nature biomedical engineering

Supplementary information

https://doi.org/10.1038/s41551-025-01488-4

A generalizable pathology foundation model using a unified knowledge distillation pretraining framework

In the format provided by the authors and unedited

Supplementary Table 1: Average WSI classification performance of foundation models across 36 tasks. The features have been pre-extracted, and the subsequent downstream tasks are conducted using ABMIL. Best performing model for each metric is **bolded** and second-best performing model is underlined. The standard deviation is included.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.563±0.183	0.518±0.215	0.769±0.135
Phikon	0.693±0.191	0.670±0.228	0.868±0.105
Ctranspath	0.698±0.183	0.677±0.212	0.855±0.111
UNI	0.721±0.181	0.706±0.214	0.875±0.105
CONCH	0.695±0.180	0.664±0.214	0.849±0.140
PLIP	0.668±0.156	0.639±0.186	0.835±0.109
CHIEF	0.687±0.171	0.667±0.202	0.856±0.108
Prov-Gigapath	0.703±0.173	0.676±0.209	0.854±0.113
GPFM	**0.752±0.161**	**0.736±0.179**	**0.891±0.096**

Supplementary Table 2: NSCLC subtyping performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation and the trained model is from TCGA cohort.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-NSCLC	0.845 (0.800-0.893)	0.842 (0.791-0.893)	0.929 (0.896-0.959)
Phikon	TCGA-NSCLC	0.888 (0.845-0.928)	0.885 (0.843-0.924)	0.982 (0.968-0.993)
Ctranspath	TCGA-NSCLC	0.894 (0.850-0.934)	0.895 (0.851-0.933)	0.963 (0.936-0.985)
UNI	TCGA-NSCLC	0.928 (0.891-0.961)	0.928 (0.890-0.962)	0.977 (0.957-0.992)
CONCH	TCGA-NSCLC	0.924 (0.888-0.957)	0.924 (0.881-0.957)	**0.986 (0.971-0.996)**
PLIP	TCGA-NSCLC	0.865 (0.821-0.908)	0.865 (0.812-0.909)	0.942 (0.910-0.969)
CHIEF	TCGA-NSCLC	0.910 (0.874-0.951)	0.910 (0.866-0.943)	0.971 (0.951-0.988)
Prov-Gigapath	TCGA-NSCLC	0.918 (0.880-0.953)	0.919 (0.877-0.957)	0.967 (0.942-0.987)
GPFM	TCGA-NSCLC	**0.948 (0.915-0.976)**	**0.947 (0.918-0.976)**	**0.986 (0.973-0.996)**
ResNet50	CPTAC-NSCLC	0.847 (0.803-0.871)	0.847 (0.803-0.871)	0.937 (0.899-0.945)
Phikon	CPTAC-NSCLC	0.901 (0.873-0.928)	0.900 (0.880-0.925)	0.967 (0.951-0.980)
Ctranspath	CPTAC-NSCLC	0.887 (0.858-0.916)	0.887 (0.856-0.914)	0.965 (0.950-0.977)
UNI	CPTAC-NSCLC	**0.911 (0.883-0.937)**	**0.911 (0.884-0.939)**	0.960 (0.942-0.976)
CONCH	CPTAC-NSCLC	0.876 (0.844-0.903)	0.876 (0.844-0.905)	0.961 (0.944-0.975)
PLIP	CPTAC-NSCLC	0.841 (0.805-0.876)	0.841 (0.808-0.873)	0.939 (0.918-0.957)
CHIEF	CPTAC-NSCLC	0.881 (0.851-0.909)	0.882 (0.851-0.912)	0.964 (0.949-0.978)
Prov-Gigapath	CPTAC-NSCLC	0.882 (0.853-0.911)	0.883 (0.853-0.911)	0.971 (0. 957-0.982)
GPFM	CPTAC-NSCLC	0.906 (0.877-0.932)	0.906 (0.880-0.934)	**0.974 (0.961-0.985)**
ResNet50	Center-1-NSCLC*	0.492 (0.481-0.5)	0.457 (0.443-0.471)	0.611 (0.497-0.726)
Phikon	Center-1-NSCLC*	0.566 (0.497-0.644)	0.590 (0.475-0.713)	0.660 (0.526-0.770)
Ctranspath	Center-1-NSCLC*	0.695 (0.583-0.802)	0.695 (0.590-0.789)	0.722 (0.592-0.848)
UNI	Center-1-NSCLC*	**0.764 (0.677-0.857)**	**0.788 (0.703-0.864)**	0.819 (0.729-0.909)
CONCH	Center-1-NSCLC*	0.617 (0.512-0.724)	0.579 (0.502-0.658)	0.595 (0.439-0.749)
PLIP	Center-1-NSCLC*	0.682 (0.577-0.783)	0.602 (0.526-0.684)	0.770 (0.667-0.864)
CHIEF	Center-1-NSCLC*	0.693 (0.575-0.814)	0.571 (0.496-0.642)	0.746 (0.592-0.878)
Prov-Gigapath	Center-1-NSCLC*	0.725 (0.598-0.864)	0.759 (0.624-0.852)	0.727 (0.554-0.900)
GPFM	Center-1-NSCLC*	0.614 (0.534-0.689)	0.653 (0.541-0.750)	**0.823 (0.728-0.902)**

Supplementary Table 3: The lung cancer metastasis detection (2 classes) and primary site prediction (6 classes). Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. cls represents "class number". * indicates the external validation cohort.

	cls	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	2	Center-1-LMD	0.816 (0.752-0.876)	0.819 (0.759-0.880)	0.905 (0.850-0.950)
Phikon	2	Center-1-LMD	0.902 (0.853-0.945)	0.894 (0.844-0.943)	0.984 (0.967-0.996)
Ctranspath	2	Center-1-LMD	0.913 (0.863-0.956)	0.908 (0.858-0.957)	0.981 (0.962-0.994)
UNI	2	Center-1-LMD	0.908 (0.860-0.952)	0.901 (0.852-0.950)	0.981 (0.962-0.997)
CONCH	2	Center-1-LMD	0.910 (0.861-0.958)	0.908 (0.852-0.951)	0.965 (0.935-0.990)
PLIP	2	Center-1-LMD	0.838 (0.777-0.896)	0.837 (0.768-0.894)	0.940 (0.901-0.972)
CHIEF	2	Center-1-LMD	0.930 (0.882-0.972)	0.935 (0.890-0.971)	0.983 (0.965-0.996)
Prov-Gigapath	2	Center-1-LMD	0.833 (0.770-0.894)	0.841 (0.772-0.897)	0.968 (0.936-0.992)
GPFM	2	Center-1-LMD	**0.940 (0.902-0.977)**	**0.943 (0.904-0.978)**	**0.985 (0.966-0.998)**
ResNet50	2	Center-2-LMD*	0.588 (0.544-0.633)	0.559 (0.509-0.608)	0.651 (0.602-0.703)
Phikon	2	Center-2-LMD*	0.801 (0.762-0.839)	0.805 (0.763-0.839)	0.914 (0.886-0.937)
Ctranspath	2	Center-2-LMD*	0.800 (0.768-0.835)	0.782 (0.740-0.821)	0.885 (0.854-0.916)
UNI	2	Center-2-LMD*	0.820 (0.783-0.855)	0.819 (0.783-0.857)	0.918 (0.893-0.939)
CONCH	2	Center-2-LMD*	**0.859 (0.828-0.888)**	**0.849 (0.813-0.881)**	**0.950 (0.931-0.967)**
PLIP	2	Center-2-LMD*	0.690 (0.645-0.733)	0.691 (0.647-0.738)	0.764 (0.721-0.807)
CHIEF	2	Center-2-LMD*	0.748 (0.703-0.788)	0.750 (0.708-0.794)	0.881 (0.848-0.911)
Prov-Gigapath	2	Center-2-LMD*	0.666 (0.624-0.708)	0.661 (0.618-0.704)	0.744 (0.698-0.787)
GPFM	2	Center-2-LMD*	0.800 (0.763-0.838)	0.805 (0.763-0.845)	0.927 (0.903-0.948)
ResNet50	6	Center-1-LMD	0.378 (0.305-0.475)	0.365 (0.283-0.453)	0.895 (0.750-0.933)
Phikon	6	Center-1-LMD	0.537 (0.433-0.666)	0.539 (0.409-0.646)	0.955 (0.806-0.979)
Ctranspath	6	Center-1-LMD	0.640 (0.481-0.796)	0.666 (0.516-0.788)	0.959 (0.806-0.979)
UNI	6	Center-1-LMD	0.709 (0.570-0.856)	0.702 (0.564-0.821)	0.961 (0.816-0.989)
CONCH	6	Center-1-LMD	0.526 (0.472-0.637)	0.475 (0.398-0.569)	0.955 (0.811-0.985)
PLIP	6	Center-1-LMD	0.600 (0.520-0.716)	0.534 (0.436-0.638)	0.936 (0.799-0.965)
CHIEF	6	Center-1-LMD	0.640 (0.486-0.808)	0.679 (0.505-0.829)	0.959 (0.810-0.987)
Prov-Gigapath	6	Center-1-LMD	0.702 (0.568-0.850)	0.695 (0.530-0.819)	0.967 (0.815-0.989)
GPFM	6	Center-1-LMD	**0.771 (0.640-0.924)**	**0.767 (0.598-0.894)**	**0.974 (0.821-0.991)**
ResNet50	6	Center-2-LMD*	0.277 (0.234-0.330)	0.250 (0.212-0.291)	0.721 (0.674-0.765)
Phikon	6	Center-2-LMD*	0.459 (0.394-0.524)	0.429 (0.380-0.479)	0.871 (0.831-0.904)
Ctranspath	6	Center-2-LMD*	0.509 (0.441-0.569)	0.519 (0.450-0.577)	0.886 (0.851-0.917)
UNI	6	Center-2-LMD*	0.568 (0.492-0.656)	0.542 (0.487-0.592)	0.900 (0.865-0.928)
CONCH	6	Center-2-LMD*	0.525 (0.471-0.574)	0.428 (0.388-0.471)	0.906 (0.868-0.940)
PLIP	6	Center-2-LMD*	0.405 (0.346-0.468)	0.399 (0.347-0.445)	0.861 (0.816-0.896)
CHIEF	6	Center-2-LMD*	0.490 (0.419-0.557)	0.532 (0.454-0.595)	0.885 (0.854-0.916)
Prov-Gigapath	6	Center-2-LMD*	0.551 (0.496-0.602)	0.481 (0.429-0.525)	0.853 (0.809-0.888)
GPFM	6	Center-2-LMD*	**0.591 (0.526-0.653)**	**0.601 (0.541-0.655)**	**0.908 (0.881-0.932)**

Supplementary Table 4: RCC subtyping performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-RCC	0.870 (0.805-0.926)	0.858 (0.797-0.910)	0.963 (0.937-0.983)
Phikon	TCGA-RCC	**0.964 (0.936-0.987)**	**0.960 (0.937-0.982)**	0.995 (0.989-0.999)
Ctranspath	TCGA-RCC	0.883 (0.820-0.938)	0.888 (0.833-0.939)	0.981 (0.965-0.992)
UNI	TCGA-RCC	0.903 (0.847-0.954)	0.913 (0.862-0.957)	0.982 (0.966-0.995)
CONCH	TCGA-RCC	0.941 (0.891-0.979)	0.937 (0.896-0.974)	0.988 (0.977-0.997)
PLIP	TCGA-RCC	0.899 (0.845-0.947)	0.904 (0.853-0.948)	0.980 (0.963-0.993)
CHIEF	TCGA-RCC	0.889 (0.824-0.945)	0.900 (0.843-0.945)	0.981 (0.964-0.993)
Prov-Gigapath	TCGA-RCC	0.903 (838-0.958)	0.907 (0.850-0.955)	0.986 (0.975-0.996)
GPFM	TCGA-RCC	0.925 (0.874-0.967)	0.930 (0.885-0.966)	**0.996 (0.992-0.999)**
ResNet50	Center-3-RCC*	0.490 (0.415-0.562)	0.409 (0.332-0.496)	0.747 (0.662-0.824)
Phikon	Center-3-RCC*	0.433 (0.378-0.495)	0.337 (0.254-0.422)	0.831 (0.744-0.909)
Ctranspath	Center-3-RCC*	0.735 (0.641-0.817)	0.735 (0.640-0.821)	0.910 (0.853-0.953)
UNI	Center-3-RCC*	0.713 (0.614-0.799)	0.717 (0.614-0.808)	0.921 (0.868-0.959)
CONCH	Center-3-RCC*	0.560 (0.472-0.647)	0.521 (0.407-0.618)	0.822 (0.745-0.890)
PLIP	Center-3-RCC*	0.523 (0.453-0.593)	0.451 (0.365-0.530)	0.818 (0.743-0.883)
CHIEF	Center-3-RCC*	0.757 (0.671-0.841)	0.755 (0.658-0.840)	0.921 (0.869-0.960)
Prov-Gigapath	Center-3-RCC*	0.628 (0.532-0.718)	0.607 (0.503-0.704)	0.751 (0.678-0.817)
GPFM	Center-3-RCC*	**0.759 (0.667-0.835)**	**0.756 (0.666-0.843)**	**0.922 (0.868-0.963)**

