

Supplementary Information for:

Vision Foundation Models for Multi-Modal Medical Image Classification

Supplementary information

The main text reports the Matthews correlation coefficient (MCC) as the primary metric. For completeness, each experiment is repeated here under two further metrics: macro-averaged AUROC, which is less sensitive to class imbalance than MCC, and sensitivity (macro-averaged per-class recall, equivalently balanced accuracy), which restates the comparison in a clinically interpretable form as the fraction of cases in each class a model correctly identifies (§4.8). The foundation-model advantage is preserved under both, most visibly on the seven-class LLD-MMRI task.

Supplementary Note 1: Full-data performance

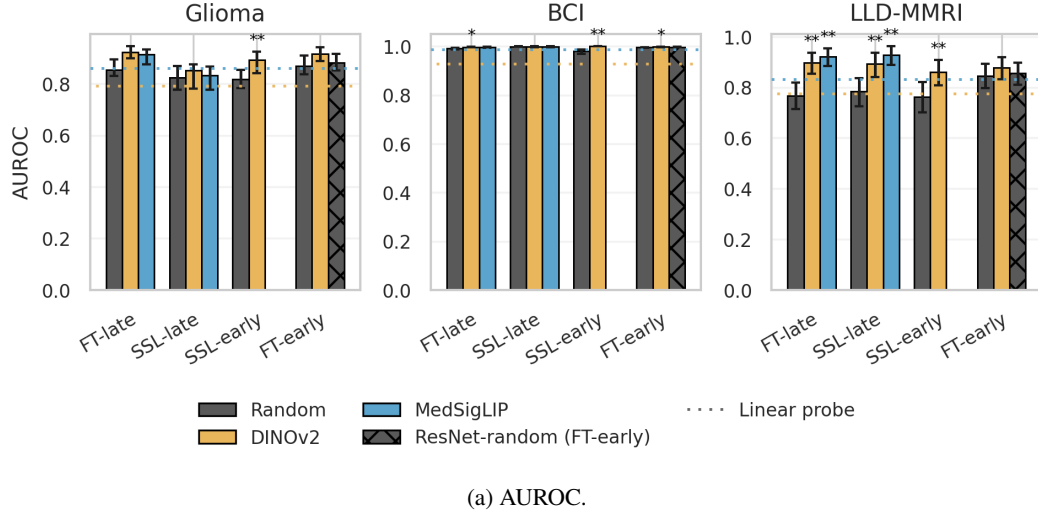
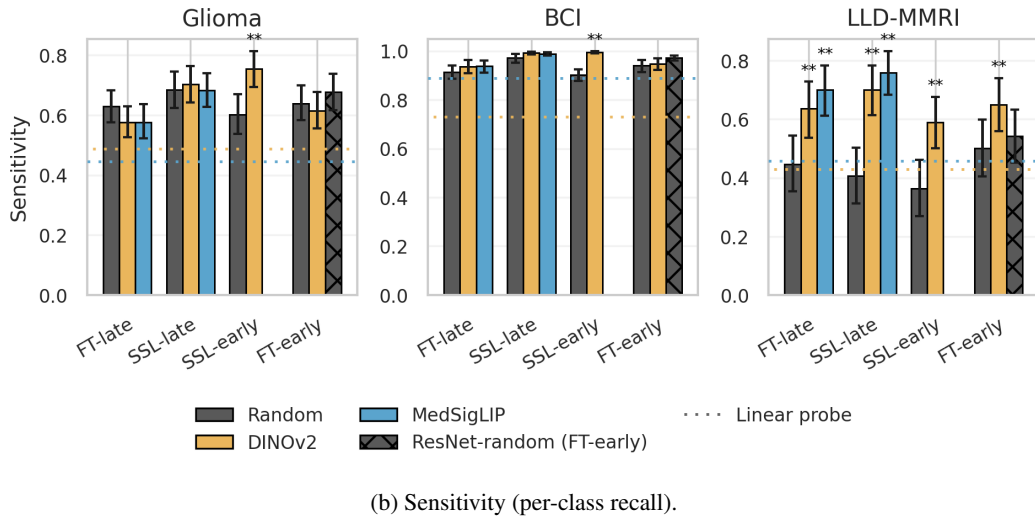
Supplementary Figure 1 gives the AUROC and sensitivity views of the full-data comparison in Figure 2.

