CARE: A Molecular-Guided Foundation Model with Adaptive Region Modeling for Whole Slide Image Analysis

Supplementary Material

A. Backbone and Pretraining Details

Hyperparameter details for CARE are reported in Tab. 6. The regional self-attention and regional cross-attention modules use a small number of layers to reduce model complexity. Tables 7 and 8 list key hyperparameters for unimodal and cross-modal pretraining, respectively. Notably, during global and local cropping we avoid fixed rectangular windows. Instead, we sample irregular crops whose areas are proportional to sub-WSI patch counts, better reflecting the inherently irregular spatial distribution of tissue in WSIs. Under a frozen augmentation strategy, we stochastically apply two sequential transforms to each crop, selected from brightness, contrast, and posterize, with sampling governed by preset probabilities. This increases augmentation diversity across crops.

Cross-modal contrastive pretraining order. We align the slide encoder with RNA embeddings before aligning it with protein embeddings. First, the RNA and protein encoders are highly parameterized, containing a very large number of parameters. Jointly training all pairwise alignments (WSI–RNA–protein) would force smaller batch sizes, weakening representation learning. Second, the RNA encoder can be initialized from scGPT [8], whereas the protein encoder has only a protein-ID embedding table for initialization. Consequently, the protein-guided stage primarily provides fine-grained refinement of CARE following the stronger RNA-guided stage.

B. Effect of Different Pretraining Stages for CARE

Why iBOT Fits CARE? iBOT's L_{MIM} objective performs self-distillation between in-view patch tokens. For each masked patch, the student's representation at that location is aligned with the teacher's representation at the same location. Unlike a standard ViT, which applies global self-attention to the entire image, CARE restricts attention within an adaptive region. Each patch interacts only with semantically related neighboring tokens, avoiding interference from distant, semantically unrelated regions. This locality has two benefits. First, CARE increases within-region consistency and suppresses cross-region noise. Second, aligning the student and the teacher within the same adaptive region context sharpens and stabilizes the L_{MIM} supervisory signal, thereby improving pretraining convergence and downstream performance.

Molecularly guided ROI salience. SPF is a lightweight aggregation module that assigns that assigns region-specific weights to adaptive regions. By inspecting these learned weights, we can localize salient ROIs. A patient's molecular profile is strongly associated with ROIs in pathology slides. Leveraging this information to guide CARE's pretraining markedly improves CARE's ability to partition WSIs into adaptive regions. This step imposes constraints at a functionally more proximal molecular level, thereby improving the consistency of image representations.

Table 6. Hyperparameter Selection for CARE.

Hyperparameter	Value
Regional self-attention depth	2
Regional self-attention heads	8
Regional cross-attention heads	8
Adaptive region self-attention depth	5
Adaptive region self-attention heads	8
Window size	8
RSL target	0.5
λ_{SPF}	0.5
Input dimension	768
Output dimension	512

C. Description of Datasets and Tasks

Datasets. In this paper, we use nine datasets to construct 33 pathology image analysis benchmarks. The datasets are detailed below.

- **CPTAC [10].** We use data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), a large-scale National Cancer Institute initiative that provides proteogenomic profiles for multiple tumor types in order to better understand the molecular basis of cancer. Our experiments cover multiple CPTAC cohorts, including glioblastoma multiforme (GBM), breast invasive carcinoma (BRCA), clear cell renal cell carcinoma (CCRCC), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSCC), lung adenocarcinoma (LUAD), and pancreatic ductal adenocarcinoma (PDA).
- **EBRAINS [28].** We evaluate our method on data obtained from EBRAINS, the European Brain Research Infrastructure for neuroscience, computing and brain-related medicine.
- **MUT-HET-RCC [1].** The MUT-HET-RCC dataset is a clear cell renal cell carcinoma (ccRCC) cohort to study intratumoral mutation heterogeneity in renal cancer.

Table 7. Hyperparameter Selection during Unimodal Self-Supervised Pretraining. We use 8×80 GB A100 GPUs for the unimodal pretraining stage.

Hyperparameter	Value
Global crop patch ratio (% per sub-WSI)	90%
Global crop number	2
Local crop patch ratio (% per sub-WSI)	18%
Local crop number	6
Prediction mask ratio	0.3
Prediction mask variance	0.15
Prediction shape	block
Prediction mask start epoch	0
Shared head for teacher	True
Shared head for student	False
Dropout	0.1
Frozen augmentation strategy for patch feature	Brightness Contrast Posterize
Batch size	8×36
Freeze last layer epochs	4
Automatic per-op mixed precision	FP16
Learning rate schedule	Cosine
Minimum learning rate	1e-6
Peak learning rate	1e-4
Number of warmup epochs	30
Initial weight decay	0.04
Final weight decay	0.4
Teacher momentum (start)	0.998
Teacher momentum (final)	1
Optimization algorithm	AdamW
Maximum gradient clipping	2
Epochs	90

Table 8. Hyperparameter Selection during molecule-slide cross-modal contrastive pretraining. We use 8×80 GB A100 GPUs for the cross-modal pretraining stage.

Hyperparameter	Value
Gene number	3999
Dropout	0.2
WSI encoder learning rate	1e-5
Minimum WSI encoder learning rate	1e-6
RNA encoder learning rate	5e-5
Minimum RNA encoder learning rate	1e-5
Other modules learning rate	1e-4
Minimum other modules learning rate	1e-5
WSI & RNA encoder freezing (epochs)	20
Epochs	150
Maximum gradient clipping	2
Batch size	8×6
Gradient accumulation steps	1
Automatic per-op mixed precision	FP16
Protein number	10
Protein selection method (per sample)	top-10
Dropout	0.2
WSI encoder learning rate	1e-5
Minimum WSI encoder learning rate	1e-6
Protein encoder learning rate	1e-4
Minimum protein encoder learning rate	5e-5
Other modules learning rate	1e-4
Minimum other modules learning rate	1e-5
WSI & Protein encoder freezing (epochs)	0
Epochs	30
Maximum gradient clipping	2
Batch size	8×10
Gradient accumulation steps	2
Automatic per-op mixed precision	FP16

- **IMP [26].** We use the IMP-CRS 2024 dataset, a large-scale collection of colorectal hematoxylin and eosin (H&E) WSIs.
- **BCNB [43].** The Early Breast Cancer Core-Needle Biopsy Whole-Slide Image (BCNB) dataset is a publicly available cohort of 1,058 early breast cancer patients, each associated with a H&E–stained core-needle biopsy whole-slide image and matched clinical data.
- **DHMC-LUAD [40].** The DHMC-LUAD dataset is a publicly available collection of 143 H&E–stained FFPE WSIs of lung adenocarcinoma (LUAD).
- **DHMC-RCC [51].** DHMC-RCC is composed of 563 publicly accessible H&E–stained FFPE whole-slide images of renal cell carcinoma (RCC).
- **SR386 [25].** We use the SR386 cohort from the SurGen dataset, a recently released colorectal cancer digital pathology resource comprising H&E-stained WSIs with linked clinical, genetic and survival information.
- **Local-LUNG.** We conduct additional validation using an internal lung cancer dataset.

