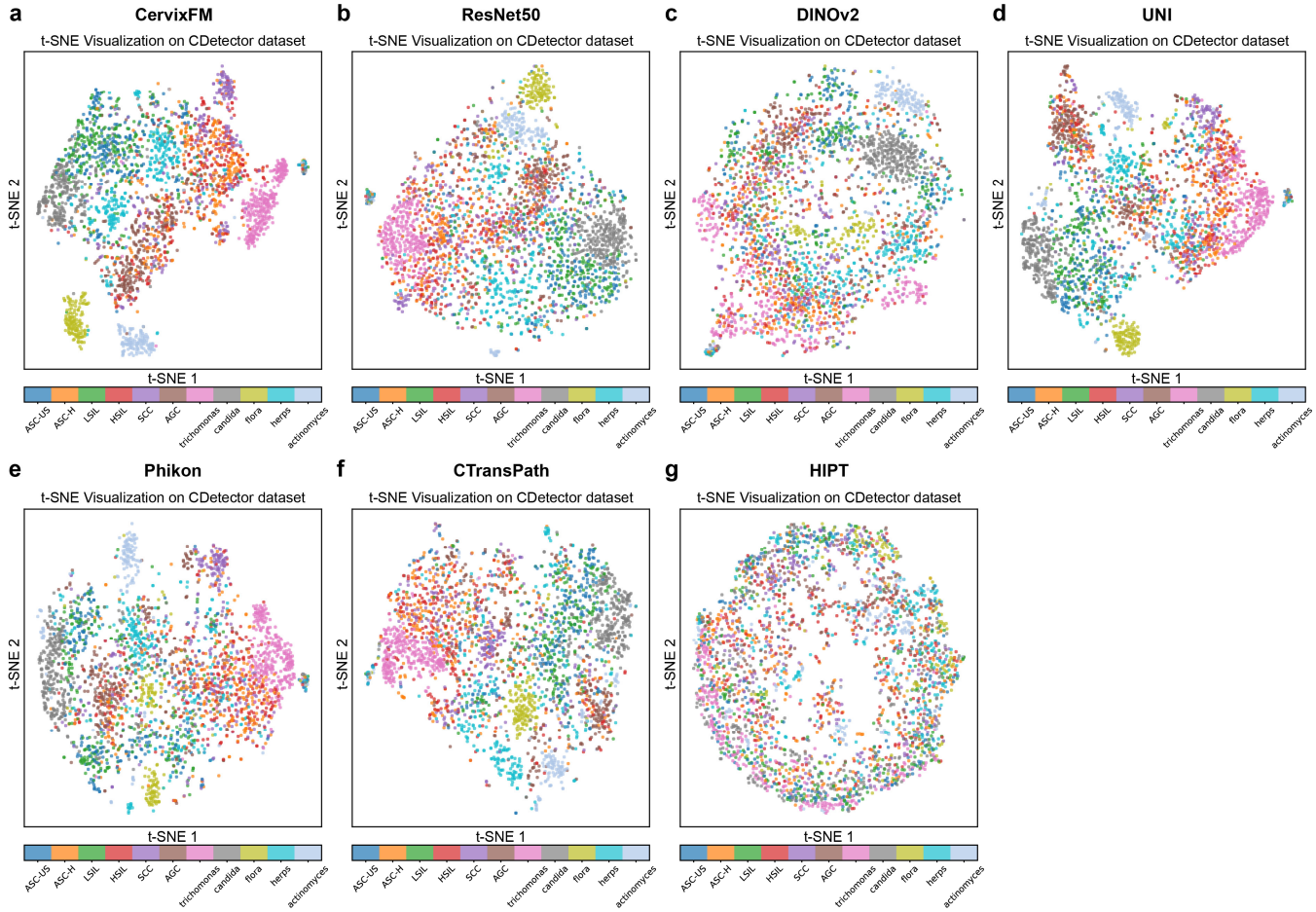
h

Encoder	AUROC on internal	AUROC on external	difference
CervixFM	**0.9098 (0.8870-0.9327)**	**0.7409 (0.7204-0.7614)**	0.1689
ResNet50	0.7075 (0.6418-0.7731)	0.6729 (0.6366-0.7092)	0.0346
DINOv2	0.7726 (0.7430-0.8021)	0.6837 (0.6299-0.7375)	0.0889
UNI	0.8879 (0.8591-0.9167)	0.6677 (0.6246-0.7107)	0.2202
Phikon	0.8771 (0.8527-0.9015)	0.4743 (0.4428-0.5059)	**0.4027**
CTransPath	0.8309 (0.8059-0.8558)	0.7152 (0.6757-0.7547)	0.1157
HIPT	0.8248 (0.8065-0.8431)	0.5462 (0.4192-0.6733)	0.2786

Supplementary Data Fig. 1: **a-g:** T-SNE visualization comparing the feature distributions of CytoFM and other models on internal versus external datasets. The feature of each WSI was obtained by max-pooling all tile features, and represented as a point in all figures. All models exhibit feature distribution differences between the internal and external datasets. Specifically, CytoFM, UNI, Phikon, and HIPT show large differences, while ResNet50, DINOv2, and CTransPath show small differences. **h:** The decline in AUROC performance for CytoFM and other models from internal to external datasets is related to the extent of the feature distribution differences: CytoFM, UNI, Phikon, and HIPT occur larger performance degradation with larger feature distribution differences, while ResNet50, DINOv2, CTransPath occur smaller degradation with smaller differences.

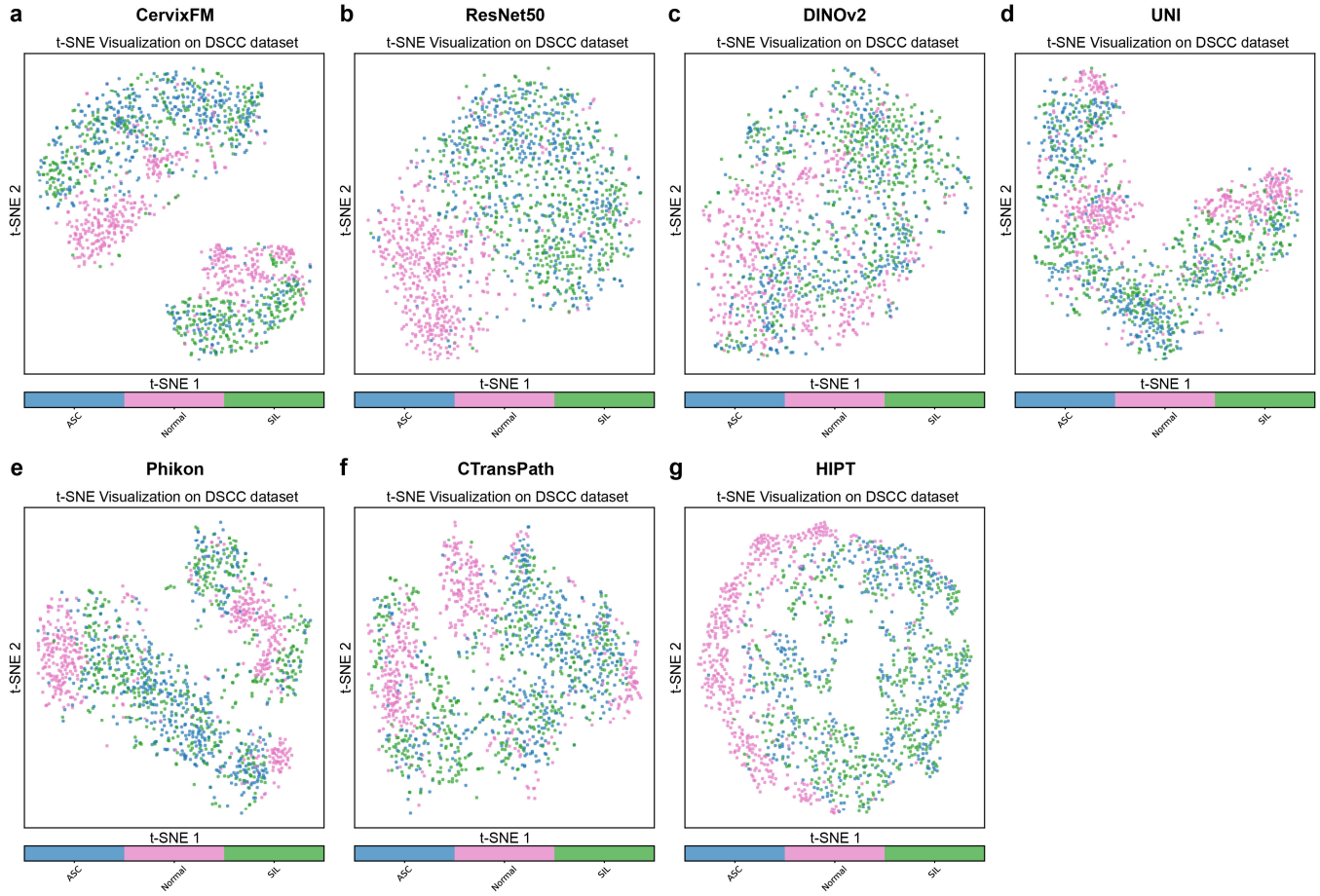


Supplementary Data Fig. 2: The label efficiency of CytoFM and comparison models on patch-level cervical cell image classification tasks across the LDCC, CDetector, HiCervix (26 classes), DSCC, CRIC, and SIPaKMeD datasets. Label efficiency was assessed by randomly sampling 1%, 5%, 10%, 20%, and 50% of data from the training set for linear probing. For each percentage, random sampling was performed five times through different seeds. The average test balanced accuracy is then summarized and reported.



Supplementary Data Fig. 3: Comparison of T-SNE visualization feature space on CDetector dataset of CytoFM versus other models. The CDetector dataset covers 11 classes of cervical cell images. For each class, 300 samples were randomly selected for

visualization. CytoFM and other comparison models extracted features of all random images normalized with the ImageNet mean and standard deviation.



Supplementary Data Fig. 4: Comparison of T-SNE visualization feature space on DSCC dataset of CytoFM versus other models. The DSCC dataset covers 3 classes of cervical cell images. For each class, 500 samples were randomly selected for visualization. CytoFM and other comparison models extracted features of all random images normalized with the ImageNet mean and standard deviation.

Center	Batch	Number of Slides
TJ1	1	44 (0:44)
	2	33 (0:33)
	3	21 (21:0)
	4	43 (15:28)
	5	53 (41:12)
	6	91 (82:9)
	7	69 (64:5)
	8	67 (27:40)
	9	174 (155:19)
	10	111 (72:39)
	11	86 (47:39)
	12	78 (44:34)
	13	79 (44:35)
	14	79 (33:46)
	15	75 (44:31)
	16	80 (47:33)
	17	80 (46:34)
TJ2	1	176 (132:44)
	2	37 (16:21)
	3	77 (65:12)
	4	38 (27:11)
	5	40 (31:9)
SFY	1	114 (52:62)
	2	245 (208:37)
	3	299 (169:130)
	4	39 (0:39)
	5	74 (0:74)
	6	52 (52:0)
	7	74 (43:31)
	8	53 (31:22)
	9	13 (0:13)
	10	33 (25:8)
	11	39 (24:15)
	12	40 (20:20)
XH	1	238 (154:84)
	2	199 (95:104)
	3	96 (0:96)
	4	421 (274:147)
Total		3,660 (2,200:1,460)

Supplementary Data Table 1: **Internal WSI dataset.** The pretraining dataset consists of 18,375 WSIs sourced from 38 batches across 4 data centers. In this study, 20% of the WSIs from every batch were randomly sampled to construct an internal dataset for weakly supervised cervical cell slide classification. The final internal dataset comprises a total of 3,660 WSIs, including 1,460 positive slides and 2,200 negative slides. The number of WSIs and the distribution of positive and negative slides from every batch of different centers are presented in the table.

Center	Number of Slides
LD	2,083 (1,208:875)
SZL	393 (391:2)
JY	103 (83:20)
Total	2,579 (1,682:897)

Supplementary Data Table 2: **External WSI dataset.** 2,579 WSIs sourced from three additional data centers were used to construct an external dataset, which serves as an external validation set for the cervical cell slide classification task. Every center contributed only one batch of data. The number of WSIs and the distribution of positive and negative slides from every center are presented in the table.

