LESIONBENCH: A 3D COMPUTED TOMOGRAPHY CANCER AGNOSTIC SEGMENTATION BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical oncology relies heavily on imaging for multiple tasks, from screening to computer-aided surgery. However, lesion segmentation is a significantly more complex task compared to multi-organ one, as lesions can appear in different body locations, exhibit highly variable sizes, shapes, and visual appearances when imaged using different scanners or protocols. To develop the next generation of AI models that can deliver acceptable performances compared to radiologists, there is a clear need to create broad, diverse, and structured 3D CT datasets with lesion annotations. Despite remarkable efforts, most available datasets are limited to a single cancer type or have scattered annotations. To address these shortcomings, we present LesionBench, a large-scale CT binary segmentation benchmark that spans more than 7 types of cancers (including primary and metastatic cancers) across the entire human body. LesionBench comprises 12,824 training CT scans and 3,035 testing CT scans from multiple centers worldwide. By collating 24 existing cancer datasets, while de-duplicating cross-dataset samples, we provide a unified, large-scale, and cancer-agnostic medical AI benchmark. We benchmark well-established segmentation models within the gold-standard nnU-Net auto-configuring framework, which includes CNNs, Transformers, and hybrid architectures on both cancer-agnostic training/evaluation settings, as well as on single-cancer settings (in- and out-of-distribution), enabling the assessment of classical segmentation metrics (DSC, NSD) and the clinically motivated SRVE. Our results show that lesion segmentation remains far from solved: the best model achieves a case-weighted DSC of 0.53 and a macro-averaged per-dataset DSC of 0.45. We further find that single-dataset specialists transfer poorly across datasets of the same organ, whereas cancer-agnostic pooling trades per-dataset precision for substantially improved robustness. To support future research in cancer-agnostic segmentation, we release the benchmark code, de-duplication pipeline, and data splits. LesionBench: `https://tinyurl.com/LesionBenchmark`

1 INTRODUCTION

Medical imaging is a cornerstone of clinical practice, and specifically in oncology, it serves as the de facto golden standard for cancer screening, diagnosis, follow-up, and imaging-guided radiotherapy and surgery. Although advancements in AI have delivered remarkable results in organ segmentation, primarily in Computed Tomography (CT) volumetric scans, only a very limited number of AI-based solutions are currently embedded in routine radiology workflows (Liu et al., 2019; Abràmoff et al., 2018; Isensee et al., 2021; Cao et al., 2023). When moving to the identification and segmentation on cancer-related lesions this number further decreases, as lesion segmentation is a far more complex task compared to organ delineation. Cancer lesions may present in very different regions of the body, can vary drastically in size, shape, and visual appearance when imaged with a CT scan. For these reasons, AI models still face significant challenges in reaching reliable and consistent results when compared to human radiologists on images coming from different hospitals, populations, and pathological conditions (Bassi et al., 2024; Huang et al., 2025; Svanera et al., 2024).

Considering the potential impact of allowing a faster delineation of regions of interest in medical scans, the medical imaging and the vision communities have come together to develop datasets and benchmarks to support advancements in the field. Initially, they mostly evaluated multi-organ segmentation tasks on primarily healthy subjects (Bassi et al., 2024; Sundar et al., 2022; Qu et al.,

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.

2023). For the multi-organ scenario, large annotated datasets have become increasingly available in the last five years. For instance, TotalSegmentator (Wasserthal et al., 2023) and AbdomenAtlas-8K (Qu et al., 2023) led the way by providing thousands of annotated CT volumes that allow for the development of AI models with promising results.