Supplementary Table 5: The breast metastasis detection performance of different foundation models on CAMELYON dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.855 (0.797-0.909)	0.857 (0.800-0.910)	0.922 (0.864-0.966)
Phikon	0.945 (0.900-0.979)	0.952 (0.918-0.982)	0.967 (0.932-0.993)
Ctranspath	0.898 (0.852-0.941)	0.908 (0.860-0.951)	0.957 (0.924-0.986)
UNI	0.963 (0.930-0.992)	**0.970 (0.940-0.994)**	0.987 (0.969-0.998)
CONCH	0.936 (0.896-0.974)	0.945 (0.910-0.977)	0.965 (0.934-0.989)
PLIP	0.882 (0.826-0.930)	0.890 (0.840-0.936)	0.929 (0.882-0.967)
CHIEF	0.902 (0.858-0.947)	0.905 (0.856-0.947)	0.944 (0.901-0.979)
Prov-Gigapath	0.941 (0.900-0.977)	0.951 (0.917-0.982)	0.969 (0.939-0.993)
GPFM	**0.964 (0.931-0.991)**	0.964 (0.932-0.988)	**0.988 (0.971-1.000)**

Supplementary Table 6: Lobular and ductal carcinoma subtyping performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation cohort.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-BRCA	0.658 (0.585-0.735)	0.691 (0.596-0.783)	0.846 (0.768-0.911)
Phikon	TCGA-BRCA	0.794 (0.718-0.865)	0.835 (0.751-0.901)	0.936 (0.887-0.977)
Ctranspath	TCGA-BRCA	0.843 (0.767-0.914)	0.859 (0.790-0.917)	0.931 (0.870-0.975)
UNI	TCGA-BRCA	0.869 (0.797-0.929)	0.879 (0.810-0.932)	0.946 (0.894-0.987)
CONCH	TCGA-BRCA	0.835 (0.750-0.905)	0.875 (0.807-0.934)	0.944 (0.902-0.979)
PLIP	TCGA-BRCA	0.823 (0.747-0.897)	0.820 (0.750-0.888)	0.893 (0.824-0.950)
CHIEF	TCGA-BRCA	0.790 (0.717-0.866)	0.829 (0.756-0.896)	0.899 (0.823-0.960)
Prov-Gigapath	TCGA-BRCA	**0.884 (0.821-0.940)**	0.877 (0.813-0.931)	0.942 (0.884-0.983)
GPFM	TCGA-BRCA	0.881 (0.813-0.947)	**0.907 (0.850-0.956)**	**0.950 (0.898-0.990)**
ResNet50	Center-3-LD*	0.441 (0.394-0.480)	0.441 (0.406-0.472)	0.562 (0.387-0.737)
Phikon	Center-3-LD*	0.824 (0.613-1.000)	**0.849 (0.597-1.000)**	**0.915 (0.726-1.000)**
Ctranspath	Center-3-LD*	0.814 (0.606-0.991)	0.814 (0.578-0.957)	0.814 (0.443-1.000)
UNI	Center-3-LD*	0.750 (0.500-1.000)	0.819 (0.491-1.000)	0.874 (0.616-1.000)
CONCH	Center-3-LD*	0.500 (0.500-0.500)	0.472 (0.447-0.491)	0.500 (0.500-0.500)
PLIP	Center-3-LD*	0.637 (0.412-0.827)	0.502 (0.374-0.635)	0.614 (0.321-0.893)
CHIEF	Center-3-LD*	0.500 (0.500-0.500)	0.472 (0.447-0.491)	0.811 (0.451-1.000)
Prov-Gigapath	Center-3-LD*	0.652 (0.404-0.898)	0.472 (0.441-0.736)	0.784 (0.542-1.000)
GPFM	Center-3-LD*	**0.887 (0.686-0.991)**	0.837 (0.648-0.966)	0.905 (0.694-1.000)

Supplementary Table 7: Coarse-grained breast carcinoma subtyping performance of different foundation. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	BRACS	0.568 (0.515-0.615)	0.522 (0.463-0.571)	0.835 (0.776-0.892)
Phikon	BRACS	0.707 (0.621-0.797)	0.701 (0.602-0.800)	0.898 (0.852-0.942)
Ctranspath	BRACS	0.674 (0.592-0.757)	0.664 (0.559-0.754)	0.908 (0.871-0.946)
UNI	BRACS	0.746 (0.660-0.840)	0.738 (0.640-0.824)	0.913 (0.865-0.956)
CONCH	BRACS	0.677 (0.606-0.752)	0.668 (0.575-0.771)	0.923 (0.883-0.958)
PLIP	BRACS	0.679 (0.596-0.773)	0.676 (0.579-0.782)	0.866 (0.805-0.917)
CHIEF	BRACS	0.717 (0.635-0.804)	0.723 (0.617-0.814)	0.922 (0.879-0.958)
Prov-Gigapath	BRACS	0.720 (0.620-0.810)	0.715 (0.613-0.802)	0.927 (0.885-0.961)
GPFM	BRACS	**0.749 (0.660-0.834)**	**0.758 (0.658-0.841)**	**0.936 (0.896-0.965)**
ResNet50	Center-3-BRCA*	0.618 (0.563-0.669)	0.521 (0.472-0.564)	0.624 (0.555-0.688)
Phikon	Center-3-BRCA*	0.636 (0.584-0.691)	0.539 (0.493-0.589)	0.659 (0.607-0.713)
Ctranspath	Center-3-BRCA*	0.583 (0.521-0.645)	0.543 (0.493-0.596)	0.670 (0.617-0.720)
UNI	Center-3-BRCA*	0.556 (0.493-0.616)	0.504 (0.454-0.552)	0.614 (0.559-0.666)
CONCH	Center-3-BRCA*	0.479 (0.450-0.503)	0.178 (0.146-0.208)	0.500 (0.498-0.501)
PLIP	Center-3-BRCA*	0.578 (0.522-0.634)	0.566 (0.514-0.618)	0.681 (0.627-0.736)
CHIEF	Center-3-BRCA*	0.568 (0.506-0.624)	0.546 (0.495-0.597)	0.670 (0.615-0.723)
Prov-Gigapath	Center-3-BRCA*	0.587 (0.530-0.642)	**0.574 (0.521-0.629)**	0.678 (0.625-0.732)
GPFM	Center-3-BRCA*	**0.675 (0.625-0.720)**	0.543 (0.490-0.594)	**0.694 (0.641-0.747)**

Supplementary Table 8: Fine-grained breast carcinoma subtyping performance of different foundation models on BRACS dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.309 (0.266-0.357)	0.250 (0.181-0.320)	0.772 (0.719-0.818)
Phikon	0.363 (0.322-0.406)	0.293 (0.251-0.332)	0.818 (0.768-0.866)
Ctranspath	**0.530 (0.450-0.626)**	**0.520 (0.407-0.615)**	0.857 (0.811-0.896)
UNI	0.433 (0.356-0.511)	0.411 (0.325-0.490)	0.855 (0.811-0.893)
CONCH	0.424 (0.352-0.505)	0.367 (0.287-0.439)	0.841 (0.797-0.884)
PLIP	0.420 (0.342-0.511)	0.414 (0.324-0.493)	0.814 (0.763-0.864)
CHIEF	0.445 (0.357-0.541)	0.445 (0.332-0.547)	0.854 (0.810-0.897)
Prov-Gigapath	0.463 (0.380-0.549)	0.433 (0.346-0.505)	0.808 (0.753-0.861)
GPFM	0.437 (0.360-0.514)	0.408 (0.326-0.493)	**0.871 (0.829-0.904)**

Supplementary Table 9: Prostate cancer grade assessment performance of different foundation models on PANDA dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.531 (0.510-0.552)	0.531 (0.508-0.553)	0.884 (0.875-0.892)
Phikon	0.731 (0.709-0.750)	0.735 (0.715-0.755)	0.943 (0.936-0.949)
Ctranspath	0.649 (0.627-0.670)	0.651 (0.629-0.671)	0.925 (0.918-0.932)
UNI	0.728 (0.707-0.749)	0.734 (0.712-0.753)	0.944 (0.937-0.950)
CONCH	0.656 (0.635-0.678)	0.657 (0.637-0.679)	0.921 (0.914-0.929)
PLIP	0.607 (0.583-0.628)	0.612 (0.591-0.635)	0.903 (0.894-0.911)
CHIEF	0.665 (0.643-0.688)	0.667 (0.643-0.689)	0.927 (0.920-0.934)
Prov-Gigapath	0.674 (0.653-0.697)	0.676 (0.653-0.699)	0.926 (0.918-0.933)
GPFM	**0.740 (0.720-0.760)**	**0.742 (0.722-0.762)**	**0.948 (0.941-0.954)**

Supplementary Table 10: Lung adenocarcinoma TP53 gene mutation prediction performance of different foundation models on TCGA-LUAD dataset. 5-fold cross validation is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.675 (0.549-0.708)	0.609 (0.493-0.714)	0.742 (0.629-0.842)
Phikon	0.783 (0.704-0.874)	0.782 (0.687-0.867)	0.841 (0.754-0.918)
Ctranspath	0.711 (0.621-0.810)	0.710 (0.601-0.806)	0.770 (0.660-0.867)
UNI	0.639 (0.553-0.749)	0.638 (0.530-0.746)	0.766 (0.667-0.867)
CONCH	0.735 (0.618-0.820)	0.730 (0.629-0.818)	0.836 (0.734-0.911)
PLIP	0.759 (0.643-0.818)	0.739 (0.629-0.832)	0.821 (0.721-0.905)
CHIEF	0.683 (0.586-0.773)	0.682 (0.570-0.783)	0.736 (0.622-0.844)
Prov-Gigapath	0.627 (0.537-0.730)	0.616 (0.505-0.731)	0.804 (0.700-0.894)
GPFM	**0.795 (0.707-0.878)**	**0.794 (0.694-0.878)**	**0.855 (0.767-0.931)**

Supplementary Table 11: WSI-level IDH1 gene mutation prediction performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-GBMLGG	0.842 (0.781-0.900)	0.849 (0.788-0.904)	0.937 (0.895-0.969)
Phikon	TCGA-GBMLGG	0.885 (0.828-0.938)	0.900 (0.846-0.947)	0.972 (0.946-0.992)
Ctranspath	TCGA-GBMLGG	0.860 (0.804-0.911)	0.838 (0.774-0.896)	0.958 (0.920-0.986)
UNI	TCGA-GBMLGG	0.917 (0.867-0.957)	0.906 (0.853-0.951)	0.976 (0.954-0.993)
CONCH	TCGA-GBMLGG	0.856 (0.795-0.915)	0.869 (0.808-0.921)	0.946 (0.894-0.981)
PLIP	TCGA-GBMLGG	0.888 (0.829-0.938)	0.889 (0.834-0.939)	0.946 (0.904-0.979)
CHIEF	TCGA-GBMLGG	0.886 (0.833-0.935)	0.883 (0.826-0.936)	0.944 (0.904-0.977)
Prov-Gigapath	TCGA-GBMLGG	0.928 (0.884-0.963)	0.920 (0.870-0.963)	**0.986 (0.970-0.997)**
GPFM	TCGA-GBMLGG	**0.936 (0.891-0.974)**	**0.934 (0.888-0.971)**	**0.986 (0.971-0.996)**
ResNet50	EBRAINS*	0.531 (0.517-0.547)	0.455 (0.428-0.485)	0.708 (0.672-0.744)
Phikon	EBRAINS*	0.749 (0.724-0.771)	0.697 (0.667-0.726)	0.920 (0.899-0.938)
Ctranspath	EBRAINS*	0.854 (0.829-0.877)	0.851 (0.826-0.875)	0.923 (0.904-0.940)
UNI	EBRAINS*	**0.882 (0.860-0.905)**	**0.875 (0.851-0.896)**	0.939 (0.922-0.955)
CONCH	EBRAINS*	0.795 (0.772-0.820)	0.758 (0.726-0.786)	0.924 (0.904-0.941)
PLIP	EBRAINS*	0.800 (0.777-0.823)	0.765 (0.738-0.792)	0.908 (0.886-0.929)
CHIEF	EBRAINS*	0.810 (0.784-0.835)	0.779 (0.750-0.805)	0.915 (0.895-0.933)
Prov-Gigapath	EBRAINS*	0.838 (0.815-0.861)	0.811 (0.785-0.839)	0.941 (0.924-0.957)
GPFM	EBRAINS*	0.875 (0.852-0.896)	0.863 (0.838-0.886)	**0.943 (0.927-0.957)**

