BEYOND AGGREGATE SCORES: REVEALING CONDITIONAL MODEL BEHAVIOR IN CHEST X-RAY NODULE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate model performance can rank systems without revealing the conditions under which their behavior diverges. We introduce PULSE, a controlled-response evaluation framework, and instantiate it in chest X-ray nodule detection. Analytic nodule-like probes vary in size, contrast, morphology and position on fixed chest radiographs, making fourteen detectors (seven families, two seeds each) directly comparable under matched conditions. The resulting behavioral atlas reveals sharply different behaviors behind similar conventional performance. At the largest tested probe size, increasing requested contrast drives Faster R-CNN toward saturation but reduces detection in both DINO seeds. Changing the intensity profile from a smooth disk to a peak-matched Gaussian also produces markedly different losses across models. These are structured differences: response-profile distances recur across discovery and held-out host banks (Spearman $\rho = 0.999$), and every model is closest to its other seed. An earlier prespecified six-model study confirms the coarse geometry and regional patterns. Measurement-design analyses show that recovering an ordering, finding disagreements and reconstructing a response field require different sampling strategies. A matched natural-lesion assay provides partial pointwise behavioral support. Controlled evaluation makes conditional model behavior visible, reproducible and testable, complementing aggregate performance.

1 INTRODUCTION

Aggregate benchmark scores can rank models while hiding the conditions under which they fail. In medical imaging, clinically important subsets can perform substantially worse than the overall benchmark suggests (Oakden-Rayner et al., 2020). Pulmonary nodules on chest radiographs illustrate the stakes: in one clinical cohort, missed nodular lung cancers were smaller and more often overlapped by superimposed structures than detected ones, and their diagnosis was substantially delayed (Quekel et al., 1999). Lung cancer was the leading cause of cancer death in GLOBOCAN 2022 (Bray et al., 2024).

Chest X-ray nodule detection therefore provides a high-stakes testbed for a broader evaluation problem: aggregate performance can rank models without revealing the conditions under which their behavior diverges. Faster R-CNN and DINO achieve numerically similar pFROC scores of 0.772 and 0.793, yet detect 0.406 and 0.077 of a common probe bank: a more than fivefold difference (Fig. 2). Aggregate metrics average over lesions, anatomical backgrounds and acquisition conditions, so more precise estimation of the average does not identify such patterns (Jacobs & Wallach, 2021; Wallach et al., 2025; Heineman et al., 2025).

Chest radiographs make this measurement controllable. Nodule-like size, contrast, morphology and position can be varied on the same host image, while the surrounding anatomy is held fixed. Anatomical structure is itself a major determinant of nodule detectability (Revesz et al., 1974; Samei et al., 1999); localized detector scores let us measure its interaction with the inserted signal. These axes are clinically motivated, but the analytic probes are not a model of the natural lesion distribution.

We call this framework PULSE (Programmable Unit-testing of Learned Systems through Experiments): a programmable approach to testing learned systems by intervening on controlled input

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

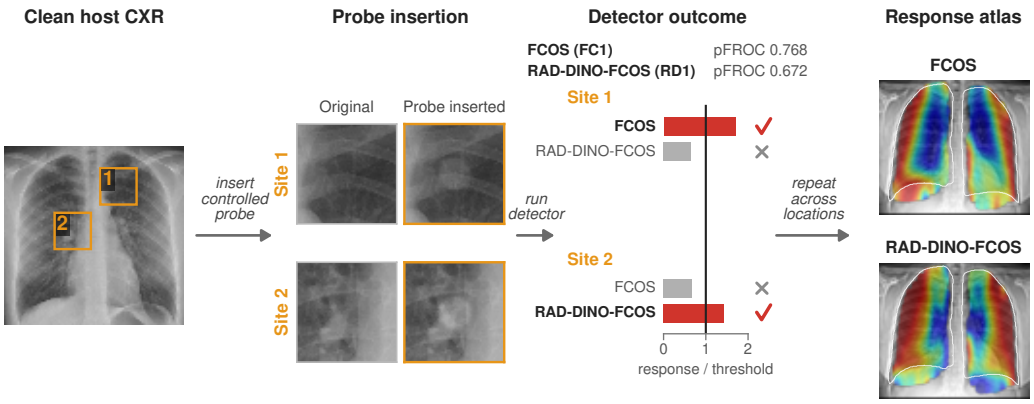


Figure 1: **The same controlled probe setting can reverse which detector succeeds when its location changes.** In this illustrative example, a clean public chest radiograph receives the same probe at site 1 or site 2 in separate trials. FCOS has the higher aggregate pFROC, yet crosses its operating threshold at site 1 but not site 2; RAD-DINO–FCOS shows the reverse pattern. Bars show each detector's localized response relative to its own paper operating threshold (Appendix Table 1); red indicates a threshold-crossing response. Extending the measurement across locations and hosts yields model-specific response maps; the rightmost panels show the paper's aggregate spatial atlas (discovery bank, medium probe; Fig. 11). Probe: smooth disk, diameter 8% of the reference width, requested CNR 1.25; detections also meet the 0.02 gain criterion.

conditions and measuring the resulting behavior. Here it yields a *behavioral atlas*: an aligned collection of conditional response measurements whose coordinates are properties of the intervention—position, size, contrast and profile—and whose entries are detector responses. Figure 1 illustrates the distinction between aggregate ranking and conditional behavior. FCOS has the higher conventional pFROC, yet the same controlled probe setting produces opposite FCOS/RAD-DINO–FCOS outcomes at two locations on one fixed radiograph. Extending the measurement across locations and hosts yields the response atlas analyzed below. We apply PULSE to fourteen detectors (seven families, two seeds each) on common host banks. Conditional response curves reveal specific differences: at the largest tested size, increasing contrast drives Faster R-CNN toward saturation but causes DINO's detection rate to fall (Fig. 3).

Response-profile geometry recurs across host sets, and every model remains nearest to its other seed. Ordering, disagreement discovery and response-field reconstruction favor different measurement strategies, while a matched natural-lesion assay provides partial pointwise support. Conditional response measurement therefore complements aggregate model evaluation.

2 Related Work

Behavioral evaluation. Measurement validity concerns what an evaluation captures, beyond the precision of its score (Jacobs & Wallach, 2021; Heineman et al., 2025). Behavioral test suites, controlled counterfactuals and controversial stimuli expose systematic differences hidden by aggregate accuracy (Ribeiro et al., 2020; Ruiz et al., 2022; Golan et al., 2020). In vision-language-action policies, disturbance-tolerance analysis provides a related example of behavior hidden by aggregate task success (Anonymous, 2026a). In medical imaging, hidden-stratification analyses find clinically meaningful failures in subsets of natural data (Oakden-Rayner et al., 2020); controlled interventions complement such subgroup evaluation by varying specified signal properties on fixed images. For object detection, Zone Evaluation studies spatial heterogeneity through partitions, maps and controlled placement (Zheng et al., 2024). PULSE extends this conditional view across position, size, contrast and morphology, then tests whether the resulting cross-model structure recurs across hosts and seeds. Controlled stimuli can reveal what evidence a model uses. Texture–shape cue conflicts, for example, distinguish recognition strategies that standard ImageNet accuracy leaves unresolved

(Geirhos et al., 2019). Our probes ask a related behavioral question in detection: how does the response vary along a specified signal axis on the same background, across trained systems?

Controlled medical-image interventions. Signal insertion supports augmentation (Pezeshk et al., 2017), while virtual imaging trials evaluate algorithmic readers under controlled conditions (Badano et al., 2018; Abadi et al., 2020; Tushar et al., 2025). Generative approaches such as RadEdit and RoentMod edit findings or acquisition characteristics to stress-test models (Pérez-García et al., 2024; Cooke et al., 2026). iTRIALSPACE composes real CT scans and lesion profiles into programmable virtual lesion trials that isolate lesion and host factors for lung CT models (Tushar et al., 2026). PULSE instead uses analytically parameterized probes to compare response geometry, readout dependence and measurement design across chest-radiograph detectors.

Target-dependent measurement. Active Testing allocates labels to estimate risk efficiently (Kossen et al., 2021), and Active Evaluation Acquisition predicts unqueried benchmark scores (Li et al., 2025). More broadly, work on evaluation design asks when a single aggregate score is insufficient and additional measurement is warranted (Anonymous, 2026b). Here the target can instead be an ordering, a local difference or a response field. Using fully measured response fields, we compare sparse estimation and adaptive acquisition for these targets.

3 Method: Controlled Response Measurement

3.1 Models, Hosts and Interventions

Seven detector families, two seeds each. The panel comprises Faster R-CNN (Ren et al., 2015), FCOS (Tian et al., 2019), RetinaNet (Lin et al., 2017), RT-DETR (Zhao et al., 2024), DINO with a ResNet-50 backbone (Zhang et al., 2023), DINOv2–FCOS (Oquab et al., 2023) and RAD-DINO–FCOS (Pérez-García et al., 2025), with two training seeds per family, denoted FR, FC, RN, RT, DN, DV and RD with suffixes 1 and 2 (e.g., FR1, FR2). All models were selected by conventional validation before their PULSE responses were examined. The DINOv2 and RAD-DINO models use frozen ViT-B/14 encoders and a common FCOS/FPN head, and every model keeps its native preprocessing and decoder, so family comparisons concern complete trained systems rather than isolated architectural components (Appendix A.2). RN2 and RT2 come from a later seed-replication extension (Appendix H).

We use 800 provider-labelled nodule-negative NODE21 radiographs (Sogancioglu et al., 2024; 2021): 500 discovery hosts and 300 held-out hosts for response evaluation, which were withheld from PULSE analysis. Every model receives identical source-image interventions. A public CC0 radiograph in Fig. 1 illustrates the measurement and contributes to no population estimate.

Parameterized probes. For host image I and condition z, the renderer forms $I_z = I + a\,p_z$ in the source intensity domain before source-range clipping, then applies each model's unchanged preprocessing. The primary probe p_z is a peak-one, smooth logistic disk. Its diameter is a fraction of lung-mask width, and its amplitude is $a = \mathrm{CNR}\,\hat{\sigma}$, where $\hat{\sigma}$ is the clean-image standard deviation in a local annulus. Morphology variants change the radial profile, eccentricity, contour or texture. Peak-matched and L_2-matched comparisons are distinguished explicitly (Appendix A.4).

The global bank has 400 conditions per host: 192 primary anchors (four sizes, 1.5–8% of lung width; four requested CNRs, 0.75–3; twelve positions), 80 morphology anchors and 128 Sobol conditions. Spatial perimetry places five size–contrast probes on a 17×17 lattice per hemithorax in 300 discovery and 300 held-out hosts.

3.2 Response Profiles and Their Summaries

Let $S^{+}_{m,h,z}$ and $S^{0}_{m,h,z}$ be the maximal nodule scores with and without the probe among boxes centered within 16 working-source pixels of its center. The primary detection readout (E0) is

$$Y_{m,h,z} = \mathbf{1}\left[S^{+}_{m,h,z} \geq \tau_m, \quad S^{+}_{m,h,z} - S^{0}_{m,h,z} \geq 0.02\right]. \tag{1}$$

Each threshold τ_m targets 0.25 false positives per image on a shared 562-image calibration set. The gain criterion requires a response above the clean background. Calibration fixes the operating point, not the scale of score increments across models.

Measuring the clean image separates the probe's effect from a high-scoring structure already present at that position. Continuous score changes retain information that E0 discards and distinguish a change in response strength from a threshold crossing.