Supplementary Note 2: Label efficiency

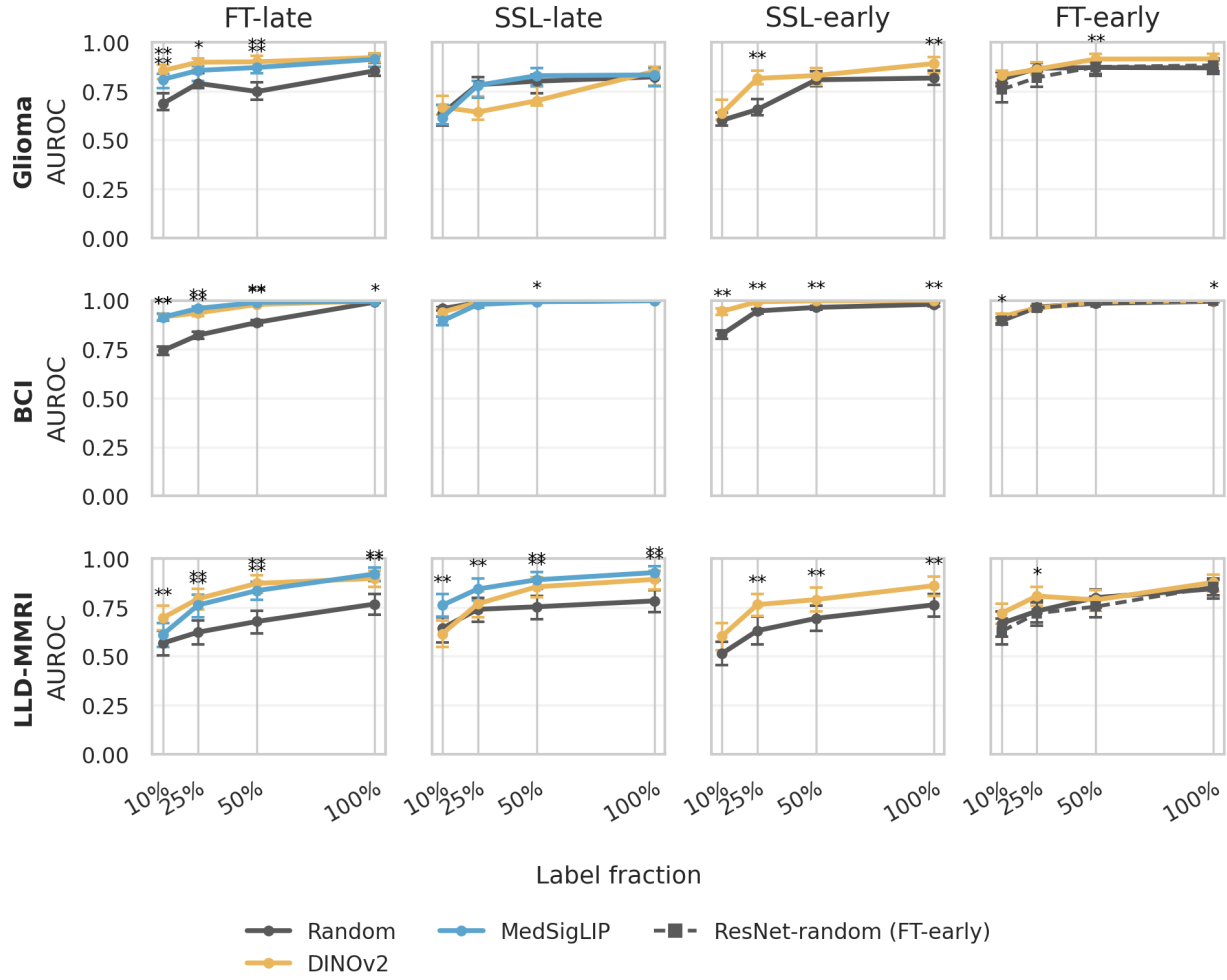
Supplementary Figures 2 and 3 repeat the label-efficiency experiment (Figure 3) in AUROC and sensitivity.

Supplementary Note 3: Missing-modality robustness

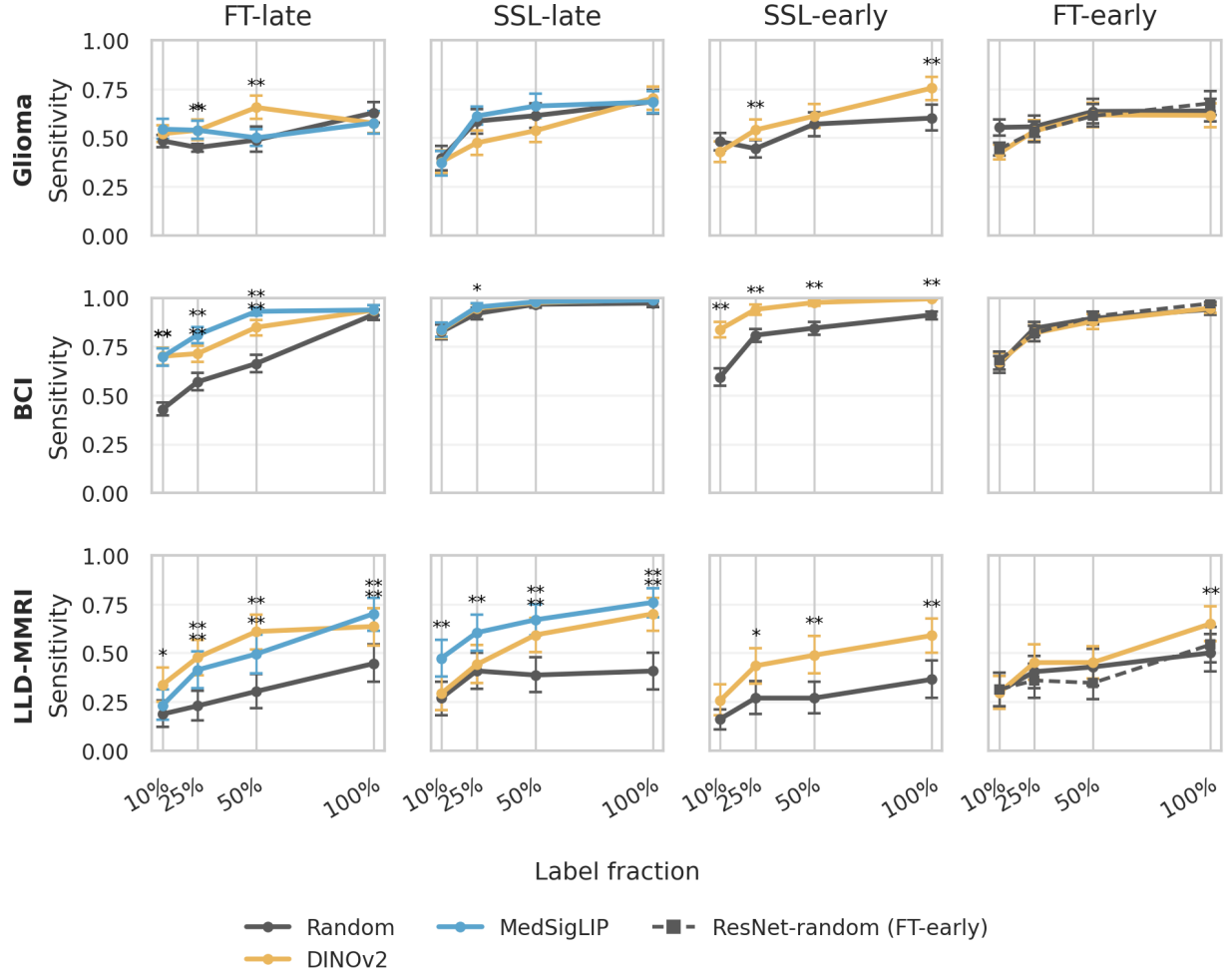
Supplementary Figures 4 and 5 repeat the missing-modality experiment (Figure 4) in AUROC and sensitivity, Supplementary Figure 6 aggregates the same comparison across all three metrics, and Supplementary Figure 7 expresses degradation in relative rather than absolute terms.



