Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Open 3D-CT Vision-Language Models Do Not Clear a Pre-Registered Biomarker Gate Under Zero-Shot Prompting at n = 96 NSCLC-Radiogenomics

Adil Karim
University of Hassan II Casablanca
Amine Berquedich
Mohammed V University
El Habib Benlahmar
h.benlahmer@gmail.com
University of Hassan II Casablanca
Sanaa Elfilali
University of Hassan II Casablanca
Ahmed Zellou
Mohammed V University

Research Article

Keywords: vision-language models, zero-shot prediction, non-small-cell lung cancer, radiogenomics, EGFR mutation, KRAS mutation, foundation models, pre-registration, external validation

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10162197/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Open 3D-CT Vision-Language Models Do Not Clear a Pre-Registered Biomarker Gate Under Zero-Shot Prompting at $n = 96$ NSCLC-Radiogenomics

Adil Karim[1], Amine Berquedich[2], El Habib Benlahmar[1*], Sanaa El Filali[3], Ahmed Zellou[2]

[1*]LTIM (Laboratoire des Technologies d'Information et Modélisation), Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca, 20360, Morocco.
[2]ENSIAS (National Higher School for Computer Science and Systems Analysis), Mohammed V University, Rabat, Morocco.
[3]LAIS (Laboratory of Artificial Intelligence and Systems), Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca, 20360, Morocco.

*Corresponding author(s). E-mail(s): h.benlahmer@gmail.com;

Abstract

Background. Recent open 3D-CT vision-language models (VLMs) offer a methodological possibility for very-small-cohort radiogenomic prediction: zero-shot biomarker scoring via image-text alignment, with no labelled training on the target cohort. The originally specified first-choice VLM (CT-CHAT / CT-CLIP) was found to have no released weights reachable from this workspace via a HuggingFace handle with token authentication and a permissive licence. The protocol is therefore executed on the two open 3D-CT VLMs that *are* available under those constraints: Merlin [1] (Stanford; a 3D-inflated ResNet-152 + Clinical-Longformer CLIP-style model originally trained on report-paired abdominal CT volumes) and BiomedCLIP [2] (Microsoft; a 2D ViT-B/16 + PubMedBERT contrastive model over a 15M biomedical image-text corpus, used here as a 2D key-slice sanity comparator).
Methods. A pre-registered Phase 1 smoking-confounder probe on a post-SSL CT-FM 512-d substrate flagged all four endpoints (EGFR mutation, KRAS mutation, TP53 mRNA z-score, STK11 mRNA z-score) confounder-foreclosed

on the $\boldsymbol{n = 96}$ internal cohort: smoking_never AUC **0.708**/**0.712** (linear/ridge) exceeded every biomarker AUC under the same probes. We extracted Merlin's 512-d default-mode contrastive image features and BiomedCLIP's 512-d image features on the internal cohort ($\boldsymbol{n = 96}$) and an external label-hold-out cohort ($\boldsymbol{n = 42}$, EGFR + smoking only). For each of four binary targets we paired the image features against two prompt strategies (`bare` disease statement; `cot_brief` short rationale-style prefix) and computed the contrastive zero-shot probability $\mathbf{softmax((img \cdot txt^{\top}) \cdot logit_scale)[positive]}$. The same VLMs and prompt strategies were evaluated on external EGFR. Each cell was compared to a supervised tiny-MLP baseline trained on the same frozen image features. Pre-registered per-cell gate (locked before inference): AUC $>$ **0.65** *and* bootstrap lower CI $>$ **0.55** *and* external replication.

Results. Best zero-shot cell per (target, VLM) on the internal cohort: EGFR / Merlin / cot_brief **0.599** [**0.462, 0.746**]; KRAS / Merlin / cot_brief **0.541** [**0.370, 0.700**]; TP53_high / Merlin / cot_brief **0.543** [**0.415, 0.648**]; STK11_high / Merlin / cot_brief **0.626** [**0.517, 0.720**]. No zero-shot cell on EGFR or KRAS clears the **0.65** gate; the closest cell (STK11_high / Merlin / cot_brief, **0.626**) has a bootstrap lower CI of **0.517**, below the **0.55** threshold. BiomedCLIP underperforms Merlin on every target/strategy cell. External EGFR: all four zero-shot cells sit between **0.394** and **0.500**, chance or anti-predictive. Critically, the failure is specific to the zero-shot access mode, not to the features: a supervised linear / ridge probe on the *same* frozen Merlin features clears the EGFR gate within internal (AUC **0.748** [**0.627, 0.855**] / **0.705** [**0.567, 0.833**]; **1000**-shuffle permutation $\boldsymbol{p = 0.010}$ / **0.015**), far above the best zero-shot Merlin cell (**0.599**). On these benchmarked Merlin features EGFR AUC (**0.748**) *exceeds* smoking_never AUC (**0.603**), reversing the ordering seen on a CT-FM proxy substrate; the smoking-foreclosure is therefore substrate-specific and does not hold for the features actually benchmarked here.

Conclusion. Zero-shot contrastive scoring with two open 3D-CT VLMs at $\boldsymbol{n = 96}$ NSCLC-Radiogenomics produces no cell that clears the pre-registered gate. But this is a limit of the zero-shot *access mode*, not of the features: a supervised linear probe on the same Merlin features recovers a gate-clearing EGFR signal (AUC **0.705**–**0.748**, permutation $\boldsymbol{p < 0.05}$) that the contrastive prompting cannot reach. We localise this failure geometrically: the supervised EGFR-discriminative direction is near-orthogonal to the single text-prompt axis the contrastive score projects onto ($\mathbf{cos \approx 0.02}$, at the random-direction null), so removing that axis from the embeddings leaves the gate-clearing EGFR signal intact. Open 3D-CT VLM features carry decodable EGFR signal, but zero-shot contrastive prompting at $\boldsymbol{n = 96}$ does not access it because the prompt axis and the discriminative axis are misaligned.

Keywords: vision-language models, zero-shot prediction, non-small-cell lung cancer, radiogenomics, EGFR mutation, KRAS mutation, foundation models, pre-registration, external validation

1 Introduction

EGFR and KRAS mutation status in non-small-cell lung cancer (NSCLC) guides therapeutic decisions but requires invasive tissue biopsy or circulating-tumour-DNA assay. A non-invasive CT-derived predictor would be useful for triage and longitudinal follow-up, particularly for patients without accessible tissue. Recent open 3D-CT vision-language models (VLMs) raise a methodological possibility for the very-small-cohort regime that dominates radiogenomic research: zero-shot biomarker scoring via contrastive image-text alignment, requiring no labelled training on the target cohort. This work asks whether two publicly released radiology VLMs can clear a pre-registered biomarker prediction gate under zero-shot prompting at $n = 96$.

