Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.913 (0.866, 0.960)**	**0.913 (0.867, 0.960)**	**0.943 (0.897, 0.979)**
PLIP	0.507 (0.500, 0.523)	0.348 (0.263, 0.438)	0.688 (0.600, 0.771)
BiomedCLIP	0.553 (0.510, 0.600)	0.465 (0.358, 0.557)	0.852 (0.784, 0.912)
OpenAICLIP	0.500 (0.500, 0.500)	0.333 (0.255, 0.418)	0.643 (0.544, 0.726)

Supplementary Data Table 1: Zero-shot classification with prompt ensembling on BRCA subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.902 (0.860, 0.938)**	**0.903 (0.863, 0.938)**	**0.975 (0.955, 0.992)**
PLIP	0.804 (0.755, 0.850)	0.797 (0.740, 0.850)	0.956 (0.932, 0.975)
BiomedCLIP	0.791 (0.737, 0.840)	0.789 (0.733, 0.842)	0.924 (0.893, 0.950)
OpenAICLIP	0.347 (0.333, 0.363)	0.194 (0.141, 0.252)	0.673 (0.626, 0.719)

Supplementary Data Table 2: Zero-shot classification with prompt ensembling on RCC subtyping ($n = 225$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.907 (0.859, 0.948)**	**0.907 (0.860, 0.947)**	**0.962 (0.933, 0.984)**
PLIP	0.787 (0.720, 0.849)	0.786 (0.720, 0.847)	0.838 (0.768, 0.899)
BiomedCLIP	0.780 (0.713, 0.843)	0.776 (0.704, 0.840)	0.877 (0.819, 0.922)
OpenAICLIP	0.553 (0.503, 0.604)	0.478 (0.382, 0.570)	0.603 (0.514, 0.689)

Supplementary Data Table 3: Zero-shot classification with prompt ensembling on NSCLC subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Cohen's κ	Balanced accuracy	Weighted F1
CONCH	**0.200 (0.128, 0.277)**	**0.425 (0.305, 0.516)**	**0.314 (0.236, 0.400)**
PLIP	0.079 (0.014, 0.142)	0.326 (0.250, 0.404)	0.198 (0.129, 0.267)
BiomedCLIP	0.009 (0.000, 0.041)	0.259 (0.204, 0.327)	0.032 (0.005, 0.071)
OpenAICLIP	0.004 (0.000, 0.021)	0.204 (0.200, 0.214)	0.045 (0.015, 0.087)

Supplementary Data Table 4: Zero-shot classification with prompt ensembling on DHMC LUAD ($n = 143$) in terms of Cohen's κ, balanced accuracy, and weighted F1 score. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Quadratic weighted κ	Balanced accuracy	Weighted F1
CONCH	**0.690 (0.667, 0.713)**	**0.600 (0.584, 0.615)**	0.398 (0.375, 0.421)
PLIP	0.180 (0.147, 0.212)	0.306 (0.293, 0.320)	0.097 (0.085, 0.109)
BiomedCLIP	0.550 (0.517, 0.582)	0.484 (0.465, 0.503)	**0.438 (0.415, 0.461)**
OpenAICLIP	0.091 (0.067, 0.118)	0.289 (0.275, 0.302)	0.183 (0.167, 0.201)

Supplementary Data Table 5: Zero-shot classification with prompt ensembling on SICAP ($n = 2,122$) in terms of quadratic weighted Cohen's κ, balanced accuracy, and weighted F1 score. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.791 (0.782, 0.800)**	**0.803 (0.794, 0.813)**	**0.979 (0.977, 0.981)**
PLIP	0.674 (0.665, 0.683)	0.687 (0.676, 0.698)	0.944 (0.941, 0.947)
BiomedCLIP	0.553 (0.542, 0.564)	0.533 (0.521, 0.545)	0.924 (0.921, 0.928)
OpenAICLIP	0.271 (0.262, 0.280)	0.247 (0.236, 0.258)	0.781 (0.777, 0.786)

Supplementary Data Table 6: Zero-shot classification with prompt ensembling on CRC100k ($n = 7,180$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.719 (0.706, 0.731)**	**0.705 (0.692, 0.718)**	**0.877 (0.870, 0.885)**
PLIP	0.624 (0.609, 0.638)	0.636 (0.622, 0.650)	0.790 (0.780, 0.801)
BiomedCLIP	0.616 (0.603, 0.628)	0.575 (0.559, 0.589)	0.824 (0.815, 0.833)
OpenAICLIP	0.296 (0.290, 0.303)	0.196 (0.183, 0.209)	0.496 (0.484, 0.508)

Supplementary Data Table 7: Zero-shot classification with prompt ensembling on WSSS4LUAD ($n = 4,693$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.620 (0.600, 0.680)**	**0.570 (0.525, 0.658)**	**0.870 (0.827, 0.890)**
PLIP	0.530 (0.520, 0.543)	0.419 (0.381, 0.461)	0.701 (0.676, 0.720)
BiomedCLIP	0.550 (0.513, 0.600)	0.465 (0.383, 0.577)	0.726 (0.688, 0.758)
OpenAICLIP	0.500 (0.500, 0.500)	0.333 (0.333, 0.333)	0.564 (0.543, 0.579)