Tasks. Moreover, using either single datasets or combinations of multiple datasets, we construct 33 computational pathology tasks, including morphological classification, molecular classification (gene mutation prediction), and survival prediction. The specific tasks are described as follows:

- **Task 1: EBRAINS-corase.** A coarse-grained 12-class brain tumor subtyping task on the EBRAINS dataset.
- **Task 2: EBRAINS-fine.** A fine-grained 30-class brain tumor subtyping benchmark derived from the EBRAINS dataset.
- **Task 3: IMP-grading.** On the public IMP-CRS 2024 dataset, each colorectal WSI is weakly labeled as non-neoplastic, low-grade, or high-grade, forming a three-class slide-level tumor grading task.

Table 9. Architectural and pretraining statistics of compared slide-level foundation models. We report only the parameters of the vision-inference module. The superscript * indicates that the model was pre-trained on the TCGA pan-cancer dataset and did not report specific numbers.

Models	Patch encoder	Slide aggregator	Output size	FLOPs/G	Para/M	Dataset size
CHIEF	Ctranspath	Variant of ABMIL	1×768	0.26	1.05	60,530
PRISM	Virchow	Perceiver	1×1280	359.36	89.21	587,196
GigaPath	GigaPath	LongNet	1×768	31.80	86.33	171,189
TANGLE	UNI	ABMIL	1×512	1.57	5.00	$24{,}000^{*}$
FEATHER	CONCH v1.5	ABMIL	1×512	0.31	0.79	$24{,}000^{*}$
TITAN	CONCH v1.5	ViT	1×768	19.21	47.36	$335,645(423{,}122/182{,}862)$
CARE	CONCH v1.5	CARE	1×512	15.77	18.79	$34{,}277(13{,}289/8{,}225)$

Table 10. Results under the linear probing (logistic regression) setting for morphologic and molecular classification tasks. The best score is in bold, and the second-best is underlined.

Task	Metric	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 1	ACC	77.8±0.1	73.0±0.1	72.2±0.2	75.8±0.1	76.4±0.1	82.3±0.1	**87.1±0.0**	85.1±0.1
	F1	87.1±0.0	83.9±0.0	82.9±0.0	87.0±0.0	86.7±0.0	90.0±0.0	**91.7±0.0**	91.2±0.0
Task 2	ACC	65.8±0.0	60.6±0.1	59.5±0.0	64.7±0.1	64.5±0.0	68.2±0.0	**74.8±0.0**	74.0±0.1
	F1	71.4±0.0	68.4±0.1	65.8±0.0	71.6±0.1	71.2±0.0	73.1±0.0	**78.8±0.0**	78.7±0.0
Task 3	ACC	90.0±0.0	91.4±0.0	91.9±0.0	91.1±0.0	91.0±0.0	91.0±0.0	**92.8±0.0**	92.6±0.0
	F1	90.6±0.0	91.4±0.0	92.2±0.0	91.3±0.0	91.3±0.0	91.3±0.0	**92.9±0.0**	92.7±0.0
Task 4	ACC	89.1±0.5	89.6±0.1	90.4±0.1	84.4±0.4	94.2±0.1	95.3±0.1	**96.5±0.0**	96.2±0.0
	F1	94.2±0.1	94.5±0.0	94.6±0.0	91.9±0.1	97.4±0.0	97.6±0.0	**98.3±0.0**	97.9±0.0
Task 5	ACC	93.8±0.1	93.5±0.2	70.7±1.6	89.7±0.2	93.3±0.1	88.9±0.3	93.3±0.4	**97.6±0.1**
	F1	98.0±0.0	98.0±0.0	92.7±0.0	97.6±0.0	98.4±0.0	97.1±0.0	98.1±0.0	**99.1±0.0**
Task 6	ACC	88.1±0.0	87.6±0.0	86.2±0.0	88.4±0.1	86.3±0.1	85.5±0.0	**89.2±0.0**	89.0±0.0
	AUC	95.4±0.0	95.6±0.0	95.2±0.0	95.6±0.0	94.7±0.0	94.1±0.0	96.6±0.0	**96.8±0.0**
Task 7	ACC	63.6±0.1	**65.7±0.1**	58.3±0.1	54.2±0.1	60.3±0.0	56.2±0.3	63.8±0.1	65.5±0.0
	F1	73.7±0.0	74.2±0.0	68.0±0.1	62.2±0.1	70.1±0.0	68.5±0.3	74.2±0.0	**74.4±0.0**
Task 8	ACC	61.0±0.0	57.2±0.1	57.5±0.1	61.4±0.1	**63.6±0.2**	54.7±0.0	59.8±0.1	61.4±0.1
	AUC	84.8±0.1	85.9±0.0	86.5±0.0	86.6±0.0	86.8±0.0	84.3±0.1	86.0±0.0	**88.9±0.0**
Task 9	ACC	69.4±0.1	70.2±0.1	68.7±0.2	69.9±0.1	**73.5±0.1**	**73.5±0.1**	70.3±0.1	71.8±0.2
	AUC	76.2±0.1	78.0±0.1	75.3±0.1	76.9±0.1	**81.1±0.1**	79.8±0.1	78.0±0.1	79.2±0.2
Task 10	ACC	70.2±0.1	72.9±0.6	**74.5±0.2**	71.1±0.6	71.6±0.1	74.0±0.2	62.9±1.0	74.4±0.1
	AUC	82.8±0.2	86.9±0.2	88.2±0.1	87.5±0.2	87.4±0.1	86.7±0.2	74.3±1.0	**88.4±0.1**
Task 11	ACC	57.7±0.1	57.7±0.0	**64.5±0.1**	57.8±0.1	61.7±0.1	59.0±0.1	60.7±0.1	59.8±0.1
	AUC	71.7±0.1	74.9±0.0	**77.9±0.0**	74.2±0.0	77.8±0.0	76.7±0.1	77.8±0.1	77.2±0.1
Task 12	ACC	65.1±0.1	66.6±0.0	**72.1±0.0**	66.5±0.1	66.8±0.1	71.7±0.0	71.8±0.0	70.5±0.1
	AUC	77.6±0.1	79.6±0.1	83.5±0.1	79.8±0.1	80.1±0.1	82.4±0.1	**84.1±0.0**	82.0±0.1
Task 13	ACC	87.8±0.1	88.9±0.0	88.7±0.0	88.2±0.0	89.0±0.1	88.5±0.1	91.4±0.0	**91.5±0.0**
	AUC	94.2±0.0	95.6±0.0	94.9±0.0	94.7±0.0	95.2±0.0	95.9±0.0	**96.7±0.0**	96.6±0.0
Task 14	ACC	**58.3±0.1**	50.8±0.0	57.6±0.1	51.6±0.2	57.6±0.2	52.0±0.0	57.4±0.2	56.9±0.1
	AUC	70.9±0.1	77.0±0.0	72.6±0.0	**82.0±0.0**	76.0±0.1	67.9±0.2	71.2±0.0	72.7±0.1

- **Task 4: Local-LUNG-subtype.** Using the Local-LUNG dataset, this task performs three-class WSI-level lung cancer subtype classification. The cohort includes 380 lung adenocarcinoma, 64 small cell lung carcinoma, and 80 lung squamous cell carcinoma cases.
- **Task 5: Combine-RCC-subtype.** We construct a three-class RCC subtype classification benchmark by combining cases from the CPTAC-CCRCC cohort and the DHMC-RCC dataset. The resulting cohort contains 713 clear cell RCC, 80 papillary RCC, and 23 chromophobe RCC WSIs.
- **Task 6: Combine-LUNG-subtype.** This binary lung

Table 11. Results on linear probing (logistic regression) setting for molecular classification tasks without a validation set. The best score is in bold, and the second-best is underlined.