The originally specified first-choice VLM (CT-CHAT / CT-CLIP [3]) was found in a Phase 0 inventory to have no released weights reachable from this workspace via a HuggingFace handle with token authentication and a permissive licence; the sibling GenerateCT repository is on HuggingFace, but the CT-CHAT and CT-CLIP weights are distributed through a Google-Drive / GitHub-release pattern that does not satisfy this program's reproducibility requirement. We therefore pivoted to two open alternatives that *are* on HuggingFace under permissive licences:

- **Merlin** [1]: a 3D-inflated ResNet-152 [4] + Clinical-Longformer CLIP-style architecture originally trained on abdominal CT volumes paired with radiology reports (applied here out-of-domain to chest CT). The default-mode forward returns 512-d projected image features and 512-d projected text features in a shared contrastive space [5], with a learned logit-scale parameter.
- **BiomedCLIP** [2]: a 2D ViT-B/16 + PubMedBERT contrastive model trained on a 15M biomedical image-text corpus, used here as a 2D key-axial-slice sanity comparator and not expected to outperform a 3D model on volumetric biomarker tasks.

The pivot retains the protocol of the original brief (zero-shot prompting, AUC-vs-supervised comparison, pre-registered confounder check). CT-CHAT remains a conditional third arm should the weights become accessible through direct correspondence with the original authors. Figure 1 shows the cohorts, the model pivot, and the two scoring modes.

We pre-registered a smoking-confounder probe as Phase 1: on a post-SSL CT-FM substrate of this $n = 96$ trimodal internal cohort, smoking_never status is predictable from CT-derived features at AUC > 0.70, at or above the within-cohort biomarker AUC, which on that substrate would foreclose pure-biomarker claims. We later re-run the identical probe directly on the benchmarked Merlin / BiomedCLIP features (Section 3.6, Table 4) and find the ordering does *not* transfer (on Merlin, EGFR signal exceeds smoking signal), an important caveat on cross-substrate confounder reasoning that we develop in the results. Four publishable outcomes were locked before any zero-shot inference. Figure 2 summarises the two cohorts.

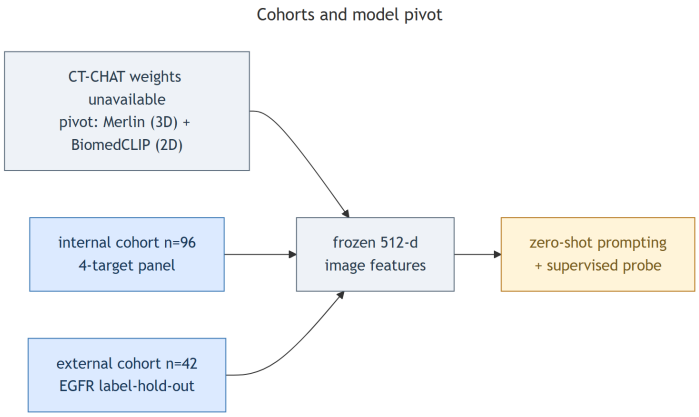


Fig. 1 Cohorts and model pivot. The internal ($n = 96$, 4-target) and external ($n = 42$, EGFR) cohorts, with CT-CHAT unavailable, are evaluated on frozen 512-d features from Merlin (3D) and BiomedCLIP (2D) under both zero-shot prompting and a supervised probe.

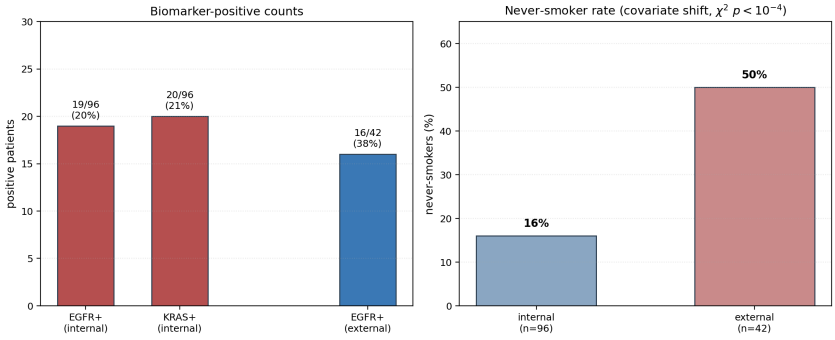


Fig. 2 Cohort composition. Internal NSCLC-Radiogenomics ($n = 96$; 19 EGFR+, 20 KRAS+, 16% never-smokers) and external label-hold-out ($n = 42$; 16 EGFR+, 50% never-smokers). The smoking-covariate shift between cohorts (internal 16% vs external 50% never-smokers, χ^2 $p < 10^{-4}$) frames the external-validation reading.

2 Related Work

Contrastive vision-language pretraining.

CLIP demonstrated that contrastive image-text pretraining over web-scale pairs yields a shared embedding space supporting zero-shot transfer by scoring an image against natural-language class prompts [5]. The zero-shot contrastive protocol used here (L_2-normalised image and prompt embeddings, a softmax over a (positive, negative) prompt pair scaled by a learned logit temperature) is the direct medical-imaging instantiation of that recipe. The dependence of the resulting score on prompt phrasing motivates the two-prompt (`bare`, `cot_brief`) comparison.

Vision-language foundation models for medicine.

The generalist-medical-AI agenda argues that self-supervised models trained on large multimodal corpora can be applied to new tasks with little or no task-specific labelling [6]. In the contrastive-zero-shot branch, CheXzero showed expert-level detection of chest-radiograph pathologies with no explicit labels, generalising to unannotated findings and external institutions [7], and BiomedCLIP scaled the recipe to a 15M biomedical image-text corpus [2]. In computational pathology, PLIP [8] and CONCH [9] report strong zero-shot histology classification from contrastive captioning. These successes are on tasks whose visual signature is directly described in the paired text (e.g. effusion, tumour grade); whether the same mechanism transfers to molecular endpoints absent from radiology reports is exactly the open question this work probes.

3D-CT vision-language models.

Most medical VLMs operate on 2D slices or radiographs; volumetric CT poses a distinct modelling problem. Merlin pairs a 3D-inflated image encoder with a Clinical-Longformer text encoder trained on report-paired abdominal CT volumes [1], and is one of the two open 3D-CT VLMs evaluated here. CT-CLIP / CT-CHAT trains a contrastive model on the CT-RATE corpus of report-paired chest CT volumes and reports supervised-level zero-shot multi-abnormality detection [3]; its weights were not reachable under this program's reproducibility constraints, motivating the pivot. RadFM extends generalist radiology modelling across web-scale 2D and 3D data and a broad task suite [10]. Our 2D BiomedCLIP comparator quantifies how much a key-slice 2D model loses against a true volumetric encoder on these targets.

Radiogenomics and CT-based mutation prediction.