The *response profile* records the output over intervention conditions. Its host-averaged form,

$$r_m^{(q)}(z) = \mathbb{E}_H[q(m, H, z)], \tag{2}$$

describes conditional behavior for a specified readout q. We summarize complementary aspects of this profile. *Response-Success Field Fraction* (RSFF) averages each host's fraction of valid conditions detected under E0, with equal host weight. *Behavioral phenotype distance* (BPD) compares the continuous changes $\Delta S = S^+ - S^0$: for each host, we standardize each model's aligned vector by its own mean and sample standard deviation, compute the RMS distance, then average across hosts. BPD therefore compares profile shape while RSFF retains coverage level. Appendix B gives definitions and inference details.

3.3 Comparisons and uncertainty

Conditional comparisons match hosts, positions and all unchanged probe parameters. Figure 3b,c shows one seed per family, and Appendix Fig. 7 all fourteen models, on the full geometry-valid discovery support. We repeat the contrast analysis on a fixed input-consistency screen (Appendix C). Requested CNR is the design axis, not a clinical conspicuity scale. We average within host before across hosts; reported 95% intervals use 2,000 paired host-bootstrap draws.

The original six-model study fixed six predictions before one-shot confirmation on the held-out hosts (Appendix D). The fourteen-model atlas is a subsequent descriptive analysis; its recurrence on the held-out bank tests stability rather than new prospective hypotheses.

The three paired comparisons change one intervention axis at a time. Increasing CNR preserves the host, location and probe profile; replacing the disk with a Gaussian preserves location, peak amplitude and nominal diameter; the regional comparison moves an otherwise identical probe between coordinate-defined regions. Each effect is a difference in host-level detection rates. The size sweep separately summarizes each size's geometry-valid placements at fixed CNR 1.25. Appendix Fig. 6 puts the paired effects on a common scale; the full maps retain the surrounding condition space (Appendix C).

Natural-lesion comparison. An exploratory assay uses 506 annotated NODE21 and VinDr-CXR evaluation images (Nguyen et al., 2022). A fixed CXR-adapted eraser estimates the lesion-removed background (ERASE); matched replacements reconstruct image-derived lesion properties on that host. Detector outputs do not select replacements. For the original six models, we compare the localized scores of the real lesion (S_R), erasure and replacement (S_M), using $|S_M - S_R|$ as response error (Appendix G).

4 A behavioral atlas of chest-radiograph nodule detectors

4.1 Similar performance, different response profiles

Conventional detection performance does not determine probe response coverage (Fig. 2). Faster R-CNN (FR1) and DINO (DN2) have similar pFROC, 0.772 and 0.793, yet their RSFF values are 0.406 and 0.077: a more than fivefold difference. Here pFROC summarizes conventional detection performance over 0.125–1.0 false positives per image.

DN1 has the highest conventional pFROC (0.835), yet covers 0.161 of the probe bank, compared with FR1's 0.406. Figure 2b shows where the difference lies: averaged over 300 hosts, FR1 responds across most of the lateral lung field, DN1 mainly near its lateral edge, and both lose response medially. On the representative radiograph, FR1 detects 137 and DN1 27 of 292 valid placements (Fig. 3a).

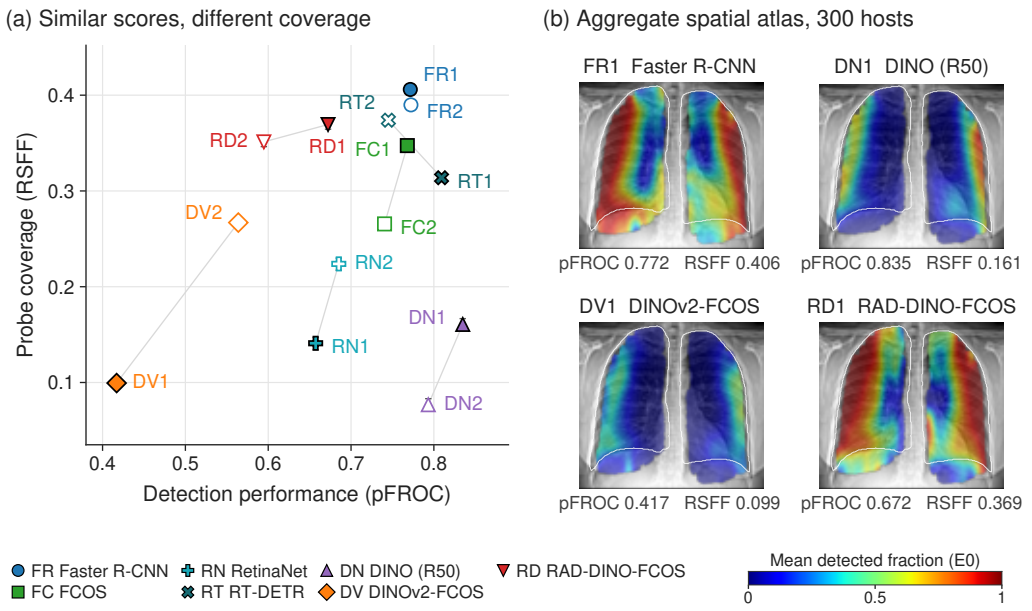


Figure 2: **Similar aggregate performance conceals different spatial response fields.** (a) pFROC on the shared 562-image calibration cohort versus equal-host probe coverage (RSFF) on the 500-host discovery bank, with 95% host-bootstrap intervals; thin lines join seeds. (b) Aggregate spatial atlas: mean E0 of the medium probe (4.75% of lung width, requested CNR 1.5) per grid cell over 300 discovery hosts, drawn on a representative NODE21 host (Appendix I). Jet colors show display-smoothed measured cell means.

Comparing the DINOv2- and RAD-DINO-based FCOS models gives a second example. DV1 and RD1 move from 0.417 to 0.672 in pFROC and from 0.099 to 0.369 in coverage; their aggregate maps differ mainly in level, and on the representative radiograph they detect 32 and 209 of 292 placements. For the second pair, the conventional scores are much closer (0.564 and 0.595), while coverage rises from 0.267 to 0.351. Thin seed-pair connectors show how much the response level changes between training runs.

Figures 3 and 4 vary requested contrast, size and intensity profile, pairing same-CXR responses with population curves for one seed per family; both seeds, paired effects and the full atlas are in Appendix Figs. 6–8, and Appendix I shows further response maps on study radiographs.

4.2 Higher requested contrast can reduce detection

Raising the requested CNR does not reliably increase detection (Fig. 3). On one radiograph (panel a), stronger and larger probes both enlarge the response field, at very different rates: at CNR 3, FR1 detects 243 of 292 placements and DN1 77. At this medium size DN1 still rises with CNR; the decline appears at the largest primary size, where DINO (DN1) falls from 0.401 at CNR 2 to 0.151 at CNR 3 on the full valid bank. DINOv2–FCOS (DV1) falls from 0.263 at CNR 1.25 to 0.037 at CNR 3. Faster R-CNN instead approaches saturation, reaching detection rates of 0.987 and 0.950 at CNR 3.

Both seeds of DINO and DINOv2–FCOS peak at an intermediate tested CNR and then decline; RAD-DINO–FCOS shows the same broad shape from a higher level (Appendix Fig. 7). RetinaNet's seeds differ substantially in level, while RT-DETR's second seed approaches a plateau.

The decline also survives matching effective-input contrast across levels. On this screened support, DN1 falls from 0.662 to 0.312, a paired change of -0.350 with 95% interval $[-0.375, -0.325]$. Negative intervals recur on the held-out host bank and on full valid support for both seeds of DN, DV and RD and for RN1 and RT1; the decline is weaker for RN2 and essentially absent for RT2 (Tables 2 and 7).

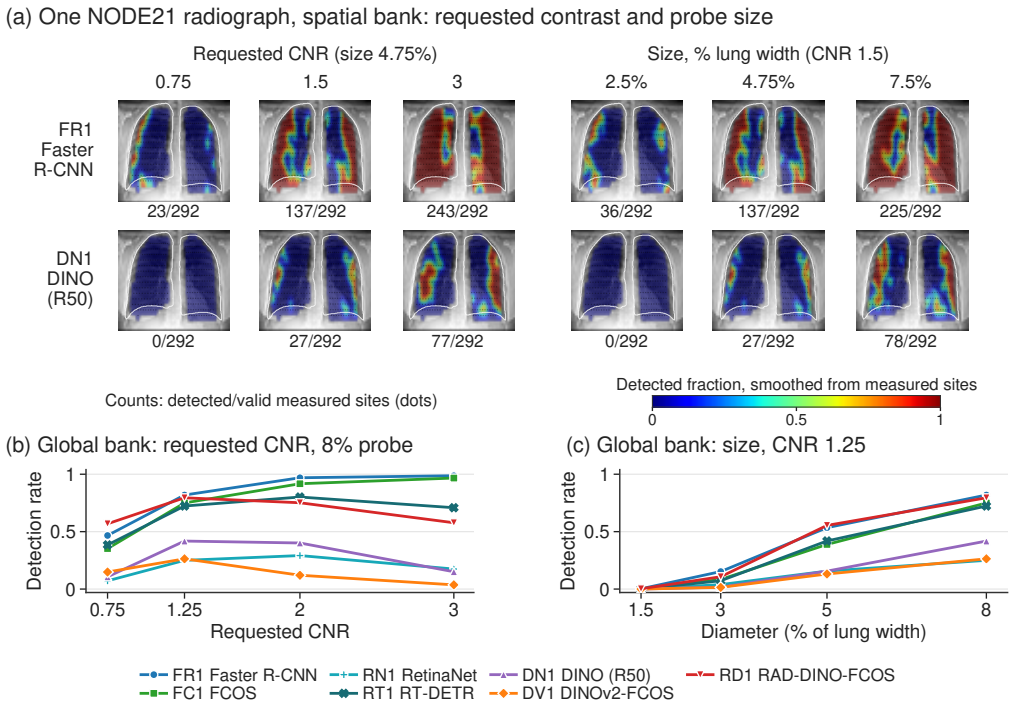


Figure 3: **Size and requested contrast separate detectors; at the largest size, higher contrast can suppress detection.** (a) Spatial-bank illustration on the radiograph of Fig. 2b: FR1 and DN1 E0 outcomes at the measured sites (dots) for three requested CNRs at size 4.75% and three sizes at CNR 1.5, Jet-smoothed for display; counts are detected/valid sites. (b, c) Global-bank population curves, the quantitative result: equal-host mean detection on each condition's valid support (500 discovery hosts) for the seed-1 model of every family, versus (b) requested CNR at size 8% and (c) size at CNR 1.25. All fourteen: Appendix Fig. 7.

The decline is visible in localized scores. For the eight models with negative CNR contrasts in Table 2A, a box remains near the probe in approximately 93% or more of discovery cells at CNR 3, while DN1's median localized score drops from 0.666 to 0.004. The phenomenon is therefore present in the underlying score, not merely in thresholded detection.

4.3 Size separates models at the same contrast

At fixed CNR 1.25, detection increases across the four tested sizes for every model on size-specific geometry-valid supports (Fig. 3c), but the response levels remain far apart. At a diameter of 5% of lung width, FR1 and RD1 reach 0.535 and 0.554, while DN2 reaches 0.058 (Appendix Fig. 7). Even at 8%, FR1 detects 0.819 of valid placements compared with DN2's 0.128. Increasing the spatial extent of the signal thus separates these models differently from increasing its contrast. The full size–contrast atlas (Appendix Fig. 8) locates both sweeps in the measured condition space.