Split	Encoder	AUROC	Balanced Accuracy	Precision	F1
split1	CytoFM	**0.9285**	**0.8523**	**0.8543**	**0.8240**
	ResNet50	0.7381	0.6651	0.7144	0.5503
	DINOv2	0.7649	0.6877	0.6943	0.6192
	UNI	0.9107	0.8175	0.8234	0.7817
	Phikon	0.9034	0.8169	0.8301	0.7803
	CTransPath	0.8197	0.7179	0.7337	0.6513
	HIPT	0.8016	0.7194	0.7423	0.6493
split2	CytoFM	**0.8883**	**0.7866**	**0.8182**	**0.7381**
	ResNet50	0.6721	0.5000	0.2981	0.0000
	DINOv2	0.8001	0.7219	0.7293	0.6622
	UNI	0.8687	0.7742	0.7920	0.7248
	Phikon	0.8660	0.7709	0.8044	0.7163
	CTransPath	0.8012	0.7172	0.7372	0.6479
	HIPT	0.8272	0.7331	0.7627	0.6651
split3	CytoFM	**0.9147**	0.8049	**0.8239**	0.7643
	ResNet50	0.7541	0.6405	0.7162	0.4885
	DINOv2	0.7759	0.6867	0.7104	0.6024
	UNI	0.8947	**0.8185**	0.8123	**0.7844**
	Phikon	0.8533	0.7618	0.8103	0.7000
	CTransPath	0.8410	0.7527	0.7661	0.6984
	HIPT	0.8264	0.7431	0.7757	0.6780
split4	CytoFM	**0.9249**	**0.8377**	**0.8571**	**0.8064**
	ResNet50	0.6321	0.5111	0.5602	0.1069
	DINOv2	0.7368	0.6432	0.7032	0.5042
	UNI	0.9067	0.8302	0.8331	0.7974
	Phikon	0.8893	0.7919	0.8041	0.7489
	CTransPath	0.8501	0.7514	0.7670	0.6957
	HIPT	0.8261	0.7273	0.7531	0.6587
split5	CytoFM	**0.8927**	**0.7844**	**0.8122**	**0.7358**
	ResNet50	0.7410	0.6788	0.6788	0.6170
	DINOv2	0.7851	0.7055	0.7128	0.6414
	UNI	0.8586	0.7614	0.7825	0.7070
	Phikon	0.8733	0.7680	0.7950	0.7139
	CTransPath	0.8425	0.7621	0.7818	0.7083
	HIPT	0.8427	0.7411	0.7837	0.6716
Total	CytoFM	**0.9098 (0.8870-0.9327)**	**0.8132 (0.7752-0.8511)**	**0.8331 (0.8070-0.8593)**	**0.7737 (0.7241-0.8234)**
	ResNet50	0.7075 (0.6418-0.7731)	0.5991 (0.4916-0.7066)	0.5935 (0.3738-0.8133)	0.3525 (0.0057-0.6994)
	DINOv2	0.7726 (0.7430-0.8021)	0.6890 (0.6525-0.7255)	0.7100 (0.6939-0.7261)	0.6059 (0.5299-0.6818)
	UNI	0.8879 (0.8591-0.9167)	0.8004 (0.7625-0.8382)	0.8087 (0.7824-0.8349)	0.7590 (0.7089-0.8091)
	Phikon	0.8771 (0.8527-0.9015)	0.7819 (0.7539-0.8100)	0.8088 (0.7925-0.8521)	0.7319 (0.6916-0.7722)
	CTransPath	0.8309 (0.8059-0.8558)	0.7403 (0.7140-0.7665)	0.7572 (0.7313-0.7830)	0.6803 (0.6450-0.7157)
	HIPT	0.8248 (0.8065-0.8431)	0.7328 (0.7207-0.7450)	0.7635 (0.7428-0.7842)	0.6645 (0.6507-0.6784)

Supplementary Data Table 3: **Weakly supervised cervical cell WSI classification on the internal dataset (2 classes).** The internal dataset was randomly split five times with different seeds. For each split, the train, validation, and test samples were drawn from every batch in a 7:1.5:1.5 ratio, while preserving the original positive-to-negative sample ratio. Based on the pre-extracted tile features from every encoder, a WSI classification model was trained and tested on each split using the ABMIL algorithm. The test performance metrics for different splits are reported separately, including AUROC, balanced accuracy, precision and binary F1 score, and the average performance across all splits is also reported with 95% confidence intervals. The best value for each metric is bolded.

Split	Encoder	AUROC	Balanced Accuracy	Precision	F1
split1	CytoFM	**0.7700**	0.6152	**0.6767**	**0.5779**
	ResNet50	0.6803	0.6128	0.6026	0.5363
	DINOv2	0.6510	0.6089	0.6666	0.5729
	UNI	0.6990	**0.6153**	0.6314	0.5677
	Phikon	0.4982	0.4887	0.4877	0.4558
	CTransPath	0.7690	0.6111	0.6341	0.5670
	HIPT	0.6359	0.5000	0.1739	0.5161
split2	CytoFM	0.7371	**0.6377**	0.6446	**0.5823**
	ResNet50	0.6599	0.5000	0.3261	0.0000
	DINOv2	**0.7410**	0.6211	**0.6495**	0.5761
	UNI	0.7068	0.6236	0.6224	0.5648
	Phikon	0.4862	0.4916	0.4923	0.4118
	CTransPath	0.6974	0.6056	0.6162	0.5574
	HIPT	0.3965	0.5000	0.1739	0.5161
split3	CytoFM	**0.7295**	0.6008	**0.6568**	0.5671
	ResNet50	0.6767	**0.6130**	0.6026	0.5300
	DINOv2	0.6881	0.6072	0.6432	**0.5677**
	UNI	0.6238	0.6025	0.5994	0.5428
	Phikon	0.4700	0.4823	0.4829	0.4286
	CTransPath	0.7023	0.6075	0.6262	0.5627
	HIPT	0.4884	0.5080	0.5536	0.5168
split4	CytoFM	**0.7333**	0.5952	**0.6523**	**0.5634**
	ResNet50	0.6338	0.5132	0.5725	0.1071
	DINOv2	0.6323	0.5831	0.5775	0.5165
	UNI	0.6541	0.6011	0.6013	0.5457
	Phikon	0.4325	0.4680	0.4686	0.4205
	CTransPath	0.6894	**0.6031**	0.6092	0.5525
	HIPT	0.5815	0.4994	0.1738	0.5157
split5	CytoFM	**0.7346**	0.6144	0.6234	0.5637
	ResNet50	0.7137	**0.6464**	0.6382	**0.5803**
	DINOv2	0.7060	0.6052	**0.6457**	0.5673
	UNI	0.6546	0.6169	0.6151	0.5581
	Phikon	0.4847	0.5099	0.5098	0.4607
	CTransPath	0.7179	0.6174	0.6204	0.5625
	HIPT	0.6289	0.5000	0.1739	0.5161
Total	CytoFM	**0.7409 (0.7204-0.7614)**	**0.6127 (0.5922-0.6331)**	**0.6508 (0.6267-0.6748)**	**0.5709 (0.5601-0.5816)**
	ResNet50	0.6729 (0.6366-0.7092)	0.5771 (0.4952-0.6590)	0.5484 (0.3914-0.7054)	0.3507 (0.0098-0.6917)
	DINOv2	0.6837 (0.6299-0.7375)	0.6051 (0.5880-0.6222)	0.6365 (0.5940-0.6790)	0.5601 (0.5295-0.5907)
	UNI	0.6677 (0.6246-0.7107)	0.6119 (0.5998-0.6239)	0.6139 (0.5969-0.6309)	0.5558 (0.5419-0.5697)
	Phikon	0.4743 (0.4428-0.5059)	0.4881 (0.4692-0.5070)	0.4883 (0.4697-0.5069)	0.4355 (0.4085-0.4624)
	CTransPath	0.7152 (0.6757-0.7547)	0.6089 (0.6020-0.6158)	0.6212 (0.6094-0.6330)	0.5604 (0.5535-0.5673)
	HIPT	0.5462 (0.4192-0.6733)	0.5015 (0.4970-0.5060)	0.2498 (0.0390-0.4606)	0.5162 (0.5157-0.5167)

Supplementary Data Table 4: **Weakly supervised cervical cell WSI classification on the external dataset (2 classes).** The external dataset was used to independently test all WSI classification models, which employed different encoders and were trained on different splits of the internal dataset. This table presents the test performance metrics for each split, including AUROC, balanced accuracy, precision and binary F1 score, as well as the average performance with 95% confidence intervals. The best value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.9269**	**0.9835**	**0.9441**
ResNet50	0.9132	0.9789	0.9327
DINOv2	0.8969	0.9734	0.9210
UNI	0.9199	0.9827	0.9390
Phikon	0.8858	0.9795	0.9237
CTransPath	0.9069	0.9776	0.9296
HIPT	0.8557	0.9468	0.8939

Supplementary Data Table 5: **Supervised patch-level cervical cell image classification on the LDCC dataset (2 classes).** The preprocessed LDCC dataset retains its official train-test split (68,190:17,049). On this split, a single linear layer was attached to every frozen encoder to perform linear probing classification. The test performance for all encoders is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score. The best value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.7448**	**0.9649**	**0.7952**
ResNet50	0.6440	0.9388	0.7318
DINOv2	0.7031	0.9498	0.7636
UNI	0.7037	0.9576	0.7752
Phikon	0.6742	0.9532	0.7663
CTransPath	0.6737	0.9527	0.7405
HIPT	0.2559	0.8711	0.5481

Supplementary Data Table 6: **Supervised patch-level cervical cell image classification on the CDetector dataset (11 classes).** The preprocessed CDetector dataset retains its official train-test split (45,897:5,057). On this split, a single linear layer was attached to every frozen encoder to perform linear probing classification. The test performance for all encoders is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score. The best value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.7683**	0.9315	**0.7643**
ResNet50	0.7212	0.9092	0.7173
DINOv2	0.7588	0.9284	0.7539
UNI	0.7679	**0.9336**	0.7610
Phikon	0.7228	0.9112	0.7211
CTransPath	0.6976	0.9006	0.6962
HIPT	0.2382	0.4836	0.2054

Supplementary Data Table 7: **Supervised patch-level cervical cell image classification on the HiCervix dataset (4 classes).** The preprocessed HiCervix dataset retains its official train-test split (56,754:8,051). On this split, a single linear layer was attached to every frozen encoder to perform linear probing classification, which adopted the 4-class scheme from the HiCervix. The test performance for all encoders is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score. The best value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.5964**	**0.9490**	**0.6031**
ResNet50	0.5120	0.9227	0.5226
DINOv2	0.5625	0.9413	0.5688
UNI	0.5945	0.9470	0.5948
Phikon	0.5554	0.9393	0.5577
CTransPath	0.5498	0.9403	0.5642
HIPT	0.0494	0.5084	0.0131

Supplementary Data Table 8: **Supervised patch-level cervical cell image classification on the HiCervix dataset (26 classes).** The preprocessed HiCervix dataset retains its official train-test split (56,754:8,051). On this split, a single linear layer was attached to every frozen encoder to perform linear probing classification, which adopted the 26-class scheme from the HiCervix. The test performance for all encoders is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score. The best value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.8157 (0.8087-0.8226)**	**0.9440 (0.9406-0.9473)**	**0.8351 (0.8289-0.8414)**
ResNet50	0.7531 (0.7467-0.7595)	0.9133 (0.9091-0.9175)	0.7800 (0.7741-0.7858)
DINOv2	0.7464 (0.7308-0.7619)	0.9079 (0.9015-0.9142)	0.7740 (0.7597-0.7883)
UNI	0.7896 (0.7806-0.7987)	0.9327 (0.9299-0.9355)	0.8115 (0.8030-0.8200)
Phikon	0.7918 (0.7853-0.7983)	0.9365 (0.9338-0.9392)	0.8151 (0.8106-0.8196)
CTransPath	0.7849 (0.7799-0.7898)	0.9347 (0.9333-0.9362)	0.8111 (0.8059-0.8163)
HIPT	0.3892 (0.1493-0.6291)	0.5246 (0.2134-0.8357)	0.3264 (0.0278-0.6250)

Supplementary Data Table 9: **Supervised patch-level cervical cell image classification on the DSCC dataset (3 classes).** Since there is no official train-test split, the preprocessed DSCC dataset was randomly divided into training and test sets in a 4:1 ratio (12,406:3,103) five times with different seeds, while maintaining consistent class distribution. On each split, a single linear layer was attached to every frozen encoder to perform linear probing classification. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.7447 (0.7225-0.7669)**	**0.9625 (0.9573-0.9676)**	**0.8596 (0.8537-0.8655)**
ResNet50	0.6882 (0.6782-0.6981)	0.9506 (0.9467-0.9546)	0.8312 (0.8287-0.8337)
DINOv2	0.6966 (0.6805-0.7128)	0.9516 (0.9499-0.9533)	0.8287 (0.8188-0.8387)
UNI	0.7352 (0.7109-0.7595)	0.9598 (0.9568-0.9628)	0.8543 (0.8447-0.8639)
Phikon	0.7052 (0.6943-0.7160)	0.9556 (0.9513-0.9598)	0.8406 (0.8312-0.8501)
CTransPath	0.6726 (0.6594-0.6859)	0.9557 (0.9509-0.9606)	0.8410 (0.8340-0.8479)
HIPT	0.4915 (0.4733-0.5098)	0.8933 (0.8866-0.9001)	0.7073 (0.6931-0.7214)

Supplementary Data Table 10: **Supervised patch-level cervical cell image classification on the CRIC dataset (6 classes).** Since there is no official train-test split, the preprocessed CRIC dataset was randomly divided into training and test sets in a 4:1 ratio (9,227:2,307) five times with different seeds, while maintaining consistent class distribution. On each split, a single linear layer was attached to every frozen encoder to perform linear probing classification. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.7414 (0.6988-0.7840)**	0.9287 (0.9159-0.9414)	**0.7022 (0.6538-0.7507)**
ResNet50	0.6716 (0.6364-0.7067)	0.9045 (0.8889-0.9202)	0.6357 (0.5896-0.6818)
DINOv2	0.6559 (0.6077-0.7041)	0.9084 (0.8892-0.9277)	0.6215 (0.5707-0.6724)
UNI	0.7144 (0.6636-0.7651)	**0.9354 (0.9184-0.9523)**	0.6766 (0.6203-0.7328)
Phikon	0.6796 (0.6314-0.7279)	0.9166 (0.8986-0.9346)	0.6387 (0.5918-0.6857)
CTransPath	0.6830 (0.6341-0.7320)	0.9194 (0.9073-0.9315)	0.6484 (0.6073-0.6894)
HIPT	0.3449 (0.1228-0.5669)	0.6956 (0.4845-0.9068)	0.2961 (0.02929-0.563)