Supplementary Data Table 8: Zero-shot classification without prompt ensembling on BRCA subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.816 (0.791, 0.831)**	**0.816 (0.783, 0.832)**	**0.964 (0.959, 0.969)**
PLIP	0.749 (0.693, 0.764)	0.741 (0.679, 0.756)	0.944 (0.939, 0.947)
BiomedCLIP	0.656 (0.644, 0.682)	0.639 (0.624, 0.668)	0.900 (0.890, 0.910)
OpenAICLIP	0.333 (0.333, 0.338)	0.167 (0.167, 0.189)	0.575 (0.562, 0.589)

Supplementary Data Table 9: Zero-shot classification without prompt ensembling on RCC subtyping ($n = 225$) in terms of balanced accuracy, weighted F1 score, and ROC AUC (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.780 (0.687, 0.807)**	**0.775 (0.671, 0.804)**	**0.917 (0.885, 0.941)**
PLIP	0.663 (0.567, 0.700)	0.640 (0.532, 0.700)	0.791 (0.763, 0.812)
BiomedCLIP	0.683 (0.620, 0.727)	0.670 (0.561, 0.721)	0.842 (0.805, 0.864)
OpenAICLIP	0.503 (0.500, 0.517)	0.371 (0.333, 0.396)	0.542 (0.513, 0.583)

Supplementary Data Table 10: Zero-shot classification without prompt ensembling on NSCLC subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Cohen's κ	Balanced accuracy	Weighted F1
CONCH	**0.111 (0.092, 0.137)**	**0.314 (0.291, 0.346)**	**0.231 (0.211, 0.258)**
PLIP	0.020 (0.014, 0.031)	0.231 (0.222, 0.246)	0.114 (0.069, 0.186)
BiomedCLIP	0.023 (0.014, 0.040)	0.253 (0.240, 0.273)	0.067 (0.044, 0.092)
OpenAICLIP	0.005 (0.000, 0.017)	0.194 (0.188, 0.205)	0.119 (0.053, 0.211)

Supplementary Data Table 11: Zero-shot classification without prompt ensembling on DHMC LUAD ($n = 143$) in terms of Cohen's κ, balanced accuracy, and weighted F1 score (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Quadratic weighted κ	Balanced accuracy	Weighted F1
CONCH	0.303 (0.248, 0.396)	0.339 (0.325, 0.396)	0.241 (0.207, 0.300)
PLIP	0.171 (0.127, 0.243)	0.287 (0.268, 0.324)	0.226 (0.185, 0.268)
BiomedCLIP	**0.396 (0.350, 0.433)**	**0.375 (0.351, 0.389)**	**0.352 (0.318, 0.371)**
OpenAICLIP	0.002 (0.000, 0.022)	0.250 (0.250, 0.256)	0.140 (0.098, 0.187)

Supplementary Data Table 12: Zero-shot classification without prompt ensembling on SICAP ($n = 2,122$) in terms of quadratic weighted Cohen's κ, balanced accuracy, and weighted F1 score (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.566 (0.533, 0.598)**	**0.542 (0.496, 0.589)**	**0.901 (0.893, 0.918)**
PLIP	0.520 (0.485, 0.540)	0.517 (0.484, 0.576)	0.879 (0.869, 0.898)
BiomedCLIP	0.422 (0.398, 0.436)	0.372 (0.346, 0.408)	0.859 (0.845, 0.868)
OpenAICLIP	0.234 (0.211, 0.250)	0.185 (0.149, 0.202)	0.727 (0.715, 0.733)

Supplementary Data Table 13: Zero-shot classification without prompt ensembling on CRC100k ($n = 7,180$) in terms of balanced accuracy, weighted F1 score, and ROC AUC (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.598 (0.560, 0.650)**	**0.590 (0.547, 0.627)**	**0.795 (0.769, 0.827)**
PLIP	0.462 (0.439, 0.492)	0.408 (0.386, 0.475)	0.710 (0.654, 0.760)
BiomedCLIP	0.512 (0.474, 0.550)	0.452 (0.408, 0.518)	0.747 (0.734, 0.775)
OpenAICLIP	0.333 (0.328, 0.333)	0.195 (0.189, 0.239)	0.551 (0.494, 0.608)

Supplementary Data Table 14: Zero-shot classification without prompt ensembling on WSSS4LUAD ($n = 4,693$) in terms of balanced accuracy, weighted F1 score, and ROC AUC (median from 50 sets of randomly sampled prompts). Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.867 (0.811, 0.914)**	**0.865 (0.805, 0.914)**	**0.941 (0.893, 0.975)**
PLIP	0.787 (0.723, 0.847)	0.783 (0.711, 0.846)	0.842 (0.771, 0.906)
BiomedCLIP	0.833 (0.781, 0.886)	0.829 (0.770, 0.889)	0.928 (0.874, 0.966)
OpenAICLIP	0.827 (0.768, 0.888)	0.824 (0.762, 0.886)	0.903 (0.852, 0.953)
ResNet50 (tr)	0.767 (0.699, 0.834)	0.764 (0.692, 0.832)	0.822 (0.750, 0.889)