Supplementary Table 12: Ovarian cancer subtyping performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	UBC-OCEAN	0.487 (0.417-0.557)	0.480 (0.385-0.563)	0.903 (0.866-0.936)
Phikon	UBC-OCEAN	0.731 (0.654-0.820)	0.751 (0.654-0.828)	0.970 (0.947-0.988)
Ctranspath	UBC-OCEAN	0.830 (0.744-0.906)	0.820 (0.727-0.891)	0.973 (0.956-0.989)
UNI	UBC-OCEAN	0.737 (0.661-0.814)	0.740 (0.634-0.826)	0.975 (0.956-0.990)
CONCH	UBC-OCEAN	**0.841 (0.748-0.918)**	**0.830 (0.728-0.905)**	0.972 (0.952-0.989)
PLIP	UBC-OCEAN	0.670 (0.596-0.748)	0.665 (0.572-0.754)	0.944 (0.912-0.972)
CHIEF	UBC-OCEAN	0.782 (0.684-0.875)	0.801 (0.689-0.882)	0.971 (0.949-0.988)
Prov-Gigapath	UBC-OCEAN	0.764 (0.671-0.857)	0.788 (0.690-0.869)	0.951 (0.915-0.981)
GPFM	UBC-OCEAN	0.809 (0.717-0.888)	0.810 (0.701-0.888)	**0.984 (0.969-0.994)**
ResNet50	Center-3-Ovary*	0.259 (0.225-0.295)	0.212 (0.179-0.252)	0.637 (0.603-0.674)
Phikon	Center-3-Ovary*	0.250 (0.215-0.285)	0.215 (0.178-0.252)	0.684 (0.657-0.709)
Ctranspath	Center-3-Ovary*	0.239 (0.206-0.275)	0.242 (0.204-0.280)	0.584 (0.557-0.612)
UNI	Center-3-Ovary*	0.267 (0.232-0.306)	0.300 (0.260-0.342)	0.685 (0.654-0.715)
CONCH	Center-3-Ovary*	0.279 (0.233-0.328)	**0.323 (0.263-0.376)**	0.593 (0.565-0.624)
PLIP	Center-3-Ovary*	**0.351 (0.302-0.404)**	0.320 (0.278-0.357)	0.627 (0.592-0.659)
CHIEF	Center-3-Ovary*	0.299 (0.257-0.341)	0.248 (0.209-0.292)	0.673 (0.643-0.703)
Prov-Gigapath	Center-3-Ovary*	0.338 (0.294-0.380)	0.311 (0.270-0.350)	0.660 (0.626-0.692)
GPFM	Center-3-Ovary*	0.276 (0.237-0.320)	0.308 (0.275-0.342)	**0.687 (0.661-0.712)**

Supplementary Table 13: Brain tumor subtyping performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-GBMLGG	0.736 (0.651-0.818)	0.732 (0.640-0.811)	0.954 (0.933-0.972)
Phikon	TCGA-GBMLGG	0.790 (0.706-0.865)	0.803 (0.715-0.871)	0.976 (0.959-0.990)
Ctranspath	TCGA-GBMLGG	0.729 (0.643-0.814)	0.740 (0.635-0.818)	0.963 (0.938-0.983)
UNI	TCGA-GBMLGG	0.795 (0.708-0.880)	0.823 (0.738-0.895)	0.981 (0.967-0.993)
CONCH	TCGA-GBMLGG	0.813 (0.723-0.886)	0.815 (0.726-0.889)	0.978 (0.963-0.990)
PLIP	TCGA-GBMLGG	0.791 (0.707-0.871)	0.806 (0.716-0.873)	0.963 (0.930-0.988)
CHIEF	TCGA-GBMLGG	0.711 (0.620-0.798)	0.718 (0.627-0.798)	0.951 (0.905-0.982)
Prov-Gigapath	TCGA-GBMLGG	**0.816 (0.727-0.901)**	**0.825 (0.738-0.895)**	0.981 (0.967-0.993)
GPFM	TCGA-GBMLGG	0.782 (0.693-0.867)	0.804 (0.706-0.882)	**0.987 (0.976-0.995)**
ResNet50	EBRAINS*	0.333 (0.333-0.333)	0.289 (0.282-0.295)	0.554 (0.506-0.600)
Phikon	EBRAINS*	0.784 (0.743-0.827)	0.726 (0.690-0.763)	0.909 (0.887-0.929)
Ctranspath	EBRAINS*	0.758 (0.718-0.795)	0.622 (0.585-0.658)	0.865 (0.836-0.890)
UNI	EBRAINS*	0.806 (0.764-0.847)	**0.767 (0.727-0.804)**	0.921 (0.900-0.941)
CONCH	EBRAINS*	0.764 (0.729-0.801)	0.642 (0.608-0.675)	0.888 (0.864-0.910)
PLIP	EBRAINS*	0.687 (0.643-0.732)	0.682 (0.634-0.724)	0.868 (0.842-0.893)
CHIEF	EBRAINS*	0.755 (0.718-0.794)	0.656 (0.617-0.694)	0.885 (0.860-0.908)
Prov-Gigapath	EBRAINS*	0.776 (0.736-0.814)	0.691 (0.654-0.727)	0.908 (0.888-0.926)
GPFM	EBRAINS*	**0.809 (0.770-0.846)**	0.713 (0.678-0.746)	**0.926 (0.906-0.946)**

Supplementary Table 14: Lesion grade classification for colon cancer. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	IMP-CRS	0.922 (0.902-0.941)	0.922 (0.904-0.939)	0.986 (0.981-0.991)
Phikon	IMP-CRS	**0.946 (0.931-0.960)**	**0.952 (0.938-0.964)**	0.994 (0.990-0.996)
Ctranspath	IMP-CRS	0.941 (0.925-0.957)	0.947 (0.933-0.960)	0.992 (0.989-0.995)
UNI	IMP-CRS	0.922 (0.902-0.941)	0.938 (0.923-0.953)	0.994 (0.991-0.996)
CONCH	IMP-CRS	0.933 (0.915-0.951)	0.943 (0.928-0.958)	0.993 (0.990-0.996)
PLIP	IMP-CRS	0.899 (0.878-0.920)	0.915 (0.895-0.934)	0.987 (0.982-0.992)
CHIEF	IMP-CRS	0.942 (0.925-0.957)	0.945 (0.931-0.958)	0.992 (0.989-0.995)
Prov-Gigapath	IMP-CRS	0.944 (0.930-0.957)	0.934 (0.918-0.948)	0.993 (0.990-0.996)
GPFM	IMP-CRS	0.916 (0.895-0.934)	0.932 (0.914-0.947)	**0.995 (0.992-0.997)**
ResNet50	Center-3-Colon-WSI*	0.584 (0.541-0.629)	0.563 (0.504-0.621)	0.874 (0.846-0.901)
Phikon	Center-3-Colon-WSI*	0.869 (0.832-0.909)	0.879 (0.842-0.914)	0.961 (0.941-0.980)
Ctranspath	Center-3-Colon-WSI*	0.827 (0.784-0.865)	0.839 (0.797-0.876)	0.955 (0.935-0.972)
UNI	Center-3-Colon-WSI*	0.917 (0.884-0.948)	0.921 (0.890-0.951)	0.979 (0.965-0.991)
CONCH	Center-3-Colon-WSI*	0.899 (0.858-0.935)	0.909 (0.873-0.940)	0.983 (0.970-0.994)
PLIP	Center-3-Colon-WSI*	0.776 (0.734-0.816)	0.788 (0.739-0.837)	0.969 (0.953-0.983)
CHIEF	Center-3-Colon-WSI*	0.835 (0.793-0.875)	0.848 (0.802-0.886)	0.967 (0.953-0.980)
Prov-Gigapath	Center-3-Colon-WSI*	0.900 (0.865-0.930)	0.904 (0.868-0.934)	0.971 (0.954-0.986)
GPFM	Center-3-Colon-WSI*	**0.925 (0.895-0.957)**	**0.937 (0.908-0.965)**	**0.984 (0.972-0.993)**

Supplementary Table 15: Primary site prediction (PSP) and T stage classification for head & neck cancers on HANCOCK dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Task	Balanced ACC	Weighted F1	AUC
ResNet50	PSP	0.555 (0.492-0.615)	0.519 (0.451-0.575)	0.844 (0.796-0.891)
Phikon	PSP	0.727 (0.648-0.800)	0.713 (0.640-0.784)	0.902 (0.858-0.951)
Ctranspath	PSP	0.674 (0.612-0.738)	0.655 (0.586-0.733)	0.859 (0.806-0.902)
UNI	PSP	0.722 (0.664-0.783)	0.725 (0.654-0.797)	0.906 (0.867-0.942)
CONCH	PSP	0.698 (0.626-0.767)	0.696 (0.605-0.775)	0.905 (0.858-0.947)
PLIP	PSP	0.672 (0.605-0.741)	0.673 (0.592-0.757)	0.884 (0.839-0.925)
CHIEF	PSP	0.632 (0.553-0.717)	0.659 (0.570-0.744)	0.861 (0.805-0.911)
Prov-Gigapath	PSP	0.667 (0.620-0.712)	0.634 (0.582-0.676)	0.910 (0.869-0.947)
GPFM	PSP	**0.754 (0.704-0.810)**	**0.736 (0.664-0.805)**	**0.920 (0.884-0.956)**
ResNet50	T Stage	0.473 (0.381-0.564)	0.487 (0.386-0.578)	0.752 (0.689-0.812)
Phikon	T Stage	0.452 (0.362-0.544)	0.456 (0.366-0.541)	0.760 (0.705-0.814)
Ctranspath	T Stage	0.405 (0.321-0.491)	0.402 (0.314-0.484)	0.746 (0.691-0.797)
UNI	T Stage	0.453 (0.381-0.540)	0.424 (0.329-0.516)	0.764 (0.709-0.819)
CONCH	T Stage	0.418 (0.327-0.502)	0.393 (0.311-0.466)	0.761 (0.709-0.813)
PLIP	T Stage	0.438 (0.342-0.527)	0.418 (0.325-0.501)	0.718 (0.659-0.771)
CHIEF	T Stage	0.433 (0.362-0.502)	0.415 (0.331-0.493)	0.745 (0.689-0.795)
Prov-Gigapath	T Stage	0.487 (0.395-0.575)	0.453 (0.362-0.544)	0.750 (0.687-0.807)
GPFM	T Stage	**0.513 (0.425-0.607)**	**0.515 (0.409-0.602)**	**0.780 (0.730-0.832)**