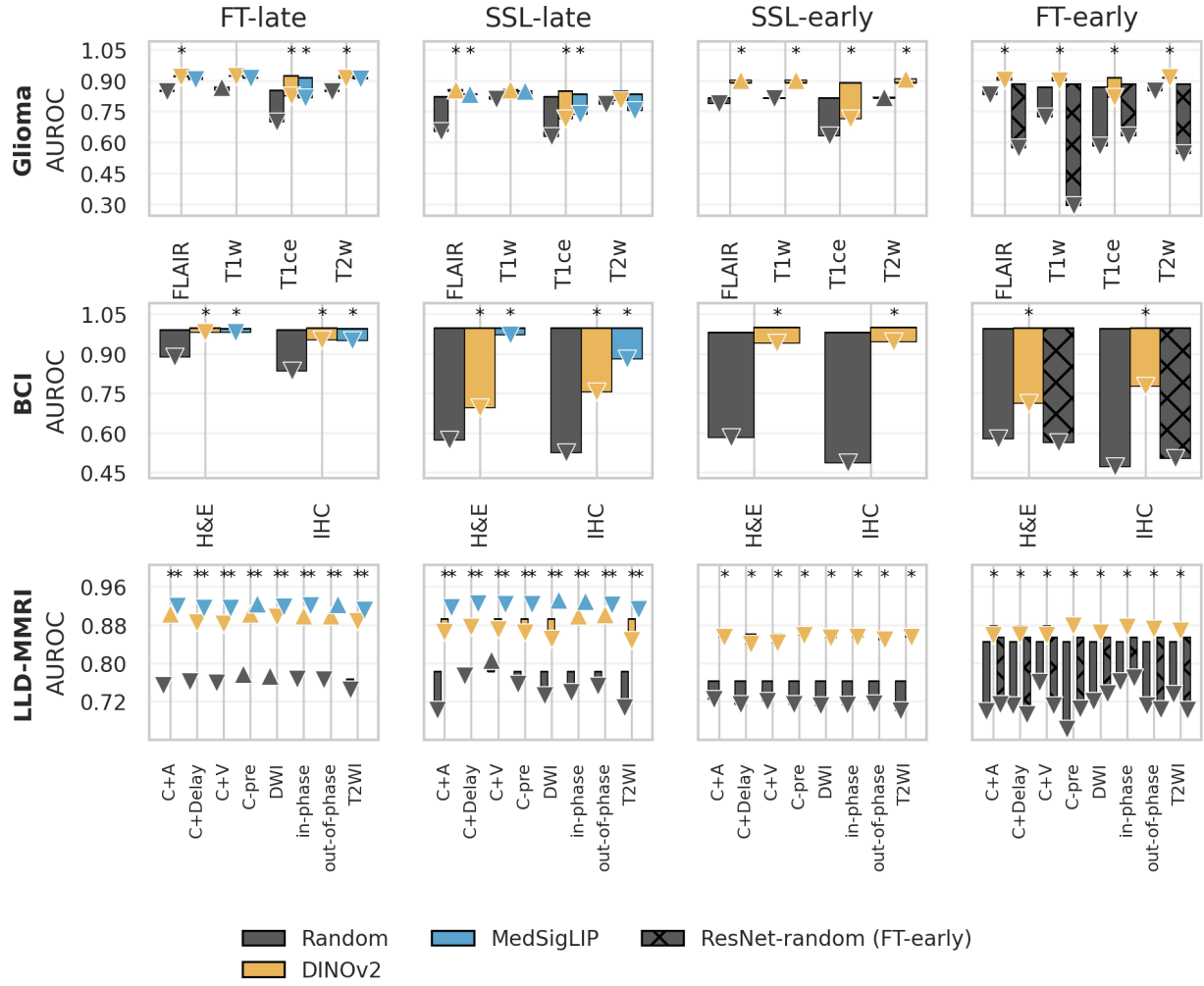
Supplementary Figure 1: **Full-data performance in AUROC and sensitivity.** Secondary-metric (a, AUROC) and clinical-metric (b, sensitivity) views of Figure 2, both most informative on the imbalanced LLD-MMRI task. Error bars are 95 % bootstrap confidence intervals, and stars mark Benjamini–Hochberg-corrected significance of each foundation-model bar against its matched random-initialisation bar (* $p < 0.05$, ** $p < 0.01$).