Supplementary Data Table 11: **Supervised patch-level cervical cell image classification on the Herlev dataset (7 classes).** Since there is no official train-test split, the preprocessed Herlev dataset was randomly divided into training and test sets in a 4:1 ratio (731:186) five times with different seeds, while maintaining consistent class distribution. On each split, a single linear layer was attached to every frozen encoder to perform linear probing classification. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.9671 (0.9625-0.9718)**	**0.9976 (0.9963-0.9989)**	**0.9669 (0.9621-0.9718)**
ResNet50	0.9328 (0.9219-0.9437)	0.9935 (0.9917-0.9952)	0.9328 (0.9217-0.9439)
DINOv2	0.9405 (0.9317-0.9493)	0.9947 (0.9936-0.9959)	0.9406 (0.9320-0.9493)
UNI	0.9525 (0.9454-0.9596)	0.9952 (0.9937-0.9967)	0.9523 (0.9457-0.9590)
Phikon	0.9544 (0.9475-0.9614)	0.9962 (0.9953-0.9971)	0.9544 (0.9472-0.9615)
CTransPath	0.9511 (0.9496-0.9525)	0.9962 (0.9956-0.9968)	0.9509 (0.9490-0.9529)
HIPT	0.4452 (0.0174-0.8731)	0.6452 (0.2813-1.0090)	0.4036 (-0.0679-0.8750)

Supplementary Data Table 12: **Supervised patch-level cervical cell image classification on the SIPaKMeD dataset (5 classes).** Since there is no official train-test split, the preprocessed SIPaKMeD dataset was randomly divided into training and test sets in a 4:1 ratio (3,237:812) five times with different seeds, while maintaining consistent class distribution. On each split, a single linear layer was attached to every frozen encoder to perform linear probing classification. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	0.9803 (0.9590-1.0020)	**0.9998 (0.9996-1.0000)**	0.9927 (0.9853-1.0000)
ResNet50	0.9627 (0.9424-0.9831)	0.9990 (0.9976-1.0000)	0.9845 (0.9780-0.9909)
DINOv2	0.9824 (0.9583-1.0070)	0.9991 (0.9980-1.0000)	0.9918 (0.9820-1.0010)
UNI	**0.9903 (0.9793-1.0010)**	**0.9998 (0.9995-1.0000)**	**0.9959 (0.9906-1.0010)**
Phikon	0.9709 (0.9403-1.0020)	0.9991 (0.9980-1.0000)	0.9876 (0.9743-1.0010)
CTransPath	0.9521 (0.9363-0.9679)	0.9978 (0.9964-0.9993)	0.9774 (0.9703-0.9846)
HIPT	0.8780 (0.8312-0.9247)	0.9918 (0.9886-0.9949)	0.9442 (0.9253-0.9632)

Supplementary Data Table 13: **Supervised patch-level cervical cell image classification on the LBC dataset (4 classes).** Since there is no official train-test split, the LBC dataset was randomly divided into training and test sets in a 4:1 ratio (768:194) five times with different seeds, while maintaining consistent class distribution. On each split, a simple MIL scheme was adopted to handle large-sized images, where 256×256 crops were extracted with a stride of 128. The crop features pre-extracted by every frozen encoder were max-pooled and then fed into a single linear layer for linear probing. For all encoders, the average test

performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.8667 (0.8228-0.9106)**	**0.9672 (0.9509-0.9835)**	0.8669 (0.8224-0.9114)
ResNet50	**0.8667 (0.8316-0.9018)**	0.9657 (0.9495-0.9819)	**0.8674 (0.8327-0.9022)**
DINOv2	0.8200 (0.8107-0.8293)	0.9399 (0.9242-0.9556)	0.8208 (0.8122-0.8294)
UNI	0.8300 (0.7895-0.8705)	0.9464 (0.9248-0.9681)	0.8309 (0.7908-0.8710)
Phikon	0.8233 (0.7674-0.8792)	0.9468 (0.9270-0.9666)	0.8246 (0.7694-0.8797)
CTransPath	0.8250 (0.7698-0.8802)	0.9420 (0.9198-0.9642)	0.8254 (0.7703-0.8806)
HIPT	0.5917 (0.3164-0.8669)	0.7190 (0.4235-1.0150)	0.5530 (0.2098-0.8961)

Supplementary Data Table 14: **Supervised patch-level cervical cell image classification on the BMT dataset (3 classes).** Since there is no official train-test split, the preprocessed BMT dataset was randomly divided into training and test sets in a 4:1 ratio (480:120) five times with different seeds, while maintaining consistent class distribution. On each split, a simple MIL scheme was adopted to handle large-sized images, where 256×256 crops were extracted with a stride of 128. The crop features pre-extracted by every frozen encoder were max-pooled and then fed into a single linear layer for linear probing. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Balanced Accuracy	AUROC	F1
CytoFM	**0.9100 (0.8548-0.9652)**	**0.9918 (0.9790-1.0040)**	**0.9111 (0.8429-0.9793)**
ResNet50	0.7852 (0.6773-0.8932)	0.9539 (0.9296-0.9782)	0.7954 (0.7142-0.8765)
DINOv2	0.7857 (0.6932-0.8782)	0.8751 (0.7999-0.9504)	0.7948 (0.6793-0.9103)
UNI	0.7876 (0.6903-0.8849)	0.9576 (0.9302-0.9851)	0.8115 (0.7287-0.8943)
Phikon	0.7462 (0.5930-0.8994)	0.9430 (0.9015-0.9844)	0.7495 (0.6399-0.8590)
CTransPath	0.6519 (0.5314-0.7725)	0.8782 (0.7488-1.0080)	0.6852 (0.5602-0.8102)
HIPT	0.4462 (0.3336-0.5588)	0.6872 (0.6496-0.7247)	0.4343 (0.2843-0.5843)

Supplementary Data Table 15: **Supervised patch-level cervical cell image classification on the CERVIX93 dataset (3 classes).** Although an official split is available, the CERVIX93 dataset was still randomly divided into training and test sets in a 4:1 ratio (72:21) five times due to its limited data, while maintaining consistent class distribution. On each split, a simple MIL scheme was adopted to handle large-sized images, where 256×256 crops were extracted with a stride of 128. The crop features pre-extracted by every frozen encoder were max-pooled and then fed into a single linear layer for linear probing. For all encoders, the average test performance across five splits is reported using three metrics: balanced accuracy, AUROC, and weighted F1 score, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	**0.8720 (0.8714-0.8727)**	**0.8905 (0.8903-0.8906)**	0.8932 (0.8932-0.8933)	**0.9146 (0.9143-0.9149)**	**0.9230 (0.9226-0.9233)**	**0.9269**
ResNet50	0.8652 (0.8651-0.8653)	0.8829 (0.8826-0.8832)	0.8842 (0.8841-0.8842)	0.8956 (0.8954-0.8958)	0.9072 (0.9070-0.9073)	0.9132
DINOv2	0.8302 (0.8302-0.8302)	0.8569 (0.8567-0.8571)	0.8701 (0.8699-0.8702)	0.8873 (0.8873-0.8874)	0.8972 (0.8906-0.9038)	0.8969
UNI	0.8627 (0.8622-0.8632)	0.8856 (0.8854-0.8859)	0.8951 (0.8949-0.8954)	0.9072 (0.9071-0.9073)	0.9169 (0.9168-0.9170)	0.9199
Phikon	0.8675 (0.8672-0.8677)	0.8825 (0.8824-0.8827)	**0.8953 (0.8951-0.8955)**	0.9031 (0.9030-0.9032)	0.9122 (0.9120-0.9124)	0.8858
CTransPath	0.8688 (0.8686-0.8690)	0.8867 (0.8866-0.8868)	0.8937 (0.8936-0.8938)	0.8984 (0.8983-0.8985)	0.9039 (0.9039-0.9040)	0.9069
HIPT	0.8205 (0.8184-0.8227)	0.8391 (0.8388-0.8394)	0.8356 (0.8355-0.8358)	0.8405 (0.8402-0.8408)	0.8496 (0.8495-0.8497)	0.8557

Supplementary Data Table 16: **Label efficiency of patch-level cervical cell image classification on the LDCC dataset (2 classes).** The preprocessed LDCC dataset contains 68,190 official training images, from which 1%, 5%, 10%, 20%, and 50% of data were randomly sampled to construct subsets that preserve the original class distribution. For each percentage, random sampling was conducted five times with different seeds. On each subset, a single linear layer was attached to every frozen encoder for linear probing training. All trained models were then evaluated on the official test set (17,049). The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	**0.4197 (0.4127-0.4267)**	**0.6180 (0.6175-0.6184)**	**0.6185 (0.6149-0.6221)**	**0.6745 (0.6720-0.6771)**	**0.7167 (0.7162-0.7171)**	**0.7448**
ResNet50	0.2721 (0.2712-0.2730)	0.4230 (0.4229-0.4232)	0.5103 (0.5101-0.5106)	0.5683 (0.5682-0.5684)	0.6142 (0.6125-0.6160)	0.6440
DINOv2	0.3293 (0.3265-0.3320)	0.5263 (0.5261-0.5264)	0.5858 (0.5852-0.5864)	0.6505 (0.6465-0.6545)	0.6837 (0.6819-0.6854)	0.7031
UNI	0.3424 (0.3405-0.3444)	0.5323 (0.5301-0.5345)	0.5917 (0.5911-0.5924)	0.6358 (0.6354-0.6361)	0.6725 (0.6705-0.6746)	0.7037
Phikon	0.3391 (0.3339-0.3443)	0.5153 (0.5122-0.5184)	0.5624 (0.5622-0.5625)	0.6031 (0.6028-0.6034)	0.6547 (0.6531-0.6564)	0.6742
CTransPath	0.2959 (0.2911-0.3008)	0.5191 (0.5190-0.5192)	0.5479 (0.5476-0.5482)	0.5998 (0.5995-0.6001)	0.6235 (0.6235-0.6236)	0.6737
HIPT	0.1801 (0.1773-0.1829)	0.1686 (0.1686-0.1686)	0.2170 (0.2170-0.2171)	0.2171 (0.2171-0.2171)	0.2524 (0.2523-0.2525)	0.2559

Supplementary Data Table 17: **Label efficiency of patch-level cervical cell image classification on the CDetector dataset (11 classes).** The preprocessed CDetector dataset contains 45,897 official training images, from which 1%, 5%, 10%, 20%, and 50% of data were randomly sampled to construct subsets that preserve the original class distribution. For each percentage, random sampling was conducted five times with different seeds. On each subset, a single linear layer was attached to every frozen encoder for linear probing training. All trained models were then evaluated on the official test set (5,057). The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	**0.3852 (0.3847-0.3856)**	0.4583 (0.4576-0.4591)	0.4844 (0.4839-0.4849)	**0.5206 (0.5202-0.5209)**	**0.5649 (0.5647-0.5651)**	**0.5964**
ResNet50	0.3305 (0.3302-0.3307)	0.3871 (0.3867-0.3875)	0.4245 (0.4243-0.4247)	0.4502 (0.4500-0.4505)	0.4823 (0.4821-0.4825)	0.5120
DINOv2	0.3722 (0.3712-0.3732)	0.4338 (0.4336-0.4340)	0.4679 (0.4673-0.4685)	0.4895 (0.4891-0.4900)	0.5298 (0.5295-0.5300)	0.5625
UNI	0.3851 (0.3846-0.3855)	**0.4659 (0.4651-0.4667)**	**0.5047 (0.5035-0.5058)**	0.5056 (0.5055-0.5057)	0.5483 (0.5476-0.5490)	0.5945
Phikon	0.3436 (0.3429-0.3443)	0.4174 (0.4169-0.4179)	0.4517 (0.4507-0.4528)	0.4739 (0.4724-0.4755)	0.5266 (0.5262-0.5270)	0.5554
CTransPath	0.3369 (0.3368-0.3371)	0.4398 (0.4393-0.4403)	0.4740 (0.4739-0.4741)	0.5090 (0.5089-0.5091)	0.5307 (0.5307-0.5308)	0.5498
HIPT	0.0358 (0.0357-0.0359)	0.0359 (0.0357-0.0361)	0.0208 (0.0205-0.0212)	0.0378 (0.0374-0.0381)	0.0544 (0.0535-0.0554)	0.0494

Supplementary Data Table 18: **Label efficiency of patch-level cervical cell image classification on the HiCervix dataset (26 classes).** The preprocessed HiCervix dataset contains 56,754 official training images, from which 1%, 5%, 10%, 20%, and 50% of data were randomly sampled to construct subsets that preserve the original class distribution. For each percentage, random sampling was conducted five times with different seeds. On each subset, a single linear layer was attached to every frozen encoder for linear probing training, which adopted the 26-class scheme. All trained models were then evaluated on the official test set (8,051). The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	**0.6707 (0.6533-0.6881)**	**0.7370 (0.7257-0.7482)**	**0.7554 (0.7487-0.7622)**	**0.7732 (0.7633-0.7831)**	**0.7967 (0.7909-0.8025)**	**0.8157 (0.8087-0.8226)**
ResNet50	0.6555 (0.6455-0.6655)	0.7029 (0.6981-0.7076)	0.7148 (0.7043-0.7253)	0.7342 (0.7190-0.7494)	0.7438 (0.7288-0.7588)	0.7531 (0.7467-0.7595)
DINOv2	0.6378 (0.6197-0.6559)	0.6830 (0.6766-0.6894)	0.6977 (0.6930-0.7023)	0.7097 (0.6933-0.7260)	0.7284 (0.7175-0.7393)	0.7464 (0.7308-0.7619)
UNI	0.6411 (0.6251-0.6571)	0.7115 (0.6940-0.7291)	0.7310 (0.7250-0.7370)	0.7533 (0.7481-0.7584)	0.7719 (0.7586-0.7851)	0.7896 (0.7806-0.7987)
Phikon	0.6525 (0.6362-0.6688)	0.7153 (0.7028-0.7277)	0.7361 (0.7213-0.7510)	0.7622 (0.7533-0.7710)	0.7817 (0.7706-0.7927)	0.7918 (0.7853-0.7983)
CTransPath	0.6521 (0.6387-0.6656)	0.7191 (0.7116-0.7266)	0.7443 (0.7434-0.7452)	0.7568 (0.7448-0.7688)	0.7733 (0.7689-0.7777)	0.7849 (0.7799-0.7898)
HIPT	0.3865 (0.1664-0.6066)	0.3872 (0.1651-0.6093)	0.4079 (0.1895-0.6262)	0.4091 (0.2011-0.6170)	0.3908 (0.1571-0.6244)	0.3892 (0.1493-0.6291)