Supplementary Table 16: Lauren subtyping of gastric cancer. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	TCGA-STAD	0.320 (0.282-0.358)	0.227 (0.175-0.291)	0.751 (0.661-0.832)
Phikon	TCGA-STAD	0.557 (0.449-0.666)	0.572 (0.443-0.688)	0.830 (0.748-0.908)
Ctranspath	TCGA-STAD	0.373 (0.316-0.437)	0.311 (0.226-0.390)	0.791 (0.717-0.861)
UNI	TCGA-STAD	0.575 (0.461-0.693)	0.594 (0.469-0.706)	0.820 (0.738-0.892)
CONCH	TCGA-STAD	0.614 (0.501-0.715)	0.619 (0.500-0.720)	0.843 (0.770-0.908)
PLIP	TCGA-STAD	0.641 (0.521-0.750)	0.635 (0.521-0.739)	0.831 (0.749-0.901)
CHIEF	TCGA-STAD	0.595 (0.475-0.711)	0.596 (0.473-0.703)	0.819 (0.745-0.889)
Prov-Gigapath	TCGA-STAD	0.534 (0.432-0.633)	0.512 (0.396-0.619)	0.791 (0.707-0.873)
GPFM	TCGA-STAD	**0.702 (0.582-0.803)**	**0.696 (0.578-0.800)**	**0.873 (0.802-0.931)**
ResNet50	Center-4*	0.333 (0.333-0.333)	0.147 (0.127-0.166)	0.620 (0.573-0.664)
Phikon	Center-4*	0.333 (0.333-0.333)	0.147 (0.127-0.166)	0.723 (0.683-0.762)
Ctranspath	Center-4*	0.333 (0.333-0.333)	0.147 (0.127-0.166)	0.660 (0.618-0.700)
UNI	Center-4*	0.340 (0.333-0.349)	0.161 (0.136-0.185)	0.719 (0.683-0.756)
CONCH	Center-4*	0.477 (0.443-0.510)	0.391 (0.347-0.438)	**0.749 (0.711-0.787)**
PLIP	Center-4*	0.487 (0.454-0.518)	0.395 (0.354-0.437)	0.729 (0.688-0.768)
CHIEF	Center-4*	0.399 (0.377-0.421)	0.265 (0.226-0.305)	0.722 (0.686-0.756)
Prov-Gigapath	Center-4*	0.347 (0.338-0.359)	0.176 (0.148-0.207)	0.717 (0.684-0.747)
GPFM	Center-4*	**0.550 (0.520-0.579)**	**0.450 (0.417-0.486)**	0.729 (0.692-0.768)
ResNet50	Center-5*	0.333 (0.333-0.333)	0.120 (0.086-0.147)	0.623 (0.555-0.690)
Phikon	Center-5*	0.333 (0.333-0.333)	0.120 (0.086-0.150)	0.740 (0.687-0.793)
Ctranspath	Center-5*	0.333 (0.333-0.333)	0.120 (0.090-0.150)	0.702 (0.634-0.764)
UNI	Center-5*	0.333 (0.333-0.333)	0.120 (0.090-0.150)	0.739 (0.689-0.790)
CONCH	Center-5*	0.390 (0.363-0.417)	0.226 (0.170-0.284)	0.754 (0.702-0.812)
PLIP	Center-5*	0.368 (0.347-0.392)	0.189 (0.136-0.240)	0.747 (0.696-0.798)
CHIEF	Center-5*	0.342 (0.333-0.357)	0.138 (0.100-0.178)	0.745 (0.688-0.797)
Prov-Gigapath	Center-5*	0.333 (0.333-0.333)	0.120 (0.086-0.150)	0.743 (0.703-0.785)
GPFM	Center-5*	**0.481 (0.442-0.518)**	**0.349 (0.294-0.405)**	**0.791 (0.733-0.846)**

Supplementary Table 17: Vascular invasion detection of gastric cancer. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	Center-1-GC	0.600 (0.492-0.710)	0.594 (0.486-0.707)	0.625 (0.490-0.754)
Phikon	Center-1-GC	0.700 (0.603-0.795)	0.697 (0.589-0.799)	0.760 (0.653-0.862)
Ctranspath	Center-1-GC	0.725 (0.628-0.816)	0.721 (0.612-0.807)	0.798 (0.693-0.889)
UNI	Center-1-GC	**0.750 (0.647-0.844)**	**0.749 (0.644-0.837)**	0.787 (0.680-0.881)
CONCH	Center-1-GC	0.675 (0.580-0.768)	0.665 (0.559-0.770)	0.764 (0.664-0.866)
PLIP	Center-1-GC	0.675 (0.582-0.768)	0.665 (0.557-0.762)	0.756 (0.640-0.859)
CHIEF	Center-1-GC	0.637 (0.530-0.740)	0.636 (0.525-0.735)	0.712 (0.589-0.813)
Prov-Gigapath	Center-1-GC	0.650 (0.554-0.747)	0.639 (0.531-0.745)	0.765 (0.661-0.862)
GPFM	Center-1-GC	0.737 (0.655-0.814)	0.725 (0.612-0.813)	**0.806 (0.707-0.892)**
ResNet50	Center-4*	0.554 (0.519-0.589)	0.420 (0.368-0.472)	0.612 (0.548-0.677)
Phikon	Center-4*	**0.637 (0.588-0.684)**	0.592 (0.535-0.643)	0.734 (0.673-0.786)
Ctranspath	Center-4*	0.620 (0.568-0.669)	0.576 (0.523-0.623)	0.684 (0.625-0.739)
UNI	Center-4*	0.636 (0.600-0.674)	0.543 (0.488-0.597)	0.717 (0.658-0.773)
CONCH	Center-4*	0.618 (0.565-0.670)	**0.602 (0.549-0.653)**	0.668 (0.612-0.731)
PLIP	Center-4*	0.545 (0.519-0.572)	0.392 (0.341-0.447)	0.639 (0.571-0.700)
CHIEF	Center-4*	0.577 (0.541-0.612)	0.458 (0.404-0.509)	0.625 (0.564-0.683)
Prov-Gigapath	Center-4*	0.612 (0.562-0.660)	0.586 (0.531-0.639)	0.674 (0.616-0.729)
GPFM	Center-4*	0.617 (0.582-0.652)	0.517 (0.463-0.573)	**0.736 (0.680-0.785)**
ResNet50	Center-5*	0.548 (0.504-0.592)	0.507 (0.445-0.572)	0.604 (0.527-0.682)
Phikon	Center-5*	**0.677 (0.612-0.740)**	0.678 (0.610-0.742)	**0.733 (0.664-0.801)**
Ctranspath	Center-5*	0.659 (0.606-0.714)	0.659 (0.589-0.723)	0.727 (0.657-0.796)
UNI	Center-5*	0.674 (0.615-0.734)	0.679 (0.614-0.740)	0.722 (0.649-0.797)
CONCH	Center-5*	0.640 (0.577-0.705)	0.635 (0.567-0.694)	0.691 (0.611-0.766)
PLIP	Center-5*	0.599 (0.550-0.651)	0.579 (0.513-0.645)	0.678 (0.606-0.746)
CHIEF	Center-5*	0.599 (0.555-0.647)	0.579 (0.510-0.643)	0.693 (0.621-0.761)
Prov-Gigapath	Center-5*	0.633 (0.573-0.690)	0.634 (0.566-0.697)	0.668 (0.590-0.739)
GPFM	Center-5*	**0.677 (0.627-0.732)**	**0.682 (0.610-0.743)**	0.727 (0.654-0.792)

Supplementary Table 18: Perineural invasion detection in gastric cancer. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	Center-1-GC	0.516 (0.467-0.578)	0.453 (0.365-0.561)	0.870 (0.781-0.946)
Phikon	Center-1-GC	0.887 (0.812-0.947)	0.868 (0.782-0.937)	0.960 (0.909-0.995)
Ctranspath	Center-1-GC	0.915 (0.853-0.971)	0.895 (0.821-0.958)	0.949 (0.893-0.990)
UNI	Center-1-GC	0.900 (0.823-0.963)	0.903 (0.832-0.963)	0.963 (0.916-0.995)
CONCH	Center-1-GC	0.907 (0.840-0.963)	0.893 (0.813-0.959)	0.979 (0.947-0.999)
PLIP	Center-1-GC	0.824 (0.739-0.907)	0.813 (0.716-0.899)	0.904 (0.830-0.964)
CHIEF	Center-1-GC	0.872 (0.782-0.953)	0.875 (0.788-0.947)	0.938 (0.877-0.983)
Prov-Gigapath	Center-1-GC	**0.964 (0.907-1.000)**	**0.972 (0.926-1.000)**	**0.996 (0.986-1.000)**
GPFM	Center-1-GC	0.909 (0.830-0.972)	0.916 (0.847-0.974)	0.966 (0.912-1.000)
ResNet50	Center-4*	0.535 (0.511-0.557)	0.347 (0.299-0.393)	0.608 (0.545-0.667)
Phikon	Center-4*	0.664 (0.627-0.706)	0.581 (0.527-0.629)	0.743 (0.690-0.795)
Ctranspath	Center-4*	0.655 (0.603-0.702)	0.608 (0.557-0.660)	0.751 (0.698-0.802)
UNI	Center-4*	0.656 (0.617-0.695)	0.563 (0.505-0.617)	0.763 (0.708-0.811)
CONCH	Center-4*	0.676 (0.634-0.721)	**0.611 (0.555-0.668)**	0.761 (0.711-0.810)
PLIP	Center-4*	0.570 (0.538-0.605)	0.433 (0.382-0.485)	0.712 (0.652-0.764)
CHIEF	Center-4*	0.635 (0.600-0.674)	0.535 (0.482-0.593)	0.751 (0.698-0.803)
Prov-Gigapath	Center-4*	**0.680 (0.636-0.723)**	**0.611 (0.551-0.658)**	0.758 (0.705-0.807)
GPFM	Center-4*	0.652 (0.616-0.691)	0.560 (0.507-0.614)	**0.768 (0.716-0.819)**
ResNet50	Center-5*	0.536 (0.503-0.574)	0.486 (0.426-0.553)	0.669 (0.596-0.737)
Phikon	Center-5*	0.742 (0.681-0.802)	0.748 (0.687-0.806)	0.844 (0.787-0.898)
Ctranspath	Center-5*	0.730 (0.669-0.793)	0.726 (0.662-0.783)	0.816 (0.755-0.869)
UNI	Center-5*	0.758 (0.695-0.814)	0.764 (0.701-0.820)	0.840 (0.783-0.891)
CONCH	Center-5*	0.742 (0.689-0.797)	0.696 (0.636-0.754)	0.840 (0.781-0.889)
PLIP	Center-5*	0.622 (0.573-0.673)	0.621 (0.550-0.694)	0.779 (0.706-0.835)
CHIEF	Center-5*	**0.787 (0.728-0.847)**	**0.797 (0.736-0.851)**	0.842 (0.787-0.896)
Prov-Gigapath	Center-5*	0.735 (0.676-0.795)	0.744 (0.684-0.804)	0.819 (0.755-0.878)
GPFM	Center-5*	0.775 (0.713-0.835)	0.778 (0.714-0.830)	**0.846 (0.791-0.897)**

Supplementary Table 19: Average C-Index of Foundation Models Across 15 Survival Analysis Tasks. The best-performing and second-best-performing models are highlighted in **bold** and underlined, respectively.

Models	C-Index
ResNet50	0.619±0.080
Phikon	0.636±0.088
Ctranspath	0.631±0.071
UNI	0.643±0.079
CONCH	0.637±0.077
PLIP	0.623±0.073
CHIEF	0.638±0.070
Prov-Gigapath	0.642±0.072
GPFM	**0.665±0.071**

Supplementary Table 20: Performance of Survival Analysis on TCGA-BRCA, TCGA-BLCA, TCGA-KIRC, and TCGA-KIRP Datasets. The 95% CI is included in parentheses. The best and second-best performed models are **bolded** and underlined.

	TCGA-BRCA	TCGA-BLCA	TCGA-KIRC	TCGA-KIRP
ResNet50	0.569 (0.451-0.699)	0.576 (0.450-0.677)	0.731 (0.642-0.820)	0.784 (0.472-0.966)
Phikon	0.730 (0.631-0.820)	**0.664 (0.548-0.763)**	0.764 (0.673-0.848)	0.697 (0.385-0.923)
Ctranspath	0.658 (0.535-0.776)	0.550 (0.416-0.680)	0.726 (0.628-0.813)	0.788 (0.595-0.979)
UNI	0.613 (0.446-0.757)	0.564 (0.449-0.680)	0.755 (0.646-0.851)	**0.863 (0.659-1.000)**
CONCH	0.666 (0.538-0.795)	0.620 (0.519-0.720)	0.760 (0.675-0.835)	0.743 (0.462-0.919)
PLIP	0.601 (0.471-0.739)	0.555 (0.446-0.664)	0.739 (0.622-0.827)	0.784 (0.574-0.959)
CHIEF	0.632 (0.483-0.763)	0.601 (0.486-0.711)	0.717 (0.615-0.816)	0.793 (0.559-0.965)
Prov-Gigapath	0.655 (0.514-0.798)	0.627 (0.501-0.739)	0.733 (0.639-0.824)	0.772 (0.468-0.959)
GPFM	**0.739 (0.643-0.837)**	0.633 (0.523-0.732)	**0.774 (0.694-0.854)**	0.797 (0.531-0.991)

Supplementary Table 21: Performance of Survival Analysis on TCGA-STAD, TCGA-CESC, TCGA-LUAD, and CPTAC-LUAD Datasets. The 95% CI is included in parentheses. Models trained on the TCGA-LUAD dataset were directly applied to the CPTAC-LUAD dataset for testing. The best and second-best performed models are **bolded** and underlined.