Task	Metric	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 15	ACC	59.1±0.3	56.2±0.2	51.9±0.2	57.1±0.2	58.1±0.2	**59.8±0.2**	57.9±0.2	57.2±0.3
	AUC	61.7±0.4	63.3±0.2	53.9±0.3	62.8±0.2	**64.8±0.3**	62.5±0.3	62.3±0.4	61.3±0.2
Task 16	ACC	**75.8±0.7**	72.8±0.8	67.9±0.6	71.8±1.0	75.0±0.8	62.7±1.2	72.9±0.8	75.3±0.8
	AUC	83.5±0.9	81.4±1.1	78.9±0.6	81.7±0.9	**83.6±0.9**	70.3±1.4	79.9±0.9	81.2±0.8
Task 17	ACC	63.1±1.5	64.1±1.4	54.8±1.2	62.0±0.9	55.1±0.7	59.6±1.2	62.8±1.0	**64.5±0.8**
	AUC	69.1±1.5	70.6±1.8	59.3±1.5	**72.2±1.4**	67.0±1.2	63.9±1.2	68.1±1.1	66.6±1.1
Task 18	ACC	55.6±1.1	57.8±1.0	61.9±0.8	61.9±0.6	**67.9±1.0**	56.8±0.8	58.9±1.1	62.4±1.1
	AUC	56.8±1.3	67.7±1.2	68.1±1.5	62.2±0.9	**76.3±1.8**	60.8±1.6	68.0±1.1	69.6±1.5
Task 19	ACC	57.5±0.8	56.4±0.7	54.0±1.1	55.2±0.7	56.4±0.8	58.1±1.6	**61.2±1.4**	58.7±1.0
	AUC	62.6±2.9	65.0±2.3	59.5±2.8	60.1±2.8	**73.8±1.6**	58.2±3.7	73.2±2.1	65.8±2.6
Task 20	ACC	47.2±0.9	47.2±0.7	47.3±0.8	46.6±0.6	53.4±0.7	50.5±0.9	49.5±0.7	**56.0±0.7**
	AUC	45.7±1.3	46.5±1.3	45.7±1.1	46.1±1.4	55.5±1.2	49.4±1.4	48.5±1.3	**58.0±1.0**
Task 21	ACC	**66.8±1.8**	54.1±0.7	59.9±2.0	63.4±1.6	58.7±0.6	57.5±1.2	60.8±1.2	66.6±1.4
	AUC	81.7±1.2	65.2±1.5	73.9±1.6	71.0±1.4	72.1±1.4	62.5±1.8	77.6±1.1	**84.2±0.5**
Task 22	ACC	77.4±1.1	70.4±0.9	72.5±0.7	77.1±1.1	69.6±1.2	66.9±1.0	**79.6±1.1**	75.7±0.9
	AUC	90.3±0.4	86.8±0.8	83.0±0.7	91.8±0.6	91.5±0.4	77.6±1.3	91.2±0.5	**92.1±0.3**
Task 23	ACC	58.6±1.1	53.6±0.5	57.0±0.9	**65.2±1.3**	50.6±0.4	54.9±1.0	56.8±0.9	60.0±1.1
	AUC	64.6±1.7	64.9±1.4	62.9±0.9	**69.2±1.9**	57.7±1.2	61.7±1.4	64.0±0.9	64.0±1.2
Task 24	ACC	62.3±1.7	54.4±0.7	63.8±2.0	64.3±2.0	58.6±1.3	55.9±1.0	61.0±1.8	**65.3±1.8**
	AUC	79.4±1.9	75.7±3.3	68.5±4.0	**81.4±1.7**	75.0±2.4	69.0±2.5	78.3±1.4	78.4±1.7
Task 25	ACC	49.3±0.7	47.5±0.1	**54.6±1.2**	46.8±0.1	47.9±0.1	51.8±0.6	49.0±0.3	51.7±1.0
	AUC	50.7±3.4	37.3±4.5	49.2±5.9	38.1±3.5	50.5±2.4	52.6±5.8	53.6±4.4	**58.7±4.8**
Task 26	ACC	70.0±0.5	65.4±0.7	69.3±0.5	**74.2±0.7**	68.4±0.8	69.5±0.5	68.9±0.6	72.2±0.5
	AUC	76.8±0.7	71.2±1.0	77.3±0.6	**81.1±0.6**	77.6±0.8	76.5±0.6	76.6±0.6	79.8±0.5
Task 27	ACC	65.9±0.7	59.9±0.9	60.5±0.3	69.1±1.0	61.2±0.7	63.8±0.9	65.6±0.5	**69.5±0.7**
	AUC	72.1±0.6	68.0±1.2	65.5±0.5	77.1±1.0	69.0±1.0	71.6±1.0	74.1±0.6	**78.7±0.7**
Task 28	ACC	69.6±0.8	63.2±1.0	71.9±1.0	64.9±1.1	63.6±0.9	66.6±1.1	70.4±1.2	**73.1±1.2**
	AUC	82.9±1.0	82.9±1.0	78.7±1.7	83.1±0.6	**86.9±0.5**	77.1±1.8	85.1±0.6	85.3±0.6

cancer subtype task pools WSIs from Local-LUNG, CPTAC-LUAD, CPTAC-LSCC, and DHMC-LUNG into a LUAD-versus-LUSC slide-level benchmark, with 1,660 lung adenocarcinoma (LUAD) and 1,159 lung squamous cell carcinoma (LUSC) WSIs.

- **Task 7: Cross-LUNG-fine.** A three-class lung adenocarcinoma growth pattern classification task, with acinar, lepidic, and solid as the predominant histologic patterns. Models are trained and validated on Local-LUNG and evaluated on DHMC-LUNG as an external test set to assess cross-cohort generalization.
- **Task 8: MUT-BAP1.** On the MUT-HET-RCC cohort, this binary task predicts BAP1 mutation status from H&E-stained RCC WSIs, with 162 wild-type and 1,130 mutant slides among 1,292 cases.
- **Task 9: MUT-PBRM1.** Also on the MUT-HET-RCC cohort, this binary slide-level task predicts PBRM1 mutation status from H&E-stained RCC WSIs, including 670 wild-type and 622 mutant slides.
- **Task 10: BCNB-ER.** A binary slide-level ER status prediction task on the BCNB cohort, with 219 ER-negative and 809 ER-positive H&E-stained WSIs.
- **Task 11: BCNB-HER2.** On the same cohort, BCNB-HER2 predicts HER2 status at the slide level, comprising 758 HER2-negative and 270 HER2-positive WSIs.
- **Task 12: BCNB-PR.** BCNB-PR frames progesterone receptor (PR) status prediction from H&E-stained WSIs as a binary slide-level task (260 PR-negative and 768 PR-positive slides).
- **Task 13: EBRIANS-IDH.** On H&E-stained glioma WSIs from the EBRAINS dataset, this binary slide-level task predicts isocitrate dehydrogenase (IDH) mutation status (540 IDH–wild-type and 333 IDH–mutant slides).
- **Task 14: Cross-MUT-BAP1.** Cross-MUT-BAP1 evaluates cross-cohort generalization for BAP1 mutation prediction by training and validating on MUT-HET-RCC and testing on CPTAC-CCRCC for binary slide-level BAP1 status.