4.4 Intensity profile further separates detectors

Replacing the smooth disk with a Gaussian of equal peak and matched diameter lowers coverage in all fourteen models (Fig. 4c and Appendix Fig. 6b; RN2 and RT2 values in Table 7). The reduction ranges from 0.012 to 0.233. The ten ResNet-50-based detectors lose 43–87% of matched coverage, compared with 19% and 23% for the two RAD-DINO models. DV1's small absolute change occurs at low coverage for both profiles (smooth-disk coverage 0.141).

The profile comparison asks which distribution of intensity supports detection. The smooth disk has a relatively uniform interior and a soft boundary; the Gaussian redistributes intensity across the center and surrounding tail (Fig. 4a). This changes radial structure and integrated energy together.

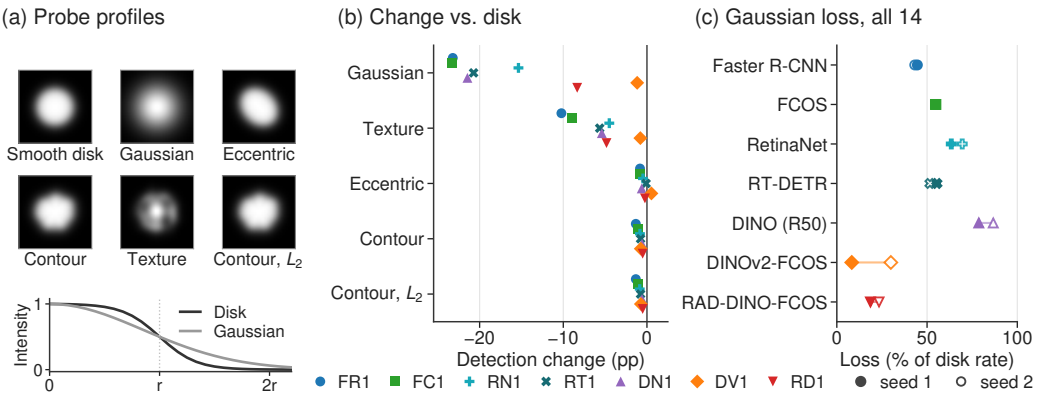


Figure 4: **Intensity profile separates detectors more than the tested geometric perturbations.** (a) Display-only renderer templates on a blank field (not radiographs; illustrative texture seed). The smooth disk is the primary probe; the Gaussian comparison variant shares its peak and FWHM diameter but has wider tails, changing radial profile and integrated energy together; the last contour variant matches the disk's L_2 energy. (b) Change in E0 detection from the disk at size 5%, averaged over four requested CNRs (500 discovery hosts), for the seed-1 model of each family. (c) Gaussian loss as a percentage of each model's disk rate, all fourteen (open: seed 2).

In the twelve-model morphology panel, eccentricity and contour changes alter coverage by at most 0.013, while texture changes produce intermediate losses (Fig. 4b). The atlas thus separates the strong sensitivity to intensity profile from the smaller effects of the tested geometric perturbations.

4.5 Moving the same probe reveals a shared regional deficit

The spatial comparison moves the probe while preserving its size and requested contrast (Figs. 2b and 6c; both banks: Appendix Fig. 11). In normalized hemithorax coordinates, u runs from medial to lateral and v from superior to inferior. We compare a medial, mid-height partition ($u \leq 0.35$, $0.35 \leq v < 0.65$) with lower-visible positions ($v \geq 0.65$). These coordinate-defined regions give the same comparison on every host.

Faster R-CNN seed 1 detects the medium probe with rates 0.305 and 0.536 in the central and lower regions, respectively: a paired deficit of 0.231. Across all fourteen models, deficits range from 0.077 to 0.355, and every model retains an interval below zero on the second host bank (Tables 3 and 7). Thus, models with otherwise different contrast and profile responses share a reproducible regional vulnerability. Because requested CNR is locally scaled, this result concerns the probe against its host background, not a fixed-amplitude patch pasted into different locations.

4.6 Response geometry recurs across hosts and seeds

The profile differences form a reproducible geometry (Fig. 5). Across all 91 model pairs, BPD on discovery and held-out hosts has Spearman correlation 0.999 and mean absolute change 0.005. Between-family distances exceed within-seed-pair distances by 0.326 $[0.322, 0.329]$. Every model is closest to its other seed, on both global and both spatial host banks. Thus all seven families retain a recognizable response profile across training runs and host sets.

Similar profile shape does not require similar coverage. The two DINOv2–FCOS seeds remain nearest neighbors in standardized response space, yet their global coverage is 0.099 and 0.267. BPD identifies a recurring pattern across conditions, while E0 coverage records how often responses reach the operating criterion. Reporting both reveals stable behavioral structure alongside substantial variation in response level.

The original six-model study provides prospective confirmation. All six prespecified predictions passed, including global and spatial geometry, model ordering, regional sign patterns and a separate

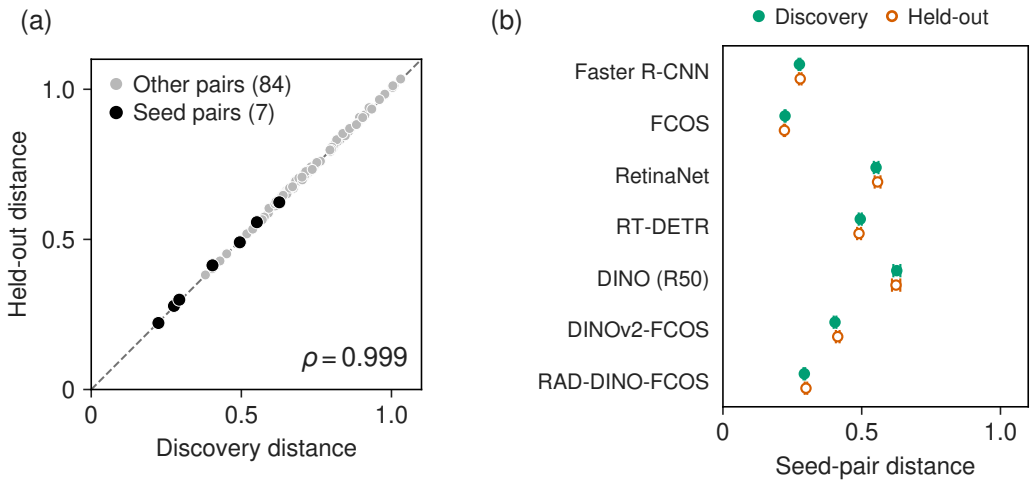


Figure 5: **Standardized score-response geometry recurs on new hosts.** (a) BPD on discovery versus held-out hosts for all 91 model pairs; the diagonal denotes equality and the seven seed pairs are black. (b) Distances for all seven seed pairs on both banks, with 95% host-bootstrap intervals. Both panels use the same distance units and limits. All fourteen models retain their seed partner as nearest neighbor. Distances are computed within host before aggregation. The two banks contain 500 and 300 hosts, respectively.

ten-host targeted test (Table 4). The regional deficit was 0.246 in discovery and 0.252 $[0.246, 0.257]$ in confirmation after matching requested and achieved contrast.

Geometry and local operating behavior answer different questions. Changing the localization radius or gain criterion can change coverage and local rankings, while the FR1/DN2 coverage separation persists throughout the tested menu (Appendix E). These results support reporting profile shape alongside coverage at a stated operating point.

4.7 The measurement budget should follow the question

A full atlas need not be remeasured for every question. In the original six-model dense bank, continuous-score ordering and the family-minus-seed BPD sign are recovered in all eight paired resamples at every tested budget, from 1% to 20%. Local signs and full-field reconstruction demand more of the sampling design (Appendix F).

These targets use different information. Recovering an ordering requires the sign of a model comparison to be correct, not an accurate response estimate at every coordinate. Finding disagreements requires measurements in informative parts of the field. Reconstructing a threshold region additionally requires its boundary to be located. We therefore evaluate claim recovery, disagreement discovery and region reconstruction separately, using the corresponding dense reference for each study.

The appropriate acquisition objective also changes with the target. At 281 queried coordinates, mean-directed acquisition recovers 0.461 of top-disagreement mass, versus 0.124 for grid sampling. Level-set acquisition instead achieves the best region intersection-over-union, 0.638, versus 0.365 for mean-directed acquisition. Thus finding where models disagree and reconstructing the region of disagreement are distinct tasks, even on the same response field (Fig. 9).

4.8 Pointwise support from natural lesions

Matched natural-lesion replacements connect controlled measurement to detector responses on real nodules. For each target lesion, the assay holds the host image fixed and compares three localized responses: the observed lesion, its estimated erasure, and a replacement fitted to image-derived lesion properties. A useful replacement should recover the response to the observed lesion more

closely than erasure does. This tests the behavioral fidelity of the replacement, independently of its pixel-level resemblance.

Anatomy-aware procedural fitting reduces response error relative to same-host erasure by 0.191 $[0.172, 0.209]$ on its valid paired support. On 463 hosts shared with an exemplar hybrid, it also has lower response error despite the hybrid's better image fit (paired difference -0.024 $[-0.032, -0.016]$). Image resemblance and behavioral fidelity are therefore distinct objectives. The advantage over erasure persists on the VinDr evaluation images (Appendix G). This assay supports pointwise response matching; it does not establish that a controlled contrast trend transfers to natural lesions.

5 Discussion

A scalar metric can order systems without characterizing the conditions that produce their errors. PULSE measures response as a function of controlled conditions, making visible when, where and under what evidence models diverge. The two views are complementary: pFROC establishes conventional performance, while the atlas exposes the fivefold coverage difference and opposite contrast responses that similar pFROC conceals. Controlled interventions make these differences observable; matched host comparisons and repeated response geometry make them reproducible.

The practical unit of evaluation is a conditional claim. Conditional comparisons identify a response regime of interest, such as declining detection with increasing contrast or reduced detection in a spatial partition. Localized scores characterize the observed change, and a second host bank tests whether it recurs. This turns an unexpected response pattern into a specific evaluation target without reconstructing the full response field. In medical imaging, where and on what appearance errors occur can matter beyond their frequency, so such regimes deserve evaluation on natural lesions and in clinical workflows. The atlas makes them explicit; it does not determine which detector should be deployed, and patient-level benefit, reader interaction and operating-point trade-offs require separate evidence. The readout is part of the claim: coverage, native-score change and standardized distance answer different questions, and a response pattern can recur even when coverage changes substantially.

A reproducible regime can also become a target for improvement: measure, diagnose the failure regime, make a targeted change to training data, augmentation or model design, then remeasure. A useful change must improve the specified regime while preserving conventional performance, and be tested on independent hosts, conditions and natural lesions to distinguish repair from overfitting to the probe bank. Here we measure and characterize failure regimes; repairing them, identifying architectural causes and demonstrating patient benefit remain separate questions. In principle, the same measure–diagnose–intervene–remeasure loop applies whenever deployment conditions can be represented by controlled evaluation axes.

6 Limitations

The atlas compares complete trained systems whose recipes, preprocessing and decoders differ, so it does not identify the causal effect of architecture or pretraining alone; its intervals resample hosts and are conditional on these models and banks. The fourteen-model atlas is descriptive, whereas prospective confirmation concerns the original six-model study, and four tested CNR levels characterize response shape only at those levels. Analytic probes provide controlled variation, not a validated model of the full natural-lesion distribution. The natural-lesion assay uses a single fixed eraser and target-informed replacements; its erasures are estimates rather than observed counterfactuals, and it tests only the original six models. The held-out host bank was withheld from PULSE analysis but overlaps earlier detector selection or qualification roles. Patient-level linkage and overlap with upstream pretraining data could not be assessed.