	TCGA-STAD	TCGA-CESC	TCGA-LUAD	CPTAC-LUAD
ResNet50	0.580 (0.459-0.679)	0.703 (0.525-0.872)	0.578 (0.464-0.685)	0.596 (0.521-0.667)
Phikon	0.601 (0.505-0.700)	**0.768 (0.655-0.866)**	**0.614 (0.498-0.727)**	0.461 (0.384-0.541)
Ctranspath	0.632 (0.540-0.720)	0.683 (0.507-0.825)	0.552 (0.439-0.664)	0.658 (0.578-0.733)
UNI	0.595 (0.481-0.705)	0.683 (0.505-0.841)	0.605 (0.468-0.728)	0.605 (0.536-0.684)
CONCH	0.624 (0.524-0.726)	0.681 (0.509-0.826)	0.603 (0.474-0.720)	0.615 (0.541-0.691)
PLIP	0.573 (0.467-0.667)	0.655 (0.489-0.814)	0.560 (0.433-0.684)	0.623 (0.547-0.702)
CHIEF	0.628 (0.516-0.730)	0.672 (0.528-0.810)	0.588 (0.472-0.705)	**0.687 (0.611-0.763)**
Prov-Gigapath	0.572 (0.462-0.676)	0.731 (0.610-0.846)	0.549 (0.426-0.670)	0.580 (0.494-0.670)
GPFM	**0.636 (0.534-0.733)**	0.683 (0.491-0.849)	0.599 (0.479-0.707)	0.669 (0.595-0.747)

Supplementary Table 22: Performance of Survival Analysis on TCGA-COADREAD, TCGA-GBM, TCGA-LGG, and TCGA-LUSC datasets. The 95% CI is included in parentheses. The best and second-best performed model are **bolded** and underlined.

	TCGA-COADREAD	TCGA-GBM	TCGA-LGG	TCGA-LUSC
ResNet50	0.632 (0.496-0.767)	0.533 (0.434-0.629)	0.723 (0.566-0.860)	0.573 (0.462-0.692)
Phikon	0.660 (0.534-0.762)	0.534 (0.452-0.613)	0.710 (0.555-0.850)	0.575 (0.462-0.677)
Ctranspath	0.656 (0.522-0.784)	0.565 (0.475-0.647)	0.693 (0.493-0.868)	0.579 (0.468-0.682)
UNI	0.662 (0.543-0.779)	0.580 (0.498-0.659)	0.687 (0.527-0.830)	0.572 (0.456-0.693)
CONCH	0.593 (0.434-0.766)	0.559 (0.474-0.651)	**0.771 (0.635-0.893)**	0.531 (0.412-0.637)
PLIP	0.673 (0.540-0.805)	0.539 (0.447-0.629)	0.684 (0.538-0.816)	0.570 (0.476-0.670)
CHIEF	**0.682 (0.557-0.803)**	0.516 (0.422-0.605)	0.694 (0.551-0.824)	0.572 (0.462-0.701)
Prov-Gigapath	0.675 (0.537-0.802)	0.581 (0.496-0.663)	0.728 (0.599-0.845)	**0.599 (0.502-0.696)**
GPFM	0.678 (0.552-0.788)	**0.590 (0.500-0.676)**	0.731 (0.562-0.872)	0.581 (0.458-0.692)

Supplementary Table 23: Performance of Survival Analysis on TCGA-HNSC, HANCOCK (external validation), and TCGA-SKCM datasets. The 95% CI is included in parentheses. The best and second-best performed model are **bolded** and underlined.

	TCGA-HNSC	HANCOCK	TCGA-SKCM
ResNet50	0.558 (0.467-0.650)	0.494 (0.452-0.537)	0.661 (0.543-0.762)
Phikon	0.638 (0.541-0.733)	0.515 (0.475-0.558)	0.604 (0.491-0.707)
Ctranspath	0.584 (0.481-0.690)	0.528 (0.484-0.572)	0.616 (0.507-0.707)
UNI	**0.663 (0.567-0.751)**	**0.560 (0.517-0.601)**	0.639 (0.536-0.736)
CONCH	0.625 (0.523-0.725)	0.502 (0.462-0.542)	**0.669 (0.585-0.749)**
PLIP	0.608 (0.519-0.694)	0.522 (0.481-0.566)	0.663 (0.567-0.755)
CHIEF	0.639 (0.543-0.716)	0.531 (0.487-0.572)	0.625 (0.527-0.716)
Prov-Gigapath	0.651 (0.564-0.739)	0.533 (0.491-0.572)	0.643 (0.528-0.749)
GPFM	0.661 (0.567-0.759)	0.535 (0.497-0.579)	0.667 (0.547-0.777)

Supplementary Table 24: Average Tissue Classification Performance of Foundation Models across 16 Patch-level Tissue tasks. The best-performing and second-best-performing models are highlighted in **bold** and underlined, respectively.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.745±0.160	0.732±0.174	0.906±0.096
Phikon	0.820±0.155	0.810±0.162	0.934±0.075
Ctranspath	0.797±0.161	0.792±0.165	0.929±0.076
UNI	0.851±0.133	0.848±0.137	0.939±0.072
CONCH	0.820±0.130	0.818±0.131	0.932±0.077
PLIP	0.756±0.150	0.743±0.164	0.899±0.116
CHIEF	0.790±0.150	0.785±0.154	0.926±0.082
Prov-Gigapath	0.856±0.132	0.856±0.135	0.944±0.065
GPFM	**0.866±0.136**	**0.865±0.142**	**0.946±0.066**

Supplementary Table 25: CRC tissue classification performance of different foundation models on CRC-100K dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.792 (0.782-0.802)	0.775 (0.765-0.785)	0.983 (0.982-0.985)
Phikon	0.867 (0.859-0.875)	0.842 (0.833-0.850)	0.992 (0.991-0.993)
Ctranspath	0.853 (0.844-0.861)	0.833 (0.825-0.843)	**0.995 (0.994-0.996)**
UNI	0.879 (0.872-0.886)	0.849 (0.841-0.858)	0.991 (0.990-0.992)
CONCH	0.855 (0.847-0.863)	0.824 (0.815-0.833)	0.993 (0.992-0.994)
PLIP	0.804 (0.796-0.813)	0.764 (0.755-0.772)	0.990 (0.989-0.992)
CHIEF	0.802 (0.795-0.810)	0.749 (0.741-0.758)	**0.995 (0.994-0.995)**
Prov-Gigapath	**0.940 (0.934-0.947)**	**0.935 (0.928-0.941)**	0.994 (0.992-0.995)
GPFM	0.896 (0.888-0.902)	0.872 (0.865-0.881)	**0.995 (0.994-0.996)**

Supplementary Table 26: CCRCC tissue classification performance of different foundation models on CCRCC-TCGA-HEL dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.930 (0.919-0.942)	0.934 (0.925-0.944)	0.993 (0.991-0.995)
Phikon	0.949 (0.936-0.960)	0.955 (0.946-0.963)	**0.997 (0.996-0.998)**
Ctranspath	0.936 (0.923-0.948)	0.938 (0.926-0.946)	0.994 (0.992-0.996)
UNI	0.946 (0.932-0.956)	0.950 (0.941-0.959)	0.996 (0.995-0.997)
CONCH	0.934 (0.920-0.946)	0.939 (0.929-0.949)	0.994 (0.992-0.995)
PLIP	0.920 (0.905-0.932)	0.919 (0.909-0.929)	0.992 (0.991-0.994)
CHIEF	0.933 (0.921-0.944)	0.935 (0.924-0.944)	0.994 (0.993-0.995)
Prov-Gigapath	0.946 (0.935-0.957)	0.948 (0.938-0.957)	**0.997 (0.995-0.997)**
GPFM	**0.953 (0.939-0.962)**	**0.956 (0.947-0.964)**	**0.997 (0.994-0.998)**

Supplementary Table 27: Breast cancer tissue classification performance of different foundation models on BACH and BreakHis dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	BACH	0.865 (0.788-0.932)	0.856 (0.776-0.928)	0.977 (0.958-0.992)
Phikon	BACH	0.918 (0.845-0.971)	0.915 (0.842-0.965)	0.988 (0.975-0.998)
Ctranspath	BACH	0.927 (0.865-0.975)	0.919 (0.861-0.965)	0.998 (0.993-1.000)
UNI	BACH	0.960 (0.915-1.000)	**0.966 (0.911-1.000)**	**1.000 (0.999-1.000)**
CONCH	BACH	0.934 (0.879-0.981)	0.933 (0.885-0.986)	0.996 (0.988-1.000)
PLIP	BACH	0.799 (0.714-0.871)	0.791 (0.698-0.880)	0.959 (0.926-0.981)
CHIEF	BACH	0.925 (0.865-0.975)	0.924 (0.859-0.975)	0.998 (0.993-1.000)
Prov-Gigapath	BACH	0.925 (0.866-0.975)	0.924 (0.859-0.974)	0.995 (0.986-1.000)
GPFM	BACH	**0.963 (0.919-1.000)**	0.965 (0.915-1.000)	0.998 (0.994-1.000)
ResNet50	BreakHis	0.937 (0.923-0.950)	0.938 (0.925-0.951)	0.986 (0.981-0.990)
Phikon	BreakHis	0.973 (0.964-0.981)	0.973 (0.965-0.982)	0.997 (0.996-0.998)
Ctranspath	BreakHis	0.962 (0.952-0.972)	0.961 (0.951-0.971)	0.995 (0.992-0.997)
UNI	BreakHis	**0.977 (0.967-0.984)**	**0.976 (0.968-0.984)**	**0.998 (0.997-0.999)**
CONCH	BreakHis	0.950 (0.935-0.961)	0.952 (0.941-0.963)	0.991 (0.986-0.994)
PLIP	BreakHis	0.943 (0.929-0.954)	0.940 (0.927-0.951)	0.989 (0.986-0.993)
CHIEF	BreakHis	0.961 (0.950-0.972)	0.961 (0.950-0.971)	0.995 (0.993-0.997)
Prov-Gigapath	BreakHis	0.974 (0.966-0.983)	0.974 (0.965-0.982)	**0.998 (0.997-0.999)**
GPFM	BreakHis	0.974 (0.965-0.984)	**0.976 (0.968-0.984)**	**0.998 (0.997-0.999)**

Supplementary Table 28: CRC polyp classification performance of different foundation models on UniToPatho datasets. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.397 (0.376-0.417)	0.384 (0.359-0.406)	0.830 (0.819-0.840)
Phikon	0.379 (0.358-0.398)	0.375 (0.356-0.395)	0.838 (0.828-0.847)
Ctranspath	0.310 (0.289-0.331)	0.302 (0.285-0.324)	0.836 (0.828-0.844)
UNI	0.462 (0.443-0.486)	0.455 (0.433-0.474)	0.840 (0.830-0.850)
CONCH	**0.522 (0.499-0.550)**	**0.527 (0.501-0.550)**	**0.865 (0.855-0.875)**
PLIP	0.437 (0.413-0.458)	0.418 (0.395-0.441)	0.823 (0.812-0.834)
CHIEF	0.394 (0.373-0.415)	0.386 (0.362-0.410)	0.830 (0.821-0.838)
Prov-Gigapath	0.442 (0.422-0.462)	0.437 (0.413-0.461)	0.845 (0.835-0.855)
GPFM	0.444 (0.420-0.463)	0.433 (0.412-0.456)	0.844 (0.834-0.851)

Supplementary Table 29: MSI screening performance of different foundation models on CRC-MSI dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.654 (0.646-0.661)	0.587 (0.581-0.592)	0.706 (0.699-0.714)
Phikon	0.695 (0.689-0.703)	0.632 (0.626-0.638)	0.772 (0.766-0.779)
Ctranspath	0.728 (0.721-0.734)	0.647 (0.641-0.652)	0.802 (0.796-0.808)
UNI	0.719 (0.713-0.727)	0.670 (0.664-0.676)	0.797 (0.790-0.803)
CONCH	0.734 (0.727-0.741)	0.669 (0.663-0.675)	0.810 (0.804-0.817)
PLIP	0.639 (0.633-0.647)	0.589 (0.583-0.595)	0.691 (0.683-0.698)
CHIEF	0.717 (0.710-0.724)	0.648 (0.642-0.653)	0.791 (0.785-0.798)
Prov-Gigapath	**0.740 (0.734-0.746)**	**0.696 (0.689-0.701)**	**0.836 (0.830-0.842)**
GPFM	0.733 (0.726-0.740)	0.672 (0.666-0.678)	0.812 (0.805-0.818)