Table 12. Results on linear probing (kNN parameter) setting for morphologic and molecular classification tasks. The best score is in bold, and the second-best is underlined.

Task	Metric	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 1	ACC	71.4±0.0	55.8±0.1	57.8±0.1	58.7±0.1	66.1±0.2	70.8±0.1	**83.5±0.1**	80.9±0.0
	F1	83.8±0.0	72.4±0.0	73.1±0.0	75.4±0.0	80.7±0.0	84.0±0.0	**90.7±0.0**	89.4±0.0
Task 2	ACC	56.6±0.0	44.8±0.1	48.0±0.0	46.4±0.0	54.1±0.0	56.9±0.0	**70.5±0.0**	66.2±0.0
	F1	63.3±0.0	53.3±0.1	54.5±0.0	55.4±0.0	62.8±0.0	65.2±0.0	**75.8±0.0**	72.0±0.0
Task 3	ACC	84.5±0.0	90.3±0.0	88.6±0.0	84.5±0.0	87.8±0.0	81.1±0.0	**90.5±0.0**	89.7±0.0
	F1	86.7±0.0	90.3±0.0	89.4±0.0	86.6±0.0	89.1±0.0	85.6±0.0	**91.3±0.0**	90.4±0.0
Task 4	ACC	77.1±0.6	85.9±0.3	90.0±0.3	73.4±0.2	90.8±0.1	78.0±0.4	97.1±0.1	**98.0±0.1**
	F1	88.3±0.1	93.7±0.1	94.7±0.1	86.7±0.0	96.0±0.0	89.5±0.1	98.1±0.0	**98.6±0.0**
Task 5	ACC	84.5±0.1	85.0±1.1	55.7±0.7	81.5±0.1	84.9±0.1	86.9±0.4	92.9±0.3	**96.1±0.1**
	F1	96.4±0.0	96.1±0.0	89.6±0.0	96.1±0.0	96.4±0.0	97.3±0.0	98.0±0.0	**98.1±0.0**
Task 6	ACC	86.6±0.1	84.7±0.1	84.8±0.0	85.3±0.1	87.7±0.0	85.2±0.0	89.0±0.0	**89.3±0.1**
	AUC	95.3±0.0	92.9±0.1	93.6±0.0	92.9±0.0	95.1±0.0	93.6±0.0	**96.2±0.0**	96.1±0.0
Task 7	ACC	53.9±0.0	53.3±0.3	57.2±0.1	45.2±0.1	55.2±0.2	46.3±0.0	**58.0±0.2**	56.0±0.1
	F1	61.3±0.1	63.6±0.4	64.0±0.1	52.8±0.1	65.4±0.1	57.5±0.1	**66.9±0.2**	66.7±0.0
Task 8	ACC	60.7±0.1	56.8±0.1	57.1±0.1	57.5±0.1	58.6±0.1	56.4±0.3	**62.8±0.2**	61.5±0.1
	AUC	71.8±0.3	75.4±0.1	70.6±0.2	73.0±0.2	72.8±0.2	73.4±0.1	75.6±0.2	**78.6±0.3**
Task 9	ACC	63.3±0.1	67.6±0.1	60.7±0.1	63.9±0.1	**68.8±0.1**	65.9±0.1	66.5±0.1	67.9±0.2
	AUC	68.8±0.1	72.2±0.1	66.5±0.3	70.4±0.1	**75.9±0.1**	71.6±0.2	73.5±0.2	74.9±0.3
Task 10	ACC	60.9±0.2	61.2±0.1	**68.2±0.1**	57.9±0.1	63.0±0.1	60.8±0.0	67.0±0.2	65.5±0.1
	AUC	72.1±0.1	70.9±0.6	**84.7±0.1**	72.0±0.9	75.8±0.2	74.7±0.1	77.3±0.3	82.5±0.2
Task 11	ACC	52.7±0.0	55.5±0.1	60.5±0.1	53.1±0.0	57.4±0.1	56.7±0.1	**64.7±0.0**	60.0±0.0
	AUC	65.7±0.2	70.5±0.1	76.8±0.0	64.9±0.1	73.8±0.1	72.0±0.1	**80.0±0.1**	73.8±0.1
Task 12	ACC	58.6±0.1	60.5±0.1	**68.0±0.0**	55.5±0.2	63.0±0.0	63.5±0.1	64.8±0.2	66.0±0.3
	AUC	66.6±0.0	73.2±0.0	**77.6±0.1**	62.4±0.3	74.7±0.1	74.9±0.2	74.5±0.1	75.9±0.2
Task 13	ACC	80.7±0.1	80.4±0.1	82.0±0.1	83.0±0.2	86.4±0.1	87.1±0.1	**90.1±0.1**	89.3±0.1
	AUC	90.6±0.1	91.5±0.1	88.9±0.1	90.2±0.1	92.1±0.1	94.5±0.1	94.3±0.1	**95.1±0.0**
Task 14	ACC	**69.2±0.1**	54.5±0.2	56.6±0.0	58.5±0.0	58.1±0.0	59.6±0.3	58.7±0.1	63.6±0.1
	AUC	**75.2±0.0**	67.8±0.0	66.9±0.0	69.0±0.0	65.1±0.1	66.2±0.2	74.1±0.1	74.3±0.0

- **Task 15: SR386-RAS.** Defined on the SR386 cohort with linked genetic markers, SR386-RAS is a binary slide-level RAS mutation prediction task, including 251 RAS wild-type and 138 RAS mutant cases.
- **Task 16: CPTAC-GBM-TP53.** Using WSIs from the CPTAC-GBM cohort, this task performs binary slide-level prediction of TP53 mutation status (162 TP53 wild-type and 80 TP53 mutant cases).
- **Task 17: CPTAC-BRCA-HER2.** On H&E-stained WSIs from the CPTAC-BRCA cohort, CPTAC-BRCA-HER2 formulates binary slide-level HER2 status prediction, with 50 HER2-negative and 39 HER2-positive breast cancer cases.
- **Task 18: CPTAC-BRCA-PIK3CA.** CPTAC-BRCA-PIK3CA uses H&E-stained WSIs from CPTAC-BRCA for binary slide-level PIK3CA mutation prediction (81 PIK3CA wild-type and 34 PIK3CA mutant cases).
- **Task 19: CPTAC-CCRCC-BAP1.** On H&E-stained WSIs from the CPTAC-CCRCC cohort, this task predicts BAP1 mutation status at the slide level, with 202 BAP1 wild-type and 43 BAP1 mutant cases.
- **Task 20: CPTAC-CCRCC-PBRM1.** CPTAC-CCRCC-PBRM1 defines binary slide-level PBRM1 mutation prediction on the same cohort, including 123 PBRM1 wild-type and 122 PBRM1–mutant cases.
- **Task 21: CPTAC-COAD-APC.** Using H&E-stained WSIs from the CPTAC-COAD cohort, CPTAC-COAD-APC defines binary slide-level APC mutation prediction (22 APC–wild-type and 76 APC–mutant cases).
- **Task 22: CPTAC-COAD-MSI.** CPTAC-COAD-MSI frames microsatellite instability (MSI) status prediction on CPTAC-COAD H&E WSIs as a binary slide-level task, with 81 MSI-stable and 25 MSI-high cases.
- **Task 23: CPTAC-COAD-PIK3CA.** On the same cohort, CPTAC-COAD-PIK3CA performs binary slide-level PIK3CA mutation prediction (73 wild-type and 25 mutant cases).
- **Task 24: CPTAC-COAD-SETD1B.** CPTAC-COAD-