Supplementary Data Table 15: Supervised classification on BRCA subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.942 (0.909, 0.969)**	**0.942 (0.909, 0.969)**	**0.992 (0.981, 0.999)**
PLIP	0.924 (0.887, 0.956)	0.924 (0.884, 0.956)	0.989 (0.979, 0.996)
BiomedCLIP	0.924 (0.886, 0.957)	0.924 (0.883, 0.956)	0.985 (0.969, 0.996)
OpenAICLIP	0.893 (0.850, 0.930)	0.894 (0.849, 0.930)	0.978 (0.957, 0.992)
ResNet50 (tr)	0.916 (0.878, 0.951)	0.916 (0.880, 0.951)	0.974 (0.951, 0.990)

Supplementary Data Table 16: Supervised classification on RCC subtyping ($n = 225$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.933 (0.892, 0.970)**	**0.933 (0.893, 0.967)**	**0.983 (0.964, 0.996)**
PLIP	0.907 (0.856, 0.947)	0.907 (0.859, 0.947)	0.963 (0.931, 0.986)
BiomedCLIP	0.927 (0.880, 0.967)	0.927 (0.880, 0.967)	0.978 (0.955, 0.993)
OpenAICLIP	0.907 (0.858, 0.948)	0.907 (0.854, 0.947)	0.961 (0.930, 0.983)
ResNet50 (tr)	0.827 (0.768, 0.885)	0.826 (0.764, 0.881)	0.916 (0.869, 0.956)

Supplementary Data Table 17: Supervised classification on NSCLC subtyping ($n = 150$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Quadratic weighted kappa	Balanced accuracy	Weighted F1
CONCH	0.833 (0.813, 0.851)	**0.711 (0.688, 0.732)**	0.745 (0.726, 0.764)
PLIP	0.762 (0.739, 0.784)	0.589 (0.570, 0.609)	0.657 (0.637, 0.678)
BiomedCLIP	0.719 (0.695, 0.740)	0.549 (0.528, 0.567)	0.628 (0.606, 0.647)
OpenAICLIP	0.704 (0.676, 0.732)	0.599 (0.579, 0.617)	0.662 (0.641, 0.682)
CTransPath	**0.835 (0.817, 0.851)**	0.678 (0.658, 0.700)	**0.747 (0.728, 0.766)**

Supplementary Data Table 18: Supervised classification on SICAP ($n = 2,122$) in terms of quadratic weighted Cohen's κ, balanced accuracy, and weighted F1 score. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.938 (0.931, 0.944)**	**0.955 (0.950, 0.960)**	**0.995 (0.994, 0.996)**
PLIP	0.879 (0.872, 0.888)	0.890 (0.884, 0.897)	0.989 (0.988, 0.991)
BiomedCLIP	0.898 (0.890, 0.905)	0.923 (0.916, 0.929)	0.991 (0.990, 0.992)
OpenAICLIP	0.884 (0.877, 0.891)	0.897 (0.890, 0.904)	0.992 (0.990, 0.993)
CTransPath	**0.938 (0.932, 0.944)**	0.950 (0.945, 0.955)	0.994 (0.993, 0.995)

Supplementary Data Table 19: Supervised classification on CRC100k ($n = 7,180$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.687 (0.646, 0.729)**	**0.717 (0.681, 0.754)**	**0.977 (0.969, 0.983)**
PLIP	0.580 (0.537, 0.625)	0.626 (0.585, 0.665)	0.966 (0.958, 0.973)
BiomedCLIP	0.543 (0.498, 0.590)	0.559 (0.515, 0.601)	0.961 (0.952, 0.969)
OpenAICLIP	0.509 (0.467, 0.555)	0.559 (0.515, 0.600)	0.947 (0.936, 0.957)
CTransPath	0.619 (0.576, 0.667)	0.616 (0.577, 0.656)	0.970 (0.963, 0.976)
KimiaNet	0.398 (0.354, 0.442)	0.446 (0.405, 0.486)	0.922 (0.907, 0.935)
ResNet50 (tr)	0.402 (0.357, 0.444)	0.381 (0.338, 0.424)	0.925 (0.911, 0.936)

Supplementary Data Table 20: Supervised classification on EBRAINS ($n = 573$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Balanced accuracy	Weighted F1	ROC AUC
CONCH	**0.371 (0.331, 0.409)**	**0.359 (0.320, 0.399)**	**0.941 (0.934, 0.948)**
PLIP	0.121 (0.088, 0.154)	0.087 (0.063, 0.113)	0.796 (0.778, 0.812)
BiomedCLIP	0.201 (0.176, 0.232)	0.124 (0.096, 0.152)	0.866 (0.851, 0.880)
OpenAICLIP	0.064 (0.054, 0.073)	0.029 (0.016, 0.045)	0.623 (0.604, 0.642)