Following this path, and recognizing the positive impact that AI-enabled accurate and reliable cancer lesion segmentation could bring to clinical practice, additional efforts have been put into curating open-source oncology-related datasets. Indeed, we have witnessed the creation of multiple cancer-specific datasets that have highly advanced the medical image analysis field (Heller et al., 2023; He et al., 2021; 2020; Shao et al., 2011; 2012; Antonelli et al., 2022; Pedrosa et al., 2019; tri, 2025; Alves et al., 2026; Saeed et al., 2025; Luo et al., 2025; Küstner et al., 2025; de Grauw et al., 2025; Armato III et al., 2015; Aerts et al., 2014; Bakr et al., 2017; Saeed et al., 2025). While recognizing these important steps forward, most datasets remain limited in size, rarely surpassing hundreds of scans, limited to a single cancer type, and, when covering more than one cancer type, only a sub-portion includes 3D labels, with the majority consisting of a single 2D slice annotation per scan (de Grauw et al., 2025). Therefore, currently available benchmarks remain limited, causing non-uniform reporting of improvements that fails to translate in real-world application (Zhong & Wang, 2023; Zhao et al., 2022; Hatamizadeh et al., 2021), as focusing on a single cancer type pushes model development to over-tailoring architecture, processing, and losses to a single pathological representation, resulting in limited translation to different pathological expressions. This raises an open question: *how can we evaluate lesion segmentation methods in pathological settings without over-tailoring models to a particular training dataset or cancer site?* To develop the next generation of cancer-agnostic segmentation models, there is a clear need for broad, mixed, and structured 3D datasets containing CT scans spanning different Field of View (FoV), image resolutions, and imaging protocols that mimic real-world conditions. The high relevance of this question is further supported by the closely related work conducted independently and in parallel with our study, which currently provides a multi-cancer dataset and runs as a challenge for pancancer model development (pan, 2025).

To address this relevant question, we present the LesionBench benchmark, an effort to provide a unified, large-scale, and cancer-agnostic medical AI benchmark to support more structured advancements in the cancer segmentation field. It collates 24 datasets for a total of 12,824 training 3D CT scans. It also includes a reproducible test set of 3,035 3D CT scans assembled from public-source datasets via released scripts and fixed splits. It complements each scan with binary volumetric 3D lesion segmentation at the voxel level from 7+ cancer types across the body (including primary tumors and metastases). We treat all annotated lesions as a single foreground class (binary lesion vs background) to enable cancer-agnostic training and evaluation across sites. We release standardized splits and a fully automated pipeline to pre-process, de-duplicate, and reproduce the train/test sets, also useful for separate studies only focusing on a subset of the included datasets. We also benchmark three model architecture families within the medical imaging gold-standard nnU-Net auto-configuring framework (Isensee et al., 2021): CNNs, Transformer-based, and hybrid solutions. We evaluate them based on classical vision metrics: Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD), and the clinically motivated Symmetric Relative Volume Error (SRVE) (Kocak et al., 2025; Reinke et al., 2024) to identify common strengths and failure modes. The LesionBench benchmark contributions can be summarized as follows:

- A large scale cancer-agnostic dataset comprising 3D CT scans covering 7+ cancer types across the body, each associated with dense 3D label masks, spanning different imaging characteristics, acquisition protocols, and centers.
- A benchmark of well-established models within the gold-standard nnU-Net framework taken from CNN, Transformer, and hybrid architecture families, identifying strengths and failure modes and allowing for fair comparison.
- An empirical characterization of the specialist-generalist trade-off in lesion segmentation. Single-dataset specialists achieve strong in-domain accuracy but transfer poorly; pooled cancer-agnostic training largely eliminates these collapses but reduces in-domain accuracy. Neither regime is sufficient on its own, indicating a capability gap.
- A codebase to easily add new datasets to LesionBench, identify duplicate scans, remove them to prevent inflated results, and support single-cancer comparison as well as cancer-agnostic, along with pre-trained model weights.

Table 1: Datasets included in the LesionBench benchmark, reporting the number of CT scans in the training and test splits, the anatomical location of the lesion, the imaging field of view, the originating disease underlying the lesion, and the number of clinical centers for to each dataset. – indicates the number of centers or countries is currently unknown. Only fully 3D-labeled samples included. For full dataset names and further details, we refer the reader to Appendix A.1 Tables 5.