Table 13. Results on linear probing (kNN parameter) setting for molecular classification tasks without a validation set. The best score is in bold, and the second-best is underlined.

Task	Metric	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 15	ACC	55.2±0.1	54.4±0.2	50.9±0.1	54.7±0.1	53.3±0.2	53.2±0.1	**55.3±0.1**	54.3±0.1
	AUC	59.4±0.3	56.1±0.3	51.2±0.4	56.7±0.3	57.5±0.4	59.2±0.4	61.3±0.2	**61.3±0.3**
Task 16	ACC	55.9±0.4	64.1±0.6	57.5±0.3	56.2±0.5	**66.8±0.9**	54.5±0.2	65.0±0.7	65.5±0.8
	AUC	67.6±1.3	**78.5±1.0**	65.0±0.7	67.1±1.8	74.8±1.4	59.5±1.0	74.5±1.4	76.9±1.0
Task 17	ACC	55.1±0.7	**61.1±0.8**	54.5±0.6	56.0±0.6	56.1±0.6	54.0±0.4	59.4±0.7	59.0±0.8
	AUC	65.5±1.0	67.5±1.1	67.5±1.4	63.8±1.1	63.4±1.2	62.3±1.9	66.9±1.1	**71.4±1.0**
Task 18	ACC	65.8±0.9	**68.3±0.8**	66.5±1.0	66.8±0.6	64.7±0.8	55.7±0.3	59.8±0.6	66.5±0.6
	AUC	72.6±1.3	**77.3±1.2**	76.4±1.4	73.7±1.6	75.9±1.5	59.0±1.8	61.6±1.3	67.9±1.4
Task 19	ACC	**51.9±0.2**	50.0±0.1	50.2±0.1	50.1±0.1	50.7±0.1	49.9±0.0	51.4±0.1	50.5±0.2
	AUC	63.7±3.5	50.5±2.9	63.8±2.0	58.5±3.4	65.4±2.8	**68.9±1.8**	64.5±2.6	63.2±2.1
Task 20	ACC	47.7±0.7	48.9±1.1	46.0±0.5	44.5±0.9	52.0±1.0	**52.2±0.8**	43.0±0.7	47.2±1.0
	AUC	45.5±1.3	48.0±1.6	48.7±0.8	41.9±1.6	49.4±1.6	**51.9±1.3**	41.5±1.3	45.5±1.5
Task 21	ACC	49.6±0.1	**51.8±0.2**	50.2±0.0	49.9±0.0	51.3±0.2	50.3±0.1	49.8±0.1	49.0±0.1
	AUC	70.8±1.0	67.5±1.3	**76.3±1.3**	66.4±1.6	72.0±1.5	69.4±2.3	67.8±1.7	65.7±1.8
Task 22	ACC	54.2±0.3	60.2±0.7	54.8±0.4	55.6±0.6	60.6±0.5	51.5±0.1	**63.1±0.4**	58.5±0.6
	AUC	79.7±1.1	87.5±0.7	70.5±1.6	83.8±1.3	**91.1±0.5**	66.9±1.3	86.7±0.8	81.2±1.5
Task 23	ACC	**53.1±0.6**	52.3±0.4	48.8±0.1	51.3±0.3	50.9±0.2	51.6±0.3	51.7±0.3	48.9±0.1
	AUC	56.0±2.0	63.3±1.4	53.0±1.5	60.3±2.3	61.4±1.8	49.7±1.3	**64.9±1.5**	59.2±1.4
Task 24	ACC	49.7±0.2	49.1±0.0	**54.3±0.6**	50.0±0.1	53.8±0.6	49.8±0.0	51.6±0.4	51.4±0.3
	AUC	65.8±2.9	74.0±2.9	63.8±3.8	71.5±2.6	**79.9±1.7**	54.4±3.3	66.9±2.4	72.2±3.1
Task 25	ACC	49.9±0.0	48.7±0.1	**51.1±0.3**	49.2±0.0	49.5±0.0	50.0±0.0	49.9±0.0	50.0±0.0
	AUC	53.1±2.1	50.7±3.8	**65.8±4.2**	47.2±3.4	50.4±2.4	57.1±3.1	58.2±2.8	55.0±2.7
Task 26	ACC	70.0±0.6	62.8±0.6	63.4±0.4	68.5±0.7	67.2±0.5	**71.5±0.5**	67.5±0.4	70.1±0.7
	AUC	75.0±0.9	68.7±1.0	70.8±0.6	74.9±0.9	74.5±0.9	75.1±0.8	**77.7±0.6**	73.7±0.8
Task 27	ACC	64.5±0.7	56.1±0.6	57.7±0.4	66.8±0.8	59.3±0.9	65.0±0.6	**68.1±0.8**	65.9±0.7
	AUC	71.5±0.9	63.9±0.9	63.4±0.7	74.3±1.0	66.5±1.3	74.2±0.7	**75.9±0.8**	75.9±0.8
Task 28	ACC	60.5±0.6	64.1±0.7	64.2±0.6	61.0±0.7	69.3±0.7	62.3±0.6	67.8±1.2	**71.7±1.4**
	AUC	82.2±1.0	80.4±0.8	80.0±1.5	82.5±0.7	86.6±0.6	76.8±1.5	84.2±1.2	**87.8±0.9**

SETD1B uses H&E-stained WSIs from CPTAC-COAD for binary slide-level SETD1B mutation status prediction, including 81 SETD1B–wild-type and 17 SETD1B–mutant cases.