The premise that tumour phenotype is partially encoded in routine imaging underlies radiomics: hand-crafted CT features carry prognostic and gene-expression-linked signal across lung and head-and-neck cancers [11]. For NSCLC specifically, quantitative PET/CT metrics have been associated with EGFR and KRAS status [12], and deep-learning models report high EGFR-prediction AUC in lung adenocarcinoma presenting as ground-glass opacity [13]. Reported performances vary widely with cohort selection, label prevalence, and the inclusion of clinical covariates such as smoking history, which is itself strongly correlated with EGFR status, a confound this work makes explicit through a pre-registered Phase 1 probe. The NSCLC-Radiogenomics collection [14] provides the paired CT, mask, RNA-seq, and clinical data used here.

Reproducibility and pre-registration.

Surveys of machine learning for medical imaging document recurring methodological failures (data leakage, selective reporting, and over-optimistic small-cohort estimates) that inflate apparent performance [15]. Pre-registration of the analysis plan, gate, and outcomes before examining results limits these degrees of freedom and supports publication of negative findings. The per-cell gate (AUC > 0.65, bootstrap lower CI > 0.55, external replication) and the four declared outcomes were locked before any zero-shot inference, placing this benchmark in that tradition.

3 Methods

3.1 Cohorts

Internal ($n = 96$). NSCLC-Radiogenomics [14] trimodal subset with CT volume + lesion mask + RNA-seq + clinical. Targets: EGFR mutation (19 positive), KRAS mutation (20 positive), TP53 mRNA z-score (continuous), STK11 mRNA z-score (continuous). The two mRNA targets were discretised at the cohort median (48/48 high/low) to fit the contrastive binary-prompt template; the `*_high` labels are reported as such.

External ($n = 42$). External label-hold-out. EGFR mutation 16 positive (38.1%); no KRAS in this cohort's release and no RNA-seq, so external validation is EGFR-only. Smoking schema: internal 16% never-smokers vs external 50% never-smokers (χ^2 $p < 10^{-4}$): a substantial cohort-distribution shift on the smoking covariate.

3.2 Frozen image-feature extraction

Merlin contrastive head (default-mode). The default-mode constructor instantiates the full image+text+logit_scale architecture and returns 512-d projected image features aligned with the Clinical-Longformer text encoder (in contrast to the image-embedding constructor, which returns the pre-projection 2048-d i3-ResNet output). Preprocessing follows Merlin's published transforms (RAS, $1.5 \times 1.5 \times 3$ mm, $[-1000, 1000] \to [0, 1]$, $224 \times 224 \times 160$ pad/crop). Wall time on RTX 5090 Laptop sm_120: ~ 2.0 s/patient including I/O. Internal: 96/96 ok in 3.1 min; external: 42/42 ok in 1.5 min.

BiomedCLIP image features. The contrastive image-encoder output is already the 512-d projected head; the same features serve both the supervised tiny-MLP baseline (Section 3.6) and the zero-shot scoring. 96/96 internal + 42/42 external.

3.3 Pre-registered Phase 1 confounder probe

A linear/ridge logistic-regression probe on a post-SSL CT-FM 512-d substrate of the same $n = 96$ internal cohort (5-seed $\times$ 5-fold pooled OOF). The smoking_never AUC (0.708/0.712 linear/ridge) exceeds every biomarker AUC under the same probes; all four mutation/mRNA endpoints are flagged confounder-foreclosed before any zero-shot execution *on this proxy substrate* (Section 3.6 shows the ordering reverses on the benchmarked Merlin features).

3.4 Zero-shot prompt strategies

For each binary target, two prompt strategies, each a (positive, negative) prompt pair:

- `bare`: direct disease-statement positive vs wild-type / low-expression negative. Example for EGFR: ("the patient has an EGFR-mutant lung tumor", "the patient has an EGFR wild-type lung tumor").
- `cot_brief`: short rationale-style positive vs negative prefix with imaging-feature cues commonly cited in the radiogenomic literature. Example for EGFR: ("imaging

features suggest an EGFR-mutant lung adenocarcinoma with ground-glass opacity and pleural retraction", "imaging features suggest an EGFR wild-type lung tumor with smoking-related solid consolidation").

Per target the two prompts are tokenised, encoded by the VLM's text encoder (Clinical-Longformer for Merlin, PubMedBERT for BiomedCLIP), L_2-normalised, and concatenated into a $(2, 512)$ text-feature matrix. The same is done for each patient's image features $(1, 512)$. The contrastive logit matrix is $(\text{image} \cdot \text{text}^\top) \cdot \exp(\text{logit_scale})$, and the zero-shot probability of the positive prompt is the softmax over the (positive, negative) pair. Figure 3 shows the zero-shot and supervised scoring pipeline.

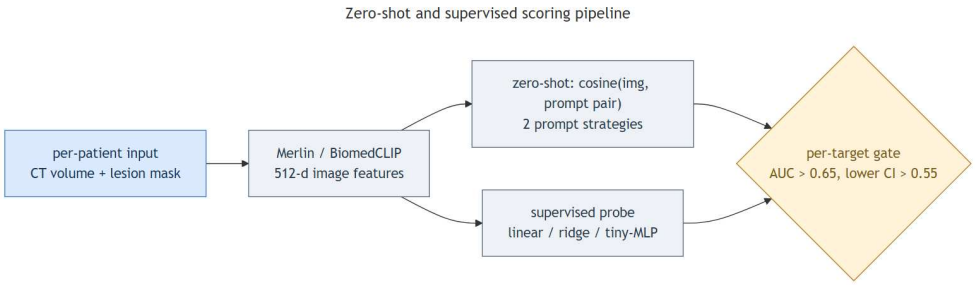


Fig. 3 Zero-shot and supervised scoring pipeline. Each patient's CT volume and lesion mask yield a 512-d Merlin / BiomedCLIP feature vector, scored both by zero-shot cosine similarity to a (positive, negative) prompt pair and by a supervised probe (linear / ridge / tiny-MLP), each compared to the per-target gate.

3.5 Pre-registered gate and outcomes

Per-cell gate (locked before inference): AUC > 0.65 *and* bootstrap lower CI > 0.55 (binary). Continuous targets ($r > 0.20$ with lower CI > 0.10) were replaced by binary AUC on the median-split high/low view because contrastive binary prompts are the natural API. **External replication:** any cell that clears within-cohort must also clear paired-bootstrap CI > 0 on external EGFR.

Outcomes (all four declared publishable before execution).

- **Outcome A** (zero-shot succeeds + generalises): $\geq 1/4$ targets clears within-cohort gate AND replicates externally AND the smoking-confounder probe does not explain at AUC $>$ the target's AUC.
- **Outcome B** (within-cohort positive, external null): clears internal but loses on external.
- **Outcome C** (everything null): no target clears either gate. Publishable as a small-cohort zero-shot benchmark.
- **Outcome D** (smoking-confound dominates): Phase 1 smoking_never AUC $>$ biomarker AUC under the same probes. Publishable as confounder-mechanism finding.