Supplementary Table 30: Pan-cancer tissue classification performance of different foundation models on PanCancer-TCGA dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. As shown in Figure 5j, the distribution of bootstrapped AUC values is highly centered. As a result, the CI for the AUC is very narrow.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.630 (0.625-0.636)	0.640 (0.636-0.646)	0.975 (0.974-0.976)
Phikon	0.924 (0.921-0.928)	0.926 (0.923-0.928)	**0.999 (0.999-0.999)**
Ctranspath	0.785 (0.780-0.790)	0.790 (0.786-0.795)	0.992 (0.991-0.992)
UNI	0.885 (0.882-0.889)	0.888 (0.885-0.892)	0.997 (0.997-0.997)
CONCH	0.784 (0.779-0.788)	0.789 (0.785-0.794)	0.991 (0.991-0.992)
PLIP	0.661 (0.656-0.667)	0.669 (0.664-0.675)	0.978 (0.978-0.979)
CHIEF	0.762 (0.757-0.767)	0.765 (0.760-0.770)	0.989 (0.989-0.990)
Prov-Gigapath	0.909 (0.905-0.912)	0.912 (0.909-0.915)	0.998 (0.998-0.998)
GPFM	**0.951 (0.949-0.954)**	**0.953 (0.950-0.955)**	**0.999 (0.999-0.999)**

Supplementary Table 31: TIL classification performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	PanCancer-TIL	0.813 (0.809-0.818)	0.843 (0.839-0.847)	0.946 (0.944-0.948)
Phikon	PanCancer-TIL	0.893 (0.889-0.896)	0.901 (0.897-0.904)	0.975 (0.974-0.977)
Ctranspath	PanCancer-TIL	0.857 (0.852-0.860)	0.880 (0.876-0.883)	0.965 (0.963-0.967)
UNI	PanCancer-TIL	**0.897 (0.893-0.900)**	0.905 (0.902-0.908)	0.977 (0.976-0.979)
CONCH	PanCancer-TIL	0.866 (0.862-0.870)	0.889 (0.885-0.892)	0.971 (0.969-0.973)
PLIP	PanCancer-TIL	0.810 (0.805-0.815)	0.843 (0.838-0.847)	0.949 (0.947-0.951)
CHIEF	PanCancer-TIL	0.845 (0.841-0.849)	0.873 (0.869-0.876)	0.961 (0.960-0.963)
Prov-Gigapath	PanCancer-TIL	0.886 (0.883-0.890)	0.894 (0.891-0.898)	0.968 (0.967-0.970)
GPFM	PanCancer-TIL	0.894 (0.890-0.897)	**0.908 (0.904-0.911)**	**0.978 (0.977-0.979)**
ResNet50	Center-3-TIL*	0.768 (0.763-0.773)	0.749 (0.743-0.755)	0.886 (0.881-0.891)
Phikon	Center-3-TIL*	0.781 (0.776-0.786)	0.762 (0.756-0.768)	0.917 (0.913-0.921)
Ctranspath	Center-3-TIL*	0.771 (0.766-0.776)	0.750 (0.744-0.757)	0.905 (0.900-0.909)
UNI	Center-3-TIL*	0.917 (0.913-0.921)	0.914 (0.910-0.918)	0.929 (0.924-0.932)
CONCH	Center-3-TIL*	0.915 (0.911-0.919)	0.912 (0.908-0.916)	0.934 (0.930-0.938)
PLIP	Center-3-TIL*	0.807 (0.802-0.812)	0.793 (0.788-0.799)	0.888 (0.884-0.893)
CHIEF	Center-3-TIL*	0.758 (0.753-0.764)	0.733 (0.727-0.740)	0.915 (0.911-0.919)
Prov-Gigapath	Center-3-TIL*	0.837 (0.832-0.842)	0.826 (0.820-0.831)	0.926 (0.922-0.930)
GPFM	Center-3-TIL*	**0.942 (0.939-0.946)**	**0.940 (0.937-0.944)**	**0.951 (0.948-0.955)**

Supplementary Table 32: ESCA subtyping performance of different foundation models on ESCA dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.601 (0.591-0.611)	0.553 (0.544-0.563)	0.886 (0.882-0.889)
Phikon	0.668 (0.662-0.676)	0.642 (0.635-0.651)	0.894 (0.890-0.897)
Ctranspath	0.642 (0.632-0.651)	0.660 (0.649-0.669)	0.899 (0.896-0.902)
UNI	**0.754 (0.744-0.761)**	**0.758 (0.749-0.765)**	**0.903 (0.901-0.904)**
CONCH	0.690 (0.682-0.698)	0.700 (0.691-0.707)	0.902 (0.899-0.904)
PLIP	0.601 (0.593-0.608)	0.552 (0.544-0.559)	0.889 (0.886-0.892)
CHIEF	0.609 (0.599-0.620)	0.628 (0.617-0.637)	0.899 (0.895-0.901)
Prov-Gigapath	0.725 (0.717-0.734)	0.738 (0.729-0.745)	0.902 (0.900-0.904)
GPFM	0.732 (0.724-0.740)	0.734 (0.725-0.740)	0.902 (0.899-0.904)

Supplementary Table 33: Metastatic tissue classification performance of different foundation models on PCAM dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.837 (0.834-0.841)	0.836 (0.832-0.840)	0.926 (0.923-0.928)
Phikon	0.898 (0.894-0.901)	0.897 (0.894-0.900)	0.969 (0.967-0.971)
Ctranspath	0.866 (0.862-0.869)	0.866 (0.862-0.869)	0.940 (0.937-0.942)
UNI	0.932 (0.929-0.934)	0.931 (0.929-0.934)	0.982 (0.981-0.983)
CONCH	0.903 (0.900-0.906)	0.903 (0.900-0.906)	0.965 (0.963-0.967)
PLIP	0.859 (0.856-0.863)	0.858 (0.854-0.862)	0.943 (0.941-0.945)
CHIEF	0.874 (0.871-0.878)	0.874 (0.870-0.877)	0.946 (0.943-0.948)
Prov-Gigapath	0.934 (0.931-0.936)	0.934 (0.931-0.936)	0.979 (0.978-0.980)
GPFM	**0.941 (0.939-0.944)**	**0.942 (0.939-0.944)**	**0.988 (0.987-0.989)**

Supplementary Table 34: Lung adenocarcinoma tissue classification performance of different foundation models on WSSS4LUAD dataset. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined.

	Balanced ACC	Weighted F1	AUC
ResNet50	0.911 (0.894-0.926)	0.910 (0.897-0.926)	0.989 (0.986-0.992)
Phikon	0.956 (0.944-0.967)	0.957 (0.944-0.966)	0.997 (0.995-0.998)
Ctranspath	0.947 (0.935-0.960)	0.949 (0.937-0.960)	0.997 (0.996-0.998)
UNI	0.951 (0.938-0.962)	0.951 (0.940-0.962)	0.997 (0.996-0.998)
CONCH	0.946 (0.933-0.960)	0.947 (0.935-0.959)	0.995 (0.993-0.997)
PLIP	0.927 (0.915-0.945)	0.934 (0.920-0.947)	0.994 (0.992-0.995)
CHIEF	0.950 (0.937-0.962)	0.951 (0.939-0.962)	0.997 (0.996-0.998)
Prov-Gigapath	0.943 (0.928-0.955)	0.941 (0.928-0.953)	0.996 (0.994-0.997)
GPFM	**0.961 (0.949-0.971)**	**0.959 (0.948-0.969)**	**0.998 (0.996-0.998)**

Supplementary Table 35: Colon tissue classification performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	Chaoyang	0.725 (0.704-0.746)	0.735 (0.715-0.757)	0.930 (0.921-0.938)
Phikon	Chaoyang	0.782 (0.762-0.804)	0.784 (0.763-0.803)	0.952 (0.945-0.958)
Ctranspath	Chaoyang	0.772 (0.752-0.793)	0.779 (0.757-0.798)	0.950 (0.943-0.957)
UNI	Chaoyang	0.790 (0.770-0.809)	0.789 (0.770-0.809)	0.952 (0.945-0.958)
CONCH	Chaoyang	0.759 (0.738-0.778)	0.762 (0.743-0.783)	0.942 (0.934-0.948)
PLIP	Chaoyang	0.747 (0.724-0.768)	0.755 (0.735-0.775)	0.941 (0.935-0.949)
CHIEF	Chaoyang	0.765 (0.743-0.785)	0.772 (0.749-0.792)	0.948 (0.942-0.956)
Prov-Gigapath	Chaoyang	**0.797 (0.776-0.816)**	0.799 (0.779-0.818)	**0.957 (0.952-0.963)**
GPFM	Chaoyang	**0.797 (0.776-0.817)**	**0.803 (0.784-0.821)**	0.956 (0.950-0.963)
ResNet50	Center-3-Colon*	0.560 (0.556-0.564)	0.495 (0.488-0.502)	0.768 (0.762-0.774)
Phikon	Center-3-Colon*	0.684 (0.679-0.689)	0.678 (0.671-0.684)	0.841 (0.835-0.846)
Ctranspath	Center-3-Colon*	0.700 (0.695-0.706)	0.701 (0.694-0.707)	0.826 (0.821-0.832)
UNI	Center-3-Colon*	0.724 (0.718-0.729)	0.724 (0.717-0.730)	0.868 (0.863-0.873)
CONCH	Center-3-Colon*	0.731 (0.725-0.737)	0.730 (0.724-0.736)	0.803 (0.796-0.809)
PLIP	Center-3-Colon*	0.626 (0.621-0.632)	0.603 (0.596-0.610)	0.770 (0.763-0.776)
CHIEF	Center-3-Colon*	0.690 (0.684-0.695)	0.688 (0.682-0.695)	0.820 (0.814-0.825)
Prov-Gigapath	Center-3-Colon*	**0.885 (0.881-0.890)**	**0.893 (0.889-0.898)**	**0.913 (0.909-0.917)**
GPFM	Center-3-Colon*	0.828 (0.823-0.833)	0.836 (0.831-0.842)	0.891 (0.886-0.896)

Supplementary Table 36: Gastric tissue classification performance of different foundation models. Non-parametric bootstrapping with 1,000 bootstrap replicates is employed for statistical analysis. The 95% CI is included in parentheses. Best performing model for each metric is **bolded** and second-best performing model is underlined. * indicates the external validation.

	Cohort	Balanced ACC	Weighted F1	AUC
ResNet50	GasHisDB	0.953 (0.947-0.958)	0.954 (0.949-0.959)	0.992 (0.990-0.993)
Phikon	GasHisDB	0.995 (0.993-0.997)	0.995 (0.994-0.997)	**1.000 (1.000-1.000)**
Ctranspath	GasHisDB	0.980 (0.976-0.983)	0.980 (0.976-0.983)	0.998 (0.998-0.999)
UNI	GasHisDB	0.996 (0.994-0.997)	0.996 (0.994-0.997)	**1.000 (1.000-1.000)**
CONCH	GasHisDB	0.981 (0.978-0.985)	0.981 (0.978-0.985)	0.998 (0.998-0.999)
PLIP	GasHisDB	0.958 (0.954-0.963)	0.958 (0.953-0.963)	0.993 (0.991-0.994)
CHIEF	GasHisDB	0.979 (0.976-0.983)	0.980 (0.976-0.983)	0.998 (0.997-0.999)
Prov-Gigapath	GasHisDB	0.995 (0.994-0.997)	0.996 (0.994-0.997)	**1.000 (1.000-1.000)**
GPFM	GasHisDB	**0.997 (0.996-0.998)**	**0.997 (0.996-0.998)**	**1.000 (1.000-1.000)**
ResNet50	Center-3-GC*	0.551 (0.539-0.563)	0.525 (0.506-0.547)	0.721 (0.696-0.742)
Phikon	Center-3-GC*	0.751 (0.730-0.769)	0.729 (0.711-0.746)	0.812 (0.789-0.834)
Ctranspath	Center-3-GC*	0.708 (0.689-0.727)	0.716 (0.696-0.736)	0.777 (0.753-0.799)
UNI	Center-3-GC*	0.819 (0.800-0.835)	0.841 (0.823-0.857)	0.796 (0.772-0.819)
CONCH	Center-3-GC*	0.623 (0.605-0.641)	0.634 (0.614-0.654)	0.763 (0.741-0.785)
PLIP	Center-3-GC*	0.555 (0.534-0.576)	0.505 (0.486-0.525)	0.598 (0.572-0.622)
CHIEF	Center-3-GC*	0.681 (0.660-0.701)	0.692 (0.673-0.713)	0.746 (0.723-0.768)
Prov-Gigapath	Center-3-GC*	0.819 (0.800-0.836)	0.841 (0.825-0.859)	0.794 (0.772-0.820)
GPFM	Center-3-GC*	**0.852 (0.837-0.870)**	**0.886 (0.871-0.900)**	**0.828 (0.804-0.851)**