- **Task 25: CPTAC-HNSCC-CASP8.** For CPTAC-HNSCC-CASP8, H&E-stained WSIs from the CPTAC-HNSCC cohort are used to predict CASP8 mutation status at the slide level in a binary setting (222 CASP8–wild-type and 33 CASP8–mutant cases).
- **Task 26: CPTAC-LUAD-EGFR.** CPTAC-LUAD-EGFR is a binary slide-level EGFR mutation prediction task on H&E-stained WSIs from the CPTAC-LUAD cohort, with 211 EGFR–wild-type and 113 EGFR–mutant cases.
- **Task 27: CPTAC-LUAD-KRAS.** Similarly, this task targets KRAS mutation status on CPTAC-LUAD H&E WSIs, formulating a binary slide-level prediction task with 200 KRAS wild-type and 112 KRAS mutant cases.
- **Task 28: CPTAC-LUAD-STK11.** CPTAC-LUAD-STK11 defines binary slide-level STK11 mutation prediction on the CPTAC-LUAD cohort, comprising 266 STK11–wild-type and 58 STK11–mutant cases.
- **Task 29: CPTAC-CCRCC-OS.** CPTAC-CCRCC-OS uses H&E-stained clear cell RCC WSIs from the CPTAC-CCRCC collection and corresponding overall survival (OS) time and event data from cBioPortal to define a slide-level OS risk prediction task for assessing model prognostic performance.
- **Task 30: CPTAC-HNSCC-OS.** On H&E-stained WSIs from the CPTAC-HNSCC cohort with associated OS time and event annotations, CPTAC-HNSCC-OS formulates a slide-level OS risk prediction benchmark to evaluate both prognostic performance and cross-cohort generalization.
- **Task 31: CPTAC-LUAD-OS.** CPTAC-LUAD-OS constructs a slide-level OS risk prediction task on H&E-stained WSIs from the CPTAC-LUAD cohort, paired with survival time and censoring status from cBioPortal, to assess model prognostic performance and robustness.
- **Task 32: CPTAC-PDA-OS.** Using H&E-stained whole-

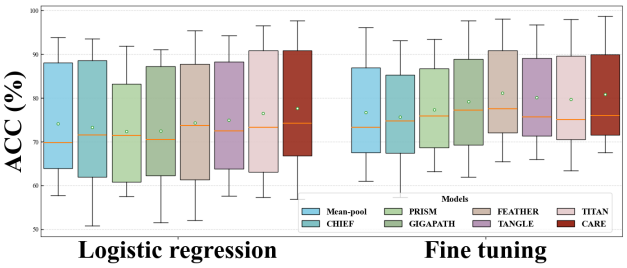

Figure 7. Experimental results of ACC box plots for different models under different experimental settings. Each box plot aggregates results from all morphology-classification and molecular-classification tasks. The horizontal line inside each box indicates the median for that model across tasks. The green circle denotes the mean.

slide images of pancreatic ductal adenocarcinoma (PDA) from the CPTAC-PDA cohort and OS time with censoring information from the corresponding cBioPortal study, CPTAC-PDA-OS defines a slide-level OS risk prediction task for evaluating prognostic performance and generalization.

- **Task 33: SR386-OS.** SR386-OS leverages 427 H&E-stained WSIs from the SR386 cohort with five-year OS follow-up and event (censoring) status to set up a slide-level OS risk prediction task, used to benchmark model prognostic accuracy and generalization ability.

In the above experiments, we group Tasks 1–7 as morphological classification tasks, which are closely related to histologic subtype information in pathology images. Tasks 8–14 are categorized as molecular classification tasks for gene mutation prediction. Tasks 15–28 are molecular classification tasks without a validation set, since the corresponding datasets contain relatively few samples. Tasks 29–33 are categorized as survival analysis tasks for overall survival risk prediction.

D. Description of Slide-Level Baseline Models

In this paper, we compare our method with several state-of-the-art slide-level foundation models. Their details are described below.

- **CHIEF [39].** CHIEF is an attention-based multiple instance learning slide encoder, which fused with a CLIP text embedding of the WSI's anatomical site. It is weakly supervised pretrained on 60,530 H&E WSIs from 19 anatomical sites, using slide-level labels such as anatomical site, cancer type, genomic alterations and prognostic outcomes as multi-task targets.
- **PRISM [30].** PRISM is a multimodal vision–language slide-level foundation model that takes H&E WSIs and pairs them with clinical pathology reports, using a perceiver-based slide encoder in a CoCa-style architecture to jointly encode image and text. It is pretrained on 587,196 WSIs, using contrastive and generative objectives that align slide embeddings with free-text diagnostic reports.
- **GigaPath [44].** GigaPath is a whole-slide vision transformer that treats a WSI as a very long sequence of image tiles. It is a LongNet-based slide encoder with masked-autoencoder pretraining models global context across 171,189 WSIs.
- **FEATHER [31].** FEATHER is a lightweight image-only slide-level models that use an attention-based multiple instance learning (ABMIL) framework on top of frozen patch encoders such as CONCH v1.5 to aggregate patch embeddings into WSI representations. It is supervised-pretrained on the PC-108 pan-cancer morphological classification task about 24,000 H&E WSIs—using 108-way slide-level morphology labels as targets.
- **TANGLE [15].** TANGLE is a multimodal foundation model that uses a slide encoder together with a gene-expression encoder, and aligns the two modalities via CLIP-style contrastive learning between WSI embeddings and bulk RNA-seq expression embeddings. It is pan-cancer pretrained on slide–expression pairs from all TCGA cohorts, where the main supervision signal comes from paired bulk transcriptomic profiles rather than discrete diagnostic or mutation labels. Notably, in this paper we use the weights of TANGLE v2.
- **TITAN [9].** TITAN is a multimodal whole-slide vision–language foundation model that uses a transformer slide encoder, together with a text encoder in a CoCa-style image–text framework to jointly represent H&E WSIs and pathology reports. It is pretrained on 335,645 WSIs using a three-stage scheme: visual self-supervised learning on slides, followed by vision–language contrastive and captioning objectives that align slide embeddings with 423,122 synthetic captions.

Moreover, we provide a detailed comparison between the above models and CARE in terms of their architectures and pretraining data, as shown in Tab. 9. Here, "FLOPs" denotes the vision-inference FLOPs of each model computed using the fvcore toolkit, and "Para" denotes the number of trainable parameters involved in a single forward pass from patch features to WSI features, excluding the parameters of the patch encoder. On the one hand, models that use ABMIL or variants of ABMIL as their backbone are lightweight, with low parameter counts and FLOPs, and therefore do not fall into the same regime as foundation models whose defining characteristic is large-scale pretraining. Among the remaining models, CARE has the lowest number of parameters and FLOPs, which accelerates the computation of slide embeddings. On the other hand, CARE is pretrained on only about one tenth of the data used by strong baselines such as TITAN and PRISM, yet

Table 14. Survival prediction experimental results. All results are obtained under the linear-probing (Linear) setting. The best score is in bold, and the second-best is underlined.