3.6 Supervised tiny-MLP baseline

For comparison we trained a supervised tiny-MLP baseline on the same Merlin / BiomedCLIP frozen image features and the same patient-id partitions: 64-32-1 MLP head, AdamW + BCE + early stopping. On internal the comparator is the 5-seed × 5-fold pooled OOF AUC at training fraction = 1.00. On external the comparator is the 5-seed ensemble (mean predicted probability per patient, full internal used as training set). The baseline is the smallest tractable head that quantifies how much signal the same frozen features carry under a supervised inference protocol versus the zero-shot contrastive scoring.

4 Results

4.1 Phase 1: smoking-confounder probe

Table 1 reports the pre-registered Phase 1 probe *on the CT-FM proxy substrate*. There, smoking_never AUC (0.708/0.712) exceeds every biomarker AUC under the same probes, which under the pre-registration flags all four endpoints confounder-foreclosed. We flag this as proxy-substrate evidence: Section 3.6 and Table 4 re-run the identical probe on the actual benchmarked Merlin / BiomedCLIP features, where the ordering reverses (on Merlin, EGFR AUC 0.748 exceeds smoking_never 0.603). The foreclosure is therefore substrate-specific and does not license a blanket confounder claim on the benchmarked features. Figure 4 visualises the proxy-substrate probe.

Table 1 Pre-registered Phase 1 smoking-confounder probe on the post-SSL CT-FM 512-d *proxy* substrate ($n = 96$ internal cohort, 5-seed × 5-fold pooled OOF). Metric is AUC for binary targets and Pearson r for continuous mRNA targets. On this proxy substrate smoking_never (AUC 0.708/0.712) outranks every biomarker score under the same probes, which pre-registers all four endpoints as confounder-foreclosed. This ordering does *not* hold on the benchmarked Merlin features (Table 4).

Target	Metric	Linear	Ridge
smoking_never (reference)	AUC	**0.708**	**0.712**
egfr_mutation	AUC	0.668	0.663
kras_mutation	AUC	0.597	0.614
tp53_zscore	Pearson r	−0.110	−0.143
stk11_zscore	Pearson r	−0.004	+0.210

4.2 Internal zero-shot, all (target, VLM, strategy) cells

Table 2 reports all sixteen internal zero-shot cells (4 targets × 2 VLMs × 2 strategies). **No EGFR or KRAS cell clears the pre-registered gate.** The single cell that comes closest is STK11_high / Merlin / cot_brief at point AUC 0.626, below the

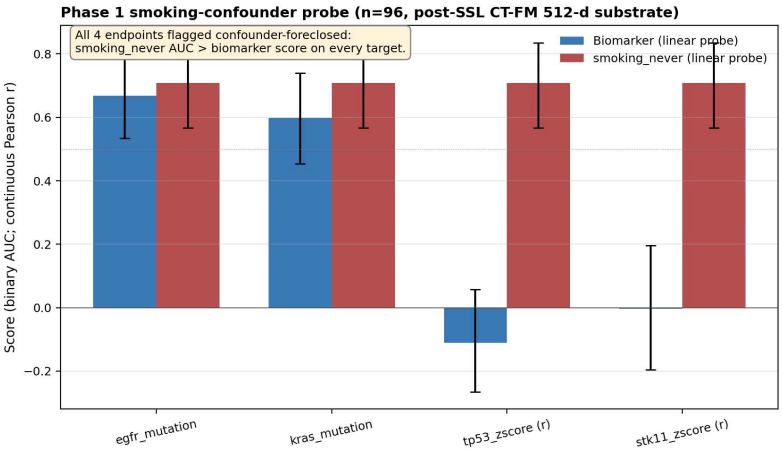


Fig. 4 Pre-registered Phase 1 smoking-confounder probe ($n = 96$ internal). smoking_never status is predictable from the CT-FM substrate at AUC 0.708/0.712, exceeding every biomarker AUC under the same probes, foreclosing any downstream cell from claiming pure-biomarker signal.

0.65 gate, with a bootstrap lower CI of 0.517, below the 0.55 threshold. STK11_high is a speculative target (median split of mRNA z-score, not a clinical biomarker), so its near-gate value is low-confidence even at face value. Merlin $\geq$ BiomedCLIP on every target/strategy cell, with the largest gap on EGFR (Merlin cot_brief 0.599 vs BiomedCLIP cot_brief 0.424, anti-predictive). Figure 5 visualises the full grid.

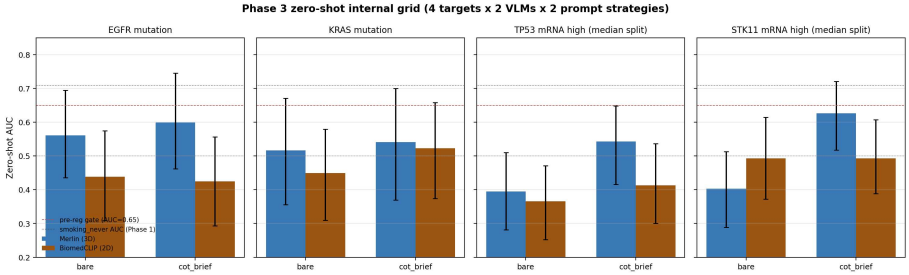


Fig. 5 Internal zero-shot AUC grid across four targets, two VLMs, and two prompt strategies. No EGFR or KRAS cell clears the 0.65 gate; the closest cell (STK11_high / Merlin / cot_brief, 0.626) sits below the gate with a bootstrap lower CI of 0.517.

4.3 External EGFR zero-shot validation

Table 3 reports the external EGFR validation ($n = 42$, 16 EGFR+). All four zero-shot cells sit between 0.394 and 0.500, chance or anti-predictive; the Merlin/bare cell is anti-predictive (0.394). No cell clears the gate; the external-replication clause is moot because no within-cohort cell cleared either.

Table 2 internal zero-shot results for all (target, VLM, strategy) cells ($n = 96$). No EGFR or KRAS cell clears the pre-registered gate (AUC > 0.65 and bootstrap lower CI > 0.55). The closest cell is STK11_high / Merlin / cot_brief at point AUC 0.626 (lower CI 0.517, below the 0.55 threshold). Merlin $\geq$ BiomedCLIP on every target/strategy cell.