Dataset	# CT Scans Included		Origin	Lesion Location	Field of View	Disease Present	# Centers
	Train	Test					
KiTS23 Heller et al. (2023)	391	98	Challenge	Kidney	Abdominal	Renal Cancer	> 10
KiPA22 He et al. (2021; 2020) Shao et al. (2011; 2012)	56	14	Challenge	Kidney	Kidney Crop	Renal Cancer	–
MSD-Lu Antonelli et al. (2022)	50	13	Challenge	Lung	Thorax	NSCLC	–
NSCLC-RG Bakr et al. (2017)	115	29	TCIA	Lung	Thorax	NSCLC	2
NSCLC-RM Aerts et al. (2014)	289	71	TCIA	Lung	Thorax	NSCLC	1
LIDC Armato III et al. (2015)	685	173	TCIA	Lung	Thorax	Lung Cancer	7
LNDb Pedrosa et al. (2019)	150	41	Challenge	Lung	Thorax	Lung Cancer	1
MSD-Lv Antonelli et al. (2022)	93	25	Challenge	Liver	Abdominal	Liver/Colorectal/Breast/Lung Cancer	8
TRIALS tri (2025)	119	31	Challenge	Liver	Abdominal	Liver Cancer	2
CLiMet Simpson et al. (2024)	157	40	TCIA	Liver	Abdominal	Colon Cancer	1
HCC-TS Moawad et al. (2021)	70	18	TCIA	Liver	Abdominal	Liver Cancer	1
MSD-Co Antonelli et al. (2022)	100	26	Challenge	Colon	Abdominal	Colon Cancer	1
MSD-Pa Antonelli et al. (2022)	224	57	Challenge	Pancreas	Abdominal	Pancreatic Cancer/Mass	1
PANO Alves et al. (2026)	534	142	Challenge	Pancreas	Abdominal	Pancreatic Cancer	5
HECKTOR Saeed et al. (2025)	544	136	Challenge	Head & Neck	Head & Neck	Head & Neck	10
SegRap Luo et al. (2025)	96	24	Challenge	Head & Neck	Head & Neck	Head & Neck	1
HN-Cet Bosch et al. (2015)	25	7	TCIA	Head & Neck	Head & Neck	Head & Neck	–
Head-HN1 Wee & Dekker (2019)	109	28	TCIA	Head & Neck	Head & Neck	Head & Neck	2
RADCURE Welch et al. (2023)	2372	584	TCIA	Head & Neck	Head & Neck	Head & Neck	1
HNSCC Grossberg et al. (2020)	384	96	TCIA	Head & Neck	Head & Neck	Head & Neck	1
HN-PET Vallieres et al. (2017)	136	38	TCIA	Head & Neck	Head & Neck	Head & Neck	4
AdACC Moawad et al. (2023)	42	11	TCIA	Adrenal Gland	Abdominal	Adrenal Cancer	1
Auto-PET Küstner et al. (2025)	829	211	Challenge	Multi-Location	Total-Body	Skin/Lymphatic/Lung Cancer	2
ULS23 de Grauw et al. (2025)	5100	1276	Challenge	Multi-Location	Tight Lesion Crop	High Mix	> 10
LesionBench (ours)	**12824**	**3035**	**Challenge/TCIA**	**Multi-Location**	**Mixed**	**High Mix**	> 10

2 LESIONBENCH BENCHMARK

2.1 DATASETS INCLUDED - STATISTICS, ANNOTATIONS, AND DUPLICATE HANDLING

We have selected 24 datasets covering more than 7 lesions locations and cancer types by searching among open-source or available for research purposes datasets. Table 1 shows the complete list of included datasets and their characteristics; Appendix A.1 Table 5 reports the detailed references for data access and licenses. As detailed in Table 1, LesionBench offers 15,859 fully 3D-labeled CT scans with variable Fields of View (FoV), lesion locations, and cancer types, which well reflect the real clinical scenario, where different imaging protocols, areas of interest, and scanners can be used daily. To support single-task benchmarking, we have created controlled splits for each of the selected datasets (Appendix A.5), and then collated them in the overall LesionBench benchmark dataset (12,824/3,035 train/test, respectively), which is used here to benchmark models on cancer-agnostic lesion segmentation.

Where available, patient demographic information was collected and summarized in Figure 1. Of the 24 datasets comprising LesionBench, 14 contained some form of patient-level metadata. Across these, age and sex were the most consistently reported variables; racial demographic information and scanner manufacturer details are additionally reported where available. As shown in Figure 1, most of the included patients' age range between 50 and 80, with more than twice as many male as female patients. When looking at less represented metadata parameters, Siemens and Philips are the most commonly employed scans. For ethnicity, although rarely reported, white is the most common.