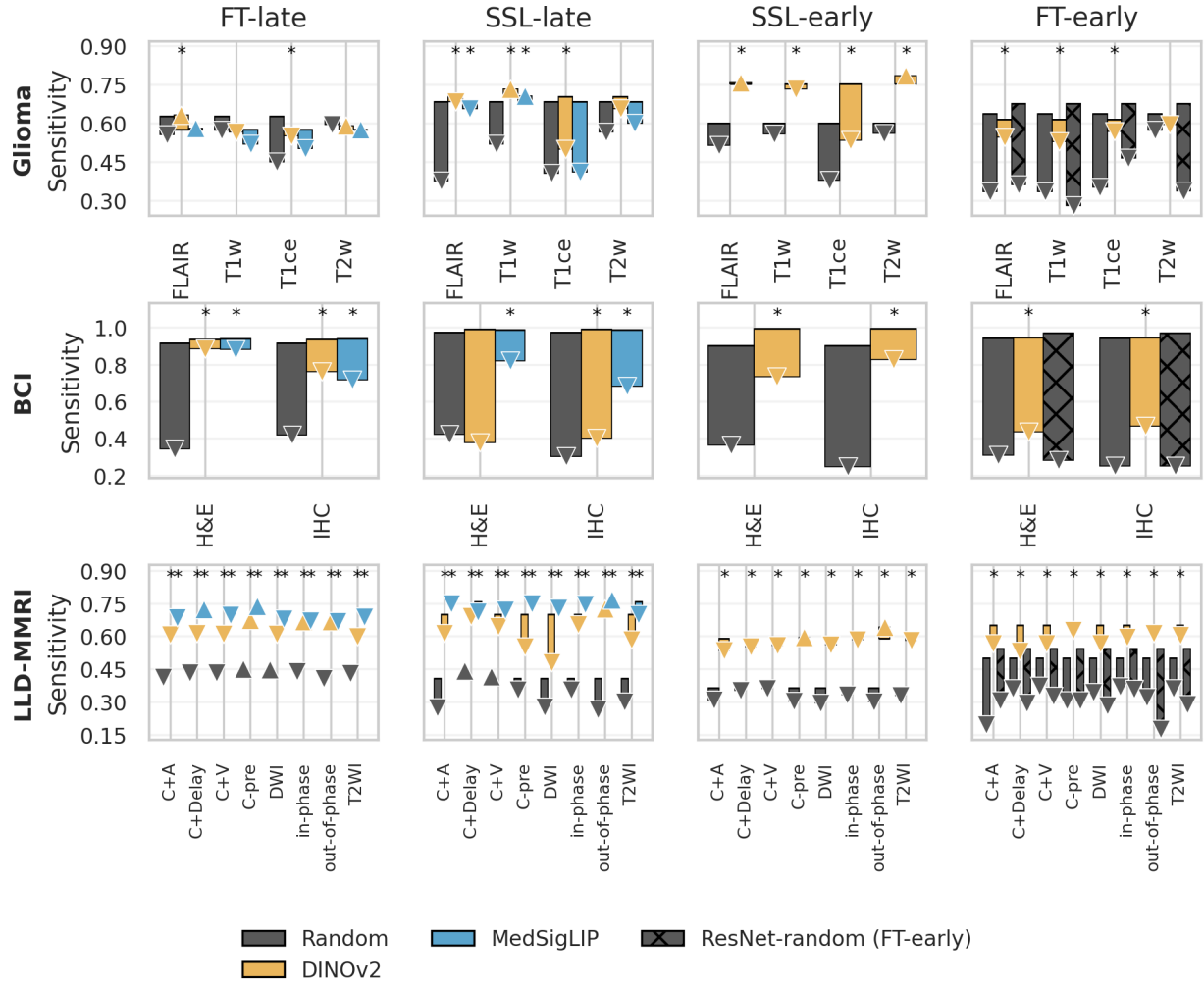
Supplementary Figure 2: **Label efficiency in AUROC.** Secondary-metric view of Figure 3, with macro-averaged test AUROC on the y-axis. Error bars are 95 % bootstrap confidence intervals, and stars mark Benjamini–Hochberg-corrected significance of each foundation-model curve against the randomly initialised curve at the same labelled fraction (* $p < 0.05$, ** $p < 0.01$).