AI USE STATEMENT

We used generative AI tools to assist with research ideation and experiment design, implementation and analysis, literature search and synthesis, and the drafting, editing, and polishing of figures and text. The authors reviewed the AI-assisted work and take responsibility for the final content of the submission.

ETHICS STATEMENT

The study uses deidentified retrospective chest radiographs under the access terms of NODE21 and VinDr-CXR and evaluates model behavior only. It does not estimate treatment effects or make clinical deployment recommendations, and no patient identifiers or raw clinical records are redistributed. The held-out host cohort overlaps some earlier detector-development roles and is therefore described as held out for response evaluation rather than fully development-independent. Displayed study radiographs are NODE21 images (Sogancioglu et al., 2021) (CC BY-NC-ND 4.0); the demonstration radiographs in Fig. 1 and Appendix Fig. 7 are CC0 images from Wikimedia Commons (Mikael Häggström and Stillwaterising, respectively).

REPRODUCIBILITY STATEMENT

The appendix specifies the data roles, controlled interventions, response measures, confirmation analyses and measurement-design studies. The supplementary package provides code, paper configurations, aggregate results and scripts for reproducing the reported figures and tables. Restricted study radiographs and model checkpoints are not redistributed; rerunning inference requires access to the corresponding datasets and trained models.

REFERENCES

Ehsan Abadi, William P. Segars, Benjamin M. W. Tsui, Paul E. Kinahan, Nick Bottenus, Alejandro F. Frangi, Andrew Maidment, Joseph Lo, and Ehsan Samei. Virtual clinical trials in medical imaging: a review. *Journal of Medical Imaging*, 7(4):042805, 2020. doi: 10.1117/1.JMI.7.4.042805.

Anonymous. What Success Rate Hides: Mapping Disturbance Tolerance in Vision-Language-Action Policies. In *Submitted to The Fifteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=Ks45FcVqtO. under review.

Anonymous. When Should AI Evaluation Measure More? In *Submitted to The Fifteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=SXWnGa2DDw. under review.

Aldo Badano, Christian G. Graff, Andreu Badal, Diksha Sharma, Rongping Zeng, Frank W. Samuelson, Stephen J. Glick, and Kyle J. Myers. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Network Open*, 1 (7):e185474, 2018. doi: 10.1001/jamanetworkopen.2018.5474.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: 10.3322/caac.21834.

Lauren H. Cooke, Matthias Jung, Jan M. Brendel, Nora M. Kerkovits, Borek Foldyna, Michael T. Lu, and Vineet K. Raghu. RoentMod: a synthetic chest X-ray modification model to identify and correct image interpretation model shortcuts. *npj Digital Medicine*, 9(1):324, 2026. doi: 10.1038/s41746-026-02497-6.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1811.12231.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020. doi: 10.1073/pnas.1912334117.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah Smith, Hanna Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17073–17114. Curran Associates, Inc., 2025. doi: 10.52202/085713-0576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/18b3b2947174ae9b1a8fc91ff89f4eb6-Abstract-Conference.html.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25bp.html.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324. URL https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. doi: 10.1038/s41597-022-01498-w.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159. Association for Computing Machinery, 2020. doi: 10.1145/3368555.3384468.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria T. A. Wetscherek, Hannah Richardson, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. RadEdit: Stress-testing biomedical vision models via diffusion image editing. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 358–376, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-73254-6. doi: 10.1007/978-3-031-73254-6_21.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine*

Intelligence, 7(1):119–130, January 2025. doi: 10.1038/s42256-024-00965-w. URL https://doi.org/10.1038/s42256-024-00965-w.

Aria Pezeshk, Nicholas Petrick, Weijie Chen, and Berkman Sahiner. Seamless lesion insertion for data augmentation in CAD training. *IEEE Transactions on Medical Imaging*, 36(4):1005–1015, 2017. doi: 10.1109/TMI.2016.2640180.

L. G. Quekel, A. G. Kessels, R. Goei, and J. M. van Engelshoven. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, 115(3):720–724, 1999. doi: 10.1378/chest.115.3.720.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

George Revesz, Harold L. Kundel, and Martin A. Graber. The influence of structured noise on the detection of radiologic abnormalities. *Investigative Radiology*, 9(6):479–486, 1974. doi: 10.1097/00004424-197411000-00009.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442.

Nataniel Ruiz, Sarah Bargal, Cihang Xie, Kate Saenko, and Stan Sclaroff. Finding differences between transformers and ConvNets using counterfactual simulation testing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14403–14418. Curran Associates, Inc., 2022. doi: 10.52202/068431-1047. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5ce3a49415f78db65a714b4f05c62f4e-Abstract-Conference.html.

Ehsan Samei, Michael J. Flynn, and William R. Eyler. Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs. *Radiology*, 213(3):727–734, 1999. doi: 10.1148/radiology.213.3.r99dc19727.

Ecem Sogancioglu, Keelin Murphy, and Bram van Ginneken. NODE21. Zenodo, 2021. URL https://zenodo.org/records/5548363. Dataset. Concept DOI 10.5281/zenodo.4725880; CC BY-NC-ND 4.0.

Ecem Sogancioglu, Bram van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, Samuele Papa, Jonas Teuwen, Ernst Th. Scholten, Steven Schalekamp, Nils Hendrix, Colin Jacobs, Ward Hendrix, Clara I. Sánchez, and Keelin Murphy. Nodule detection and generation on chest X-rays: NODE21 challenge. *IEEE Transactions on Medical Imaging*, 43(8):2839–2853, 2024. doi: 10.1109/TMI.2024.3382042.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, January 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Suvorov_Resolution-Robust_Large_Mask_Inpainting_With_Fourier_Convolutions_WACV_2022_paper.html.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, 2019. doi: 10.1109/ICCV.2019.00972. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Tian_FCOS_Fully_Convolutional_One-Stage_Object_Detection_ICCV_2019_paper.html.

Fakrul Islam Tushar, Liesbeth Vancoillie, Cindy McCabe, Amareswararao Kavuri, Lavsen Dahal, Brian Harrawood, Milo Fryling, Mojtaba Zarei, Saman Sotoudeh-Paima, Fong Chi Ho, Dhrubajyoti Ghosh, Michael R. Harowicz, Tina D. Tailor, Sheng Luo, W. Paul Segars, Ehsan Abadi, Kyle J. Lafata, Joseph Y. Lo, and Ehsan Samei. Virtual lung screening trial (VLST): An in silico study inspired by the national lung screening trial for lung cancer detection. *Medical Image Analysis*, 103:103576, July 2025. doi: 10.1016/j.media.2025.103576. URL https://doi.org/10.1016/j.media.2025.103576.

Fakrul Islam Tushar, Umme Hafsa Momy, Joseph Y. Lo, and Geoffrey D. Rubin. iTRIALSPACE: Programmable virtual lesion trials for controlled evaluation of lung CT models, 2026. URL https://arxiv.org/abs/2605.05761v2.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J. Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82232–82251. PMLR, 2025. URL https://proceedings.mlr.press/v267/wallach25a.html.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.03605.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, 2024. doi: 10.1109/CVPR52733.2024.01605. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_DETRs_Beat_YOLOs_on_Real-time_Object_Detection_CVPR_2024_paper.html.

Zhaohui Zheng, Yuming Chen, Qibin Hou, Xiang Li, Ping Wang, and Ming-Ming Cheng. Zone evaluation: Revealing spatial bias in object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8636–8651, 2024. doi: 10.1109/TPAMI.2024.3409416. URL https://doi.org/10.1109/TPAMI.2024.3409416.

A DATA, MODELS AND CONTROLLED INTERVENTIONS

A.1 HOSTS AND ROLES

The controlled panel contains 800 provider-labelled nodule-negative frontal NODE21 radiographs (Sogancioglu et al., 2024; 2021): 500 discovery hosts and 300 hosts held out for response evaluation. Within discovery, the global bank uses 500 hosts, spatial perimetry 300 and the dense measurement-design bank 12. Role membership preceded the corresponding response analysis. No controlled host belongs to a detector fitting split, but some served in earlier checkpoint selection or qualification; confirmation hosts are therefore held out for response evaluation rather than development-unseen. Hosts are the sampling units.

A.2 MODELS AND TRAINING

Table 1 summarizes the fourteen models. Faster R-CNN, FCOS and RetinaNet use ImageNet-initialized ResNet-50 FPN backbones at 1024-pixel input; RT-DETR (R50vD, 640 pixels) and DINO (ResNet-50, 1024 pixels) start from COCO detectors. The DINOv2 and RAD-DINO models keep their named ViT-B/14 encoders frozen at 896-pixel input and train an identical FCOS/FPN head; the two encoders of each seed index share initial head weights. Every model was selected by validation pFROC on conventional development data before any PULSE response was examined. Family comparisons therefore concern complete trained systems. Full recipes and qualification rules are in the companion.

Table 1: **Fourteen models under common calibration.** Init.: ImageNet (IN), COCO, or a frozen named ViT encoder. τ: native-score threshold at 0.25 false positives per image on the shared 562-image calibration set. P: pFROC; S_τ: sensitivity at τ. R_{500}, R_{300}: global RSFF on the discovery and held-out banks.

ID	Detector (backbone)	Init.	Input	τ	P	S_τ	R_{500}	R_{300}
FR1	Faster R-CNN (R50-FPN)	IN	1024	0.7885	0.772	0.707	0.406	0.405
FR2	Faster R-CNN (R50-FPN)	IN	1024	0.7530	0.772	0.689	0.390	0.387
FC1	FCOS (R50-FPN)	IN	1024	0.4005	0.768	0.725	0.347	0.346
FC2	FCOS (R50-FPN)	IN	1024	0.4585	0.740	0.658	0.266	0.262
RN1	RetinaNet (R50-FPN)	IN	1024	0.3468	0.657	0.554	0.141	0.139
RN2	RetinaNet (R50-FPN)	IN	1024	0.2811	0.685	0.622	0.224	0.222
RT1	RT-DETR (R50vD)	COCO	640	0.009134	0.809	0.784	0.314	0.313
RT2	RT-DETR (R50vD)	COCO	640	0.03721	0.745	0.730	0.374	0.372
DN1	DINO detector (R50)	COCO	1024	0.1372	0.835	0.797	0.161	0.157
DN2	DINO detector (R50)	COCO	1024	0.2543	0.793	0.730	0.077	0.075
DV1	DINOv2 ViT-B/14 + FCOS	frozen	896	0.4707	0.417	0.342	0.099	0.098
DV2	DINOv2 ViT-B/14 + FCOS	frozen	896	0.4028	0.564	0.491	0.267	0.266
RD1	RAD-DINO ViT-B/14 + FCOS	frozen	896	0.4121	0.672	0.595	0.369	0.364
RD2	RAD-DINO ViT-B/14 + FCOS	frozen	896	0.4201	0.595	0.514	0.351	0.350

A.3 COMMON CALIBRATION

All fourteen models are thresholded on the same 562 NODE21 calibration images, with center-in-box matching and pFROC over 0.125–1.0 false positives per image. This calibration set is disjoint from detector fitting and selection but shares some images with the controlled host banks, so it fixes operating points rather than providing independent validation. All models are evaluated on identical source interventions, geometric validity is fixed independently of the model, and native decoders are kept unchanged; alternative readouts are exact replays of the retained decoded outputs.

A.4 PROBE DESIGN

A probe is added in the source intensity domain, $I_z = \mathrm{clip}_{[\ell,u]}(I + a\,p_z)$ with $a = \kappa\,\hat{\sigma}_{\mathrm{ann}}$, where κ is the requested CNR and $\hat{\sigma}_{\mathrm{ann}}$ the clean-image standard deviation in the annulus $0.75d \le r \le 1.5d$ around the center, for probe diameter d. Each model's own preprocessing is then applied unchanged.