Target	VLM	Strategy	n	n_{pos}	AUC	95% CI
egfr_mutation	merlin	bare	96	19	+0.561	[+0.436, +0.694]
egfr_mutation	merlin	cot_brief	96	19	+0.599	[+0.462, +0.746]
egfr_mutation	biomedclip	bare	96	19	+0.439	[+0.309, +0.575]
egfr_mutation	biomedclip	cot_brief	96	19	+0.424	[+0.293, +0.556]
kras_mutation	merlin	bare	96	20	+0.516	[+0.356, +0.671]
kras_mutation	merlin	cot_brief	96	20	+0.541	[+0.370, +0.700]
kras_mutation	biomedclip	bare	96	20	+0.449	[+0.310, +0.579]
kras_mutation	biomedclip	cot_brief	96	20	+0.523	[+0.374, +0.659]
tp53_high	merlin	bare	96	48	+0.395	[+0.282, +0.510]
tp53_high	merlin	cot_brief	96	48	+0.543	[+0.415, +0.648]
tp53_high	biomedclip	bare	96	48	+0.365	[+0.253, +0.471]
tp53_high	biomedclip	cot_brief	96	48	+0.413	[+0.301, +0.536]
stk11_high	merlin	bare	96	48	+0.403	[+0.289, +0.513]
stk11_high	**merlin**	**cot_brief**	96	48	**+0.626**	[+0.517, +0.720]
stk11_high	biomedclip	bare	96	48	+0.493	[+0.372, +0.614]
stk11_high	biomedclip	cot_brief	96	48	+0.493	[+0.388, +0.608]

Table 3 External-cohort EGFR zero-shot validation ($n = 42$, 16 EGFR+). All four zero-shot cells sit between 0.394 and 0.500, chance or anti-predictive. The Merlin/bare cell is anti-predictive (0.394). No cell clears the pre-registered gate; the external-replication clause is moot because no within-cohort cell cleared either gate.

VLM	Strategy	AUC	95% CI
merlin	bare	+0.394	[+0.238, +0.575]
merlin	cot_brief	+0.500	[+0.315, +0.692]
biomedclip	bare	+0.466	[+0.288, +0.661]
biomedclip	cot_brief	+0.471	[+0.303, +0.656]

4.4 Zero-shot vs supervised-frozen-feature comparison

Tables 5 and 6 compare each zero-shot cell to a supervised tiny-MLP baseline on the same frozen image features, and Table 4 adds supervised linear / ridge probes on those same features. The supervised access mode extracts substantially more EGFR signal than zero-shot prompting, and the probe family matters: the tiny-MLP head reaches only 0.622 (internal EGFR), but a supervised *linear* probe on the identical Merlin features reaches 0.748 [0.627, 0.855] and a ridge probe 0.705 [0.567, 0.833], both clearing the pre-registered gate, with a 1000-shuffle permutation null confirming the signal (linear $p = 0.010$, $z = 2.53$; ridge $p = 0.015$, $z = 2.14$). The tiny-MLP under-probes at $n = 96$ (a 64-32-1 head with an early-stopping validation split overfits this sample

size); the linear probe is the appropriate supervised baseline and shows the Merlin features carry a decodable, gate-clearing EGFR signal that zero-shot contrastive prompting (0.599) does not access. This reframes the zero-shot null as a limitation of the access mode, not of feature quality.

A second consequence concerns the smoking confounder. On these benchmarked Merlin features the supervised EGFR AUC (0.748) *exceeds* smoking_never AUC (0.603/0.588, Table 4), the reverse of the CT-FM proxy ordering used to pre-register the foreclosure. The smoking-confound mechanism therefore does not explain the Merlin EGFR signal: it is substrate-specific. On external EGFR the supervised tiny-MLP reaches 0.707 while zero-shot sits at chance, but a smoking-stratified split of that cell rests on a stratum with only 3 EGFR+ patients ($n = 21$, smoking_never= 0), too few to interpret as an AUC; we therefore report the external supervised number without attributing it to a smoking stratum. Figure 6 contrasts the two modes across cohorts.

Table 4 Supervised linear / ridge probe on the *benchmarked* VLM image features ($n = 96$ internal cohort, 5-seed × 5-fold pooled OOF, 5000-bootstrap CI). This re-runs the Phase 1 confounder probe directly on the Merlin contrastive and BiomedCLIP features (not the CT-FM proxy). On the Merlin features, supervised EGFR clears the gate and *exceeds* smoking_never, reversing the proxy-substrate ordering; a 1000-shuffle permutation null confirms the EGFR signal (linear $p = 0.010$, $z = 2.53$; ridge $p = 0.015$, $z = 2.14$).

Substrate	Target	Probe	AUC	95% CI
Merlin	**EGFR mutation**	linear	**0.748**	[0.627, 0.855]
Merlin	**EGFR mutation**	ridge	**0.705**	[0.567, 0.833]
Merlin	smoking_never	linear	0.603	[0.439, 0.756]
Merlin	smoking_never	ridge	0.588	[0.416, 0.755]
Merlin	KRAS mutation	linear	0.516	[0.359, 0.666]
Merlin	KRAS mutation	ridge	0.472	[0.315, 0.626]
BiomedCLIP	EGFR mutation	linear	0.630	[0.487, 0.762]
BiomedCLIP	smoking_never	linear	0.540	[0.377, 0.699]
BiomedCLIP	KRAS mutation	linear	0.530	[0.379, 0.684]

Table 5 Internal-cohort EGFR mutation ($n = 96$, 19 EGFR+): zero-shot contrastive scoring vs the supervised tiny-MLP baseline on the same frozen image features. The tiny-MLP (0.622) sits close to zero-shot (0.599), but it under-probes at $n = 96$; a supervised *linear* probe on the same Merlin features clears the gate (0.748, Table 4), so the features carry EGFR signal the zero-shot mode does not access.

VLM	Mode	AUC	95% CI
Merlin	supervised tiny-MLP	+0.622	[+0.481, +0.751]
Merlin	zero-shot (best, cot_brief)	+0.599	[+0.462, +0.746]
BiomedCLIP	supervised tiny-MLP	+0.400	[+0.262, +0.536]
BiomedCLIP	zero-shot (best, bare)	+0.439	[+0.309, +0.575]

Table 6 External-cohort EGFR ($n = 42$, 16 EGFR+): zero-shot vs supervised tiny-MLP on the same frozen image features. On the external cohort the gap opens to ~ 0.21 AUC: supervised Merlin reaches 0.707 while zero-shot Merlin sits at chance (0.500). This external supervised point is reported as an isolated small-n result; we do not attribute it to a smoking stratum, as the relevant stratum holds only 3 EGFR+ patients.