Supplementary Data Table 21: Zeroshot classification on Ebrains ($n = 573$) in terms of balanced accuracy, weighted F1 score, and ROC AUC. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.400 (0.362, 0.435)**	**0.776 (0.748, 0.803)**	**0.888 (0.865, 0.910)**	**0.688 (0.665, 0.710)**
PLIP	0.056 (0.042, 0.072)	0.198 (0.170, 0.226)	0.307 (0.276, 0.341)	0.187 (0.165, 0.210)
BiomedCLIP	0.161 (0.137, 0.189)	0.405 (0.368, 0.441)	0.553 (0.517, 0.591)	0.373 (0.346, 0.402)
OpenAICLIP	0.018 (0.009, 0.027)	0.046 (0.032, 0.061)	0.084 (0.065, 0.103)	0.049 (0.038, 0.062)

Supplementary Data Table 22: Zero-shot text-to-image retrieval performance for Source A ($n = 797$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.402 (0.369, 0.433)**	**0.792 (0.764, 0.818)**	**0.883 (0.859, 0.905)**	**0.692 (0.670, 0.714)**
PLIP	0.065 (0.049, 0.082)	0.222 (0.192, 0.251)	0.320 (0.287, 0.351)	0.202 (0.179, 0.225)
BiomedCLIP	0.171 (0.146, 0.197)	0.439 (0.403, 0.472)	0.582 (0.547, 0.615)	0.397 (0.373, 0.423)
OpenAICLIP	0.010 (0.004, 0.018)	0.049 (0.035, 0.064)	0.073 (0.055, 0.092)	0.044 (0.033, 0.056)

Supplementary Data Table 23: Zero-shot image-to-text retrieval performance for Source A ($n = 797$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.171 (0.151, 0.188)**	**0.437 (0.412, 0.462)**	**0.562 (0.539, 0.587)**	**0.390 (0.372, 0.409)**
PLIP	0.020 (0.013, 0.027)	0.075 (0.063, 0.089)	0.132 (0.116, 0.148)	0.076 (0.066, 0.086)
BiomedCLIP	0.100 (0.086, 0.115)	0.264 (0.242, 0.284)	0.353 (0.328, 0.375)	0.239 (0.221, 0.256)
OpenAICLIP	0.009 (0.005, 0.014)	0.036 (0.027, 0.044)	0.052 (0.042, 0.063)	0.032 (0.026, 0.040)

Supplementary Data Table 24: Zero-shot text-to-image retrieval performance for Source B ($n = 1,755$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.164 (0.146, 0.181)**	**0.403 (0.379, 0.426)**	**0.530 (0.508, 0.553)**	**0.366 (0.347, 0.384)**
PLIP	0.024 (0.017, 0.031)	0.077 (0.065, 0.090)	0.124 (0.110, 0.140)	0.075 (0.065, 0.085)
BiomedCLIP	0.099 (0.085, 0.113)	0.251 (0.233, 0.272)	0.359 (0.336, 0.380)	0.236 (0.220, 0.252)
OpenAICLIP	0.005 (0.002, 0.008)	0.027 (0.019, 0.035)	0.043 (0.034, 0.053)	0.025 (0.019, 0.031)

Supplementary Data Table 25: Zero-shot image-to-text retrieval performance for Source B ($n = 1,755$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.087 (0.047, 0.133)**	**0.247 (0.180, 0.313)**	**0.387 (0.313, 0.460)**	**0.240 (0.189, 0.293)**
PLIP	0.027 (0.007, 0.053)	0.080 (0.040, 0.127)	0.180 (0.120, 0.247)	0.096 (0.060, 0.136)
BiomedCLIP	0.053 (0.020, 0.093)	0.193 (0.133, 0.267)	0.313 (0.240, 0.387)	0.187 (0.140, 0.240)
OpenAICLIP	0.020 (0.000, 0.047)	0.060 (0.027, 0.100)	0.107 (0.060, 0.160)	0.062 (0.033, 0.098)

Supplementary Data Table 26: Zero-shot text-to-image retrieval performance for TCGA LUAD ($n = 165$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Recall@1	Recall@5	Recall@10	Mean Recall
CONCH	**0.067 (0.030, 0.103)**	**0.188 (0.133, 0.248)**	**0.291 (0.224, 0.364)**	**0.182 (0.135, 0.230)**
PLIP	0.018 (0.000, 0.042)	0.079 (0.042, 0.121)	0.115 (0.067, 0.164)	0.071 (0.040, 0.101)
BiomedCLIP	0.048 (0.018, 0.085)	0.158 (0.103, 0.212)	0.291 (0.224, 0.364)	0.166 (0.125, 0.210)
OpenAICLIP	0.006 (0.000, 0.018)	0.042 (0.012, 0.079)	0.097 (0.055, 0.139)	0.048 (0.026, 0.075)