Supplementary Data Table 19: **Label efficiency of patch-level cervical cell image classification on the DSCC dataset (3 classes).** Since there is no official train-test split, the preprocessed DSCC dataset was randomly divided into training and test sets

in a 4:1 ratio (12,406:3,103) five times, while maintaining consistent class distribution. For each split, 1%, 5%, 10%, 20%, and 50% of training data were randomly sampled to construct subsets that preserve the class distribution of training set. On each subset, a single linear layer was attached to every frozen encoder for linear probing training. All trained models were then evaluated on the corresponding test set. The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	0.3848 (0.3599-0.4097)	**0.5868 (0.5565-0.6171)**	0.6079 (0.5897-0.6261)	**0.6640 (0.6377-0.6904)**	**0.7120 (0.6795-0.7444)**	**0.7447 (0.7225-0.7669)**
ResNet50	0.3407 (0.3050-0.3764)	0.4651 (0.4548-0.4754)	0.5046 (0.4974-0.5119)	0.5684 (0.5342-0.6025)	0.6396 (0.6099-0.6693)	0.6882 (0.6782-0.6981)
DINOv2	0.3542 (0.3271-0.3812)	0.5081 (0.4770-0.5392)	0.5332 (0.5115-0.5549)	0.5962 (0.5821-0.6103)	0.6479 (0.6382-0.6575)	0.6966 (0.6805-0.7128)
UNI	**0.3986 (0.3537-0.4436)**	0.5737 (0.5594-0.5880)	**0.6203 (0.6058-0.6348)**	0.6617 (0.6486-0.6748)	0.7098 (0.6875-0.7321)	0.7352 (0.7109-0.7595)
Phikon	0.3931 (0.3607-0.4256)	0.5661 (0.5468-0.5853)	0.5935 (0.5846-0.6024)	0.6429 (0.6265-0.6592)	0.6919 (0.6649-0.7189)	0.7052 (0.6943-0.7160)
CTransPath	0.3535 (0.3311-0.3759)	0.5162 (0.4956-0.5369)	0.5499 (0.5316-0.5682)	0.6016 (0.5865-0.6167)	0.6436 (0.6176-0.6695)	0.6726 (0.6594-0.6859)
HIPT	0.3131 (0.2925-0.3338)	0.3754 (0.3482-0.4026)	0.4037 (0.3696-0.4377)	0.4373 (0.4114-0.4631)	0.4702 (0.4509-0.4896)	0.4915 (0.4733-0.5098)

Supplementary Data Table 20: **Label efficiency of patch-level cervical cell image classification on the CRIC dataset (6 classes).** Since there is no official train-test split, the preprocessed CRIC dataset was randomly divided into training and test sets in a 4:1 ratio (9,227:2,307) five times, while maintaining consistent class distribution. For each split, 1%, 5%, 10%, 20%, and 50% of training data were randomly sampled to construct subsets that preserve the class distribution of training set. On each subset, a single linear layer was attached to every frozen encoder for linear probing training. All trained models were then evaluated on the corresponding test set. The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	1%	5%	10%	20%	50%	100%
CytoFM	**0.8124 (0.7934-0.8314)**	**0.9119 (0.8944-0.9295)**	**0.9323 (0.9221-0.9425)**	**0.9511 (0.9412-0.9609)**	**0.9600 (0.9560-0.9640)**	**0.9671 (0.9625-0.9718)**
ResNet50	0.7048 (0.6674-0.7423)	0.8600 (0.8426-0.8775)	0.8807 (0.8711-0.8904)	0.9035 (0.8901-0.9168)	0.9232 (0.9164-0.9300)	0.9328 (0.9219-0.9437)
DINOv2	0.6670 (0.6388-0.6952)	0.8499 (0.8364-0.8634)	0.8863 (0.8699-0.9027)	0.9126 (0.9006-0.9246)	0.9329 (0.9226-0.9432)	0.9405 (0.9317-0.9493)
UNI	0.8016 (0.7914-0.8118)	0.9087 (0.9012-0.9161)	0.9142 (0.9031-0.9254)	0.9356 (0.9210-0.9501)	0.9505 (0.9434-0.9576)	0.9525 (0.9454-0.9596)
Phikon	0.7805 (0.7727-0.7883)	0.8885 (0.8703-0.9068)	0.9035 (0.8927-0.9144)	0.9285 (0.9166-0.9404)	0.9466 (0.9409-0.9523)	0.9544 (0.9475-0.9614)
CTransPath	0.7705 (0.7410-0.8001)	0.8679 (0.8579-0.8780)	0.8878 (0.8654-0.9103)	0.9171 (0.9048-0.9294)	0.9386 (0.9287-0.9484)	0.9511 (0.9496-0.9525)
HIPT	0.3470 (0.0940-0.5999)	0.3465 (-0.0412-0.7341)	0.3862 (0.0088-0.7635)	0.4044 (-0.0205-0.8293)	0.4181 (-0.0178-0.8541)	0.4452 (0.0174-0.8731)

Supplementary Data Table 21: **Label efficiency of patch-level cervical cell image classification on the SIPaKMeD dataset (5 classes).** Since there is no official train-test split, the preprocessed SIPaKMeD dataset was randomly divided into training and test sets in a 4:1 ratio (3,237:812) five times, while maintaining consistent class distribution. For each split, 1%, 5%, 10%, 20%, and 50% of training data were randomly sampled to construct subsets that preserve the class distribution of training set. On each subset, a single linear layer was attached to every frozen encoder for linear probing training. All trained models were then evaluated on the corresponding test set. The label efficiency of all encoders at each percentage was assessed by the average test balanced accuracy across five sampling. The best average performance for each percentage is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.8942**	**0.9694**	**0.9840**	**0.9101**	**0.9125**
ResNet50	0.8810	0.9597	0.9761	0.8947	0.8996
DINOv2	0.8453	0.9548	0.9773	0.8667	0.8713
UNI	0.8714	0.9611	0.9785	0.8864	0.8901
Phikon	0.8827	0.9645	0.9807	0.8967	0.9010
CTransPath	0.8732	0.9598	0.9787	0.8892	0.8958
HIPT	0.8621	0.9490	0.9680	0.8842	0.8866

Supplementary Data Table 22: **Patch-level cervical cell image retrieval on the LDCC dataset (2 classes).** The preprocessed LDCC dataset retains its official train-test split (68,190:17,049), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in LDCC, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.7538**	**0.8984**	**0.9371**	**0.7983**	**0.8050**
ResNet50	0.6360	0.8311	0.8865	0.6905	0.7054
DINOv2	0.6417	0.8406	0.8910	0.6992	0.7109
UNI	0.7020	0.8752	0.9195	0.7550	0.7676
Phikon	0.6957	0.8717	0.9098	0.7542	0.7645
CTransPath	0.6791	0.8539	0.9029	0.7328	0.7423
HIPT	0.5452	0.7544	0.8191	0.6146	0.6284

Supplementary Data Table 23: **Patch-level cervical cell image retrieval on the CDetector dataset (11 classes).** The preprocessed CDetector dataset retains its official train-test split (45,897:5,057), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in CDetector, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.7378**	**0.8947**	**0.9385**	**0.7643**	**0.7738**
ResNet50	0.6281	0.8354	0.9045	0.6650	0.6723
DINOv2	0.6525	0.8611	0.9226	0.6920	0.7023
UNI	0.7184	0.8824	0.9333	0.7436	0.7544
Phikon	0.6781	0.8659	0.9240	0.7074	0.7163
CTransPath	0.6556	0.8563	0.9180	0.6882	0.6900
HIPT	0.2622	0.4203	0.5208	0.2787	0.2791

Supplementary Data Table 24: **Patch-level cervical cell image retrieval on the HiCervix dataset (4 classes).** The preprocessed HiCervix dataset retains its official train-test split (56,754:8,051), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in HiCervix, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels (4 classes) of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.5348**	**0.7061**	**0.7787**	**0.5692**	**0.5808**
ResNet50	0.4156	0.5948	0.6866	0.4526	0.4669
DINOv2	0.4444	0.6285	0.7179	0.4822	0.4978
UNI	0.5156	0.6880	0.7661	0.5504	0.5583
Phikon	0.4716	0.6528	0.7341	0.5071	0.5141
CTransPath	0.4424	0.6255	0.7102	0.4755	0.4929
HIPT	0.0738	0.0852	0.1115	0.0723	0.0712

Supplementary Data Table 25: **Patch-level cervical cell image retrieval on the HiCervix dataset (26 classes).** The preprocessed HiCervix dataset retains its official train-test split (56,754:8,051), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in HiCervix, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels (26 classes) of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.7260 (0.7127-0.7394)**	**0.9149 (0.9106-0.9191)**	**0.9624 (0.9577-0.9670)**	**0.7590 (0.7548-0.7633)**	**0.7644 (0.7554-0.7733)**
ResNet50	0.7056 (0.6954-0.7157)	0.9048 (0.8992-0.9104)	0.9571 (0.9543-0.9599)	0.7373 (0.7325-0.7421)	0.7422 (0.7348-0.7496)
DINOv2	0.6585 (0.6502-0.6667)	0.8817 (0.8747-0.8886)	0.9418 (0.9345-0.9491)	0.6949 (0.6824-0.7075)	0.7069 (0.6934-0.7203)
UNI	0.6889 (0.6783-0.6994)	0.8962 (0.8900-0.9025)	0.9497 (0.9447-0.9548)	0.7183 (0.7098-0.7267)	0.7273 (0.7198-0.7348)
Phikon	0.7064 (0.6907-0.7221)	0.9021 (0.8991-0.9051)	0.9523 (0.9487-0.9559)	0.7397 (0.7316-0.7479)	0.7504 (0.7416-0.7593)
CTransPath	0.7006 (0.6863-0.7149)	0.8972 (0.8862-0.9082)	0.9462 (0.9387-0.9538)	0.7242 (0.7167-0.7317)	0.7335 (0.7280-0.7390)
HIPT	0.3575 (0.1041-0.6109)	0.7769 (0.6403-0.9135)	0.8319 (0.6622-1.0020)	0.4397 (0.2386-0.6408)	0.4625 (0.2728-0.6521)

Supplementary Data Table 26: **Patch-level cervical cell image retrieval on the DSCC dataset (3 classes).** Since there is no official train-test split, the preprocessed DSCC dataset was randomly divided into candidate and query sets in a 4:1 ratio (12,406:3,103) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in DSCC, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.8920 (0.8875-0.8965)**	0.9525 (0.9478-0.9572)	0.9679 (0.9631-0.9728)	0.8840 (0.8785-0.8895)	0.8749 (0.8627-0.8872)
ResNet50	0.8716 (0.8682-0.8751)	0.9415 (0.9399-0.9431)	0.9606 (0.9570-0.9643)	0.8637 (0.8566-0.8707)	0.8500 (0.8428-0.8571)
DINOv2	0.8439 (0.8371-0.8506)	0.9323 (0.9295-0.9352)	0.9541 (0.9512-0.9571)	0.8430 (0.8372-0.8488)	0.8318 (0.8267-0.8370)
UNI	0.8902 (0.8842-0.8962)	**0.9534 (0.9506-0.9562)**	**0.9701 (0.9680-0.9721)**	**0.8917 (0.8890-0.8943)**	**0.8878 (0.8854-0.8902)**
Phikon	0.8892 (0.8805-0.8980)	0.9509 (0.9444-0.9573)	0.9669 (0.9647-0.9691)	0.8877 (0.8798-0.8955)	0.8822 (0.8768-0.8876)
CTransPath	0.8780 (0.8706-0.8853)	0.9453 (0.9402-0.9504)	0.9610 (0.9571-0.9648)	0.8761 (0.8700-0.8823)	0.8650 (0.8568-0.8731)
HIPT	0.6825 (0.6717-0.6933)	0.8626 (0.8536-0.8716)	0.9073 (0.8996-0.9151)	0.7041 (0.6942-0.7139)	0.7069 (0.6950-0.7189)

Supplementary Data Table 27: **Patch-level cervical cell image retrieval on the CRIC dataset (6 classes).** Since there is no official train-test split, the preprocessed CRIC dataset was randomly divided into candidate and query sets in a 4:1 ratio (9,227:2,307) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in CRIC, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	0.6430 (0.6292-0.6569)	**0.8634 (0.8330-0.8939)**	**0.9172 (0.8938-0.9406)**	**0.6624 (0.6172-0.7075)**	**0.6452 (0.6029-0.6874)**
ResNet50	0.5462 (0.5110-0.5814)	0.7914 (0.7741-0.8087)	0.8731 (0.8630-0.8832)	0.5957 (0.5506-0.6408)	0.6054 (0.5699-0.6409)
DINOv2	0.4839 (0.4441-0.5236)	0.7194 (0.6680-0.7707)	0.8194 (0.7786-0.8601)	0.4978 (0.4605-0.5352)	0.4957 (0.4498-0.5416)
UNI	0.6097 (0.5867-0.6326)	0.8269 (0.8159-0.8379)	0.8989 (0.8747-0.9232)	0.6269 (0.5856-0.6682)	0.6333 (0.5897-0.6770)
Phikon	**0.6484 (0.6010-0.6957)**	0.8376 (0.7835-0.8917)	0.9075 (0.8637-0.9514)	0.6366 (0.5952-0.6779)	0.6355 (0.6010-0.6700)
CTransPath	0.5516 (0.5017-0.6015)	0.8032 (0.7759-0.8306)	0.8871 (0.8701-0.9041)	0.5763 (0.5591-0.5936)	0.5731 (0.5409-0.6053)
HIPT	0.3280 (0.1677-0.4882)	0.5871 (0.4190-0.7552)	0.6796 (0.5184-0.8407)	0.3473 (0.1646-0.5300)	0.3516 (0.1961-0.5071)