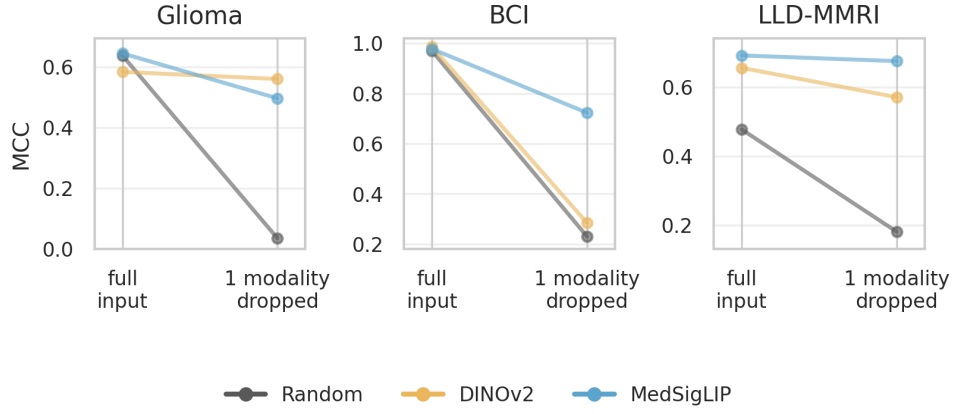
Supplementary Figure 3: **Label efficiency in sensitivity.** Clinical-metric view of Figure 3, with macro-averaged test sensitivity (per-class recall) on the y-axis. Error bars and significance stars follow Supplementary Figure 2.



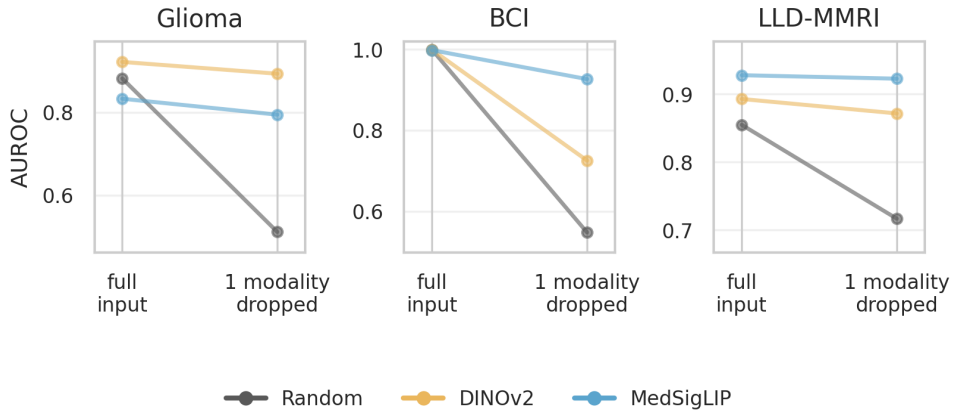
Supplementary Figure 4: **Missing-modality robustness in AUROC.** Secondary-metric view of Figure 4, with macro-averaged test AUROC on the y-axis. Stars in the top row of each panel mark Benjamini-Hochberg-corrected significance of each foundation-model initialisation against the randomly initialised reference for that masked modality (* $p < 0.05$, ** $p < 0.01$).



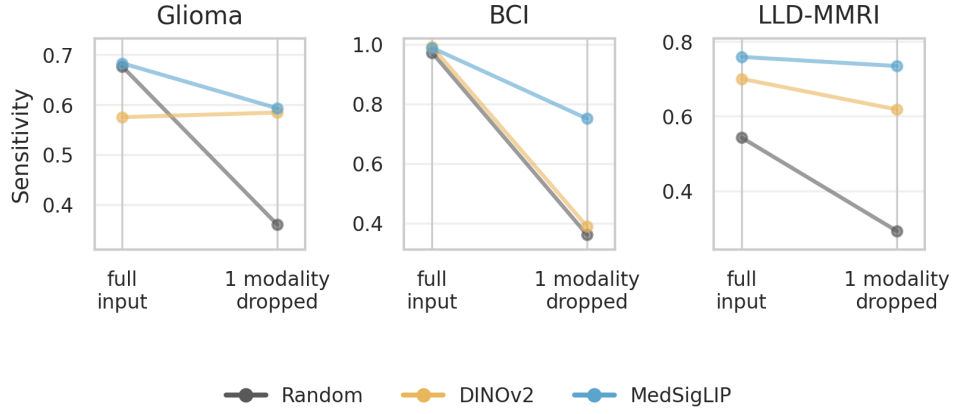
Supplementary Figure 5: **Missing-modality robustness in sensitivity.** Clinical-metric view of Figure 4, with sensitivity (per-class recall) on the y-axis. Floating bars span each method's full-input value down to its masked-modality value, and significance stars follow Supplementary Figure 4.



(a) MCC.