Supplementary Data Table 27: Zero-shot image-to-text retrieval performance for TCGA LUAD ($n = 165$) in terms of Recall@K for $K \in \{1, 5, 10\}$ and mean recall over K. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Dice score	Precision	Recall
CONCH	**0.601 (0.530, 0.675)**	**0.672 (0.630, 0.722)**	**0.751 (0.696, 0.803)**
PLIP	0.549 (0.496, 0.605)	0.605 (0.556, 0.656)	0.644 (0.595, 0.694)
BiomedCLIP	0.484 (0.452, 0.520)	0.536 (0.505, 0.569)	0.557 (0.519, 0.598)
OpenAICLIP	0.367 (0.314, 0.426)	0.599 (0.573, 0.629)	0.605 (0.571, 0.639)

Supplementary Data Table 28: Zero-shot segmentation performance on SICAP ($n = 31$) in terms of the macro-averaged Dice score as well as precision and recall. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model Name	Dice	Precision	Recall
CONCH	**0.615 (0.597, 0.633)**	**0.663 (0.650, 0.675)**	**0.709 (0.693, 0.726)**
PLIP	0.426 (0.411, 0.443)	0.526 (0.513, 0.537)	0.541 (0.528, 0.554)
BiomedCLIP	0.446 (0.430, 0.462)	0.581 (0.569, 0.592)	0.601 (0.588, 0.615)
OpenAICLIP	0.367 (0.351, 0.381)	0.492 (0.481, 0.503)	0.511 (0.498, 0.524)

Supplementary Data Table 29: Zero-shot segmentation performance on DigestPath ($n = 250$) in terms of the macro-averaged Dice score as well as precision and recall. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	METEOR	ROUGE
CONCH	**0.195 (0.182, 0.211)**	**0.214 (0.199, 0.230)**
GIT-base	0.122 (0.115, 0.130)	0.135 (0.125, 0.145)
GIT-large	0.125 (0.117, 0.134)	0.153 (0.143, 0.163)

Supplementary Data Table 30: Captioning performance with fine-tuning on Source A (train $n = 558$, validation $n = 77$, test $n = 162$) in terms of METEOR and ROUGE. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Model name	Quadratic weighted kappa	Balanced accuracy	Weighted F1
100% Training Labels			
CONCH	**0.872 (0.868, 0.876)**	**0.739 (0.732, 0.746)**	**0.831 (0.826, 0.835)**
CTransPath	0.813 (0.807, 0.818)	0.690 (0.683, 0.697)	0.786 (0.781, 0.791)
ResNet50	0.806 (0.800, 0.812)	0.679 (0.673, 0.686)	0.776 (0.772, 0.781)
ViT-B/16	0.820 (0.815, 0.826)	0.676 (0.669, 0.682)	0.781 (0.776, 0.786)
ViT-L/16	0.823 (0.818, 0.828)	0.667 (0.660, 0.673)	0.782 (0.777, 0.787)
KimiaNet	0.826 (0.820, 0.832)	0.711 (0.704, 0.718)	0.800 (0.795, 0.805)
10% Training Labels			
CONCH	**0.826 (0.821, 0.831)**	**0.679 (0.674, 0.685)**	**0.778 (0.773, 0.782)**
CTransPath	0.722 (0.715, 0.729)	0.625 (0.620, 0.632)	0.707 (0.702, 0.713)
ResNet50	0.673 (0.664, 0.682)	0.653 (0.645, 0.659)	0.682 (0.676, 0.688)
ViT-B/16	0.758 (0.751, 0.764)	0.641 (0.634, 0.647)	0.727 (0.722, 0.732)
ViT-L/16	0.768 (0.762, 0.775)	0.656 (0.650, 0.663)	0.736 (0.731, 0.741)
KimiaNet	0.730 (0.723, 0.737)	0.620 (0.614, 0.626)	0.711 (0.706, 0.717)
1% Training Labels			
CONCH	**0.662 (0.654, 0.669)**	**0.441 (0.439, 0.443)**	**0.557 (0.551, 0.564)**
CTransPath	0.595 (0.587, 0.604)	0.433 (0.430, 0.435)	0.522 (0.515, 0.529)
ResNet50	0.516 (0.506, 0.525)	0.414 (0.409, 0.419)	0.493 (0.487, 0.500)
ViT-B/16	0.571 (0.562, 0.579)	0.423 (0.420, 0.427)	0.531 (0.525, 0.538)
ViT-L/16	0.533 (0.524, 0.542)	0.404 (0.401, 0.407)	0.513 (0.506, 0.520)
KimiaNet	0.396 (0.387, 0.405)	0.383 (0.378, 0.388)	0.463 (0.456, 0.469)

Supplementary Data Table 31: End-to-end finetuned classification result on Gleason grading (AGGC + PANDA + SICAP) ($n = 29,039$) in terms of quadratic weighted Cohen's κ, balanced accuracy, and weighted F1 score. Best performing model for each metric is bolded. 95% CI is included in parentheses.