Supplementary Data Table 28: **Patch-level cervical cell image retrieval on the Herlev dataset (7 classes).** Since there is no official train-test split, the preprocessed Herlev dataset was randomly divided into candidate and query sets in a 4:1 ratio (731:186) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in Herlev, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.9650 (0.9596-0.9705)**	**0.9884 (0.9818-0.9951)**	**0.9936 (0.9883-0.9989)**	**0.9660 (0.9579-0.9741)**	**0.9633 (0.9570-0.9696)**
ResNet50	0.9143 (0.9024-0.9262)	0.9591 (0.9471-0.9711)	0.9749 (0.9697-0.9800)	0.9175 (0.9055-0.9295)	0.9165 (0.9016-0.9314)
DINOv2	0.8586 (0.8501-0.8671)	0.9522 (0.9401-0.9643)	0.9739 (0.9658-0.9819)	0.8697 (0.8576-0.8818)	0.8599 (0.8499-0.8698)
UNI	0.9456 (0.9370-0.9542)	0.9798 (0.9734-0.9863)	0.9857 (0.9792-0.9923)	0.9478 (0.9401-0.9555)	0.9421 (0.9359-0.9483)
Phikon	0.9500 (0.9431-0.9569)	0.9793 (0.9743-0.9843)	0.9869 (0.9819-0.9920)	0.9424 (0.9307-0.9540)	0.9355 (0.9234-0.9475)
CTransPath	0.9397 (0.9312-0.9481)	0.9734 (0.9690-0.9778)	0.9845 (0.9813-0.9877)	0.9313 (0.9197-0.9429)	0.9197 (0.9061-0.9333)
HIPT	0.4424 (0.0314-0.8533)	0.4857 (0.0010-0.9704)	0.4985 (-0.0079-1.0050)	0.4426 (0.0313-0.8540)	0.4379 (0.0344-0.8415)

Supplementary Data Table 29: **Patch-level cervical cell image retrieval on the SIPaKMeD dataset (5 classes).** Since there is no official train-test split, the preprocessed SIPaKMeD dataset was randomly divided into candidate and query sets in a 4:1 ratio

(3,237:812) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in SIPaKMeD, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.9987 (0.9946-1.0030)**	**1.0000 (1.0000-1.0000)**	**1.0000 (1.0000-1.0000)**	0.9897 (0.9733-1.0060)	0.9936 (0.9780-1.0090)
ResNet50	0.9928 (0.9821-1.0030)	0.9969 (0.9934-1.0000)	0.9990 (0.9961-1.0020)	**0.9948 (0.9870-1.0030)**	0.9928 (0.9855-1.0000)
DINOv2	0.9897 (0.9796-0.9998)	0.9979 (0.9944-1.0010)	**1.0000 (1.0000-1.0000)**	0.9866 (0.9780-0.9952)	0.9753 (0.9638-0.9867)
UNI	0.9948 (0.9884-1.0010)	0.9990 (0.9961-1.0020)	**1.0000 (1.0000-1.0000)**	0.9938 (0.9824-1.0050)	**0.9938 (0.9868-1.0010)**
Phikon	0.9918 (0.9810-1.0020)	**1.0000 (1.0000-1.0000)**	**1.0000 (1.0000-1.0000)**	0.9928 (0.9871-0.9985)	0.9876 (0.9819-0.9934)
CTransPath	0.9907 (0.9854-0.9961)	0.9969 (0.9934-1.0000)	**1.0000 (1.0000-1.0000)**	0.9825 (0.9767-0.9882)	0.9845 (0.9695-0.9995)
HIPT	0.9103 (0.8924-0.9282)	0.9691 (0.9590-0.9792)	0.9835 (0.9721-0.9950)	0.8907 (0.8703-0.9112)	0.8753 (0.8487-0.9018)

Supplementary Data Table 30: **Patch-level cervical cell image retrieval on the LBC dataset (4 classes).** Since there is no official train-test split, the LBC dataset was randomly divided into candidate and query sets in a 4:1 ratio (768:194) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all large-sized images in LBC, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Every large-sized image feature was obtained through max pooling the features of all 256×256 crops. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.8517 (0.8203-0.8831)**	**0.9500 (0.9257-0.9743)**	**0.9700 (0.9498-0.9902)**	**0.8483 (0.8226-0.8741)**	**0.8350 (0.8082-0.8618)**
ResNet50	0.7800 (0.7502-0.8098)	0.9117 (0.8978-0.9255)	0.9583 (0.9390-0.9777)	0.8067 (0.7563-0.8570)	0.8067 (0.7698-0.8435)
DINOv2	0.7617 (0.7337-0.7896)	0.9150 (0.8591-0.9709)	0.9483 (0.9115-0.9852)	0.7683 (0.7169-0.8198)	0.7733 (0.7428-0.8039)
UNI	0.8233 (0.7928-0.8539)	0.9350 (0.9138-0.9562)	0.9617 (0.9429-0.9805)	0.8233 (0.7976-0.8491)	0.8050 (0.7702-0.8398)
Phikon	0.7900 (0.7441-0.8359)	0.9150 (0.8812-0.9488)	0.9417 (0.9051-0.9782)	0.8117 (0.7725-0.8508)	0.7850 (0.7403-0.8297)
CTransPath	0.7717 (0.7217-0.8216)	0.9100 (0.8629-0.9571)	0.9517 (0.9318-0.9716)	0.7683 (0.7083-0.8284)	0.7550 (0.7126-0.7974)
HIPT	0.5233 (0.2962-0.7504)	0.7567 (0.5516-0.9617)	0.8750 (0.7793-0.9707)	0.5383 (0.3132-0.7634)	0.5517 (0.3181-0.7853)

Supplementary Data Table 31: **Patch-level cervical cell image retrieval on the BMT dataset (3 classes).** Since there is no official train-test split, the preprocessed BMT dataset was randomly divided into candidate and query sets in a 4:1 ratio (480:120) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all large-sized images in BMT, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Every large-sized image feature were obtained through max pooling features of all 256×256 crops. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	0.8190 (0.6909-0.9472)	**1.0000 (1.0000-1.0000)**	**1.0000 (1.0000-1.0000)**	0.8381 (0.7852-0.8910)	0.8286 (0.7492-0.9079)
ResNet50	0.8286 (0.7060-0.9512)	0.9810 (0.9486-1.0130)	0.9810 (0.9486-1.0130)	**0.9143 (0.8495-0.9791)**	**0.8476 (0.7981-0.8971)**
DINOv2	0.6667 (0.5732-0.7602)	0.9238 (0.8709-0.9767)	0.9810 (0.9486-1.0130)	0.6571 (0.6077-0.7066)	0.6762 (0.5991-0.7533)
UNI	0.7619 (0.6684-0.8554)	0.9810 (0.9281-1.0340)	**1.0000 (1.0000-1.0000)**	0.8286 (0.7612-0.8960)	0.7905 (0.7376-0.8434)
Phikon	**0.8476 (0.7981-0.8971)**	0.9810 (0.9486-1.0130)	**1.0000 (1.0000-1.0000)**	0.7905 (0.7111-0.8698)	0.7429 (0.6354-0.8503)
CTransPath	0.7048 (0.6076-0.8019)	0.9524 (0.8800-1.0250)	0.9810 (0.9281-1.0340)	0.7238 (0.6361-0.8115)	0.7524 (0.6876-0.8172)
HIPT	0.4952 (0.3468-0.6436)	0.6667 (0.4893-0.8440)	0.8190 (0.7133-0.9248)	0.5333 (0.4839-0.5828)	0.5238 (0.4514-0.5962)

Supplementary Data Table 32: **Patch-level cervical cell image retrieval on the CERVIX93 dataset (3 classes).** Since there is no official train-test split, the CERVIX93 dataset was randomly divided into candidate and query sets in a 4:1 ratio (72:21) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all large-sized images in CERVIX93, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Every large-sized image feature was obtained through max pooling the features of all 256×256 crops. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.6538**	**0.8230**	**0.8891**	**0.6838**	**0.7002**
ResNet50	0.6148	0.8173	0.8880	0.6431	0.6640
DINOv2	0.5854	0.7992	**0.8891**	0.6261	0.6555
UNI	0.6086	0.7958	0.8756	0.6380	0.6488
Phikon	0.6284	0.8071	0.8846	0.6544	0.6816
CTransPath	0.6007	0.7992	0.8761	0.6335	0.6550
HIPT	0.5707	0.7986	0.8744	0.6216	0.6623

Supplementary Data Table 33: **Patch-level other cell image retrieval on the UFSC OCPap v1 dataset (3 classes).** The preprocessed UFSC OCPap v1 dataset retains its official train-test split (11,580:1,768), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in UFSC OCPap v1, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.9481**	**0.9942**	**0.9981**	**0.9635**	**0.9596**
ResNet50	0.9077	0.9808	0.9962	0.9250	0.9288
DINOv2	0.9269	0.9904	0.9942	0.9327	0.9442
UNI	0.9346	0.9846	0.9923	0.9558	0.9558
Phikon	0.8981	0.9635	0.9731	0.9231	0.9231
CTransPath	0.9346	0.9750	0.9904	0.9288	0.9385
HIPT	0.6962	0.8692	0.9077	0.7212	0.7154

Supplementary Data Table 34: **Patch-level other cell image retrieval on the UMID dataset (3 classes).** The preprocessed UMID dataset retains its official train-test split (2,459:520), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in UMID, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	0.9481	0.9481	0.9610	0.9351	0.9351
ResNet50	0.8312	0.9351	0.9481	0.8442	0.8571
DINOv2	0.8442	0.8701	0.8831	0.8571	0.8312
UNI	0.9221	0.9221	0.9610	0.9221	0.9351
Phikon	0.9351	0.9610	**0.9870**	0.9351	0.9351
CTransPath	**0.9610**	**0.9740**	0.9740	**0.9740**	**0.9740**
HIPT	0.7143	0.7662	0.7792	0.7143	0.7403

Supplementary Data Table 35: **Patch-level other cell image retrieval on the LISC dataset (3 classes).** The preprocessed LISC dataset retains its official train-test split (2,186:77), with the training set as candidate set and the test set as query set. Every frozen encoder extracted features of all images in LISC, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. The retrieval accuracy for all encoders is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$. The best value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	**0.5668 (0.5636, 0.5700)**	**0.7816 (0.7796, 0.7836)**	**0.8537 (0.8518, 0.8557)**	**0.6386 (0.6347, 0.6425)**	**0.6572 (0.6531, 0.6613)**
ResNet50	0.4017 (0.3995, 0.4040)	0.6388 (0.6378, 0.6399)	0.7369 (0.7362, 0.7377)	0.4676 (0.4648, 0.4703)	0.4933 (0.4915, 0.4952)
DINOv2	0.4053 (0.4036, 0.4069)	0.6524 (0.6495, 0.6553)	0.7508 (0.7486, 0.7530)	0.4812 (0.4789, 0.4835)	0.5096 (0.5078, 0.5113)
UNI	0.4818 (0.4793, 0.4843)	0.7089 (0.7057, 0.7121)	0.7957 (0.7930, 0.7984)	0.5518 (0.5490, 0.5545)	0.5727 (0.5695, 0.5759)
Phikon	0.5001 (0.4975, 0.5026)	0.7253 (0.7232, 0.7274)	0.8070 (0.8047, 0.8092)	0.5699 (0.5685, 0.5714)	0.5897 (0.5862, 0.5932)
CTransPath	0.4858 (0.4844, 0.4872)	0.7089 (0.7068, 0.7111)	0.7940 (0.7920, 0.7960)	0.5504 (0.5485, 0.5523)	0.5691 (0.5673, 0.5709)
HIPT	0.2333 (0.1508, 0.3159)	0.4171 (0.2071, 0.6272)	0.5167 (0.2435, 0.7899)	0.2659 (0.1607, 0.3710)	0.2835 (0.1662, 0.4008)