2.1.1 ANNOTATIONS RECONCILIATION

To construct a unified benchmark, we need to consider the heterogeneity in the number and composition of segmentation targets across single datasets. The number of annotated structures per volume ranges from a single lesion to up to 30 targets in datasets such as SegRap2023 (Luo et al., 2025). As this benchmark focuses exclusively on lesion segmentation performance, retaining auxiliary segmentation targets during training introduces the risk that a model optimizes for non-lesion structures at the expense of lesion detection accuracy, thereby confounding the evaluation of the target task. Therefore, for each patient, all segmentation masks were standardized (per-dataset lesion retention strategy in Appendix A.3 Table 7) by removing non-lesion structures, thus containing only lesion-

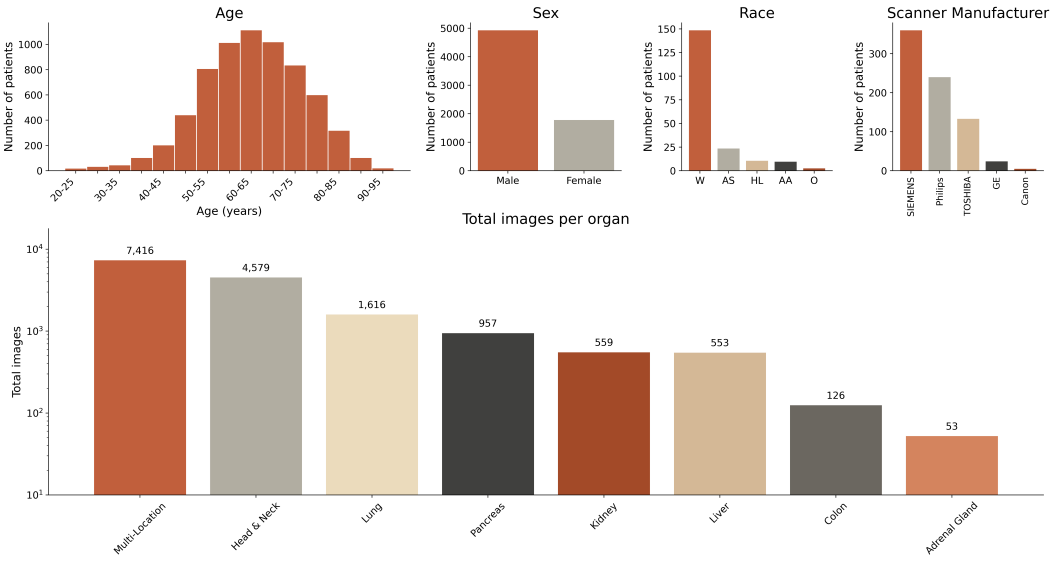


Figure 1: Demographic and dataset characteristics, where available. Age and sex were highly available across datasets. Race and scanner manufacturer were considerably scarcer. *Top row, left to right:* Age across all datasets in 5-year bins. Sex distribution; note one patient in KiTS23 is transgender female and has been excluded from the chart but is included in all analyses. Racial breakdown where: (W) White, (AS) Asian, (HL) Hispanic/Latino, (AA) African American, (O) Other. CT scanner manufacturer. *Bottom:* Total number of images per anatomical region across all datasets.

relevant targets prior to training and evaluation. Additionally, as lesion label values are not consistent across datasets, reflecting differing annotation conventions, all retained lesion voxels were remapped to a consistent foreground value of 1, with all remaining voxels assigned to background class 0. This standardization ensures consistent binary labeling across all datasets, eliminating label ambiguity.

Where each volume has been independently annotated by multiple clinicians inter-observer variability may manifest at two distinct levels: (i) disagreement on the presence or number of target lesions within a given volume; (ii) where annotators agree on the presence of a lesion, the precise delineation of its boundary may be inconsistent, introducing spatial ambiguity in the final segmentation. Both effects compound the uncertainty in what constitutes the correct reference mask. For datasets exhibiting this behavior, a conservative consensus strategy was adopted with the aim of maximizing true positive retention. Lesion targets were retained only where a majority of annotators agreed on their presence, reducing the influence of spurious or uncertain detections whilst preserving lesions with high inter-rater concordance. To reconcile the remaining per-annotator delineations into a single reference mask, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Warfield et al., 2004) was applied. A threshold of 0.5 was applied to the resulting probabilistic map to yield a binary consensus mask used as the reference standard for all subsequent evaluations.