Hyperparameter	Value
Automatic mixed precision	fp16
Batch size	384
Gradient accumulation	4
Weight decay	0.2
AdamW β	(0.9, 0.999)
Temperature	Learned
Peak learning rate	1e-4
Learning rate schedule	Cosine
Warmup steps	250
Epochs	40

Supplementary Data Table 32: Hyperparameters used in visual-language pretraining. $8 \times$ 80GB NVIDIA A100 GPUs were used for training. Effective batch size used for optimization is batch size $\times$ gradient accumulation steps. The maximum sequence length for captions is set to 128.

Hyperparameter	Value
Layers	12
Heads	12
Patch size	16
Head activation	GELU
Embedding dimension	768
Drop path rate	0.1
Global crop scale	0.32, 1.0
Global crop number	2
Local crop scale	0.05, 0.32
Local crop number	10
Partial prediction shape	Block
Partial prediction ratio	0.3
Partial prediction variance	0.2
Gradient clipping max norm	0.3
Normalize last layer	✓
Shared head	✓
AdamW β	(0.9, 0.999)
Batch size	1024
Freeze last layer epochs	3
Warmup epochs	10
Warmup teacher temperature epochs	30
Max epochs	80
Learning rate schedule	Cosine
Learning rate (start)	0
Learning rate (post warmup)	2e-3
Learning rate (final)	2e-6
Teacher temperature (start)	0.04
Teacher temperature (final)	0.4
Teacher momentum (start)	0.996
Teacher momentum (final)	1.000
Weight decay (start)	0.04
Weight decay (end)	0.4
Automatic mixed precision	fp16

Supplementary Data Table 33: Hyperparameters used in pretraining the vision model. $4 \times$ 80GB NVIDIA A100 GPUs were used for training. Batch size refers to the total batch size across GPUs.

Hyperparameter	Value
Layers	24
Heads	12
Embedding dimension	768
Hidden dimension	3,072
Max. sequence length	512
Pos. embedding	Absolute
Vocabulary size	32,000
Automatic mixed precision	fp16
Batch size	64
Gradient accumulation	8
Weight decay	0.01
AdamW β	(0.9, 0.999)
Peak learnng rate	1e-3
Learning rate schedule	Linear
Warmup steps	500
Training steps	15,000

Supplementary Data Table 34: Hyperparameters used in pretraining the language model. In-house pathology reports were first de-identified using regex pattern matching before tokenization. 4 × 80GB NVIDIA A100 GPUs were used for training. Batch size refers to the total batch size across GPUs. Effective batch size used for optimization is batch size × gradient accumulation steps. The sequence length of training examples was set to the maximum sequence length supported by the model (*i.e.* 512).

```
CLASSNAME.
a photomicrograph showing CLASSNAME.
a photomicrograph of CLASSNAME.
an image of CLASSNAME.
an image showing CLASSNAME.
an example of CLASSNAME.
CLASSNAME is shown.
this is CLASSNAME.
there is CLASSNAME.
a histopathological image showing CLASSNAME.
a histopathological image of CLASSNAME.
a histopathological photograph of CLASSNAME.
a histopathological photograph showing CLASSNAME.
shows CLASSNAME.
presence of CLASSNAME.
CLASSNAME is present.
an H&E stained image of CLASSNAME.
an H&E stained image showing CLASSNAME.
an H&E image showing CLASSNAME.
an H&E image of CLASSNAME.
CLASSNAME, H&E stain.
CLASSNAME, H&E.
```

Supplementary Data Table 35: Prompt templates used for all tasks involving prompts. The name of the class replaces `CLASSNAME`. See **Tables 38-44** for class prompts of each task.

Hyperparameter	Value
Batch size	1
Weight decay	1e-5
AdamW β	(0.9, 0.999)
Peak learning rate	1e-4
Learning rate schedule	Cosine
Epochs	20

Supplementary Data Table 36: Hyperparameters used in slide-level weakly-supervised classification. A single 24GB NVIDIA GeForce RTX 3090 GPU was used for each ABMIL model using weakly-supervised learning and slide-level labels. Note for EBRAINS, given the availability of a validation set, we use the validation loss to select the optimal model and perform early stopping with a patience of 10 epochs and train for up to 40 epochs (all other hyperparameters are unchanged).

Hyperparameter	Value
Batch size	16
Weight decay	0.2
AdamW β	(0.9, 0.999)
Learning rate	1e-4
Warmup steps	10
Early stopping patience	10
Epochs	40

Supplementary Data Table 37: Hyperparameters used in caption fine-tuning. A single 24GB NVIDIA GeForce RTX 3090 GPU was used for training. The maximum sequence length for captions is set to 128. Top-K sampling with $K = 50$ was used as decoding strategy at generation time.