The diameter is a fraction of the host's lung-mask width. The primary profile is a peak-one logistic soft-edged disk, $1/(1 + \exp[(r - d/2)/\sigma_e])$, where r is the distance from the center and σ_e the edge-transition width in pixels. The morphology variants add eccentricity, a bounded radial contour perturbation, or internal texture; the Gaussian variant has the same peak and FWHM diameter but wider tails; one contour variant is matched to the disk's L_2 energy. Achieved source CNR is the realized peak over $\hat{\sigma}_{\mathrm{ann}}$ and effective-input CNR the corresponding ratio at each detector's input.

A.5 Global and spatial banks

The global bank has 400 conditions per host: 192 primary anchors (twelve positions; sizes 1.5, 3, 5 and 8% of lung width; requested CNRs 0.75, 1.25, 2 and 3), 80 morphology anchors (four mid-lung positions, five variants, four CNRs) and a 128-point Sobol design over position, size, CNR and variant. The spatial bank places five probes on a 17×17 lattice per hemithorax: sizes 2.5, 4.75 and 7.5% at CNR 1.5, and CNRs 0.75 and 3 at size 4.75%. Placements respect probe size, boundary clearance and the noise annulus; invalid placements stay explicit and enter no denominator. Atlas statistics are computed on measured cells without interpolation. The NODE21 overlays apply display-only Gaussian smoothing to those measured cell values ($\sigma = 0.75$ grid spacing), which enters no estimate.

B Response measures and inference

Readout. For each inserted center, S^+ and S^0 are the maximum inserted and clean nodule scores among boxes centered within 16 working-source pixels; the two maxima may come from different boxes. The primary readout is $Y_{m,h,z} = \mathbf{1}\{S^+_{m,h,z} \geq \tau_m,\ S^+_{m,h,z} - S^0_{m,h,z} \geq 0.02\}$. Decoders, score floors and output caps are fixed parts of this operating contract. Alternative readouts are defined in the companion.

Coverage and host weighting. For valid conditions $\mathcal{Z}_h$ on host h,

$$\widehat{\mathrm{RSFF}}_m = \frac{1}{n}\sum_h \frac{1}{|\mathcal{Z}_h|}\sum_{z\in\mathcal{Z}_h} Y_{m,h,z}.$$

Invalid placements enter neither numerator nor denominator, and RSFF gives each host equal weight.

Profile distance. Let $s(x)$ center a host vector and divide by its sample standard deviation. The host-paired distance and its host average are

$$d_{a,b,h} = \left(\frac{1}{K_h}\sum_k [s(\Delta S_{a,h})_k - s(\Delta S_{b,h})_k]^2\right)^{1/2}, \qquad \mathrm{BPD}_{a,b} = \frac{1}{n}\sum_h d_{a,b,h}.$$

Global profiles align by condition; spatial profiles by probe, side, row and column. BPD is thus the host-averaged RMS distance between standardized, aligned score-response profiles. Because profiles are compared within each host before averaging, BPD is not determined by the two mean responses.

Regions. The central-coordinate region is $u \leq 0.35$, $0.35 \leq v < 0.65$ in normalized hemithorax coordinates; lower-visible lung has $v \geq 0.65$. These are coordinate partitions, not anatomical segmentations. Regional effects subtract the matched lower-visible mean.

Inference. Intervals use 2,000 whole-host bootstrap draws unless stated otherwise, keeping all models and conditions of a host together. They describe these models and banks rather than a population of architectures.

C Atlas analyses

Evidence status. The atlas slices were chosen after the discovery banks had been examined, so they are descriptive; agreement between the 500- and 300-host banks measures stability. The pre-specified tests are in Appendix D. Summaries are computed within host and then averaged.

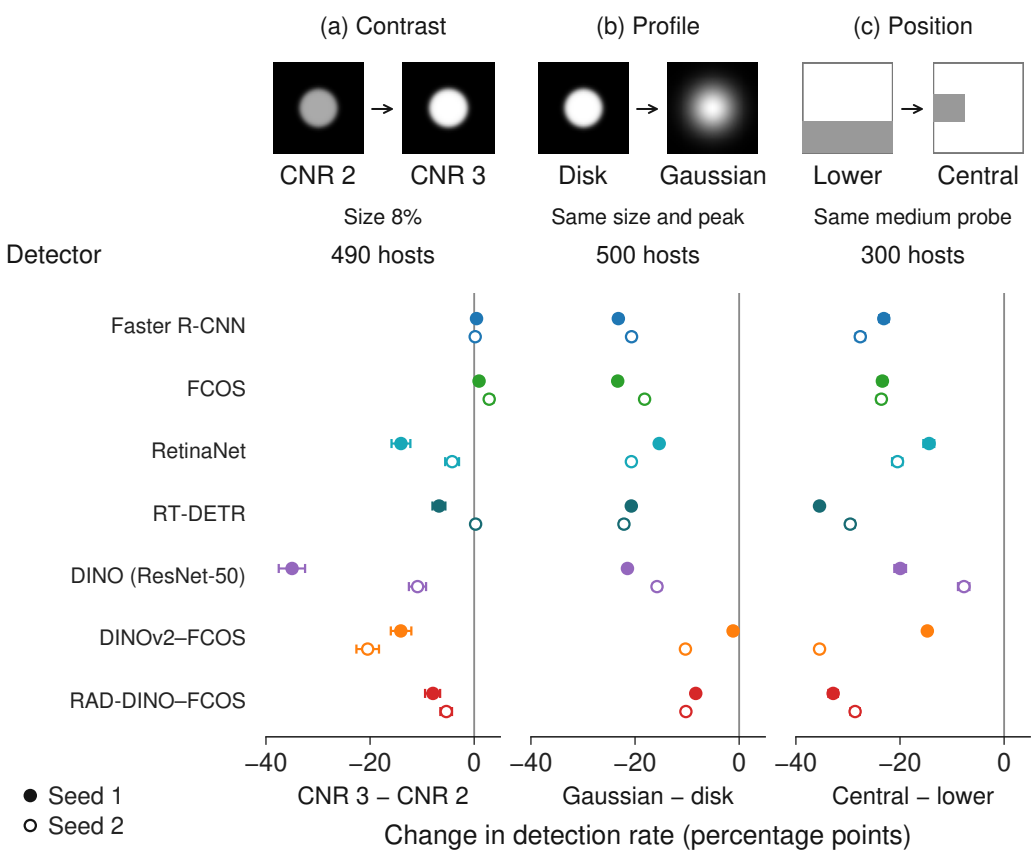


Figure 6: **Three interventions reveal distinct detector sensitivities.** Paired changes in E0 detection rate (percentage points) for all fourteen models (filled/open: seeds 1/2), with 95% host-bootstrap intervals. Schematics show the intervention, not measured radiographs. (a) CNR 2 to 3 at size 8%, on shared support (490 hosts). (b) Smooth disk to peak-matched Gaussian at size 5%, averaged over four CNRs (500 hosts). (c) Lower-visible to central-coordinate positions at size 4.75% and CNR 1.5 (300 hosts). Numerical and held-out values: Tables 2 (a) and 3 (c) for the original twelve, Table 7 for the added seeds, and the companion morphology table for (b).

Matched support. The shared-support CNR comparison keeps a host–position cell only if, for every compared level and model, effective-input CNR is within 25% of achieved source CNR and the source-range clipping burden is at most 1%. The screen was fixed on the original twelve models and reused unchanged for the two added seeds. It retains 0.38 of cells at size 8%; Table 2 also reports full valid support. Morphology comparisons match size, position and requested CNR; regional comparisons use the medium spatial probe.

Supporting comparisons. Table 2 shows that the contrast decline is not a localization failure: a box usually remains near the probe while its native score falls. Table 3 gives the paired regional effects and, separately, variation between hosts. The profile-by-CNR grid, morphology table, full five-probe spatial atlas and a ten-model geometry check are in the companion.

D PROSPECTIVE CONFIRMATION

The original six-model study fixed six hypotheses, estimators and decision rules before any response on the confirmation hosts was accessed. C1–C5 use the 300 held-out hosts; C6 uses a separate ten-host targeted test set. C1 and C2 test global and spatial profile geometry (between-family minus within-seed BPD), C3 the coverage ordering, C4 the central-coordinate deficit, C5 the projected-field sign pattern, and C6 the separation of Faster R-CNN from RetinaNet at a demanding visible-round

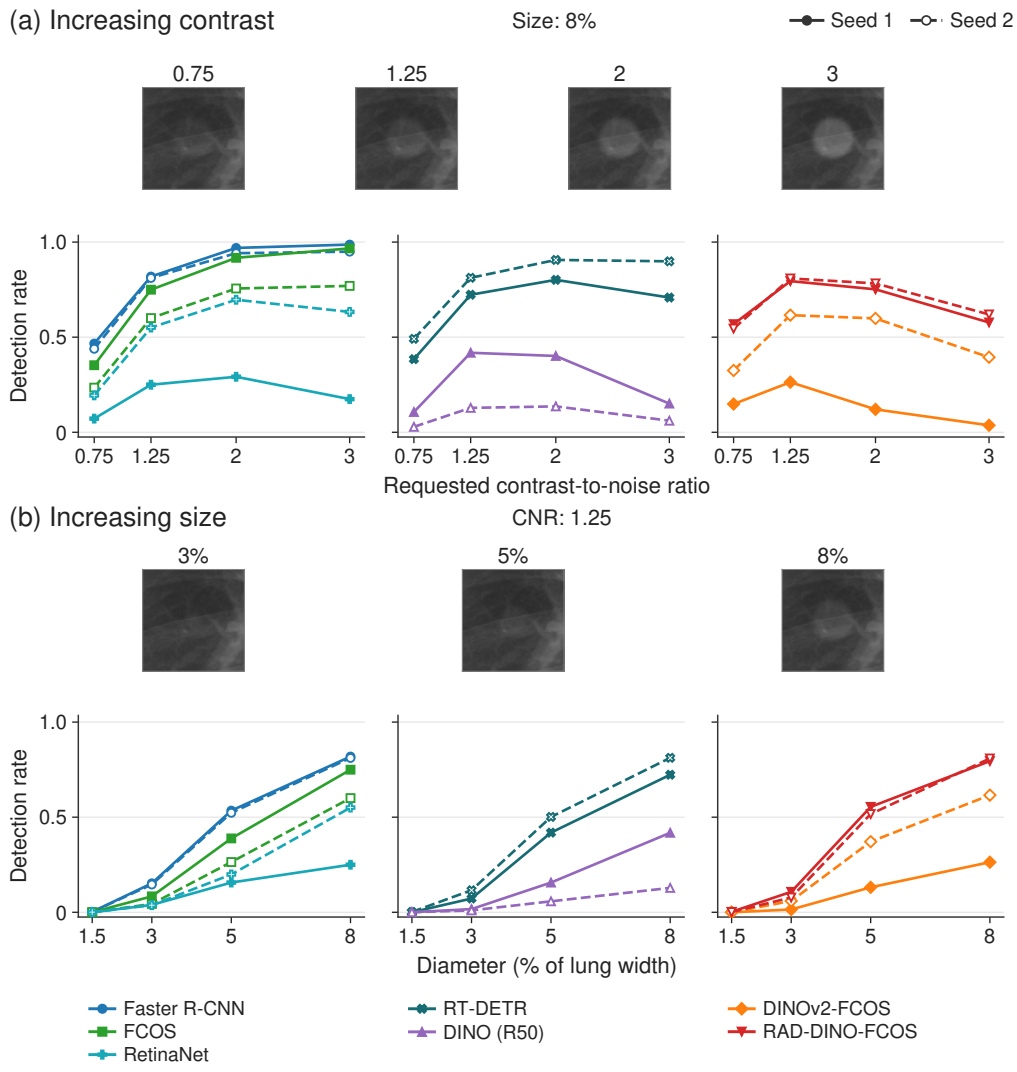


Figure 7: **Size and requested-contrast sweeps for all fourteen models.** (a) CNR sweep at size 8%. (b) Size sweep at CNR 1.25. Seed 1 solid and filled, seed 2 dashed and open. Curves are equal-host mean detection rates on each condition's geometry-valid support (500 discovery hosts). Probe crops on a public radiograph illustrate the conditions and contribute to no estimate. Figure 3b,c shows the seed-1 subset.

condition. Each claim has its own rule; all six passed (Table 4). Confirmation changes response hosts, not model identities, and the fourteen-model atlas and measurement-design studies were not part of this prespecified test.