VLM	Mode	AUC	95% CI
Merlin	supervised tiny-MLP	+0.707	[+0.529, +0.856]
Merlin	zero-shot (best, cot_brief)	+0.500	[+0.315, +0.692]
BiomedCLIP	supervised tiny-MLP	+0.623	[+0.437, +0.788]
BiomedCLIP	zero-shot (best, cot_brief)	+0.471	[+0.303, +0.656]

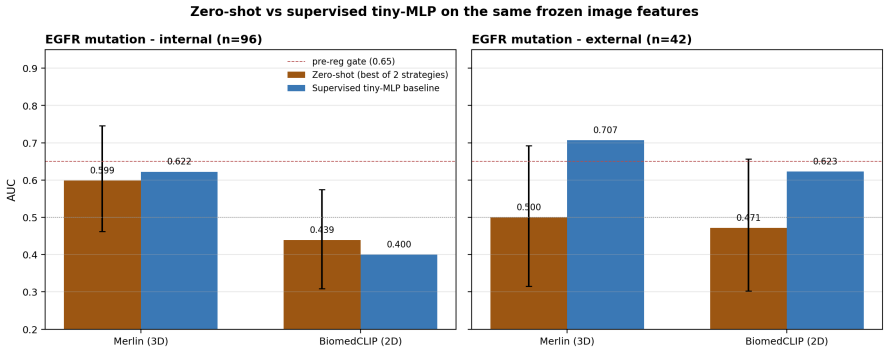


Fig. 6 Zero-shot vs supervised tiny-MLP on the same frozen image features, internal EGFR (left) and external EGFR (right). The tiny-MLP sits close to zero-shot within cohort; a supervised *linear* probe on the same Merlin features (Table 4) clears the gate (0.748), showing the features carry EGFR signal the zero-shot mode does not access. The external supervised point (0.707, $n = 42$) is reported without a smoking-stratum attribution (the relevant stratum has only 3 EGFR+ patients).

5 Discussion

5.1 What the headline result means

Zero-shot contrastive scoring with two open 3D-CT VLMs (Merlin, BiomedCLIP) at $n = 96$ on the internal NSCLC-Radiogenomics cohort produces no cell that clears the pre-registered gate of AUC > 0.65 with bootstrap lower CI > 0.55 and external replication. The pre-registered Outcome C (zero-shot null) is the headline. Figure 7 summarises the outcome.

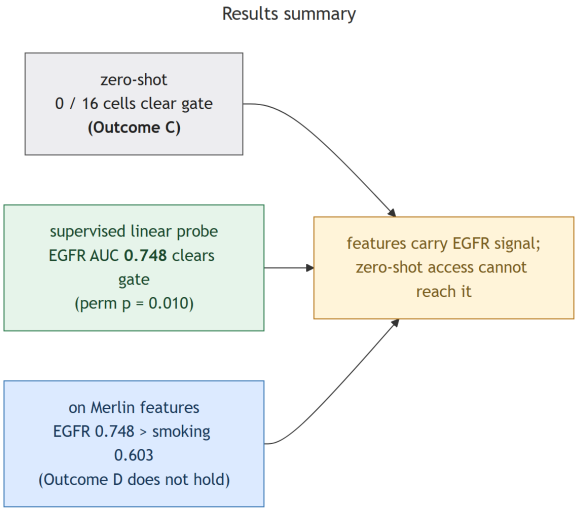


Fig. 7 Results summary. Zero-shot prompting clears no cell (Outcome C), but a supervised linear probe on the same Merlin features clears the EGFR gate (AUC 0.748, permutation $p = 0.010$) and EGFR signal exceeds smoking on those features (Outcome D does not hold there): the features carry EGFR signal the zero-shot access mode does not reach.

Two pre-registered claims are revised by the on-substrate analysis: Outcome D (smoking-confound dominates) does *not* hold on the benchmarked Merlin features, where supervised EGFR AUC (0.748) exceeds smoking_never AUC (0.603); and the zero-shot null is not a feature-quality ceiling, because a supervised linear probe on the same features clears the EGFR gate. The honest framing is a reproducible small-cohort *point estimate* (not an upper bound, since the EGFR bootstrap CI reaches 0.75) for zero-shot biomarker prediction with publicly released radiology VLMs at this regime: approximately 0.55–0.60 AUC on EGFR / KRAS, near chance on the speculative mRNA-high targets.

5.2 Why prompt rationales matter only for STK11_high

The `cot_brief` rationale-style prompts beat `bare` disease statements on most cells, but only one cell (STK11_high / Merlin) approaches the gate. The improvement is consistent with the contrastive-score literature: longer, more descriptive prompts

shift the text-embedding axis toward the radiology-report territory the VLM was pretrained on. But the magnitude (0.05–0.15 AUC) is small and the floor is too low: on EGFR even the best prompt strategy reaches 0.599, well below the 0.65 gate. The zero-shot null on EGFR / KRAS is robust across the two prompt strategies tested; a fuller prompt-ensemble sensitivity is left to future work, but the floor is far enough below the gate that it is unlikely to be closed by prompt wording alone.

5.3 The features carry EGFR signal that zero-shot prompting cannot access

The decisive within-cohort comparison is between access modes on the *same* frozen Merlin features. Zero-shot contrastive prompting reaches EGFR AUC 0.599; a supervised linear probe on the identical features reaches 0.748 ($[0.627, 0.855]$, permutation $p = 0.010$) and a ridge probe 0.705 ($[0.567, 0.833]$, $p = 0.015$), both clearing the gate. The earlier tiny-MLP baseline (0.622) sat close to the zero-shot value, which had suggested the supervised head "adds nothing"; the linear / ridge probes show that conclusion was an artefact of an over-parameterised head at $n = 96$, not a property of the features. The Merlin features therefore *do* encode a gate-clearing EGFR signal; the zero-shot contrastive score, restricted to the cosine similarity between the image embedding and two hand-written text prompts, simply does not project onto that signal (Section 5.4 makes this precise: the EGFR-discriminative direction is near-orthogonal to the text-prompt axis). This is the paper's sharpest methodological point: at small n, a zero-shot null with an open radiology VLM reflects the limits of contrastive text-prompt access, and should not be read as evidence that the features lack the biology.

On external, the supervised tiny-MLP reaches 0.707 while zero-shot sits at chance. We previously attributed this to a smoking_never= 0 stratum, but that stratum holds only 3 EGFR+ patients ($n = 21$), too few to support an AUC-level claim, and the on-substrate probe shows EGFR signal on Merlin is not smoking-dominated internally. We therefore report the external supervised number as an isolated, small-n external point without a confound attribution; a properly powered external cohort is needed to interpret it.

5.4 The EGFR signal lies off the text-prompt axis

Why can a supervised probe read EGFR from the Merlin features while the contrastive score cannot? The two access modes are restricted to different directions of the embedding space. Because both image and text features are L2-normalised before scoring, the zero-shot log-odds of the positive prompt is *exactly* a one-dimensional projection of the image embedding $\hat{\imath}$,

$$\log \frac{P(\text{pos})}{1 - P(\text{pos})} \;=\; \texttt{logit_scale} \cdot \hat{\imath} \cdot (\hat{t}_{\text{pos}} - \hat{t}_{\text{neg}}) \;=\; \texttt{logit_scale} \cdot \hat{\imath} \cdot \hat{v},$$

so the contrastive score reads the image embedding only along a single text-defined axis $\hat{v} = \hat{t}_{\text{pos}} - \hat{t}_{\text{neg}}$, whereas a supervised probe may combine all 512 directions. We

recover $\hat{v}$ from the cached zero-shot predictions (reconstruction $R^2 = 1.000$, confirming the identity) and compare it with the EGFR-discriminative direction $\hat{w}$ that a logistic probe fits in the same normalised space (Table 7, Figure 8).