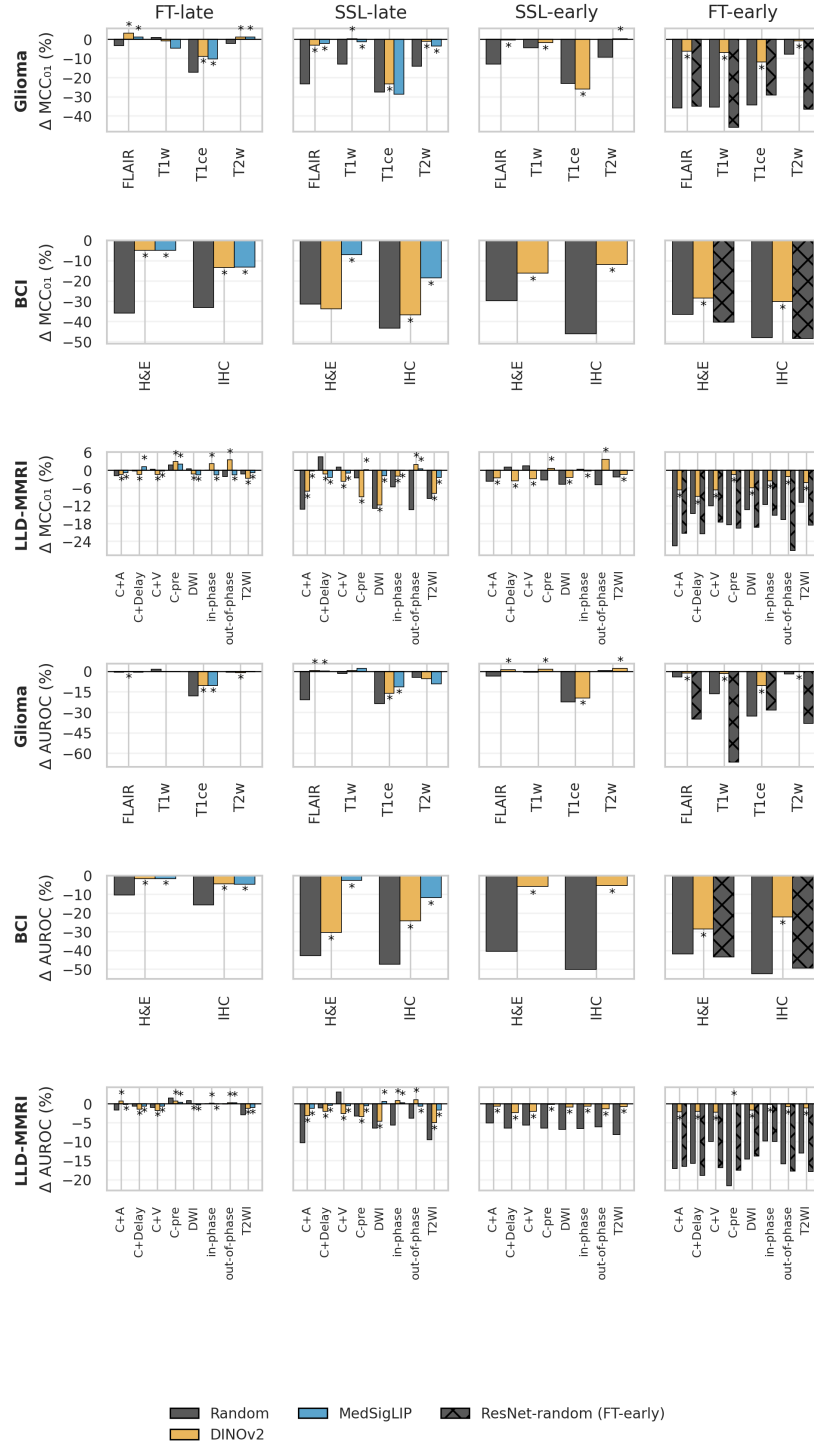


(b) AUROC.



(c) Sensitivity (per-class recall).

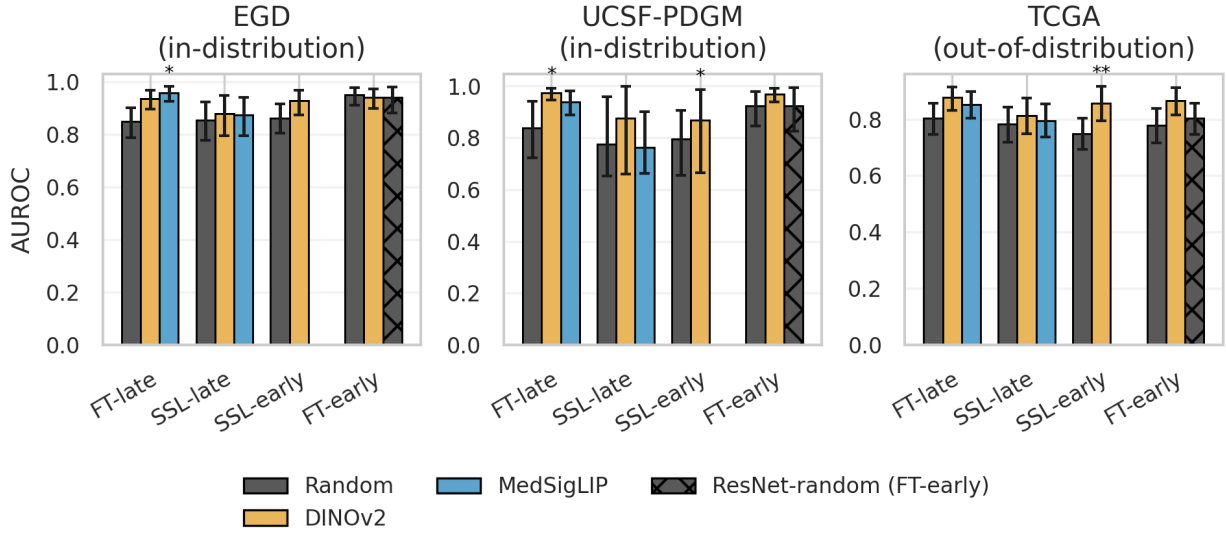
Supplementary Figure 6: **Missing-modality robustness, aggregated across metrics.** Each line is one initialisation family's best configuration going from its full-input test score (left) to the mean of its single-modality-masked scores (right), in (a) MCC, (b) AUROC, and (c) sensitivity. The best configuration per family is selected by full-input validation MCC, the same selector used for the main comparisons. Foundation-model initialisation stays close to its full-input score while random initialisation collapses, and the gap is widest in sensitivity: on LLD-MMRI, MedSigLIP late fusion holds at 0.73 (from 0.76) while the random baseline falls to 0.29 (from 0.54).