Supplementary Table 37: CRC Tissue Retrieval Performance on CRC-100K Dataset. The table reports the Top-1, Top-3, and Top-5 ACC of different foundation models on the CRC-100K dataset for CRC tissue retrieval. Non-parametric bootstrapping with 1,000 bootstrap replicates is used for statistical analysis. The 95% CI is included in parentheses. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	ACC@1	ACC@3	ACC@5
ResNet50	0.777 (0.767-0.787)	0.940 (0.934-0.946)	0.958 (0.954-0.962)
Phikon	0.884 (0.876-0.892)	0.964 (0.960-0.968)	0.966 (0.962-0.970)
Ctranspath	0.825 (0.817-0.833)	0.910 (0.906-0.914)	0.915 (0.911-0.919)
UNI	0.911 (0.903-0.919)	0.981 (0.977-0.985)	0.983 (0.981-0.985)
CONCH	0.879 (0.871-0.887)	0.974 (0.970-0.978)	0.976 (0.972-0.980)
PLIP	0.798 (0.790-0.806)	0.909 (0.905-0.913)	0.915 (0.911-0.919)
CHIEF	0.820 (0.814-0.826)	0.882 (0.880-0.884)	0.885 (0.883-0.887)
Prov-Gigapath	**0.925 (0.917-0.933)**	0.988 (0.986-0.990)	0.993 (0.991-0.995)
GPFM	0.906 (0.900-0.912)	**0.993 (0.991-0.995)**	**0.995 (0.993-0.997)**

Supplementary Table 38: VQA performance of different foundation models on PathVQA dataset. The open-ended, closed-ended and overall ACC are reported. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	Open ACC	Closed ACC	Overall ACC
ResNet50	28.17%(26.63%-29.70%)	86.52%(85.43%-87.61%)	57.32% (56.41%-58.28%)
Phikon	30.78%(29.28%-32.29%)	87.20%(86.13%-88.27%)	58.97% (58.10%-59.93%)
Ctranspath	31.11%(29.58%-32.65%)	87.51%(86.44%-88.58%)	59.35% (58.42%-60.28%)
UNI	33.85%(32.28%-35.42%)	**88.69%(87.64%-89.74%)**	61.28% (60.39%-62.23%)
CONCH	**37.08%(35.40%-38.77%)**	88.51%(87.49%-89.53%)	**62.84% (61.84%-63.81%)**
PLIP	30.83%(29.29%-32.37%)	88.02%(86.94%-89.09%)	59.42% (58.48%-60.35%)
CHIEF	32.11% (30.49%-33.60%)	88.36% (87.28%-89.46%)	60.23% (59.26%-61.18%)
Prov-Gigapath	33.46% (31.80%-35.04%)	88.35% (87.26%-89.40%)	60.91% (59.88%-61.90%)
GPFM	34.26%(32.67%-35.84%)	88.41%(87.32%-89.49%)	61.39% (60.39%-62.30%)

Supplementary Table 39: Performance of WSI-level VQA on WSI-VQA dataset. The best performing model for each metric is **bolded** and the second-best performing model is underlined. CE ACC represents Close-Ended accuracy.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CE ACC
ResNet50	0.386	0.324	0.301	0.157	0.230	0.456	0.482
Phikon	0.359	0.323	0.322	0.189	0.227	0.450	**0.536**
Ctranspath	0.386	0.333	0.316	0.162	0.238	0.459	0.462
UNI	0.381	0.322	0.315	0.202	0.231	0.458	0.482
CONCH	0.386	0.332	0.314	0.177	0.234	0.456	0.487
PLIP	0.388	0.317	0.288	0.148	0.225	0.457	0.474
CHIEF	**0.400**	**0.350**	**0.335**	0.206	**0.245**	**0.474**	0.497
Prov-Gigapath	0.381	0.322	0.303	0.179	0.234	0.470	0.526
GPFM	0.395	0.345	0.326	**0.214**	0.240	0.470	0.503

Supplementary Table 40: Performance of foundation models in WSI report generation on TCGA WSI-Report dataset. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ResNet50	0.252±0.003	0.113±0.003	0.062±0.003	0.039±0.003	0.093±0.001	0.179±0.002
Phikon	**0.404±0.005**	**0.290±0.005**	**0.225±0.005**	**0.181±0.005**	**0.178±0.003**	**0.336±0.005**
Ctranspath	0.254±0.004	0.131±0.003	0.079±0.003	0.052±0.003	0.097±0.002	0.189±0.003
UNI	0.363±0.005	0.250±0.005	0.189±0.005	0.151±0.004	0.156±0.003	0.298±0.005
CONCH	0.246±0.005	0.149±0.004	0.104±0.004	0.077±0.003	0.110±0.002	0.208±0.004
PLIP	0.265±0.004	0.135±0.003	0.080±0.003	0.053±0.003	0.102±0.002	0.188±0.003
CHIEF	0.278±0.003	0.147±0.003	0.088±0.003	0.057±0.002	0.105±0.002	0.201±0.003
Prov-Gigapath	0.325±0.005	0.216±0.005	0.159±0.004	0.125±0.004	0.140±0.002	0.265±0.005
GPFM	0.384±0.005	0.271±0.005	0.210±0.005	0.169±0.005	0.168±0.003	0.320±0.005

Supplementary Table 41: Performance of foundation models in WSI report generation on PatchGastricADC22 dataset. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ResNet50	0.596±0.019	0.496±0.021	0.424±0.022	0.369±0.023	0.301±0.011	0.564±0.021
Phikon	**0.655±0.025**	**0.577±0.027**	**0.522±0.029**	**0.481±0.030**	**0.347±0.015**	**0.623±0.026**
Ctranspath	0.643±0.020	0.556±0.022	0.495±0.023	0.450±0.025	0.334±0.012	0.598±0.022
UNI	0.609±0.023	0.533±0.025	0.482±0.027	0.444±0.028	0.327±0.014	0.596±0.024
CONCH	0.641±0.019	0.555±0.023	0.495±0.025	0.450±0.026	0.337±0.013	0.599±0.022
PLIP	0.650±0.021	0.560±0.024	0.498±0.026	0.451±0.027	0.338±0.013	0.599±0.023
CHIEF	0.594±0.021	0.494±0.023	0.425±0.025	0.372±0.026	0.298±0.013	0.561±0.025
Prov-Gigapath	0.637±0.022	0.555±0.025	0.497±0.026	0.454±0.027	0.338±0.013	0.601±0.023
GPFM	0.651±0.021	0.569±0.023	0.512±0.025	0.470±0.026	0.343±0.013	0.606±0.025

Supplementary Table 42: Performance of foundation models in WSI report generation on TCGA WSI-Report dataset, split by cancer types. Report generation results on breast, lung, and kidney are reported respectively. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
			Breast			
ResNet50	0.228±0.007	0.079±0.004	0.034±0.003	0.016±0.002	0.081±0.002	0.157±0.004
Phikon	**0.416±0.006**	**0.312±0.005**	**0.251±0.004**	**0.208±0.004**	**0.194±0.003**	**0.364±0.005**
Ctranspath	0.254±0.007	0.131±0.004	0.078±0.002	0.047±0.002	0.097±0.004	0.192±0.004
UNI	0.361±0.006	0.255±0.005	0.198±0.004	0.161±0.003	0.162±0.004	0.306±0.005
CONCH	0.265±0.006	0.163±0.005	0.113±0.004	0.084±0.003	0.117±0.003	0.226±0.005
PLIP	0.269±0.005	0.148±0.003	0.093±0.003	0.061±0.002	0.106±0.003	0.201±0.002
CHIEF	0.272±0.010	0.151±0.006	0.093±0.004	0.060±0.003	0.117±0.004	0.209±0.005
Prov-Gigapath	0.334±0.007	0.230±0.004	0.175±0.005	0.139±0.003	0.148±0.004	0.278±0.005
GPFM	0.390±0.007	0.289±0.004	0.231±0.005	0.192±0.004	0.182±0.003	0.346±0.006
			Lung			
ResNet50	0.224±0.008	0.085±0.005	0.035±0.003	0.016±0.002	0.078±0.002	0.159±0.004
Phikon	**0.405±0.009**	**0.284±0.006**	**0.211±0.005**	**0.162±0.004**	**0.179±0.004**	**0.346±0.007**
Ctranspath	0.150±0.011	0.066±0.006	0.032±0.004	0.015±0.002	0.058±0.004	0.131±0.006
UNI	0.329±0.009	0.220±0.006	0.158±0.004	0.119±0.003	0.140±0.004	0.279±0.006
CONCH	0.229±0.008	0.134±0.005	0.088±0.004	0.061±0.003	0.091±0.003	0.188±0.005
PLIP	0.198±0.007	0.079±0.005	0.035±0.004	0.014±0.003	0.072±0.003	0.139±0.004
CHIEF	0.209±0.012	0.098±0.007	0.044±0.004	0.019±0.003	0.079±0.004	0.167±0.007
Prov-Gigapath	0.308±0.009	0.199±0.006	0.140±0.004	0.102±0.003	0.132±0.004	0.260±0.006
GPFM	0.349±0.008	0.235±0.006	0.173±0.005	0.132±0.004	0.152±0.004	0.300±0.007
			Kidney			
ResNet50	0.426±0.006	0.281±0.004	0.202±0.003	0.153±0.003	0.187±0.002	0.320±0.004
Phikon	0.500±0.007	0.375±0.006	0.300±0.005	0.247±0.005	0.225±0.004	0.406±0.005
Ctranspath	0.415±0.011	0.269±0.007	0.193±0.005	0.147±0.004	0.184±0.005	0.318±0.007
UNI	0.450±0.006	0.333±0.005	0.267±0.005	0.222±0.004	0.201±0.004	0.364±0.005
CONCH	0.420±0.006	0.280±0.005	0.203±0.004	0.156±0.004	0.185±0.003	0.318±0.004
PLIP	0.400±0.006	0.259±0.004	0.185±0.003	0.141±0.002	0.171±0.002	0.303±0.003
CHIEF	0.384±0.004	0.233±0.005	0.153±0.004	0.106±0.003	0.153±0.004	0.280±0.005
Prov-Gigapath	0.416±0.006	0.292±0.005	0.224±0.005	0.179±0.004	0.184±0.004	0.329±0.005
GPFM	**0.504±0.008**	**0.381±0.006**	**0.307±0.005**	**0.255±0.004**	**0.226±0.004**	**0.407±0.005**

Supplementary Table 43: Human-based blind evaluation of foundation models in WSI report generation on TCGA WSI-report dataset, where the generated reports of breast, lung, and kidney cancers are used for evaluation. The number of reports in each score rated by the pathologist is listed and the average score is reported. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	Score: 0	Score: 0.3	Score: 0.7	Score: 1	Avg.
Breast					
ResNet50	184	4	0	0	0.01
Phikon	18	136	34	0	0.34
Ctranspath	80	88	20	0	0.21
UNI	9	134	45	0	**0.38**
CONCH	36	125	27	0	0.30
PLIP	118	69	1	0	0.11
CHIEF	25	139	24	0	0.31
Prov-Gigapath	24	127	37	0	0.34
GPFM	15	126	47	0	**0.38**
Lung					
ResNet50	153	3	1	0	0.01
Phikon	7	109	41	0	0.39
Ctranspath	19	87	51	0	0.39
UNI	13	109	35	0	0.36
CONCH	11	120	26	0	0.35
PLIP	53	104	0	0	0.20
CHIEF	15	113	29	0	0.35
Prov-Gigapath	13	111	33	0	0.36
GPFM	6	100	51	0	**0.42**
Kidney					
ResNet50	175	0	0	0	0.00
Phikon	5	86	84	0	0.48
Ctranspath	32	76	67	0	0.39
UNI	2	91	82	0	0.48
CONCH	11	118	46	0	0.39
PLIP	172	3	0	0	0.01
CHIEF	9	100	66	0	0.44
Prov-Gigapath	5	86	82	2	0.49
GPFM	2	83	90	0	**0.50**

Supplementary Table 44: Performance Comparison of DINOv2 and GPFM Pretraining Methods Across 12 Tasks. DINOv2 represents the pretrainined foundation model without using Expert Knowledge Distillation compared with GPFM. Overall, the Expert Knowledge Distillation module shows an average improvement across balanced ACC, weighted F1 score, and AUC.