Supplementary Data Table 36: **Patch-level other cell image retrieval on the Bone Marrow Cytomorphology MLL Helmholtz Fraunhofer (BM) dataset (21 classes).** Since there is no official train-test split, the preprocessed BM dataset was randomly divided into candidate and query sets in a 4:1 ratio (137,091:34,281) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in BM, and measured feature similarity using L2 distance to retrieve several

most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	0.8749 (0.8737, 0.8762)	0.9440 (0.9415, 0.9465)	0.9572 (0.9548, 0.9597)	**0.9078 (0.9053, 0.9103)**	**0.9074 (0.9056, 0.9091)**
ResNet50	0.8373 (0.8348, 0.8398)	0.9185 (0.9173, 0.9196)	0.9373 (0.9358, 0.9389)	0.8690 (0.8671, 0.8708)	0.8680 (0.8663, 0.8698)
DINOv2	0.8343 (0.8323, 0.8363)	0.9192 (0.9169, 0.9215)	0.9389 (0.9366, 0.9411)	0.8685 (0.8669, 0.8702)	0.8677 (0.8658, 0.8696)
UNI	0.8633 (0.8607, 0.8659)	0.9392 (0.9379, 0.9404)	0.9541 (0.9522, 0.9560)	0.8986 (0.8964, 0.9007)	0.8988 (0.8969, 0.9007)
Phikon	0.8549 (0.8531, 0.8567)	0.9344 (0.9324, 0.9364)	0.9505 (0.9482, 0.9528)	0.8894 (0.8879, 0.8910)	0.8886 (0.8866, 0.8906)
CTransPath	**0.8799 (0.8788, 0.8811)**	**0.9462 (0.9448, 0.9476)**	**0.9605 (0.9584, 0.9625)**	0.9042 (0.9027, 0.9057)	0.9030 (0.9011, 0.9050)
HIPT	0.7022 (0.4664, 0.9380)	0.7849 (0.4917, 1.0780)	0.8148 (0.5361, 1.0940)	0.7332 (0.4759, 0.9906)	0.7357 (0.4767, 0.9948)

Supplementary Data Table 37: **Patch-level other cell image retrieval on the CSF dataset (16 classes).** Since there is no official train-test split, the CSF dataset was randomly divided into candidate and query sets in a 4:1 ratio (98,540:24,641) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in CSF, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Acc@1	Acc@3	Acc@5	MVAcc@5	MVAcc@10
CytoFM	0.9523 (0.9477, 0.9569)	0.9786 (0.9728, 0.9845)	0.9838 (0.9788, 0.9888)	0.9564 (0.9520, 0.9608)	0.9528 (0.9475, 0.9582)
ResNet50	0.8578 (0.8473, 0.8684)	0.9397 (0.9315, 0.9478)	0.9566 (0.9480, 0.9652)	0.8794 (0.8668, 0.8919)	0.8792 (0.8735, 0.8849)
DINOv2	0.9089 (0.9016, 0.9162)	0.9665 (0.9638, 0.9693)	0.9760 (0.9729, 0.9791)	0.9240 (0.9145, 0.9335)	0.9269 (0.9159, 0.9378)
UNI	**0.9690 (0.9627, 0.9754)**	**0.9847 (0.9822, 0.9872)**	**0.9888 (0.9866, 0.9909)**	**0.9660 (0.9576, 0.9745)**	**0.9626 (0.9575, 0.9678)**
Phikon	0.9580 (0.9522, 0.9638)	0.9820 (0.9808, 0.9832)	0.9879 (0.9856, 0.9902)	0.9587 (0.9541, 0.9634)	0.9557 (0.9516, 0.9598)
CTransPath	0.9432 (0.9396, 0.9469)	0.9710 (0.9660, 0.9760)	0.9778 (0.9739, 0.9816)	0.9502 (0.9468, 0.9536)	0.9480 (0.9436, 0.9525)
HIPT	0.7851 (0.7669, 0.8032)	0.9112 (0.8956, 0.9269)	0.9429 (0.9347, 0.9510)	0.8315 (0.8156, 0.8474)	0.8475 (0.8280, 0.8670)

Supplementary Data Table 38: **Patch-level other cell image retrieval on the NeuNN dataset (3 classes).** Since there is no official train-test split, the preprocessed NeuNN dataset was randomly divided into candidate and query sets in a 4:1 ratio (4,481:1,124) five times, while maintaining consistent class distribution. Every frozen encoder extracted features of all images in NeuNN, and measured feature similarity using L2 distance to retrieve several most similar candidate images for each query. Retrieval success was determined by comparing the class labels of the query and candidate images. For all encoders, the average retrieval accuracy across five splits is reported using five metrics: Acc@K for $K \in \{1,3,5\}$ and MVAcc@N for $N \in \{5,10\}$, along with 95% confidence intervals. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	0.6550 (0.6525-0.6575)	0.5101 (0.5072, 0.5131)	**0.6032 (0.5949, 0.6116)**	0.7755 (0.7665, 0.7844)
DINOv2	0.6313 (0.6267-0.6359)	0.4851 (0.4805, 0.4897)	0.5668 (0.5592, 0.5744)	0.7845 (0.7786, 0.7905)
UNI	**0.6595 (0.6551-0.6640)**	**0.5149 (0.5101, 0.5197)**	0.6007 (0.5913, 0.6102)	**0.7871 (0.7759, 0.7982)**
Phikon	0.6555 (0.6508-0.6603)	0.5099 (0.5050, 0.5148)	0.5972 (0.5877, 0.6068)	0.7870 (0.7795, 0.7946)
HIPT	0.6171 (0.6136-0.6206)	0.4729 (0.4707, 0.4750)	0.5696 (0.5608, 0.5783)	0.7551 (0.7386, 0.7716)

Supplementary Data Table 39: **Cervical cell image segmentation on the APACS23 dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. The APACS23 dataset follows an official train-test split (2,227:1,338), based on which five repeated experiments were conducted. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	**0.7115 (0.7105, 0.7125)**	**0.5541 (0.5528, 0.5554)**	**0.6257 (0.6231, 0.6283)**	0.8314 (0.8297, 0.8331)
DINOv2	0.6915 (0.6893, 0.6937)	0.5308 (0.5282, 0.5334)	0.5940 (0.5905, 0.5976)	0.8358 (0.8342, 0.8375)
UNI	0.7055 (0.7040, 0.7070)	0.5470 (0.5454, 0.5487)	0.6180 (0.6151, 0.6209)	0.8298 (0.8252, 0.8344)
Phikon	0.7039 (0.7026, 0.7053)	0.5452 (0.5437, 0.5467)	0.6154 (0.6123, 0.6186)	0.8294 (0.8255, 0.8333)
HIPT	0.6868 (0.6853, 0.6884)	0.5254 (0.5235, 0.5272)	0.5845 (0.5823, 0.5867)	**0.8414 (0.8396, 0.8431)**

Supplementary Data Table 40: **Cervical cell image segmentation on the CNSeg (PatchSeg) dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. The CNSeg (PatchSeg) dataset follows an official train-test split (3,010:477), based on which five repeated

experiments were conducted. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	**0.8266 (0.8260, 0.8271)**	**0.7068 (0.7059, 0.7076)**	**0.7699 (0.7693, 0.7706)**	0.8989 (0.8973, 0.9005)
DINOv2	0.8176 (0.8168, 0.8184)	0.6943 (0.6933, 0.6954)	0.7514 (0.7496, 0.7532)	0.9044 (0.9020, 0.9067)
UNI	0.8227 (0.8216, 0.8238)	0.7015 (0.6999, 0.7030)	0.7585 (0.7555, 0.7615)	0.9058 (0.9034, 0.9082)
Phikon	0.8207 (0.8201, 0.8213)	0.6986 (0.6978, 0.6994)	0.7569 (0.7544, 0.7594)	0.9038 (0.9003, 0.9074)
HIPT	0.8142 (0.8132, 0.8152)	0.6898 (0.6884, 0.6911)	0.7433 (0.7406, 0.7460)	**0.9088 (0.9056, 0.9120)**

Supplementary Data Table 41: **Cervical cell image segmentation on the CNSeg (ClusterSeg) dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. The CNSeg (ClusterSeg) dataset follows an official train-test split (1,795:567), based on which five repeated experiments were conducted. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	**0.9588 (0.9567, 0.9609)**	**0.9213 (0.9176, 0.9251)**	**0.9517 (0.9400, 0.9634)**	0.9671 (0.9562, 0.9780)
DINOv2	0.9378 (0.9246, 0.9510)	0.8840 (0.8610, 0.9071)	0.9043 (0.8632, 0.9454)	0.9771 (0.9538, 1.0004)
UNI	0.9542 (0.9495, 0.9589)	0.9129 (0.9045, 0.9213)	0.9375 (0.9189, 0.9562)	0.9725 (0.9600, 0.9850)
Phikon	0.9439 (0.9376, 0.9502)	0.8954 (0.8846, 0.9062)	0.9052 (0.8903, 0.9201)	0.9887 (0.9824, 0.9950)
HIPT	0.9238 (0.9006, 0.9469)	0.8601 (0.8198, 0.9004)	0.8668 (0.8174, 0.9162)	**0.9921 (0.9816, 1.0025)**

Supplementary Data Table 42: **Cervical cell image segmentation on the ISBI (Cytoplasm) dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. The ISBI (Cytoplasm) dataset follows an official train-test split (45:900), based on which five repeated experiments were conducted. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	0.8373 (0.8295, 0.8451)	0.7237 (0.7125, 0.7349)	0.8920 (0.8729, 0.9111)	0.7975 (0.7743, 0.8206)
DINOv2	0.7974 (0.7777, 0.8171)	0.6702 (0.6446, 0.6957)	0.8925 (0.8583, 0.9267)	0.7417 (0.6901, 0.7934)
UNI	0.8294 (0.8205, 0.8382)	0.7122 (0.6992, 0.7252)	**0.9006 (0.8850, 0.9163)**	0.7762 (0.7554, 0.7969)
Phikon	**0.8378 (0.8228, 0.8527)**	**0.7249 (0.7034, 0.7463)**	0.8618 (0.8070, 0.9165)	**0.8327 (0.7615, 0.9039)**
HIPT	0.8195 (0.8164, 0.8227)	0.6989 (0.6947, 0.7031)	0.8362 (0.8039, 0.8685)	0.8245 (0.7872, 0.8617)

Supplementary Data Table 43: **Cervical cell image segmentation on the ISBI (Nuclei) dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. The ISBI (Nuclei) dataset follows an official train-test split (45:900), based on which five repeated experiments were conducted. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	**0.6835 (0.6635, 0.7036)**	**0.5243 (0.5026, 0.5460)**	**0.7422 (0.7265, 0.7580)**	0.6516 (0.6179, 0.6853)
DINOv2	0.6621 (0.6464, 0.6779)	0.5004 (0.4843, 0.5166)	0.7006 (0.6736, 0.7277)	0.6594 (0.6230, 0.6957)
UNI	0.6734 (0.6596, 0.6872)	0.5127 (0.4992, 0.5262)	0.7350 (0.7136, 0.7564)	0.6415 (0.6058, 0.6773)
Phikon	0.6792 (0.6692, 0.6891)	0.5185 (0.5085, 0.5286)	0.7328 (0.7072, 0.7584)	0.6458 (0.6213, 0.6702)
HIPT	0.6635 (0.6504, 0.6766)	0.5021 (0.4876, 0.5166)	0.6971 (0.6622, 0.7320)	**0.6728 (0.6456, 0.7000)**

Supplementary Data Table 44: **Cervical cell image segmentation on the BTTFA dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. Because of the absence of an official train-test split, the BTTFA dataset was randomly divided into training and test sets in a 4:1 ratio (83:21) five times, and a separate segmentation experiment was conducted on each split. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	0.8272 (0.8136, 0.8409)	0.7119 (0.6926, 0.7312)	0.8071 (0.7894, 0.8248)	0.8609 (0.8523, 0.8695)
DINOv2	0.8138 (0.8009, 0.8267)	0.6933 (0.6756, 0.7110)	0.7935 (0.7764, 0.8106)	0.8501 (0.8423, 0.8579)
UNI	**0.8294 (0.8166, 0.8422)**	**0.7147 (0.6966, 0.7328)**	**0.8121 (0.7978, 0.8265)**	0.8590 (0.8491, 0.8690)
Phikon	0.8265 (0.8129, 0.8401)	0.7108 (0.6918, 0.7298)	0.8024 (0.7874, 0.8173)	**0.8646 (0.8542, 0.8749)**
HIPT	0.8037 (0.7924, 0.8150)	0.6796 (0.6648, 0.6944)	0.7892 (0.7693, 0.8091)	0.8349 (0.8271, 0.8427)

Supplementary Data Table 45: **Cervical cell image segmentation on the CCEDD dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. Because of the absence of an official train-test split, the CCEDD dataset was randomly divided into training and test sets in a 4:1 ratio (548:138) five times, and a separate segmentation experiment was conducted on each split. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	Dice	IoU	Precision	Recall
CytoFM	**0.8997 (0.8986, 0.9009)**	**0.8200 (0.8184, 0.8216)**	**0.8707 (0.8688, 0.8726)**	0.9348 (0.9325, 0.9370)
DINOv2	0.8972 (0.8963, 0.8980)	0.8159 (0.8147, 0.8171)	0.8623 (0.8613, 0.8633)	0.9392 (0.9382, 0.9402)
UNI	0.8993 (0.8985, 0.9001)	0.8192 (0.8179, 0.8204)	0.8674 (0.8654, 0.8694)	0.9378 (0.9359, 0.9397)
Phikon	0.8986 (0.8978, 0.8993)	0.8180 (0.8170, 0.8190)	0.8667 (0.8648, 0.8685)	0.9369 (0.9347, 0.9391)
HIPT	0.8955 (0.8954, 0.8957)	0.8132 (0.8130, 0.8135)	0.8578 (0.8560, 0.8596)	**0.9412 (0.9389, 0.9434)**

Supplementary Data Table 46: **Cervical cell image segmentation on the DSCC dataset.** Every frozen encoder was integrated with ViT-Adapter to extract multi-scale features, which were subsequently fed into the Mask2Former decoder for segmentation. Because of the absence of an official train-test split, the DSCC dataset was randomly divided into training and test sets in a 4:1 ratio (12,407:3,102) five times, and a separate segmentation experiment was conducted on each split. For all encoders, the average segmentation performance and 95% confidence intervals across the five experiments are reported using four metrics: Dice, IoU, Precision, and Recall. The best average value for each metric is bolded.