Supplementary Figure 7: **Relative degradation under missing modalities.** Alternative view of Figure 4 expressing single-modality dropout as relative percentage degradation rather than absolute level, for both metrics: the upper three rows show change in MCC and the lower three change in AUROC, with datasets along the rows and (training strategy, fusion) configurations along the columns. Each bar is the percentage change from a method's own full-input baseline to its masked-modality value, holding the validation-selected checkpoint fixed within each method. This view normalises for differing baselines, while the absolute-axis view in Figure 4 remains the headline because it makes the randomly initialised floor effect visible.

17 **Supplementary Note 4: Cross-site generalisation**

18 Supplementary Figure 8 gives the AUROC view of the cross-site generalisation experiment (Figure 5).



Supplementary Figure 8: **Cross-site generalisation in AUROC**. Secondary-metric view of Figure 5, with test AUROC on the y-axis. Each panel is one site: EGD and UCSF-PDGM are in-distribution and TCGA is the held-out out-of-distribution cohort. Error bars are 95 % bootstrap confidence intervals, and stars mark Benjamini–Hochberg-corrected significance against the matched random-initialisation bar (* $p < 0.05$, ** $p < 0.01$).

19 **Supplementary Note 5: Validation-set MCC for all models**

20 Throughout the main text we select the best checkpoint for each model on the validation set and report its held-out test
 21 performance. For completeness, Supplementary Tables 1–5 list the validation-set MCC that drives this selection for
 22 every model, across the three training experiments (full-data, missing-modality, and label-efficiency). These are the
 23 selection-metric values, not the test results, and we provide them for transparency rather than for ranking the methods.

Supplementary Table 1: Validation-set MCC for the full-data experiment. These are the selection-metric values used to pick the best checkpoint per model, and the corresponding held-out test results appear in the main text. Glioma values are the macro-average over the in-distribution and out-of-distribution validation cohorts. Bold marks the best-validation foundation-model configuration and underline the best-validation random-initialisation configuration, the two models entering the best-versus-best comparison.

Strategy	Init	Fusion	Glioma	BCI	LLD-MMRI
FT	Random (ResNet)	Early	<u>0.547</u>	0.923	<u>0.561</u>
	Random	Early	0.540	0.882	0.452
	Random	Late	0.432	0.871	0.468
	DINOv2	Early	0.605	0.943	0.676
	DINOv2	Late	0.656	0.928	0.688
	MedSigLIP	Late	0.671	0.943	0.644
SSL	Random	Early	0.397	0.820	0.342
	Random	Late	0.478	<u>0.953</u>	0.399
	DINOv2	Early	0.581	0.985	0.622
	DINOv2	Late	0.644	0.987	0.699
	MedSigLIP	Late	0.671	0.987	0.734
Linear probe	DINOv2	Late	0.462	0.660	0.318
	MedSigLIP	Late	0.379	0.828	0.404

Supplementary Table 2: Validation-set MCC under missing modalities: Glioma. *Full* is the all-modalities setting, and the remaining columns each mask the named modality at inference time. Bold marks the best-validation foundation-model configuration and underline the best-validation random-initialisation configuration, selected independently for each masked-modality column (the per-mask best-versus-best comparison).

Strategy	Init	Fusion	Full	T2w	T1ce	FLAIR	T1w
FT	Random (ResNet)	Early	<u>0.547</u>	0.000	0.087	0.151	-0.091
	Random	Early	0.540	0.347	-0.022	0.000	0.000
	Random	Late	0.432	<u>0.382</u>	<u>0.217</u>	<u>0.352</u>	<u>0.380</u>
	DINOv2	Early	0.605	0.584	0.421	0.561	0.504
	DINOv2	Late	0.656	0.625	0.413	0.614	0.574
	MedSigLIP	Late	0.671	0.621	0.403	0.593	0.544
SSL	Random	Early	0.397	0.342	0.039	0.341	0.354
	Random	Late	0.478	0.230	0.162	0.177	0.331
	DINOv2	Early	0.581	0.560	0.195	0.538	0.580
	DINOv2	Late	0.644	0.531	0.299	0.567	0.569
	MedSigLIP	Late	0.671	0.586	0.093	0.613	0.628
Linear probe	DINOv2	Late	0.462	0.421	0.324	0.372	0.403
	MedSigLIP	Late	0.379	0.330	0.169	0.515	0.388

Supplementary Table 3: Validation-set MCC under missing modalities: BCI. *Full* is the all-modalities setting, and the remaining columns each mask the named modality at inference time. Bold marks the best-validation foundation-model configuration and underline the best-validation random-initialisation configuration, selected independently for each masked-modality column (the per-mask best-versus-best comparison).