Task	Class	Class names
TCGA BRCA	IDC	invasive ductal carcinoma breast invasive ductal carcinoma invasive ductal carcinoma of the breast invasive carcinoma of the breast, ductal pattern breast IDC
	ILC	invasive lobular carcinoma breast invasive lobular carcinoma invasive lobular carcinoma of the breast invasive carcinoma of the breast, lobular pattern breast ILC
TCGA NSCLC	LUAD	adenocarcinoma lung adenocarcinoma adenocarcinoma of the lung LUAD
	LUSC	squamous cell carcinoma lung squamous cell carcinoma squamous cell carcinoma of the lung LUSC
TCGA RCC	CCRCC	clear cell renal cell carcinoma renal cell carcinoma, clear cell type renal cell carcinoma of the clear cell type clear cell RCC
	PRCC	papillary renal cell carcinoma renal cell carcinoma, papillary type renal cell carcinoma of the papillary type papillary RCC
	CHRCC	chromophobe renal cell carcinoma renal cell carcinoma, chromophobe type renal cell carcinoma of the chromophobe type chromophobe RCC

Supplementary Data Table 38: Class prompts for BRCA, NSCLC, and RCC subtyping.

Task	Class	Class names
DHMC LUAD	papillary	papillary pattern adenocarcinoma papillary pattern adenocarcinoma of the lung lung adenocarcinoma, papillary growth pattern lung adenocarcinoma with a predominantly papillary growth pattern lung adenocarcinoma, papillary predominant histological subtype
	solid	solid pattern adenocarcinoma solid pattern adenocarcinoma of the lung lung adenocarcinoma, solid growth pattern lung adenocarcinoma with a predominantly solid growth pattern lung adenocarcinoma, solid predominant histological subtype
	micropapillary	micropapillary pattern adenocarcinoma micropapillary pattern adenocarcinoma of the lung lung adenocarcinoma, micropapillary growth pattern lung adenocarcinoma with a predominantly micropapillary growth pattern lung adenocarcinoma, micropapillary predominant histological subtype
	acinar	acinar pattern adenocarcinoma acinar pattern adenocarcinoma of the lung lung adenocarcinoma, acinar growth pattern lung adenocarcinoma with a predominantly acinar growth pattern lung adenocarcinoma, acinar predominant histological subtype
	leipidic	leipidic pattern adenocarcinoma leipidic pattern adenocarcinoma of the lung lung adenocarcinoma, leipidic growth pattern lung adenocarcinoma with a predominantly leipidic growth pattern lung adenocarcinoma, leipidic predominant histological subtype

Supplementary Data Table 39: Class prompts for DHMC LUAD pattern classification.

Task	Class	Class names
CRC100k	ADI	adipose adipose tissue adipocytes fat fat cells
	BACK	background penmarking empty space background artifacts
	DEB	debris colorectal adenocarcinoma debris and necrosis necrosis necrotic debris
	LYM	lymphocytes lymphoid aggregate immune cells lymphoid infiltrate inflammatory cells
	MUC	mucus mucin mucus pool mucin pool
	MUS	smooth muscle smooth muscle tissue muscle muscularis propria muscularis mucosa
	NORM	normal colon mucosa uninvolved colon mucosa benign colon mucosa benign epithelium
	STR	cancer-associated stroma tumor-associated stroma stromal cells stromal tissue stroma
	TUM	colorectal adenocarcinoma epithelium colorectal adenocarcinoma tumor adenocarcinoma malignant epithelium

Supplementary Data Table 40: Class prompts for CRC100k. The Class column refers to the original class names used in the CRC100k dataset.

Task	Class	Class names
WSSS4LUAD	normal	non-tumor normal tissue non-cancerous tissue
	stroma	tumor-associated stroma cancer-associated stroma tumor-associated stromal tissue cancer-associated stromal tissue
	tumor	tumor tissue tumor epithelial tissue cancerous tissue
SICAP	NC	non-cancerous tissue non-cancerous prostate tissue benign tissue benign glands benign prostate tissue benign prostate glands
	G3	gleason grade 3 gleason pattern 3 prostate cancer, gleason grade 3 prostate cancer, gleason pattern 3 prostate adenocarcinoma, well-differentiated well-differentiated prostatic adenocarcinoma
	G4	gleason grade 4 gleason pattern 4 prostate cancer, gleason grade 4 prostate cancer, gleason pattern 4 prostate adenocarcinoma, moderately differentiated moderately differentiated prostatic adenocarcinoma
	G5	gleason grade 5 gleason pattern 5 prostate cancer, gleason grade 5 prostate cancer, gleason pattern 5 prostate adenocarcinoma, poorly differentiated poorly differentiated prostatic adenocarcinoma
	Tumor	prostatic adenocarcinoma adenocarcinoma prostate cancer tumor tissue cancerous tissue
DigestPath	Benign	benign tissue benign colon tissue benign colorectal tissue benign rectal tissue
	Malignant	malignant tissue malignant colon tissue malignant colorectal tissue malignant rectal tissue

Supplementary Data Table 41: Class prompts for WSSS4LUAD, SICAP, and DigestPath. SICAP ROI-level classification uses the prompts for NC through G5. SICAP slide-level tumor segmentation uses prompts for NC and Tumor.