Twelve-model geometry. On the original twelve models, discovery and held-out global BPD over the 66 model pairs have Spearman correlation 0.9981 [0.9943,0.9990], and every model retains its nearest neighbor. The full matrix, spatial counterpart and detailed decision mechanics are in the companion.

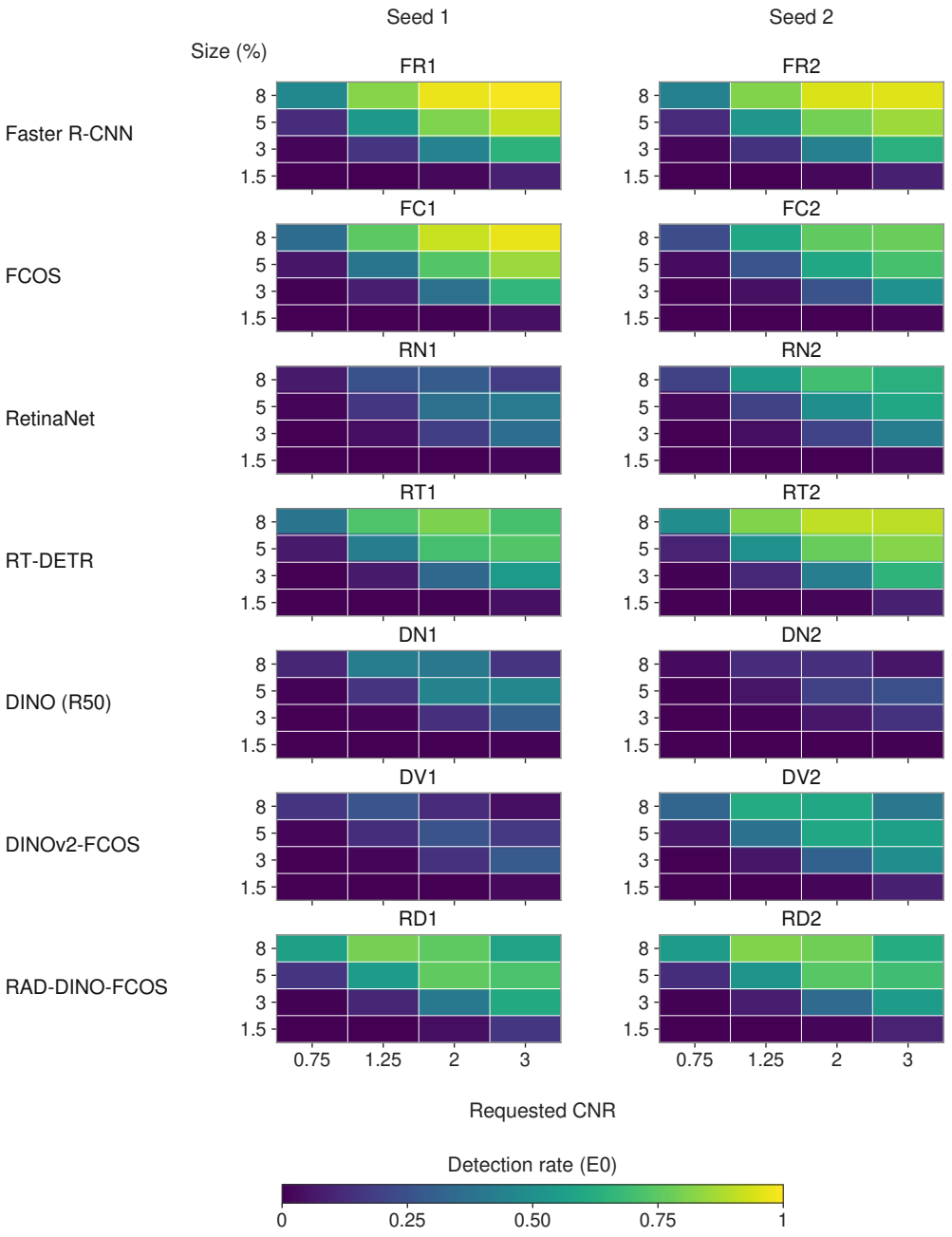


Figure 8: **The size–contrast atlas: fourteen models, one coordinate system.** Rows group detector families, with their two seeds side by side. Each cell is the equal-host mean E0 response across valid primary positions on 500 discovery hosts; requested CNR increases left to right and diameter bottom to top, on a common 0–1 scale. Figures 3 and 7 extract the size-8% row and CNR-1.25 column as curves.

Table 2: **Contrast decline and localized outputs at size 8%.** (A) Paired E0 differences, CNR 3 minus CNR 2, with 95% whole-host intervals. Shared support retains 0.38 of matched cells on each bank, with 490 of 500 discovery and 293 of 300 held-out hosts contributing; full valid support retains 0.96/0.97. (B) Discovery/shared-support outputs: fraction of CNR-3 cells with a box inside the localization radius, median native score at CNR 2→3, and unified threshold τ.

A. Detection change

	Shared support		Full valid support	
Model	Discovery	Held-out	Discovery	Held-out
FR1	0.005 [0.002, 0.007]	0.004 [0.000, 0.008]	0.018	0.019
FR2	0.002 [−0.002, 0.006]	0.006 [−0.003, 0.014]	0.008	0.004
FC1	0.009 [0.006, 0.013]	0.013 [0.008, 0.019]	0.049	0.055
FC2	0.029 [0.021, 0.037]	0.025 [0.013, 0.038]	0.014	0.009
RN1	−0.141 [−0.159, −0.123]	−0.154 [−0.179, −0.129]	−0.117	−0.125
RT1	−0.067 [−0.080, −0.055]	−0.063 [−0.080, −0.048]	−0.093	−0.100
DN1	−0.350 [−0.375, −0.325]	−0.360 [−0.395, −0.327]	−0.250	−0.247
DN2	−0.109 [−0.125, −0.092]	−0.096 [−0.117, −0.076]	−0.075	−0.070
DV1	−0.141 [−0.160, −0.121]	−0.163 [−0.192, −0.133]	−0.084	−0.077
DV2	−0.205 [−0.226, −0.183]	−0.221 [−0.250, −0.191]	−0.204	−0.209
RD1	−0.079 [−0.094, −0.065]	−0.100 [−0.120, −0.080]	−0.174	−0.194
RD2	−0.054 [−0.065, −0.043]	−0.075 [−0.090, −0.059]	−0.163	−0.179

B. Localization and native scores

Model	Localized	Median score	τ	Model	Localized	Median score	τ
FR1	1.00	0.988→0.987	0.789	DN1	0.94	0.666→0.004	0.137
FR2	1.00	0.979→0.974	0.753	DN2	1.00	0.050→0.021	0.254
FC1	1.00	0.697→0.684	0.401	DV1	0.93	0.422→0.344	0.471
FC2	1.00	0.653→0.653	0.459	DV2	0.98	0.582→0.489	0.403
RN1	0.98	0.235→0.127	0.347	RD1	1.00	0.614→0.602	0.412
RT1	1.00	0.058→0.051	0.009	RD2	1.00	0.559→0.557	0.420

E READOUT SENSITIVITY

Because decoded outputs are retained, gain, radius and threshold changes are exact replays on the same geometry. Table 5 summarizes the cases that define the measurement contract: FR1 is relatively stable, RT1 is sensitive to the gain criterion, DV1 is sensitive to the localization radius, and the large FR1/DN2 separation persists at every tested setting. Some local comparisons change with the readout; for example, a gain of 0.05 places RT1 below FC2 and DV2. Coverage is therefore reported at a stated operating point; the companion gives the complete gain, radius and false-positive-target table and the alternative-readout menus.

F MEASUREMENT DESIGN

The dense-bank study asks which claims or fields can be recovered from subsamples; the acquisition study asks where two models disagree and whether that region can be reconstructed. They use different hosts, supports and query units (Fig. 9).

Fixed-bank claims. The dense reference measures six models on every valid cell of twelve hosts, and subsamples keep 1–20% of each host's valid support. Global ordering and the family-versus-seed geometry sign were recovered at every tested budget, whereas local regional signs and full-field reconstruction were more sensitive to the sampling procedure: some sampling designs miss the regional sign even at 20%, and no tested surrogate reaches a held-out field MSE of 0.01 at any budget. The companion gives the designs, estimators and failure cases.

Disagreement discovery and region reconstruction. The acquisition bank measures visible-lung coordinates at six CNR levels on eight hosts. For each of fifteen model pairs, discovery is disagree-

Table 3: **Regional effects and variation between hosts.** (A) Medium-probe E0 at central-coordinate and lower-visible positions; differences are central minus lower, with 95% host-bootstrap intervals (300 hosts per bank). (B) Global per-host coverage distribution (500 hosts). The percentiles describe variation between hosts, not uncertainty in the mean.

A. Paired regional contrast

Model	Central	Lower	Difference (discovery)	Difference (held-out)
FR1	0.305	0.536	−0.231 [−0.240, −0.222]	−0.234 [−0.242, −0.225]
FR2	0.245	0.522	−0.276 [−0.285, −0.267]	−0.287 [−0.295, −0.278]
FC1	0.224	0.458	−0.234 [−0.242, −0.226]	−0.243 [−0.250, −0.235]
FC2	0.113	0.349	−0.236 [−0.244, −0.228]	−0.243 [−0.250, −0.235]
RN1	0.024	0.168	−0.144 [−0.155, −0.134]	−0.138 [−0.148, −0.128]
RT1	0.123	0.478	−0.355 [−0.361, −0.348]	−0.366 [−0.373, −0.360]
DN1	0.029	0.229	−0.199 [−0.211, −0.188]	−0.199 [−0.210, −0.188]
DN2	0.008	0.085	−0.077 [−0.089, −0.067]	−0.078 [−0.090, −0.068]
DV1	0.002	0.150	−0.148 [−0.153, −0.142]	−0.143 [−0.148, −0.138]
DV2	0.038	0.392	−0.355 [−0.362, −0.348]	−0.355 [−0.361, −0.349]
RD1	0.162	0.490	−0.328 [−0.338, −0.319]	−0.342 [−0.352, −0.333]
RD2	0.159	0.446	−0.286 [−0.295, −0.277]	−0.299 [−0.309, −0.289]

B. Distribution of host-level coverage

Model	Mean	10th percentile	25th–75th percentiles	90th percentile
FR1	0.406	0.335	0.371–0.442	0.477
FR2	0.390	0.318	0.352–0.426	0.466
FC1	0.347	0.274	0.307–0.385	0.416
FC2	0.266	0.197	0.226–0.301	0.340
RN1	0.141	0.045	0.090–0.188	0.234
RT1	0.314	0.266	0.289–0.339	0.363
DN1	0.161	0.075	0.115–0.203	0.247
DN2	0.077	0.005	0.027–0.106	0.163
DV1	0.099	0.058	0.076–0.122	0.143
DV2	0.267	0.222	0.244–0.291	0.317
RD1	0.369	0.294	0.332–0.404	0.439
RD2	0.351	0.278	0.314–0.387	0.419

Table 4: **All six prespecified predictions passed.** Brackets are 95% host-bootstrap intervals, not a joint familywise band. The full prespecified decision rules are given in the companion.