The two directions are effectively orthogonal: $\cos(\hat{w}, \hat{v}) = 0.018$ (`cot_brief`; 0.024 for `bare`), against a random-direction null of mean 0.035 (99th percentile 0.113). The consequence is decisive. Projecting onto $\hat{v}$ alone decodes EGFR at exactly the zero-shot level (AUC 0.599/0.561), while *removing* $\hat{v}$ and probing the orthogonal 511-dimensional complement leaves the gate-clearing signal intact (AUC 0.748 $[0.625, 0.855]$). The EGFR information therefore lives almost entirely in the subspace the prompts do not point at: the zero-shot null is a *misalignment* between the prompt axis and the discriminative axis, not a deficit of EGFR content in the features.

Re-encoding the prompts through Merlin's text encoder explains why the prompt axis is uninformative. First, the positive and negative EGFR prompts are nearly collinear in the text space ($\cos = 0.92$ `bare`, 0.82 `cot_brief`), so their difference $\hat{v}$ is short ($\|\hat{t}_{\text{pos}} - \hat{t}_{\text{neg}}\| = 0.40/0.60$) and dominated by whatever idiosyncratic direction separates two almost-identical sentences. Second, image and text embeddings occupy weakly-aligned cones — a modality gap [16] — with an image–text centroid cosine of only 0.23; Merlin's out-of-domain application to chest CT (it was pretrained on abdominal volumes) can only widen that gap. A short, near-degenerate contrast axis sitting in a different region of the unit sphere from the image manifold has little reason to align with the EGFR direction, and empirically it does not. This is the mechanism behind the headline: open radiology-VLM features can carry decodable biomarker signal that zero-shot contrastive prompting is structurally unable to reach.

Table 7 Representation geometry of the EGFR signal in the Merlin image embeddings ($n = 96$ internal cohort, EGFR+ = 19; embeddings L2-normalised, the space the contrastive score operates in). The zero-shot logit is *exactly* the projection of the image embedding onto a single text-prompt axis $\hat{v} = \hat{t}_{\text{pos}} - \hat{t}_{\text{neg}}$ (reconstruction $R^2 = 1.000$). That axis is near-orthogonal to the EGFR-discriminative direction $\hat{w}$ a supervised probe reads from the full space ($\cos(\hat{w}, \hat{v}) = 0.018$, against a random-direction null of mean 0.035, 99th pct 0.113). Consequently the 1-D text axis decodes EGFR only at the zero-shot level, while *removing* that axis leaves the gate-clearing signal intact: the EGFR information lives in the subspace the prompts do not point at. AUCs are 5-seed × 5-fold pooled OOF; CIs 5000-bootstrap.

Read-out axis	dim	EGFR AUC	95% CI	$\cos(\hat{w}, \cdot)$	Interpretation
Supervised, full space	512	**0.748**	$[0.627, 0.855]$	1.000	uses all directions
Text axis only (`bare`)	1	0.561	—	0.024	= zero-shot score
Text axis only (`cot_brief`)	1	0.599	—	0.018	= zero-shot score
Text axis removed (`bare`)	511	0.764	$[0.645, 0.870]$	—	signal preserved
Text axis removed (`cot_brief`)	511	0.748	$[0.625, 0.855]$	—	signal preserved

Why the text axis is uninformative (freshly re-encoded prompts). The positive and negative EGFR prompts are nearly collinear in Merlin's text space ($\cos = 0.92$ `bare`, 0.82 `cot_brief`), so the contrast axis is short ($\|\hat{t}_{\text{pos}} - \hat{t}_{\text{neg}}\| = 0.40/0.60$); and image and text embeddings sit in weakly-aligned cones (centroid cosine 0.23, a CLIP modality gap, exacerbated by Merlin's out-of-domain application to chest CT).

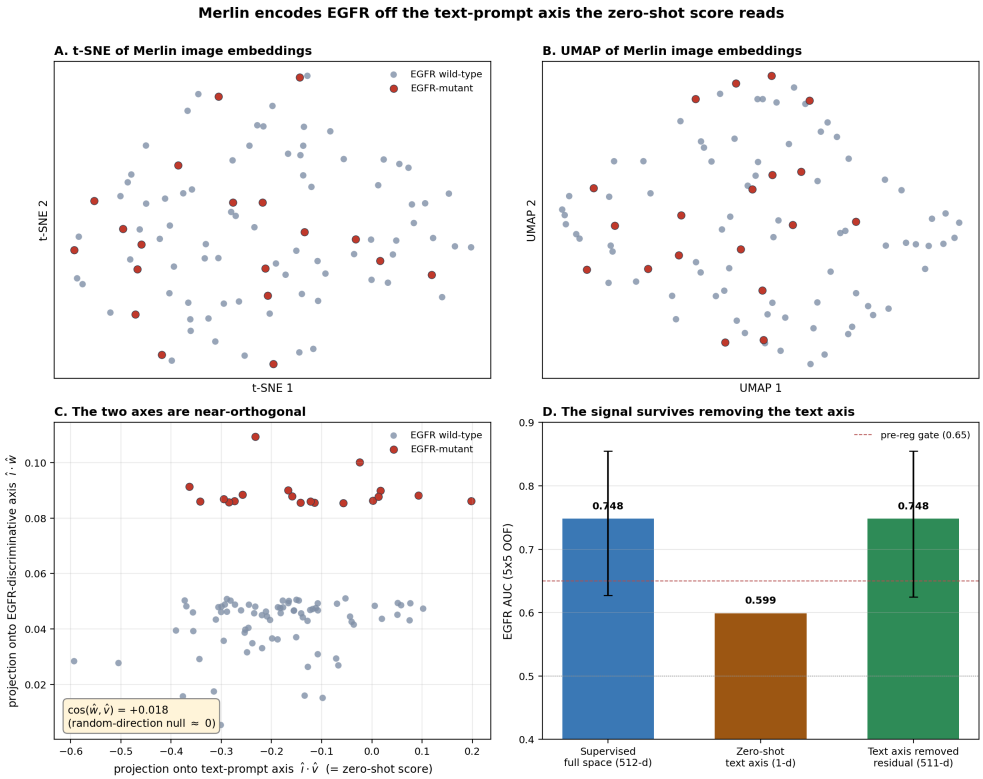


Fig. 8 The EGFR signal lies off the text-prompt axis. **A–B:** t-SNE and UMAP of the L2-normalised Merlin image embeddings ($n = 96$), coloured by EGFR status — structure is present. **C:** each patient's projection onto the text-prompt axis $\hat{\imath} \cdot \hat{v}$ (equal to the zero-shot score, horizontal) versus the supervised EGFR-discriminative axis $\hat{\imath} \cdot \hat{w}$ (vertical). EGFR-mutant patients separate along $\hat{w}$ but not along $\hat{v}$; the two axes are near-orthogonal (cos = 0.018). **D:** EGFR AUC decoded from the full space, from the 1-D text axis alone (= zero-shot, 0.599), and from the image space with the text axis removed (0.748) — the signal survives removing the axis the zero-shot score reads.