Class	Class names
Glioblastoma, IDH-wildtype	glioblastoma, IDH-wildtype glioblastoma without IDH mutation glioblastoma with retained IDH glioblastoma, IDH retained
Transitional meningioma	transitional meningioma meningioma, transitional type meningioma of transitional type meningioma, transitional
Anaplastic meningioma	anaplastic meningioma meningioma, anaplastic type meningioma of anaplastic type meningioma, anaplastic
Pituitary adenoma	pituitary adenoma adenoma of the pituitary gland pituitary gland adenoma pituitary neuroendocrine tumor neuroendocrine tumor of the pituitary neuroendocrine tumor of the pituitary gland
Oligodendroglioma, IDH-mutant and 1p/19q codeleted	oligodendroglioma, IDH-mutant and 1p/19q codeleted oligodendroglioma oligodendroglioma with IDH mutation and 1p/19q codeletion
Haemangioma	hemangioma haemangioma of the CNS hemangioma of the CNS haemangioma of the central nervous system hemangioma of the central nervous system
Ganglioglioma	gangliocytoma glioneuronal tumor circumscribed glioneuronal tumor
Schwannoma	schwannoma Antoni A Antoni B neurilemoma
Anaplastic oligodendroglioma, IDH-mutant, 1p/19q codeleted	anaplastic oligodendroglioma, IDH-mutant and 1p/19q codeleted anaplastic oligodendroglioma anaplastic oligodendroglioma with IDH mutation and 1p/19q codeletion
Anaplastic astrocytoma, IDH-wildtype	anaplastic astrocytoma, IDH-wildtype anaplastic astrocytoma without IDH mutation anaplastic astrocytoma, IDH retained anaplastic astrocytoma with retained IDH

Supplementary Data Table 42: Class prompts for EBRAINS subtyping.

Class	Class names
Pilocytic astrocytoma	pilocytic astrocytoma juvenile pilocytic astrocytoma spongioblastoma pilomyxoid astrocytoma
Angiomatous meningioma	angiomatous meningioma meningioma, angiomatous type meningioma of angiomatous type meningioma, angiomatous
Haemangioblastoma	haemangioblastoma capillary hemangioblastoma lindau tumor angioblastoma
Gliosarcoma	gliosarcoma gliosarcoma variant of glioblastoma
Adamantinomatous craniopharyngioma	adamantinomatous craniopharyngioma craniopharyngioma
Anaplastic astrocytoma, IDH-mutant	anaplastic astrocytoma, IDH-mutant anaplastic astrocytoma with IDH mutation anaplastic astrocytoma with mutant IDH anaplastic astrocytoma with mutated IDH
Ependymoma	ependymoma subependymoma myxopapillary ependymoma
Anaplastic ependymoma	anaplastic ependymoma ependymoma, anaplastic ependymoma, anaplastic type
Glioblastoma, IDH-mutant	glioblastoma, IDH-mutant glioblastoma with IDH mutation glioblastoma with mutant IDH glioblastoma with mutated IDH
Atypical meningioma	atypical meningioma meningioma, atypical type meningioma of atypical type meningioma, atypical

Supplementary Data Table 43: Class prompts for EBRAINS subtyping. Continued.

Class	Class names
Metastatic tumours	metastatic tumors metastases to the brain metastatic tumors to the brain brain metastases brain metastatic tumors
Meningothelial meningioma	meningothelial meningioma meningioma, meningothelial type meningioma of meningothelial type meningioma, meningothelial
Langerhans cell histiocytosis	langerhans cell histiocytosis histiocytosis X eosinophilic granuloma Hand-Schüller-Christian disease Hashimoto-Pritzker disease Letterer-Siwe disease
Diffuse large B-cell lymphoma of the CNS	diffuse large B-cell lymphoma of the CNS DLBCL DLBCL of the CNS DLBCL of the central nervous system
Diffuse astrocytoma, IDH-mutant	diffuse astrocytoma, IDH-mutant diffuse astrocytoma with IDH mutation diffuse astrocytoma with mutant IDH diffuse astrocytoma with mutated IDH
Secretory meningioma	secretory meningioma meningioma, secretory type meningioma of secretory type meningioma, secretory
Haemangiopericytoma	haemangiopericytoma solitary fibrous tumor hemangiopericytoma angioblastic meningioma
Fibrous meningioma	fibrous meningioma meningioma, fibrous type meningioma of fibrous type meningioma, fibrous
Lipoma	lipoma CNS lipoma lipoma of the CNS lipoma of the central nervous system
Medulloblastoma, non-WNT/non-SHH	medulloblastoma, non-WNT/non-SHH medulloblastoma medulloblastoma group 3 medulloblastoma group 4

Supplementary Data Table 44: Class prompts for EBRAINS subtyping. Continued.