Task	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 29	48.4±16.5	53.2±15.4	38.0±20.0	55.8±4.7	46.6±8.3	<u>56.6±15.3</u>	40.5±18.9	**63.0±8.9**
Task 30	53.8±15.8	<u>58.7±8.7</u>	39.9±9.9	57.3±11.1	48.6±23.6	57.6±12.4	38.9±12.6	**65.9±18.0**
Task 31	48.4±13.8	<u>56.3±9.6</u>	38.0±21.8	**56.9±16.1**	50.8±9.2	46.4±16.0	54.6±20.6	47.5±11.9
Task 32	42.2±7.6	50.3±10.2	48.3±8.7	46.8±8.9	52.2±8.1	**61.3±10.5**	43.5±6.6	<u>56.6±10.1</u>
Task 33	49.7±4.8	53.0±7.3	46.2±10.2	<u>62.0±7.5</u>	50.8±6.0	47.4±6.2	**58.5±6.5**	57.1±3.1

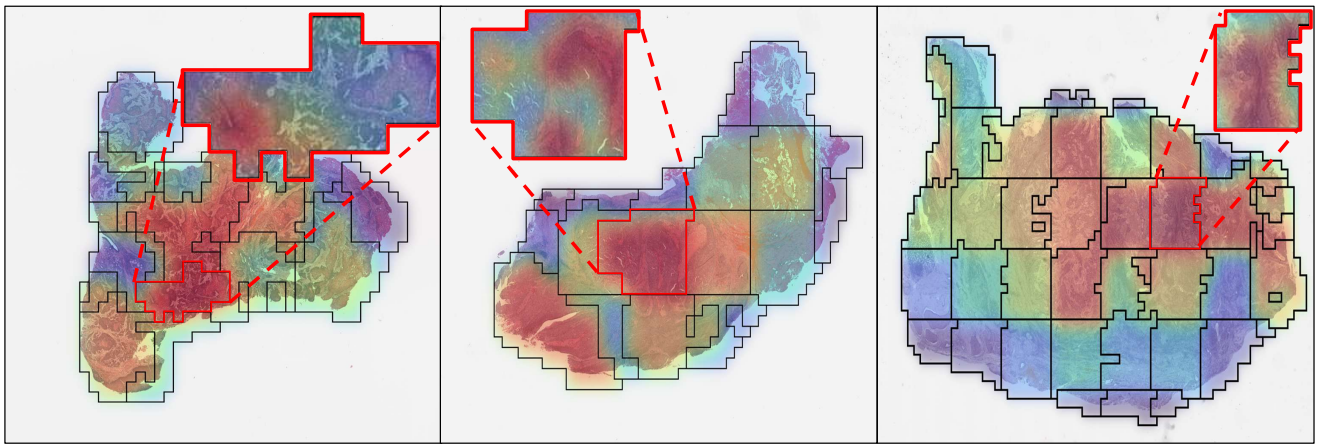

Figure 8. Heatmap visualization of CARE patches within the ROI. The WSIs are sourced from the CPTAC-HNSCC dataset. Warmer colors (red) indicate stronger patch attention.

still achieves impressive performance, highlighting the advantages of its architecture and pretraining strategy.

E. Implementation Details for Downstream Task Evaluation

Implementation Details. All datasets are preprocessed with the CLAM toolkit [22] using a unified pipeline. For Tasks 1–14, we evaluate three settings: logistic regression, k-nearest neighbors (kNN), and fine tuning. For Tasks 15–28, the datasets are relatively small. We therefore split the data only into training and test sets without a separate validation set for hyperparameter search, and report results for logistic regression and KNN with fixed hyperparameters. For Tasks 29–33 (survival analysis), we again use a train–test split only and apply a single linear layer on top of slide features to map them to logits.

In the logistic regression setting, we use the "lbfgs" solver. When a validation set is available, we perform a logarithmic grid search over 45 log–uniform values of the regularization parameter (C) in the range $[10^{-6}, 10^{5}]$. When no validation set is available, we fix $C = 0.5$. In the kNN setting, if a validation set exists, we search over ($k \in \{3, 5, 7, 9, 11, 13, 15, 17, 19, 23\}$). If no validation set is available, we choose $k = \sqrt{N}$, where N is the number of training samples. In the fine-tuning setting for Tasks 1–14, we attach a linear layer on top of each foundation model to map slide features to logits. The slide-level backbone is initialized from its corresponding pretrained weights, while the linear layer is randomly initialized, and all parameters are fine tuned with a learning rate of 2×10^{-5}. All downstream evaluation tasks are trained and tested on a single NVIDIA RTX 4090 GPU.

Evaluation Metric. For Tasks 1–28, we report balanced accuracy (ACC) as the primary metric. For Tasks 1–5 and Task 7, which are multi-class classification problems, we additionally report the macro-F1 (F1) score. For Task 6 and Tasks 8–28, which are binary classification problems, we additionally report the AUROC (AUC). We use F1 for multi-class tasks because it directly reflects per-class precision–recall trade-offs under class imbalance, whereas for binary tasks AUROC is preferred as it summarizes the model's overall threshold-independent discrimination between positive and negative cases. For Tasks 29–33, which are survival analysis tasks, we report the concordance index (C-index). We run 50 Monte Carlo cross-validation splits for Tasks 15–28 and 5 splits for the remaining tasks.

Table 15. Results on fine tuning setting. Since Tasks 15 to 33 do not have a validation set, the fine-tuning experiments were only evaluated on Tasks 1 to 14. The best score is in bold, and the second-best is underlined.