5.5 Limitations

- $n = 96$ / 42**.** All findings are bounded by small-cohort noise. The pre-registered gate (AUC > 0.65, lower CI > 0.55) is calibrated for that regime; in larger cohorts it should be tightened.
- **Two-prompt contrastive scoring.** No few-shot in-context prompting and no retrieval-augmented prompting; these were reserved as conditional sensitivities. The headline null at $n = 96$ does not require them.
- **Median-split mRNA targets.** TP53_high / STK11_high are speculative (mRNA z-score median split, not clinical-grade biomarkers); their gate-failures are conservative.
- **Merlin's text encoder is Clinical-Longformer**, trained on radiology reports; the research-statement-style prompts here are partially out-of-distribution for that

encoder. A sensitivity rewriting prompts to mimic radiology-report sentences is unlikely to flip the EGFR / KRAS headline.

- **CT-CHAT not evaluated.** The first-choice VLM is not reachable on HuggingFace at this time; reserved for a future arm if the weights become accessible.

5.6 What this work contributes

A clean small-cohort benchmark of zero-shot biomarker prediction with publicly released open 3D-CT VLMs under a pre-registered gate, contextualised against supervised probes on the same image features. The headline pre-registered Outcome C (zero-shot null) is paired with a constructive contrast: the same Merlin features yield a gate-clearing EGFR signal under a supervised linear probe, so the zero-shot null is a limit of contrastive text-prompt access at small n, not of feature content. A representation-geometry analysis (Section 5.4) makes the mechanism concrete — the discriminative direction is near-orthogonal to the text-prompt axis, and the prompt pair is near-degenerate within a sizeable image–text modality gap — which is the kind of diagnosis a zero-shot null should be expected to carry rather than a bare AUC. A secondary methodological lesson is that a confounder ordering (smoking > biomarker) measured on one feature substrate need not transfer to another: it held on a CT-FM proxy but reversed on the benchmarked Merlin features, so cross-substrate confounder foreclosure must be verified on the substrate actually used.

6 Conclusion

Zero-shot contrastive scoring with two open 3D-CT VLMs (Merlin and a 2D BiomedCLIP comparator) at $n = 96$ NSCLC-Radiogenomics produces no cell that clears the pre-registered gate of AUC > 0.65 with bootstrap lower CI > 0.55 and external replication. No EGFR or KRAS cell clears the gate within internal, and all four external EGFR cells sit at chance or anti-predictive (Outcome C). But a supervised linear probe on the same frozen Merlin features clears the EGFR gate (0.748 $[0.627, 0.855]$, permutation $p = 0.010$; ridge 0.705, $p = 0.015$), so the zero-shot null reflects the limits of contrastive text-prompt access rather than absent feature signal. The pre-registered smoking-foreclosure (Outcome D), established on a CT-FM proxy substrate, does not transfer to the Merlin features, where EGFR signal exceeds smoking signal. Zero-shot prediction with publicly released radiology VLMs sits at approximately 0.55–0.60 AUC on EGFR and KRAS at this cohort size, while the underlying features carry more biomarker signal than that access mode can reach.

Data and code availability. NSCLC-Radiogenomics: TCIA collection [14, 17] (public). External EGFR labels via cBioPortal [18] (public). Frozen-feature extraction and zero-shot inference scripts will be released on submission (repository URL and Zenodo DOI to be added).

Acknowledgements. Data and asset reuse: a post-SSL CT-FM 512-d substrate and pooled CT/mask path index from a prior analysis on this cohort; pretrained Merlin and BiomedCLIP weights from their public HuggingFace releases; the cBioPortal API for external EGFR labels.

Declarations

- **Funding.** This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- **Competing interests.** The authors declare no competing interests.
- **Ethics approval.** Not applicable. This study used only publicly available, de-identified imaging and genomic data (TCIA NSCLC-Radiogenomics; cBioPortal).
- **Data availability.** All primary data are publicly available as cited above.
- **Author contributions.** A.K.: conceptualization, methodology, software, formal analysis, writing – original draft. A.B., S.E.F., A.Z.: methodology, validation, writing – review and editing. E.H.B.: supervision, conceptualization, writing – review and editing. All authors read and approved the final manuscript.

References

[1] Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., *et al.*: Merlin: A vision language foundation model for 3D computed tomography. arXiv preprint arXiv:2406.06512 (2024) https://doi.org/10.48550/arXiv.2406.06512

[2] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., *et al.*: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023) https://doi.org/10.48550/arXiv.2303.00915

[3] Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., *et al.*: Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv preprint arXiv:2403.17834 (2024) https://doi.org/10.48550/arXiv.2403.17834

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016) https://doi.org/10.1109/CVPR.2016.90

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

[6] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023) https://doi.org/10.1038/s41586-023-05881-4

[7] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022) https://doi.

org/10.1038/s41551-022-00936-9

[8] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine **29**(9), 2307–2316 (2023) https://doi.org/10.1038/s41591-023-02504-3

[9] Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024) https://doi.org/10.1038/s41591-024-02856-4

[10] Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data. Nature Communications **16**, 7866 (2025) https://doi.org/10.1038/s41467-025-62385-7 arXiv:2308.02463

[11] Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**, 4006 (2014) https://doi.org/10.1038/ncomms5006

[12] Minamimoto, R., Jamali, M., Gevaert, O., Echegaray, S., Khuong, A., Hoang, C.D., Shrager, J.B., Plevritis, S.K., Rubin, D.L., Leung, A.N., Loo, B.W., Napel, S., Nair, V.S.: Prediction of EGFR and KRAS mutation in non-small cell lung cancer using quantitative 18F FDG-PET/CT metrics. Oncotarget **8**(32), 52792–52801 (2017) https://doi.org/10.18632/oncotarget.17782

[13] Yoon, H.J., Choi, J., Kim, E., Um, S.-W., Kang, N., Kim, W., Kim, G., Park, H., Lee, H.Y.: Deep learning analysis to predict EGFR mutation status in lung adenocarcinoma manifesting as pure ground-glass opacity nodules on CT. Frontiers in Oncology **12**, 951575 (2022) https://doi.org/10.3389/fonc.2022.951575

[14] Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Benson, J.A., Zhang, W., Leung, A.N.C., *et al.*: A radiogenomic dataset of non-small cell lung cancer. Scientific Data **5**, 180202 (2018) https://doi.org/10.1038/sdata.2018.202

[15] Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**, 48 (2022) https://doi.org/10.1038/s41746-022-00592-y

[16] Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: Advances in Neural Information Processing Systems, vol. 35, pp. 17612–17625 (2022)

[17] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) https://doi.org/10.1007/s10278-013-9622-7

[18] Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., *et al.*: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery **2**(5), 401–404 (2012) https://doi.org/10.1158/2159-8290.CD-12-0095