Encoder	AP	AP75	AP50
CytoFM	**20.36 (20.20, 20.51)**	**19.02 (17.88, 20.16)**	40.41 (40.07, 40.75)
ResNet50	16.99 (16.74, 17.25)	12.98 (11.83, 14.13)	37.34 (36.20, 38.48)
DINOv2	19.68 (19.45, 19.91)	15.81 (14.79, 16.83)	41.47 (40.56, 42.37)
UNI	19.93 (19.41, 20.44)	17.82 (16.68, 18.96)	40.49 (39.64, 41.35)
Phikon	17.78 (16.95, 18.61)	14.59 (13.09, 16.09)	37.66 (36.71, 38.61)
CTransPath	19.92 (19.23, 20.62)	16.02 (14.86, 17.19)	**41.69 (40.58, 42.79)**
HIPT	19.24 (18.69, 19.80)	17.51 (15.77, 19.24)	38.88 (38.39, 39.38)

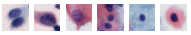
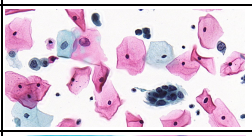
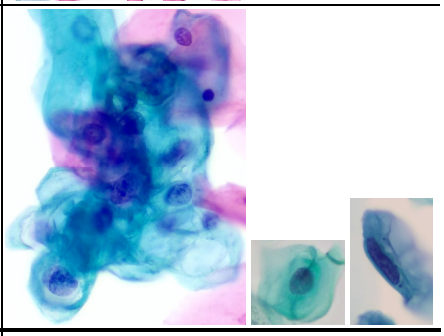
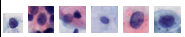
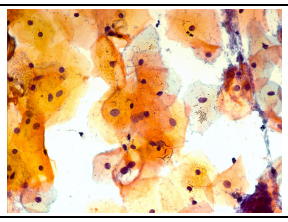
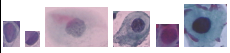
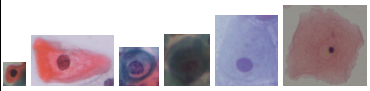
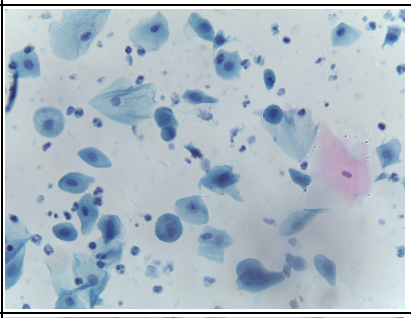
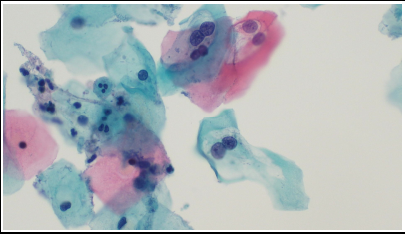
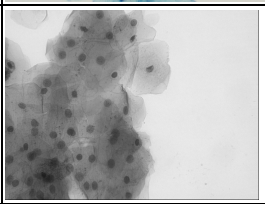
Supplementary Data Table 47: **Cervical cell image detection on the CDetector dataset.** Detection was based on the YOLOF framework, where every frozen encoder served as backbone to extract single-level features. The CDetector dataset follows an official train-test split (6,666:744), based on which five repeated experiments were conducted. For all encoders, the average detection precision and 95% confidence intervals across the five experiments are reported using three metrics: AP, AP75, and AP50. The best average value for each metric is bolded.

Encoder	AP	AP75	AP50
CytoFM	**20.61 (17.43, 23.80)**	**23.14 (16.65, 29.63)**	**37.10 (31.35, 42.86)**
ResNet50	13.26 (12.68, 13.84)	12.01 (9.904, 14.12)	27.41 (24.71, 30.10)
DINOv2	18.10 (16.82, 19.39)	17.55 (14.93, 20.16)	35.37 (32.03, 38.71)
UNI	17.95 (16.41, 19.48)	19.78 (15.91, 23.64)	32.44 (30.05, 34.83)
Phikon	15.52 (13.65, 17.39)	14.56 (12.56, 16.57)	30.81 (24.56, 37.07)
CTransPath	15.88 (13.79, 17.97)	13.72 (11.65, 15.80)	34.21 (27.02, 41.40)
HIPT	19.46 (17.93, 20.99)	20.03 (16.59, 23.46)	34.17 (31.48, 36.87)

Supplementary Data Table 48: **Cervical cell image detection on the HMCHH-TCT-CellDet dataset.** Detection was based on the YOLOF framework, where every frozen encoder served as backbone to extract single-level features. Because of the absence of an official train-test split, the HMCHH-TCT-CellDet dataset was randomly divided into training and test sets in a 4:1 ratio (6,429:1,608) five times, and a separate detection experiment was conducted on each split. For all encoders, the average detection precision and 95% confidence intervals across the five experiments are reported using three metrics: AP, AP75, and AP50. The best average value for each metric is bolded.

Hyperparameter	Value
Image size	224
Patch size	16
Depth of transformer	24
Transformer block class	NestedTensorBlock
Attention class	MemEffAttention
Number of attention heads	16
FFN layer	swiglufused
Embedding dimension	1024
Ratio of MLP hidden dim to embedding dim	4.0
MLP activation layer	GELU
Number of splitting block sequence into block chunk units	4
Stochastic depth rate	0.0
Global crop size	224
Global crop scale	0.5-1.0
Number of global crops	2
Local crop size	96
Local crop scale	0.25-0.5
Number of local crops	8
Mask ratio range	0.1-0.5
Separate head	True
DINO loss weight for global crops	1.0
DINO loss weight for local crops	0.5
IBOT loss weight	0.2
Koleo loss weight	0.05
Centering	centering
Batch size	114
Max iterations	564,813
Warmup iterations	56,481
Warmup teacher temperature iterations	169,443
Freeze last layer iterations	45,185
Warmup learning rate (start)	0.0
Warmup learning rate (summit)	5e-4
Learning rate schedule	Cosine
Learning rate (end)	1e-6
Weight decay schedule	Cosine
Weight decay (start)	0.04
Weight decay (end)	0.2
Teacher momentum schedule	Cosine
Teacher momentum (start)	0.994
Teacher momentum (end)	1.0
Warmup teacher temperature (start)	0.04
Teacher temperature (summit and end)	0.07
Gradient clipping max norm	3.0
AdamW β	(0.9, 0.999)
Automatic mixed precision	FP16

Supplementary Data Table 49: **DINOv2 hyperparameters used for CytoFM pretraining.** The batch size refers to the total batch size across GPUs.

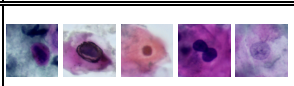
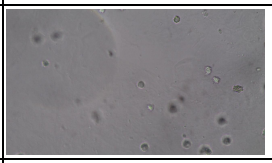
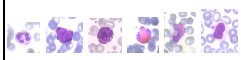
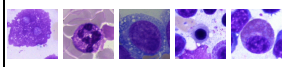
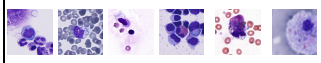
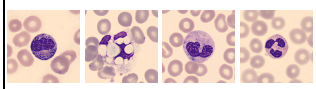
Dataset	Size	Image size	Class label	Link	Example
LDCC	198,952 train/test	128×128	**3 classes** (positive/ negative/ junk)	https://biod.whu.edu.cn/sjj.htm	
CDetector	7,410 train/test	unfixed	**11 classes** (ASC-US/ ASC-H/ LSIL/ HSIL/ SCC/ AGC/ trichomonas/ candida/ flora/ herps/ actinomyces)	https://github.com/CVIU-CSU/ComparisonDetector	
HiCervix	40,229 train/val/ test	unfixed	**4 classes** (negative/ ASC/ AGC/ microbe) **26 classes** (Normal/ ECC/ RPC/ MPC/ PG/ Atrophy/ EMC/ HCG/ ASC-US/ LSIL/ ASC-H/ HSIL/ SCC/ AGC/ AGC-NOS/ AGC-FN/ ADC/ AGC-ECC-NOS/ AGC-EMC-NOS/ ADC-ECC/ ADC-EMC/ FUNGI/ ACTINO/ TRI/ HSV/ CC)	https://zenodo.org/records/11087263	
DSCC	15,509	128×128	**3 classes** (NORMAL/ SIL/ ASC)	https://biod.whu.edu.cn/sjj.htm	
CRIC	400	1,376×1,020	**6 classes** (NILM/ ASC-US/ LSIL/ ASC-H/ HSIL/ SCC)	https://figshare.com/collections/CRIC_Cervix_Cell_Classification/4960286/2	
Herlev	917	unfixed	**7 classes** (Squamous cell carcinoma in situ intermediate/ Mild squamous non-keratinizing dysplasia/ Moderate squamous non-keratinizing dysplasia/ Columnar epithelial/ Intermediate squamous epithelial/ Superficial squamous epithelial/ Severe squamous non-keratinizing dysplasia)	https://mde-lab.aegean.gr/index.php/downloads/	
SIPaKMeD	4,049	unfixed	**5 classes** (Superficial-Intermediate/ Parabasal/ Koilocytotic/ Metaplastic/ Dyskeratotic)	https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed	
LBC	962	2,048×1,536	**4 classes** (NILM/ LSIL/ HSIL/ SCC)	https://data.mendeley.com/datasets/zddtpgzv63/4	
BMT	600	3,264×1,840 or 1,920×1,080	**3 classes** (NILM/ LSIL/ HSIL)	https://doi.org/10.7303/syn55259257	
CERVIX93	93 train/test	1,280×960	**3 classes** (Negative/ LSIL/ HSIL)	https://github.com/parham-ap/cytology_dataset	

Supplementary Data Table 50: **Summary of publicly available cervical cell image classification and retrieval datasets.**

Column “Size” reports the actual size of the downloaded dataset. Column “Example” displays sample images scaled down to 10% of their original size to visually illustrate the differences in resolution between different datasets. The CDetector dataset is originally designed for multi-class detection tasks, but it can be cropped to obtain a classification dataset through the cell detection boxes. The HiCervix dataset provides three levels of class labels, including the simplest 4-class and the most complex 26-class schemes. The CRIC dataset is also designed for multi-class detection tasks, but it can be cropped to obtain a classification dataset through the nucleus center points. The LDCC, CDetector, HiCervix, and CERVIX93 datasets all provide their official data splits, including train, validation (for HiCervix), and test sets, but the remaining datasets provide no official splits.

Dataset	Preprocessing	New size	Split	Class label
LDCC	Remove images from the junk class and then resize all images to 224×224.	85,239	train (68,190) test (17,049)	**2 classes** (positive/ negative)
CDetector	Crop square crops from the original images based on the detection box labels, with the side length equal to the longer side of each detection box plus 10 pixels, and then resize them to 224×224.	50,954	train (45,897) test (5,057)	**11 classes** (ASC-US/ ASC-H/ LSIL/ HSIL/ SCC/ AGC/ trichomonas/ candida/ flora/ herps/ actinomyces)
HiCervix	For each image, if either side is smaller than 800 pixels, retain its aspect ratio and resize it to 224×224 with white padding. Otherwise, randomly crop four 800×800 crops for the training and validation sets, or center-crop one 800×800 crop for the test set, and then resize them to 224×224.	64,805	train (56,754) test (8,051)	**4 classes** (negative/ ASC/ AGC/ microbe) **26 classes** (Normal/ ECC/ RPC/ MPC/ PG/ Atrophy/ EMC/ HCG/ ASC-US/ LSIL/ ASC-H/ HSIL/ SCC/ AGC/ AGC-NOS/ AGC-FN/ ADC/ AGC-ECC-NOS/ AGC-EMC-NOS/ ADC-ECC/ ADC-EMC/ FUNGI/ ACTINO/ TRI/ HSV/ CC)
DSCC	Resize all images to 224×224.	15,509	4:1(train:test)	**3 classes** (NORMAL/ SIL/ ASC)
CRIC	Crop square crops from the original images based on the nucleus center point labels, with the side length equal to 90 pixels, and then resize them to 224×224.	11,534	4:1(train:test)	**6 classes** (NILM/ ASC-US/ LSIL/ ASC-H/ HSIL/ SCC)
Herlev	Retain the aspect ratio of all images and resize them to 224×224 with white padding.	917	4:1(train:test)	**7 classes** (Squamous cell carcinoma in situ intermediate/ Mild squamous non-keratinizing dysplasia/ Moderate squamous non-keratinizing dysplasia/ Columnar epithelial/ Intermediate squamous epithelial/ Superficial squamous epithelial/ Severe squamous non-keratinizing dysplasia)
SIPaKMeD	Re-crop square crops from the original large-sized images based on the original small crops, with the side length equal to the longer side of each original small crops, and then resize them to 224×224.	4,049	4:1(train:test)	**5 classes** (Superficial-Intermediate/ Parabasal/ Koilocytotic/ Metaplastic/ Dyskeratotic)
LBC	—	962	4:1(train:test)	**4 classes** (NILM/ LSIL/ HSIL/ SCC)
BMT	Resize all images to 1,920×1,080.	600	4:1(train:test)	**3 classes** (NILM/ LSIL/ HSIL)
CERVIX93	—	93	4:1(train:test)	**3 classes** (Negative/ LSIL/ HSIL)