2.1.2 Duplicate Handling

Within the medical imaging community, dataset challenges frequently sample from pre-existing collections; for example, the Medical Segmentation Decathlon liver task (Antonelli et al., 2022) is derived from LiTS (Bilic et al., 2023). As each respective dataset in our benchmark was split independently at the patient level, it is plausible that a scan appears in the training partition of one dataset and the test partition of another. Left unaddressed, such data leakage would artificially inflate performance metrics under both in-distribution and cancer-agnostic evaluation settings. To mitigate this, a voxel-level de-duplication pipeline was applied across all datasets prior to evaluation.

All volumes were first standardized into a common space: RAS canonical oriented, trilinear interpolation resampled to 2mm isotropic resolution, and intensity-normalized to $[0, 1]$ through clipping at the 0.5th and 99.5th percentiles. Two complementary representations were then extracted from

each volume. A 13-dimensional *statistical fingerprint* was computed, comprising 9 global intensity statistics (mean, std, median, min, max, and the 5th, 25th, 75th, and 95th percentiles) concatenated with a four-bin normalized intensity histogram. A *spatial hash* was obtained by downsampling each volume to a fixed 128^3 voxel grid and flattening the result into a vector of $\sim$ 2.1M values; to avoid exhausting system memory, spatial hashes were written to a memory-mapped file and loaded on demand during comparison. Duplicate detection proceeded in three cascaded stages of increasing computational cost. First, statistical fingerprints were z-score normalized, and a ball tree was used to retrieve all cross-split candidate pairs (one training and one testing volume) within a Euclidean radius of 2.0, reducing the search from $\mathcal{O}(n^2)$ to a tractable candidate set. Candidate pairs were subjected to a shape pre-filter, discarding any pair whose resampled dimensions differed by more than five voxels along any axis. Surviving pairs were finally confirmed by computing the cosine similarity between their spatial hash vectors; pairs with similarity ≥ 0.98 were flagged as duplicates.

Our de-duplication pipeline identified over 3,000 candidate duplicate images at a similarity threshold of 0.98 (Appendix A.2 Table 6), with head and neck and total-body images disproportionately represented owing to their high structural similarity under the spatial hash (Appendix A.2 Figures 2 and 3). Manual inspection was performed on all candidate images that appear in at least one train-test pair with a similarity ≥ 0.995 (963 candidate images; full methodology in Appendix A.2), of which 154 images were confirmed as leaks and reassigned from the test set to the training set to prevent data leakage.

2.2 Evaluation Framework - Selected Models, Experiments, and Metrics

We surveyed the architectures employed by state-of-the-art medical image segmentation benchmarks (Isensee et al., 2024; Bassi et al., 2024) and selected a subset spanning three architectural families: convolutional, transformer-based, and hybrid approaches. All models were implemented within the medical imaging gold-standard auto-configuring nnU-Net framework for fair comparison. We included nnU-Net (Isensee et al., 2021) in its standard configuration, alongside the nnU-Net ResEnc M, L, and XL variants, which incorporates residual connections in the encoder as CNNs. We evaluate Primus-v3 in M and L presets as transformer (Wald et al., 2025), and MedNeXt (Roy et al., 2023) in M and L configurations as hybrid, which adapt ConvNeXt blocks within a U-Net-style architecture to combine convolutions inductive biases with transformer-inspired principles. We benchmarked each model in the LesionBench cancer-agnostic settings, but we also evaluated each model separately trained on each of the 24 datasets to set performance upper bounds. Inference experiments were conducted on a high-performance cluster. To comply with data governance restrictions, the LesionBench dataset was held in an isolated silo. All experiments were executed for reproducibility within a Docker container, on an NVIDIA Ampere A100 GPU with 40GB of memory, with 12 CPU cores and 128GB of system memory per experiment. We evaluate segmentation performance using two complementary metrics: the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD), described in Appendix A.4. While DSC captures regional agreement, it is relatively insensitive to localized boundary errors; NSD reports the fraction of each surface that lies within a given distance of the other. Although both DSC and NSD are widely employed computer vision metrics, they are known to hide clinical failure modes when looked in isolation. Therefore, we also evaluated volume variations through the Symmetric Relative Volume Error (SRVE): $\mathrm{SRVE} = \frac{|V_{\mathrm{pred}} - V_{\mathrm{GT}}|}{V_{\mathrm{pred}} + V_{\mathrm{GT}}} \times 100\%$, where V_{pred} and V_{GT} are the predicted and ground-truth lesion volumes. SRVE is bounded in $[0, 100]$, with 0 meaning a perfect volume match and 100 a complete miss. Since every ground-truth label contains a lesion, an SRVE of 100 indicates the model produced an empty prediction.