Claim	Discovery	Held-out	Decision
C1	0.365	0.365 [0.361, 0.371]	PASS
C2	0.390	0.387 [0.383, 0.392]	PASS
C3	Prespecified rank	Same rank; 3/3 relations	PASS
C4	0.246	0.252 [0.246, 0.257]	PASS
C5	0.147	0.139 [0.132, 0.147]	PASS
C6	Range 0.967	2/2 directions; range 0.800	PASS

Table 5: **Readout sensitivity is model-specific.** Global RSFF on 500 hosts. Primary readout: radius 16 working-source pixels, gain 0.02, 0.25 false positives per image.

Model	Primary	Perturbation	Coverage
FR1	0.406	gain and radius menu	0.397–0.407
RT1	0.314	gain $0.02 \rightarrow 0.05$	0.171
DV1	0.099	radius $8 \rightarrow 24$	$0.021 \rightarrow 0.174$
DN2	0.077	full tested menu	0.061–0.113 (below FR1)

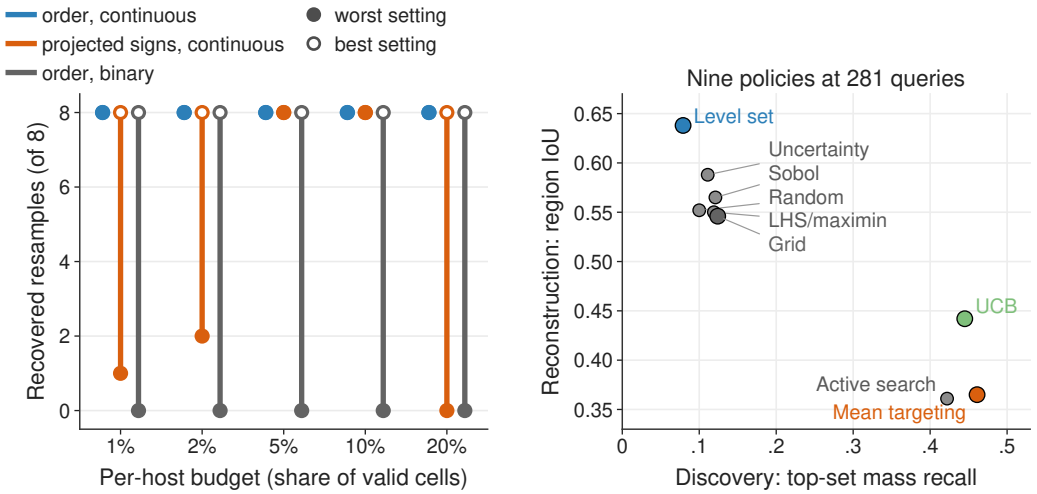


Figure 9: **Different questions favor different measurements.** Left: fixed-bank recovery on twelve hosts and six models; filled/open markers show the lowest/highest recovered count out of eight paired resamples across four designs and five estimators. Right: disagreement discovery versus threshold-region reconstruction at 281 queries for nine policies, averaged over fifteen model pairs and 100 acquisition seeds. Mean targeting leads discovery; level set leads region reconstruction.

ment mass recall: the fraction of total absolute E0-difference mass in the top-decile set that is captured by queried coordinates. Reconstruction is the IoU of the predicted and true regions where the absolute difference is at least 0.15. Mean targeting best identified high-disagreement locations, whereas level-set acquisition best reconstructed the disagreement region: at 281 queries, mean targeting recovers 0.461 of the top-set mass (grid: 0.124) but reaches a region IoU of only 0.365 (grid: 0.546), while level set gives the highest IoU, 0.638. Mean targeting exceeds grid discovery on all four prospectively locked pairs and all fifteen offline pair means. Discovery and reconstruction should therefore be evaluated as separate objectives.

G NATURAL-LESION EVIDENCE

Assay. The original six models are evaluated on 506 annotated NODE21 and VinDr-CXR evaluation images (149 and 357); replacement donors come only from a separate training split. ERASE is a lesion-removed estimate from a fixed CXR-adapted Big-LaMa eraser (Suvorov et al., 2022), so REAL−ERASE can contain reconstruction residuals. A4 fits procedural profiles, AR1 adds anatomy-aware fitting and AR2 admits training-bank residual exemplars; REAL-derived descriptors inform the fit, but detector outputs never select candidates. The local score is the maximum nodule score with box center inside the slightly expanded annotation box, and a hit crosses the native threshold without a gain criterion, so this readout differs from E0. Response error is $|S_M - S_R|$, and recovery is the replacement-hit rate among each model's REAL-hit/ERASE-miss hosts. A fixed image-only subset of 166 hosts, defined without detector outputs, compares all methods on common support.

Pointwise support. All three replacements reduce response error relative to erasure (Fig. 10); Table 6 reports AR1, including the VinDr images, which overlap no detector-development role. Image fit and response fit are distinct: AR1 has lower response error than AR2 on their 463 shared hosts (-0.024 $[-0.032, -0.016]$) despite AR2's better image fit.

Directional transport. Recovering a response at matched lesions does not show that changing an analytic probe property predicts the effect of the same change in natural lesions. Natural support was insufficient for a directional transport test, so natural directional transport is non-estimable. The

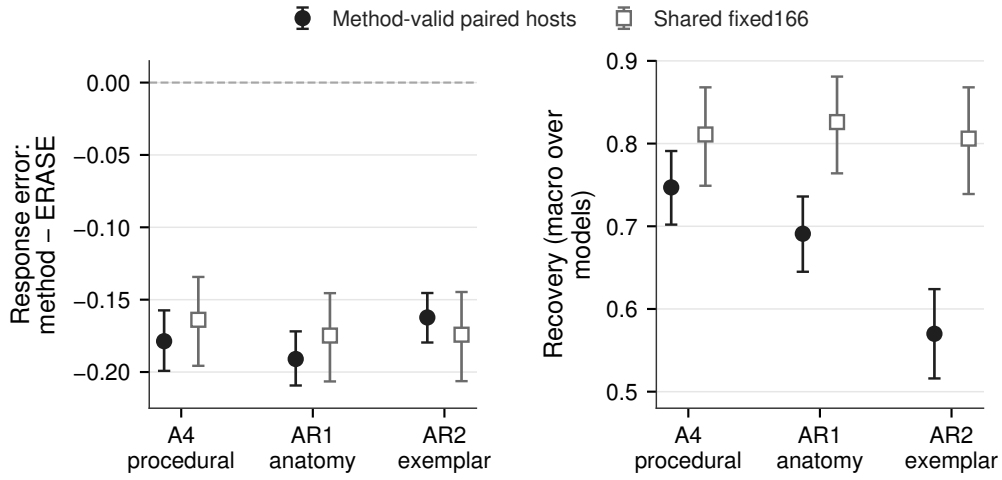


Figure 10: **Matched replacement recovers part of the natural-lesion response.** A4: procedural baseline; AR1: anatomy-aware procedural fitting; AR2: exemplar hybrid. Left: paired replacement-minus-ERASE response-error difference, averaged over six models; negative values favor replacement. Right: macro-averaged recovery among each model's REAL-hit/ERASE-miss hosts. Filled markers use method-specific valid support; open markers use the same 166-host image-only subset for all methods. Intervals are 95% whole-host bootstrap intervals.

Table 6: **Matched AR1 replacement reduces response error relative to erasure.** Six-model macro means for anatomy-aware procedural fitting (AR1). Valid/target: hosts with a valid AR1 replacement over the cohort size. AR1−ERASE uses paired valid hosts, with 95% whole-host intervals. Recovery is macro-averaged over model-specific REAL-hit/ERASE-miss subsets, so the valid count is not a common recovery denominator. Fixed: the common 166-host image-only subset. VinDr images do not overlap detector development. The full A4/AR1/AR2 table is in the companion.

Cohort	Valid/target	AR1 error	Recovery	AR1−ERASE [95% CI]
All	463/506	0.078	0.691	$-0.191\,[-0.209, -0.172]$
Fixed	166/166	0.047	0.826	$-0.175\,[-0.206, -0.145]$
VinDr	330/357	0.059	0.717	$-0.125\,[-0.144, -0.106]$

companion gives the support audit, image-agreement measures, eligibility rules, overlap counts and full method table.

H SECOND SEEDS FOR RETINANET AND RT-DETR

RN2 and RT2 repeat the original RetinaNet and RT-DETR recipes, splits, pretraining and decoders, changing only the training seed; epoch selection preceded any PULSE response access. Both complete all four banks on the same hosts, conditions and valid support, forming the fourteen-model panel.

Response geometry replicates: each new model and its original partner are mutual nearest neighbors on every bank, with global seed-pair distances of 0.552 (RetinaNet) and 0.495 (RT-DETR), within the range of the other seed pairs (0.224–0.626). Gaussian-profile losses and the central-coordinate deficit recur for both. Contrast sensitivity varies between seeds: the size-8% decline is much weaker for RN2, and RT2 does not reproduce RT1's decline on either bank (Table 7). The RT-DETR contrast decline therefore does not characterize both tested seeds.

Table 7: **Second-seed replication: stable geometry and model-dependent contrast sensitivity.** Global-bank RSFF and Gaussian−disk change (discovery, 500 hosts); E0 at CNR 3 minus CNR 2 at size 8% on shared support (490 discovery and 293 held-out hosts); central minus lower-visible E0 for the medium probe on the held-out spatial bank (300 hosts), whose 95% intervals all exclude zero. Brackets: 95% paired whole-host intervals.

		CNR 3 − CNR 2			
Model	RSFF	Discovery	Held-out	Gauss.−disk	Central−lower
RN1	0.141	$-0.141\,[-0.159, -0.123]$	$-0.154\,[-0.179, -0.129]$	-0.154	-0.138
RN2	0.224	$-0.042\,[-0.056, -0.029]$	$-0.038\,[-0.058, -0.020]$	-0.207	-0.208
RT1	0.314	$-0.067\,[-0.080, -0.055]$	$-0.063\,[-0.080, -0.048]$	-0.207	-0.366
RT2	0.374	$0.003\,[0.000, 0.005]$	$0.002\,[-0.001, 0.005]$	-0.221	-0.319

I SPATIAL RESPONSE GALLERY ON NODE21 HOSTS

Three figures show the paper's spatial findings on study radiographs, using only measurements already reported. The spatial banks insert each probe separately on a 17×17 grid per hemithorax, and a site is measured only where the probe is geometry-valid. Single-host maps smooth the binary E0 outcomes at measured sites for display ($\sigma = 0.75$ grid spacing, within one spacing of a measured site); faint dots mark measured sites and corner counts give detected/valid placements. The aggregate map smooths per-cell means in the same way. Jet runs from blue (0, missed) to red (1, detected); the coloured region is the placement support and the white outline the visible-lung support; smoothing is for display only, and colors are not probabilities or saliency. The display hosts were selected by fixed rules on discovery-bank responses before any figure was drawn: a representative host closest to the population atlas, and a host with similar FR1/RD1 totals but large spatial disagreement. Further single-host pages, exact anchor audits and the selection audit are in the companion.