Strategy	Init	Fusion	Full	IHC	H&E
FT	Random (ResNet)	Early	0.923	0.000	0.181
	Random	Early	0.882	0.006	0.229
	Random	Late	0.871	0.225	0.207
	DINOv2	Early	0.943	0.352	0.382
	DINOv2	Late	0.928	0.641	0.815
	MedSigLIP	Late	0.943	0.640	0.843
SSL	Random	Early	0.820	-0.006	0.291
	Random	Late	<u>0.953</u>	0.092	<u>0.336</u>
	DINOv2	Early	<u>0.985</u>	0.763	0.660
	DINOv2	Late	0.987	0.275	0.319
	MedSigLIP	Late	0.987	0.575	0.792
Linear probe	DINOv2	Late	0.660	0.124	0.378
	MedSigLIP	Late	0.828	0.039	0.432

Supplementary Table 4: Validation-set MCC under missing modalities: LLD-MMRI. *Full* is the all-modalities setting, and the remaining columns each mask the named modality at inference time. Bold marks the best-validation foundation-model configuration and underline the best-validation random-initialisation configuration, selected independently for each masked-modality column (the per-mask best-versus-best comparison).

Strategy	Init	Fusion	Full	T2WI	out-of-phase	C-pre	DWI	C+A	C+V	C+Delay	in-phase
FT	Random (ResNet)	Early	0.561	0.215	0.136	0.138	0.264	0.172	0.323	0.153	0.340
	Random	Early	0.452	0.281	0.186	0.126	0.253	0.073	0.272	0.170	0.322
	Random	Late	0.468	<u>0.445</u>	<u>0.417</u>	<u>0.424</u>	<u>0.448</u>	<u>0.404</u>	<u>0.387</u>	<u>0.369</u>	<u>0.434</u>
	DINOv2	Early	0.676	0.580	0.665	0.626	0.564	0.592	0.529	0.530	0.629
	DINOv2	Late	0.688	0.672	0.672	0.672	0.660	0.703	0.655	0.608	0.688
	MedSigLIP	Late	0.644	0.548	0.593	0.612	0.612	0.593	0.575	0.579	0.656
SSL	Random	Early	0.342	0.275	0.289	0.269	0.220	0.235	0.164	0.245	0.271
	Random	Late	0.399	0.235	0.316	0.342	0.293	0.182	0.363	0.289	0.348
	DINOv2	Early	0.622	0.512	0.621	0.559	0.606	0.573	0.523	0.575	0.559
	DINOv2	Late	0.699	0.462	0.605	0.587	0.389	0.483	0.556	0.608	0.624
	MedSigLIP	Late	0.734	0.670	0.717	0.702	0.693	0.575	0.716	0.652	0.721
Linear probe	DINOv2	Late	0.318	0.267	0.282	0.339	0.270	0.283	0.268	0.246	0.265
	MedSigLIP	Late	0.404	0.302	0.369	0.399	0.395	0.521	0.320	0.388	0.358

Supplementary Table 5: Validation-set MCC across label fractions (label-efficiency experiment). Each block is one dataset, and columns are the fraction of training labels used. Values are the selection-metric scores underlying the test-set label-efficiency curves in the main text. Bold marks the best-validation foundation-model configuration and underline the best-validation random-initialisation configuration, the two models entering the best-versus-best comparison.

Strategy	Init	Fusion	Glioma				BCI				LLD-MMRI			
			10%	25%	50%	100%	10%	25%	50%	100%	10%	25%	50%	100%
FT	Random (ResNet)	Early	0.253	<u>0.421</u>	0.514	<u>0.547</u>	0.625	0.778	0.840	0.923	0.124	0.301	<u>0.420</u>	<u>0.561</u>
	Random	Early	<u>0.361</u>	0.400	<u>0.521</u>	0.540	0.590	0.735	0.819	0.882	<u>0.232</u>	<u>0.369</u>	0.396	0.452
	Random	Late	0.250	0.282	0.330	0.432	0.283	0.407	0.521	0.871	0.148	0.224	0.291	0.468
	DINOv2	Early	0.373	0.497	0.580	0.605	0.704	0.791	0.869	0.943	0.250	0.342	0.504	0.676
	DINOv2	Late	0.429	0.584	0.623	0.656	0.669	0.753	0.842	0.928	0.272	0.420	0.558	0.688
	MedSigLIP	Late	0.380	0.512	0.604	0.671	0.708	0.804	0.875	0.943	0.113	0.348	0.489	0.644
SSL	Random	Early	0.204	0.281	0.406	0.397	0.371	0.637	0.748	0.824	0.111	0.189	0.239	0.342
	Random	Late	0.250	0.370	0.414	0.478	<u>0.740</u>	<u>0.887</u>	<u>0.906</u>	<u>0.953</u>	0.205	0.302	0.287	0.399
	DINOv2	Early	0.251	0.382	0.458	0.581	0.777	0.932	0.968	<u>0.985</u>	0.172	0.360	0.495	0.622
	DINOv2	Late	0.165	0.273	0.411	0.644	0.754	0.934	0.974	0.987	0.248	0.461	0.605	0.699
	MedSigLIP	Late	0.127	0.553	0.638	0.671	0.740	0.934	0.983	0.987	0.477	0.572	0.668	0.734
Linear probe	DINOv2	Late	0.141	0.325	0.333	0.445	0.492	0.571	0.588	0.660	0.040	-0.003	0.174	0.318
	MedSigLIP	Late	0.087	0.210	0.195	0.379	0.593	0.720	0.765	0.828	0.061	0.173	0.315	0.404