Task	Metric	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Task 1	ACC	82.6±0.0	70.6±0.1	78.9±0.1	85.2±0.0	84.5±0.1	88.0±0.0	88.3±0.1	**88.9±0.1**
	F1	83.9±0.0	75.5±0.1	85.0±0.1	90.0±0.0	88.6±0.0	89.4±0.0	**92.0±0.0**	91.7±0.0
Task 2	ACC	67.9±0.1	57.4±0.2	63.5±0.1	71.8±0.0	71.3±0.0	74.6±0.1	**75.9±0.0**	73.0±0.1
	F1	68.7±0.1	54.4±0.2	66.2±0.1	73.8±0.0	72.8±0.0	74.9±0.0	**78.3±0.0**	75.1±0.1
Task 3	ACC	88.5±0.0	93.1±0.0	93.4±0.0	**93.9±0.0**	93.5±0.0	92.8±0.0	93.2±0.0	93.2±0.0
	F1	87.7±0.0	92.9±0.0	93.4±0.0	**93.9±0.0**	93.6±0.0	92.8±0.0	92.5±0.0	92.9±0.0
Task 4	ACC	91.0±0.2	85.6±0.1	92.6±0.1	94.6±0.0	95.0±0.1	97.1±0.1	**97.9±0.0**	95.1±0.1
	F1	94.0±0.1	90.9±0.0	95.6±0.1	96.1±0.0	97.2±0.0	97.4±0.1	**98.2±0.1**	97.1±0.0
Task 5	ACC	96.1±0.1	92.7±0.1	89.0±0.2	97.7±0.1	96.8±0.1	98.1±0.1	97.2±0.1	**98.7±0.0**
	F1	96.5±0.0	97.0±0.0	97.2±0.0	99.2±0.0	98.8±0.0	**99.3±0.0**	99.0±0.0	98.6±0.0
Task 6	ACC	86.2±0.0	85.5±0.0	86.1±0.1	89.1±0.0	89.1±0.0	87.9±0.0	**89.5±0.0**	88.5±0.1
	AUC	94.1±0.0	93.7±0.0	95.0±0.0	96.0±0.0	96.2±0.0	95.3±0.0	**96.7±0.0**	96.5±0.0
Task 7	ACC	66.0±0.0	61.4±0.0	66.2±0.2	61.9±0.2	**72.3±0.1**	68.2±0.1	71.8±0.0	71.4±0.1
	F1	50.5±0.2	51.9±0.4	65.2±1.0	57.4±1.0	68.4±1.0	68.0±0.3	**71.3±0.5**	65.4±2.0
Task 8	ACC	74.1±0.0	**77.5±0.3**	71.9±0.1	74.8±0.2	70.8±0.2	75.3±0.2	67.2±0.2	71.4±0.2
	AUC	83.4±0.0	83.5±0.1	87.1±0.1	86.3±0.0	87.4±0.0	84.0±0.1	87.2±0.0	**89.0±0.0**
Task 9	ACC	67.9±0.2	66.9±0.1	71.0±0.0	68.6±0.1	71.8±0.1	71.3±0.1	70.1±0.1	**72.2±0.1**
	AUC	73.5±0.2	74.1±0.1	77.8±0.0	77.3±0.1	**80.9±0.2**	78.4±0.1	78.1±0.1	79.3±0.1
Task 10	ACC	72.6±0.2	76.4±0.0	78.1±0.1	79.7±0.6	78.9±0.2	**79.7±0.1**	73.5±0.1	76.5±0.3
	AUC	81.5±0.2	83.4±0.2	89.4±0.1	**90.0±0.2**	88.6±0.1	87.1±0.1	86.8±0.0	88.0±0.1
Task 11	ACC	61.0±0.3	67.2±0.1	67.9±0.2	68.0±0.1	**70.7±0.1**	70.4±0.1	63.5±0.1	67.5±0.2
	AUC	69.4±0.1	73.2±0.1	77.0±0.1	77.6±0.1	**79.4±0.1**	77.3±0.0	71.7±0.3	77.6±0.1
Task 12	ACC	67.4±0.1	73.0±0.1	73.6±0.3	71.2±0.6	72.4±0.3	74.8±0.1	74.1±0.1	**75.4±0.1**
	AUC	75.4±0.1	78.3±0.1	82.2±0.0	81.5±0.1	82.0±0.1	81.3±0.1	82.1±0.1	**82.9±0.1**
Task 13	ACC	87.1±0.1	85.0±0.1	86.9±0.1	88.3±0.0	89.1±0.1	**91.9±0.1**	89.7±0.0	90.2±0.1
	AUC	93.9±0.0	92.7±0.1	95.0±0.0	96.0±0.0	96.3±0.0	95.8±0.1	96.2±0.0	**96.5±0.0**
Task 14	ACC	66.0±0.0	67.9±0.6	63.2±0.2	63.8±0.7	65.9±0.2	65.5±0.1	64.1±0.2	**69.7±0.1**
	AUC	71.5±0.0	77.2±0.1	78.0±0.1	84.5±0.0	83.5±0.0	71.5±0.1	81.3±0.1	**85.0±0.1**

F. Experimental Results in Detail

We present detailed experimental results for 75 experiments on 33 tasks across 9 datasets in Tabs. 10 to 15. In the era of foundation models, no single model achieves state-of-the-art performance across all tasks, and each model tends to exhibit unique strengths on specific benchmarks. Across 145 metrics from 75 experiments spanning 33 tasks, CARE attains 39 best and 50 second-best results among all compared methods. Its advantage is particularly pronounced on molecular prediction and survival analysis tasks shown in Tabs. 1 and 16, likely due to the guidance provided by molecular features during pretraining. Intuitively, performance on molecular prediction tasks remains weaker than on morphologic classification across all models. While existing approaches already meet practical requirements for morphologic classification in clinical deployment, CARE represents a modest yet meaningful step toward deployable models for molecular prediction.

For the survival prediction tasks, we observe that all models exhibit large standard deviations and relatively low performance. On the one hand, survival analysis is intrinsically challenging when relying solely on histopathology images. On the other hand, we only report linear probing results for all foundation models rather than fully fine-tuned performance. Notably, compared with the second-best model, CARE achieves a 7.2% improvement on Task 30, highlighting its substantial potential for survival analysis. A distinctive advantage of CARE over other models is its ability to explicitly identify regions of interest (ROIs) in pathology slides that are highly associated with survival outcomes.

The box-plot results for Tasks 1–14 under different experimental settings are shown in Figs. 5 and 7. Based on the balanced accuracy, we observe that fine-tuning consistently outperforms linear probing with logistic regression. Considering the mean, median, and interquartile ranges of the box plots, CARE delivers highly competitive performance,

Table 16. Average AUC (or F1) results across task categories. The best score is in bold, and the second-best is underlined.

Task	Head	Mean-pool	CHIEF	PRISM	GigaPath	TANGLE	FEATHER	TITAN	CARE
Morph. Class.	LR	87.2±0.0	86.6±0.0	84.5±0.0	85.3±0.0	87.1±0.0	87.4±0.1	90.1±0.0	**90.1±0.0**
	kNN	82.2±0.0	80.3±0.1	79.8±0.0	78.0±0.0	83.6±0.0	81.8±0.0	**88.1±0.0**	87.3±0.0
	FT	82.2±0.1	79.5±0.1	85.4±0.2	86.6±0.2	87.9±0.1	88.1±0.1	**89.7±0.1**	88.2±0.3
Molecular Class.	LR	79.8±0.1	82.5±0.1	82.7±0.1	83.1±0.1	83.5±0.1	82.0±0.1	81.1±0.2	**83.6±0.1**
	kNN	72.5±0.1	73.8±0.1	75.7±0.1	71.3±0.2	75.5±0.1	74.5±0.1	76.7±0.2	**79.1±0.2**
	FT	78.3±0.1	80.3±0.1	83.8±0.1	84.7±0.1	85.4±0.1	82.2±0.1	83.3±0.1	**85.5±0.1**
Molecular Class.$_V$	LR	69.9±1.4	67.6±1.6	66.0±1.7	69.9±1.4	71.5±1.2	65.3±1.8	71.5±1.2	**73.1±1.3**
	kNN	65.9±1.5	66.0±1.6	64.6±1.5	65.6±1.6	**69.2±1.4**	63.7±1.6	68.5±1.4	68.2±1.4

demonstrating that its adaptive region modeling remains effective even when pretrained on only one-tenth of the data used by mainstream models.

G. Interpretability Analysis

In Fig. 6, we show a heatmap visualization of the adaptive regions to highlight which regions are more important. Each region is rendered with a uniform color (for better visual appearance, we apply Gaussian blur to smooth the region boundaries). The pathologist provided detailed annotations of the ROIs on the WSI. In the annotated regions, prominent cytologic atypia is present, with deeply staining cytoplasm and numerous mitotic figures. Intercellular bridges are scant, and focal central necrosis is observed within some tumor nests. The ROI identified by CARE (the most intensely highlighted red region) closely matches the pathologist-annotated areas. Moreover, from CARE-iBOT to CARE-RNA and finally to CARE, the ROI attention progressively converges toward these expert-annotated regions.

To further investigate the contribution of different patches within the ROI, we additionally visualize an attention-score heatmap over the patches inside the ROI, shown in Fig. 8. Each subfigure shows the adaptive region scores of the SPF module under the CARE pretrained weights. The zoomed-in inset in each subfigure displays the attention-score heatmap of patches within the ROI, indicating which patches contribute most strongly to the ROI representation.