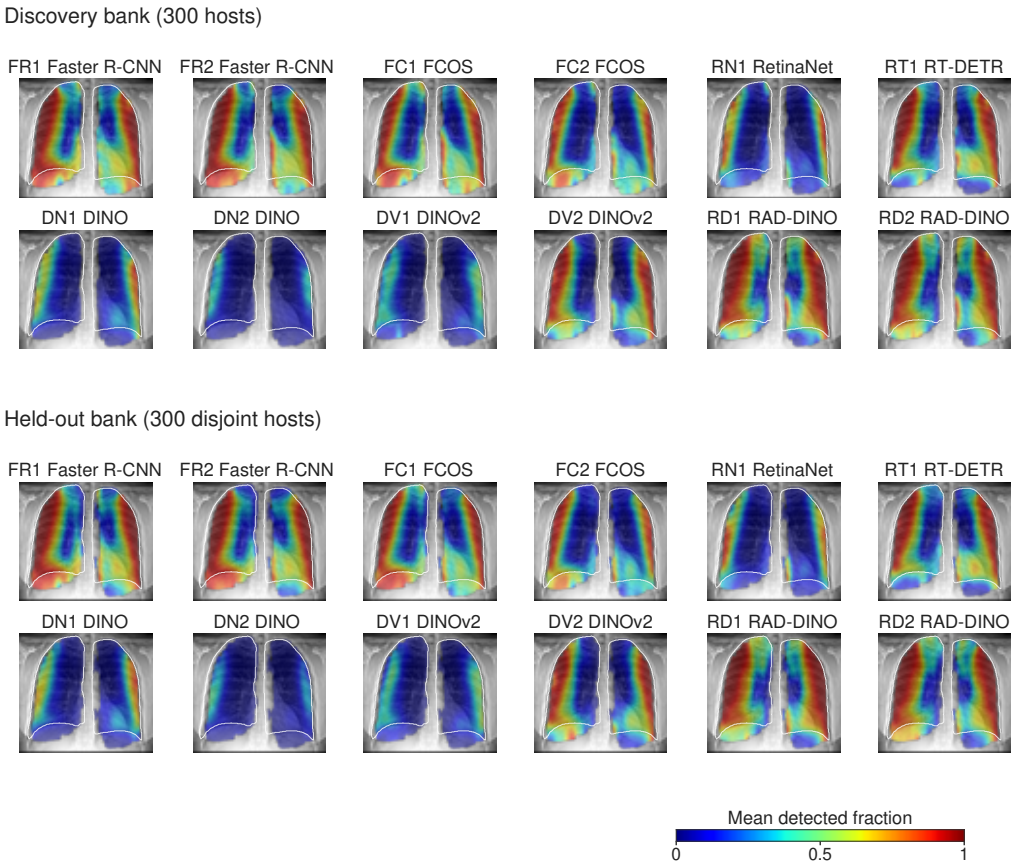


Figure 11: **The aggregate spatial pattern recurs on disjoint host banks.** Mean E0 of the medium probe (4.75%, requested CNR 1.5) per grid cell over 300 discovery and 300 held-out hosts, for the original twelve-model atlas, projected onto the representative host. RN2 and RT2 completed the same spatial banks; their spatial summaries are reported in Appendix H. The lateral-to-medial gradient and the model ordering are shared by both banks (Section 4.5). The unsmoothed per-cell discovery atlas is in the companion.

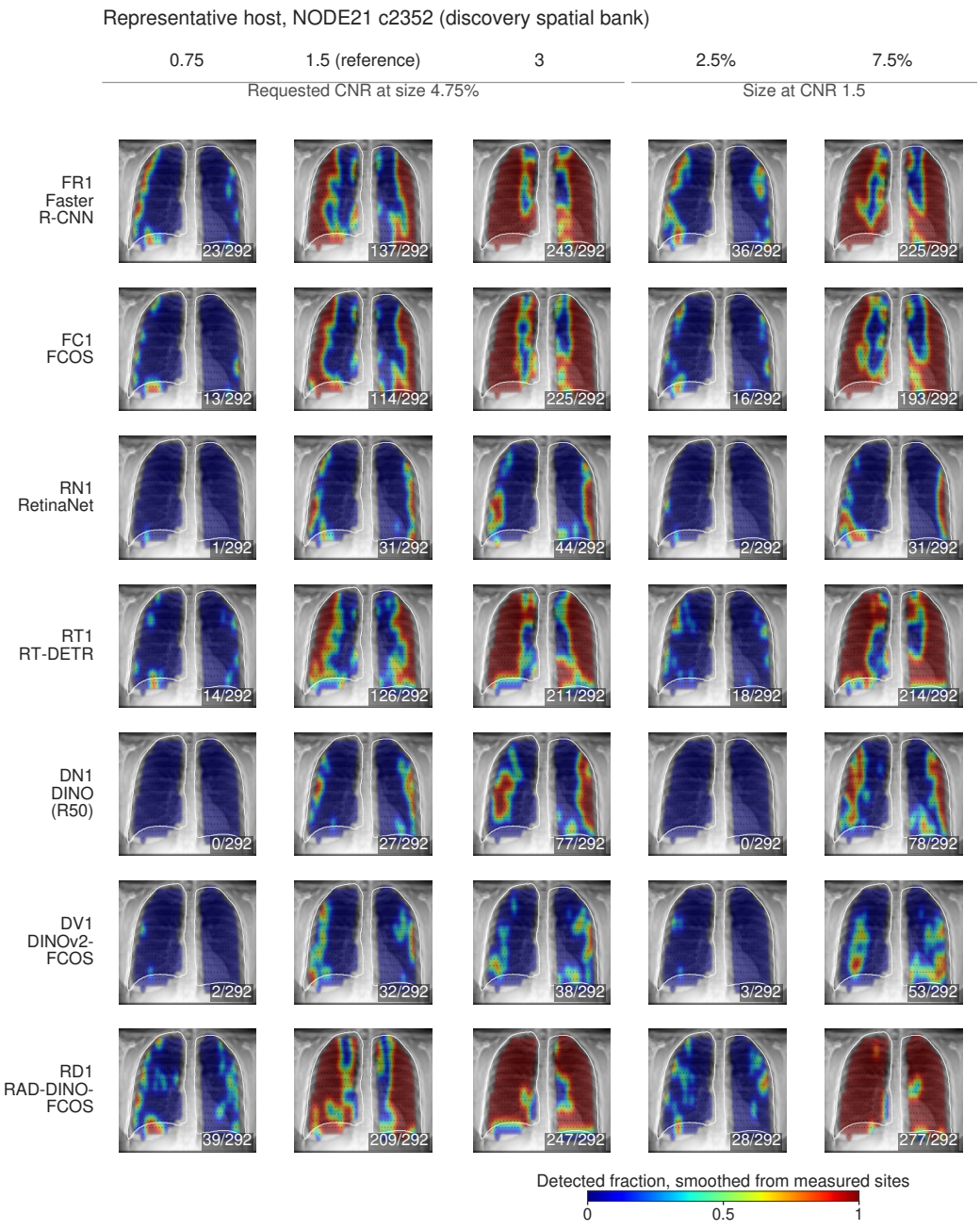


Figure 12: **Response maps grow with size and requested CNR, at different rates per family.** Representative host; seed-1 model of each family (rows) under the five spatial-bank probes: requested CNR 0.75, 1.5 and 3 at 4.75% of lung width, and sizes 2.5% and 7.5% at CNR 1.5. Fig. 3a shows the FR1 and DN1 rows.

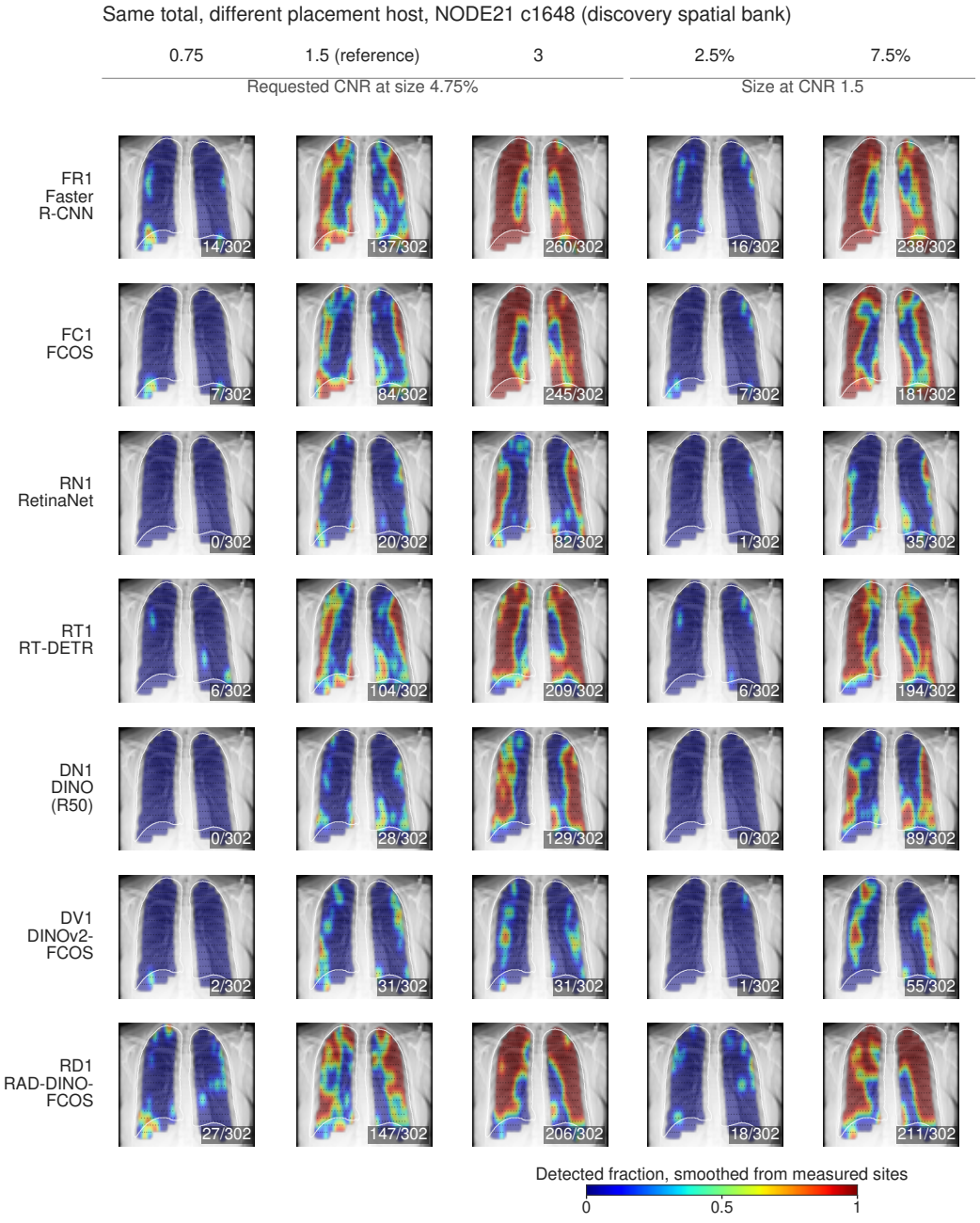


Figure 13: **Similar coverage, different response locations.** Layout as in Fig. 12. At the medium probe, Faster R-CNN and RAD-DINO–FCOS detect similar numbers of placements on this host but in different parts of the lung.