	Method	Balanced ACC	Weighted F1	AUC
CRC-100K	DINOv2	0.845	0.822	0.990
	GPFM	0.896(+0.051)	0.872(+0.050)	0.995(+0.005)
WSSS4LUAD	DINOv2	0.957	0.956	0.998
	GPFM	0.961(+0.004)	0.959(+0.003)	0.998(+0.000)
PCAM	DINOv2	0.925	0.925	0.976
	GPFM	0.941(+0.016)	0.942(+0.017)	0.988(+0.012)
PanCancer-TCGA	DINOv2	0.939	0.940	0.999
	GPFM	0.951(+0.012)	0.953(+0.013)	0.999(+0.000)
PanCancer-TIL	DINOv2	0.857	0.864	0.963
	GPFM	0.894(+0.037)	0.908(+0.044)	0.978(+0.015)
chaoyang	DINOv2	0.802	0.808	0.957
	GPFM	0.797(-0.005)	0.803(-0.005)	0.956(-0.001)
CCRCC-TCGA-HEL	DINOv2	0.945	0.951	0.996
	GPFM	0.953(+0.008)	0.956(+0.005)	0.997(+0.001)
BreakHis	DINOv2	0.984	0.982	0.999
	GPFM	0.974(-0.008)	0.976(-0.006)	0.998(-0.001)
BACH	DINOv2	0.922	0.920	0.990
	GPFM	0.963(+0.041)	0.965(+0.045)	0.998(+0.008)
UniToPatho	DINOv2	0.457	0.431	0.844
	GPFM	0.444(-0.013)	0.433(+0.002)	0.844(+0.000)
CRC-MSI	DINOv2	0.679	0.655	0.777
	GPFM	0.733(+0.054)	0.672(+0.023)	0.812(+0.035)
ESCA	DINOv2	0.705	0.705	0.900
	GPFM	0.732(+0.027)	0.734(+0.029)	0.902(+0.002)
Average	DINOv2	0.835	0.830	0.949
	GPFM	0.853(+0.018)	0.848(+0.018)	0.955(+0.006)

Supplementary Table 45: The primary site of tissues used for pretraining foundation models and downstream tasks evaluation.

Primariy Site	The Number of Slides
prostate	19,253
colon	9,870
lung	8,232
breast	7,721
female reproductive system	6,870
kidney	4,742
stomach	4,121
brain	3,283
skin	3,168
esophagus	3,100
artery	2,499
thyroid	2,064
pancreas	1,965
adipose	1,793
liver	1,681
lymph	1,660
heart	1,620
adrenal gland	1,359
head and neck	1,093
bladder	1,056
testis	1,007
muscle	1,001
nerve	975
tongue, tonsil and mouth	902
spleen	874
unknown	839
small intestine	798
soft tissue	524
peritoneum	310
larynx	303
thymus	252
minor salivary gland	247
rectosigmoid	240
eye	150
	95,572

Supplementary Table 46: The public datasets used in this study. Please note that some datasets may need permission before downloading.

Dataset	Link or Source
1. TCGA	https://portal.gdc.cancer.gov/
2. CPTAC	https://proteomic.datacommons.cancer.gov/pdc/
3. PANDA	https://www.kaggle.com/c/prostate-cancer-grade-assessment/data
4. NADT-Prostate	https://www.cancerimagingarchive.net/collection/nadt-prostate/
5. BCNB	https://bcnb.grand-challenge.org/
6. CAMELYON16	https://camelyon16.grand-challenge.org/Data/
7. CAMELYON17	https://camelyon17.grand-challenge.org/Data/
8. BRACS	https://www.bracs.icar.cnr.it/download/
9. TIGER2021	https://tiger.grand-challenge.org/
10. MIDOG2022	https://midog.deepmicroscopy.org/download-dataset/
11. AGGC2022	https://aggc22.grand-challenge.org/
12. O.B.R.	https://www.cancerimagingarchive.net/collection/ovarian-bevacizumab-response/
13. ACROBAT2023	https://acrobat.grand-challenge.org/
14. AML-C-LMU	https://www.cancerimagingarchive.net/collection/aml-cytomorphology_lmu/
15. ARCH	https://warwick.ac.uk/fac/cross_fac/tia/data/arch
16. BACH	https://zenodo.org/records/3632035
17. CAMEL	https://drive.google.com/open?id=1brr8CnU6ddzAYT157wkdXjbSzoiIDF9y
18. DiagSet	https://ai-econsilio.diag.pl/
19. DLBCL	https://github.com/stanfordmlgroup/DLBCL-Morph
20. GTEx	https://gtexportal.org/home/histologyPage
21. HunCRC	https://www.cancerimagingarchive.net/collection/hungarian-colorectal-screening/
22. Janowczyk	https://andrewjanowczyk.com/use-case-1-nuclei-segmentation/
23. LC25000	https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af
24. MIDOG2021	https://imig.science/midog2021/download-dataset/
25. OCELOT	https://zenodo.org/record/7844149
26. Oste. Tumor	https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/
27. PAIP2019	https://paip2019.grand-challenge.org/
28. PAIP2020	https://paip2020.grand-challenge.org/
29. PAIP2021	https://paip2021.grand-challenge.org/
30. Post-NAT-BRCA	https://www.cancerimagingarchive.net/collection/post-nat-brca/
31. SICAPv2	https://data.mendeley.com/datasets/9xxm58dvs3/1
32. SLN-Breast	https://www.cancerimagingarchive.net/collection/sln-breast/
33. SPIE2019	https://breastpathq.grand-challenge.org/
34. PatchGastricADC22	https://zenodo.org/records/6550925
35. UBC-OCEAN	https://www.kaggle.com/competitions/UBC-OCEAN/data
36. WSI-VQA	https://github.com/cpystan/WSI-VQA
37. CRC-100K	https://zenodo.org/records/1214456
38. CRC-MSI	https://zenodo.org/records/3832231
39. CCRCC-TCGA-HEL	https://zenodo.org/records/7898308
40. PanCancer-TCGA	https://zenodo.org/records/5889558
41. PanCancer-TIL	https://zenodo.org/records/6604094
42. ESCA	https://zenodo.org/records/7548828
43. PCAM	https://github.com/basveeling/pcam
44. BreakHis	https://www.kaggle.com/datasets/ambarish/breakhis
45. UniToPatho	https://ieee-dataport.org/open-access/unitopatho
46. Chaoyang	https://github.com/bupt-ai-cz/HSA-NRL
47. PathVQA	https://huggingface.co/datasets/flaviagiammarino/path-vqa
48. HistGen	https://github.com/dddavid4real/HistGen
49. IMP-CRS	https://rdm.inesctec.pt/dataset/nis-2023-008
50. HANCOCK	https://github.com/ankilab/HANCOCK_MultimodalDataset
51. GasHisDB	https://figshare.com/articles/dataset/GasHisSDB/15066147

Supplementary Table 47: The number of slides and processed patches of 33 datasets used for pretraining foundation models. "-" represents the dataset only providing ROIs.

Dataset Name	Number of Slides	Total Patches
TCGA	26,285	120,496,200
GTExPortal	24,467	31,892,017
CPTAC	7,164	11,768,225
CAMELYON17	841	4,612,382
HunCRC	200	3,369,925
BRACS	381	2,992,229
DiagSet	825	2,500,385
AGGC2022	286	2,130,584
CAMELYON16	288	1,706,890
DLBCL	203	1,524,388
PAIP2020	118	1,362,725
O.B.R	283	1,159,516
PAIP2021	220	1,048,840
NADT-Prostate	1,303	919,847
PANDA	7,114	905,206
PAIP2019	96	505,356
TIGER2021	174	312,835
BCNB	1,036	263,734
Post-NAT-BRCA	96	241,547
SLN-Breast	129	139,166
BACH	30	108,256
ACROBAT2023	153	76,128
MIDOG2022	395	43,342
ARCH	-	25,919
MIDOG2021	193	24,025
LC25000	-	19,678
SICAPv2	-	18,783
AML-C-LMU	-	18,365
CAMEL	-	16,744
OCELOT	-	3,201
SPIE2019	-	2,579
Janowczyk	-	2,260
Oste. Tumor	-	1,391
Total	72,280	190,212,668

Supplementary Table 48: The hyper parameters for pretraining the proposed foundation model. The pretraining is conducted on 2 DGX nodes with 16×80GB H800 GPUs.

	Hyperparamerters	Value
model	Layer number	24
	Feature dim	1024
	Patch size	14
	Heads number	16
	FFN layer	mlp
	Drop path ratio	0.4
	Layer scale	1e-5
optimization	Teacher momentum	0.992
	Total batch size	1,536
	Base learning rate	4e-4
	Minimum learning rate	1e-6
	Global crops scale	0.32, 1.0
	Global crops size	224
	Local crops scale	0.05, 0.32
	Local crops number	8
	Local crops size	98
	Gradient clip	3.0
	Warmup iterations	50,000
	Total iterations	500,000
loss weights	DINO	1.0
	iBOT	1.0
	CLS UNI	1.0
	Patch UNI	0.25
	CLS Phikon	0.5
	Patch Phikon	0.125
	CLS CONCH	1.0
	Patch CONCH	0.0

Supplementary Table 49: The configuration of different foundation models used for comparison. The details of the datasets used in GPFM are shown in Supplementary Table 48. UDK represents Unified Knowledge Distillation

Model	Data Source	WSIs	Patches	Model arch.	Model size	Pretraining
ResNet50	ImageNet	NA	NA	ResNet50	25M	Supervised
Ctranspath	TCGA+PAIP	32K	4.2M	SwinTrans.	28M	MoCoV3
Phikon	TCGA	6K	43M	ViT-B	86M	iBOT
UNI	Private+GTEx	100K	100M	ViT-L	307M	DINOv2
PLIP	OpenPath	NA	200K	ViT-B	86M	CLIP
CONCH	PMC-Path +EDU	NA	1.2M	ViT-B	86M	CoCa
CHIEF	Public+Private	60K	15M	SwinTrans.	28M	MoCoV3+CLIP
Prov-Gigapath	Private	171K	1.3B	Vit-g	1.1B	DINOv2+MAE
GPFM (our)	33 Public datasets	72K	190M	ViT-L	307M	UDK

Supplementary Table 50: The architecture of ABMIL model and training details for WSI classification and survival analysis.

Architecture	Two-layer ABMIL
Embedding Dimension	512
Hidden Dimensions	128
Dropout Rates	0.25
Optimizer	AdamW
Learning Rate	2e-4
WSI Classification Loss	Cross-entropy
Survival Analysis Loss	NLL loss
Maximum Epochs	100
Early Stopping	Yes

Supplementary Table 51: The datasets used for survival analysis.

Dataset	Cases	WSIs
TCGA-BRCA	1,023	1,089
TCGA-BLCA	376	446
TCGA-KIRC	498	504
TCGA-KIRP	261	285
TCGA-STAD	363	389
TCGA-CESC	250	260
TCGA-LUAD	455	518
TCGA-LUSC	452	484
TCGA-COADREAD	579	588
TCGA-GBM	372	856
TCGA-LGG	462	843
TCGA-SKCM	415	456
TCGA-HNSC	443	472
HANCOCK	749	1078

Supplementary Table 52: The public codes used in this study. Please note that the pretrained weights of UNI and CONCH need to be permitted before downloading.

code	source
UNI	https://huggingface.co/MahmoodLab/UNI
Phikon	https://huggingface.co/owkin/phikon
CONCH	https://huggingface.co/MahmoodLab/CONCH
CHIEF	https://github.com/hms-dbmi/CHIEF/
Prov-Gigapath	https://github.com/prov-gigapath/prov-gigapath
CLAM	https://github.com/mahmoodlab/CLAM
CTranspath	https://github.com/Xiyue-Wang/TransPath
PLIP	https://github.com/PathologyFoundation/plip
MUMC	https://github.com/pengfeiliHEU/MUMC
HistGen	https://github.com/dddavid4real/HistGen
Torchmetrics	https://github.com/Lightning-AI/torchmetrics
Scikit-learn	https://scikit-learn.org/stable/