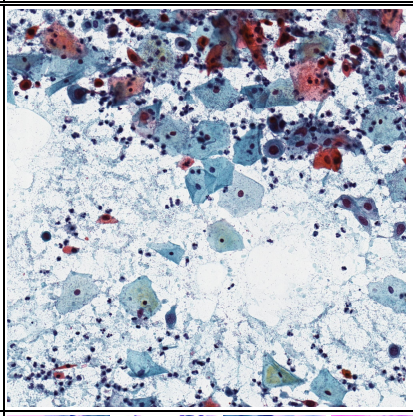
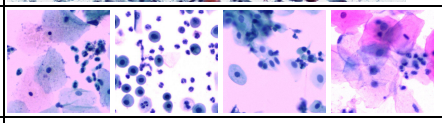
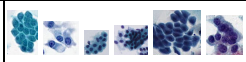
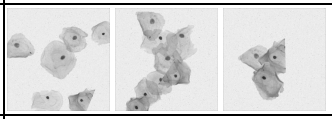
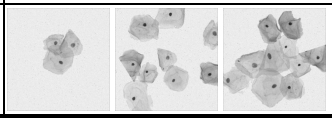
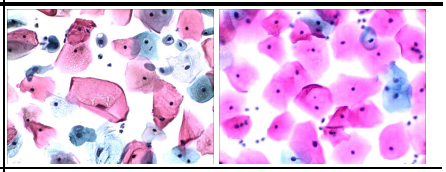
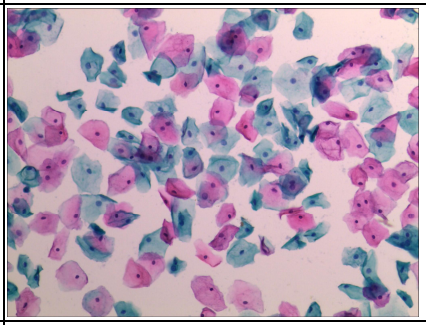
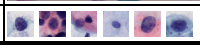
Supplementary Data Table 51: **Preprocessing of publicly available cervical cell image classification and retrieval datasets.** Due to the significant variation in image size within the HiCervix dataset, specialized preprocessing procedure was conducted for images with different sizes, which was inspired by the dataset source paper. The SIPaKMeD dataset provides not only 4,049 original small cropped images, but also the large-sized source images from which the crops were derived. To obtain a new classification dataset with more complete cellular morphological information, a new set of cropped images was created by re-cropping the large-sized images. For all datasets with official splits except for CERVIX93, their splits were retained, with the validation sets merged into their training sets. Considering that the CERVIX93 dataset contains only a limited number of images, retaining its original split would not allow for a reliable performance evaluation of every encoder. Therefore, for CERVIX93 as well as other datasets without official splits, each was randomly divided into training and test sets in a 4:1 ratio, while maintaining consistent class distribution between the two sets. Such random partitioning operation was performed five times using different random seeds. In retrieval tasks, the training set served as the candidate set and the test set served as the query set.

Dataset	Type	Size	Image size	Class label	Link	Example
UFSC OCPap v1	oral cell	17,148 train/val/ test	256×256	**5 classes** (Abnormal/ Healthy/ Blood/ Dividing/ Out of Focus)	https://arquivos.ufsc.br/d/5035aec3c24f421a95d0/	
UMID	urine cell	588 train/val/ test	1,280×720	**3 classes** (RBC/ pus/ epithelial cell)	https://github.com/dipamgoswami/UMID-Urine-Microscopic-Image-Dataset/tree/main	
LISC	WBC	2,263 train/test	unfixed	**5 classes** (Basophil/ Eosinophil/ Lymphocyte/ Monocyte/ Neutrophil)	https://github.com/nimaadmed/WBC_Feature?tab=readme-ov-file	
Bone Marrow Cytomorphology MLL Helmholtz Fraunhofer	bone marrow cell	171,372	250×250	**21 classes** (Band neutrophils/ Segmented neutrophils/ Lymphocytes/ Monocytes/ Eosinophils/ Basophils/ Metamyelocytes/ Myelocytes/ Promyelocytes/ Blasts/ Plasma cells/ Smudge cells/ Other cells/ Artefacts/ Proerythroblasts/ Not identifiable/ Erythroblasts/ Hairy cells/ Abnormal eosinophils/ Immature lymphocytes/ Faggot cells)	https://www.cancerimagingarchive.net/collection/bone-marrow-cytomorphology_mll_helmholtz_fraunhofer/	
CSF	cerebrospinal fluid cell	123,181	224×224	**16 classes** (Lymphocyte/ Activated lymphocyte/ Plasma cell/ Monocyte/ Activated monocyte/ Macrophage/ Erythrophage/ Haemosiderophage/ Neutrophilic granulocyte/ Eosinophilic granulocyte/ Tumour cell/ Cell shadow/ Shrunken cell/ Mitosis/ Haematoidin crystal/ Erythrocyte)	https://doi.org/10.5281/zenodo.6543147	
NeuNN	Neutrophil	5,605	360×363	**7 classes** (Normal neutrophils/ Hypogranulated neutrophils/ Cryoglobulins/ Döhle bodies/ Howell-Jolly body-like inclusions/ Green-blue inclusions of death/ Bacteria)	https://data.mendeley.com/datasets/rh3jw43hjs/1	

Supplementary Data Table 52: **Summary of publicly available other cell image retrieval datasets.** Column "Size" reports the actual size of the downloaded dataset. Column "Example" displays sample images scaled down to 10% of their original size to visually illustrate the differences in resolution between different datasets. The UFSC OCPap v1 dataset provides three sub-datasets for oral nuclei detection, segmentation, and classification, while this work only used the classification dataset. The UMID dataset is originally designed for multi-class detection tasks, but it can be cropped to obtain a classification dataset through the indication of cell detection boxes. The LISC dataset does not provide an official download link. Consequently, this work employed a version of the dataset generated by cropping and augmenting the original LISC dataset images, as described in the paper "New segmentation and feature extraction algorithm for classification of white blood cells in peripheral smear images". The UFSC OCPap v1, UMID, and LISC datasets provide their official data splits, including train, validation (for UFSC OCPap v1 and UMID), and test sets, but the remaining datasets provide no official splits.

Dataset	Preprocessing	New size	Split	Class label
UFSC OCPap v1	Remove images from the dividing and outoffocus classes and then resize all images to 224×224.	13,348	retrieval (11,580) query (1,768)	**3 classes** (abnormal/ healthy/ blood)
UMID	Crop rectangular crops from the original images based on the detection box labels, with 10 pixels added to each side, and then retain the aspect ratio of all crops and resize them to 224×224 with white padding.	2,979	retrieval (2,459) query (520)	**3 classes** (RBC/ pus cell or WBC/ epithelial cell)
LISC	Retain the aspect ratio of all images and resize them to 224×224 with white padding.	2,263	retrieval (2,186) query (77)	**5 classes** (Basophil/ Eosinophil/ Lymphocyte/ Monocyte/ Neutrophil)
Bone Marrow Cytomorphology MLL Helmholtz Fraunhofer	Resize all images to 224×224.	171,372	4:1(retrieval:query)	**21 classes** (Band neutrophils/ Segmented neutrophils/ Lymphocytes/ Monocytes/ Eosinophils/ Basophils/ Metamyelocytes/ Myelocytes/ Promyelocytes/ Blasts/ Plasma cells/ Smudge cells/ Other cells/ Artefacts/ Not identifiable/ Proerythroblasts/ Erythroblasts/ Hairy cells/ Abnormal eosinophils/ Immature lymphocytes/ Faggot cells)
CSF	—	123,181	4:1(retrieval:query)	**16 classes** (lymphocyte/ activated lymphocyte/ plasma cell/ monocyte/ activated monocyte/ macrophage/ erythrophage/ haemosiderophage/ neutrophilic granulocyte/ eosinophilic granulocyte/ tumour cell/ cell shadow/ shrunken cell/ mitosis/ haematoidin crystal/ erythrocyte)
NeuNN	Resize all images to 224×224.	5,605	4:1(retrieval:query)	**7 classes** (Normal neutrophils/ Hypogranulated neutrophils/ Cryoglobulins/ Döhle bodies/ Howell-Jolly body-like inclusions/ Green-blue inclusions of death/ Bacteria)

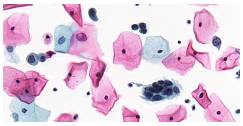
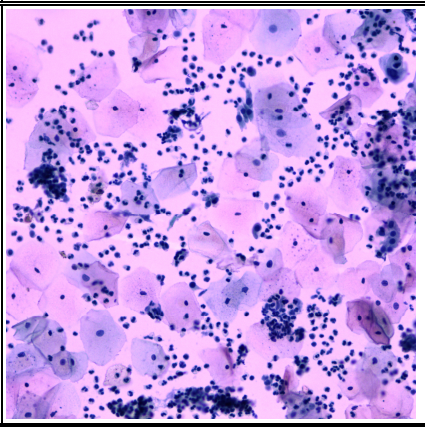
Supplementary Data Table 53: **Preprocessing of publicly available other cell image retrieval datasets.** For all datasets with official splits, their splits were retained, with the validation sets merged into their training sets. In retrieval tasks, the training set served as the candidate set, and the test set served as the query set. For all datasets without official splits, each was randomly divided into retrieval and query sets in a 4:1 ratio, while maintaining consistent class distribution between the two sets. Such random partitioning operation was performed five times using different random seeds.

Dataset	Size	Image size	Mask label	Link	Example
APACS23	3,565 train/test	2,000×2,000	Cytoplasm mask	https://osf.io/cka2f/overview	
CNSeg (PatchSeg)	3,487 train/test	512×512	Nucleus edge points	https://www.kaggle.com/datasets/zhaojing0522/cervical-nucleus-segmentation	
CNSeg (ClusterSeg)	2,362 train/test	unfixed	Nucleus edge points	https://www.kaggle.com/datasets/zhaojing0522/cervical-nucleus-segmentation	
ISBI (Cytoplasm)	945 train/test	512×512	Cytoplasm mask	https://cs.adelaide.edu.au/~carneiro/isbi14_challenge/dataset.html	
ISBI (Nuclei)	945 train/test	512×512	Nucleus mask	https://cs.adelaide.edu.au/~carneiro/isbi14_challenge/dataset.html	
BTTFA	104	1,024×768	Nucleus mask	https://data.mendeley.com/datasets/jks43dkjj7/1	
CCEDD	686	2,048×1,536	Cytoplasm edge	https://github.com/nachifur/LLPC	
DSCC	15,509	128×128	Nucleus mask	https://biod.whu.edu.cn/sjj.htm	

Supplementary Data Table 54: **Summary of publicly available cervical cell image segmentation datasets.** Column "Size" reports the actual size of the downloaded dataset. Column "Example" displays sample images scaled down to 10% of their original size to visually illustrate the differences in resolution between different datasets. The CNSeg dataset contains patch segmentation dataset (PatchSeg) and cluster segmentation dataset (ClusterSeg). This work considered these two datasets as distinct datasets. The ISBI dataset contains 945 synthetic images and provides corresponding nucleus and cytoplasm masks. This work also considered the sythetic images with different masks as distinct datasets. The APACS23, CNSeg (PatchSeg), CNSeg (ClusterSeg), ISBI (Cytoplasm) and ISBI (Nuclei) datasets provide their official data splits, including training and test sets, but the remaining datasets provide no official splits.

Dataset	Preprocessing	New size	Split	Mask label
APACS23	Resize all images and cytoplasm masks to 224×224.	3,565	train (2,227) test (1,338)	Cytoplasm mask
CNSeg (PatchSeg)	Resize all images to 224×224 and convert nucleus edge points to mask images and resize them to 224×224.	3,487	train (3,010) test (477)	Nucleus mask
CNSeg (ClusterSeg)	Resize all images to 224×224 and convert nucleus edge points to mask images and resize them to 224×224.	2,362	train (1,795) test (567)	Nucleus mask
ISBI (Cytoplasm)	Resize all images and cytoplasm masks to 224×224.	945	train (45) test (900)	Cytoplasm mask
ISBI (Nuclei)	Resize all images and nucleus masks to 224×224.	945	train (45) test (900)	Nucleus mask
BTTFA	Resize all images and nucleus masks to 224×224.	104	4:1(train:test)	Nucleus mask
CCEDD	Resize all images to 224×224 and convert cytoplasm edge to mask images and resize them to 224×224.	686	4:1(train:test)	Cytoplasm mask
DSCC	Resize all images and nucleus masks to 224×224.	15,509	4:1(train:test)	Nucleus mask

Supplementary Data Table 55: **Preprocessing of publicly available cervical cell image segmentation datasets.** For all datasets with official splits, their splits were retained. For all datasets without official splits, each was randomly divided into training and test sets in a 4:1 ratio. Such random partitioning operation was performed five times using different random seeds.

Dataset	Size	Image size	detection label	Link	Example
CDetector	7,410 train/test	unfixed	cell detection box with cell class (ASC-US/ ASC-H/ LSIL/ HSIL/ SCC/ AGC/ trichomonas/ candida/ flora/ herps/ actinomyces)	https://github.com/CVIU-CSU/ComparisonDetector	
HMCHH-TCT-CellDet	8,037	2,048×2,048	cell detection box with cell class (abnormal/ dysbacteriosis/ candida/ actinomycetes/ trichomonad)	https://springernature.figshare.com/articles/dataset/A_large_annotated_cervical_cytology_images_dataset_for_AI_models_to_aid_cervical_cancer_screening/27901206	

Supplementary Data Table 56: **Summary of publicly available cervical cell image detection datasets.** Column "Size" reports the actual size of the downloaded dataset. Column "Example" displays sample images scaled down to 10% of their original size to visually illustrate the differences in resolution between different datasets. The CDetector dataset provides official training set and test set, while the HMCHH-TCT-CellDet dataset provides no official splits. Therefore, in the preprocessing stage, the split of CDetector dataset was retained and the HMCHH-TCT-CellDet dataset was divided five times with different random seeds in a 4:1 ratio. The size of the images in both datasets remained unchanged, only the detection box labels were converted to COCO format.