3 Benchmark Results

3.1 Cancer Agnostic Evaluation

Table 2 reports the cancer-agnostic models' global and macro-averaged per-dataset DSC scores on the LesionBench test set, along with their breakdown into anatomical regions and single datasets (boxplots in Appendix A.7 Figures 4 and 5). Both global and macro DSC values confirm the difficulty of the lesion segmentation task when compared to the near-ceiling performance for organ segmentation (high and low DSC visual examples in Appendix A.7 Figures 13 and 14) as models'

Table 2: **Evaluation on LesionBench (#3,035) DSC** ($\mu \pm \sigma$) averaged over LesionBench test set, macro-averaged per dataset, and the DSC breakdown per cancer location and the per dataset one. All the evaluated models have been trained on the LesionBench training set (#12,824) for 2k epochs.

Model	Params	Epoch Speed (s)	Global	Macro	Kidney	Lung	Liver	Colon	Pancreas	Head & Neck	Adrenal	Multi-Loc
nnUNet	88.43M	41.54	0.23 ± 0.28	0.27 ± 0.17	0.41 ± 0.33	0.21 ± 0.26	0.20 ± 0.27	0.12 ± 0.18	0.11 ± 0.16	0.28 ± 0.23	0.58 ± 0.28	0.21 ± 0.30
ResEnc-M	383.24M	38.51	0.31 ± 0.29	0.34 ± 0.16	0.54 ± 0.31	0.31 ± 0.30	0.27 ± 0.29	0.19 ± 0.22	0.19 ± 0.22	0.30 ± 0.23	0.64 ± 0.27	0.32 ± 0.32
ResEnc-L	527.4M	96.67	0.48 ± 0.31	0.43 ± 0.17	0.58 ± 0.29	0.35 ± 0.30	0.38 ± 0.32	0.22 ± 0.25	0.28 ± 0.26	0.45 ± 0.28	0.79 ± 0.19	0.56 ± 0.30
ResEnc-XL	527.4M	163.36	0.53 ± 0.30	0.45 ± 0.17	0.63 ± 0.29	0.44 ± 0.31	0.39 ± 0.33	0.26 ± 0.26	0.33 ± 0.27	0.47 ± 0.28	0.80 ± 0.18	0.60 ± 0.28
MedNeXt-M	17.55M	468.32	0.29 ± 0.29	0.32 ± 0.16	0.39 ± 0.33	0.29 ± 0.28	0.26 ± 0.30	0.12 ± 0.17	0.14 ± 0.19	0.28 ± 0.23	0.60 ± 0.28	0.31 ± 0.31
MedNeXt-L	61.78M	481.12	0.31 ± 0.29	0.33 ± 0.15	0.41 ± 0.32	0.30 ± 0.28	0.29 ± 0.31	0.17 ± 0.20	0.17 ± 0.20	0.29 ± 0.23	0.59 ± 0.32	0.33 ± 0.32
Primus-M	244.51M	53.92	0.25 ± 0.29	0.30 ± 0.15	0.44 ± 0.33	0.25 ± 0.30	0.21 ± 0.28	0.11 ± 0.18	0.17 ± 0.21	0.34 ± 0.26	0.52 ± 0.35	0.21 ± 0.30
Primus-L	428.29M	81.94	0.26 ± 0.29	0.30 ± 0.15	0.43 ± 0.32	0.26 ± 0.29	0.21 ± 0.28	0.11 ± 0.18	0.15 ± 0.19	0.32 ± 0.26	0.51 ± 0.30	0.22 ± 0.30

Model	KiTS23	KiPA22	MSD Lung	NSCLC Radgen	NSCLC Rad	LIDC-IDRI	LNDb	MSD Liver	TRIALS	CLiMet	HCC-TS	MSD Co
nnUNet	0.36 ± 0.30	0.78 ± 0.23	0.35 ± 0.22	0.42 ± 0.22	0.39 ± 0.31	0.12 ± 0.20	0.05 ± 0.11	0.17 ± 0.27	0.10 ± 0.16	0.19 ± 0.26	0.39 ± 0.33	0.12 ± 0.18
ResEnc-M	0.50 ± 0.30	0.82 ± 0.24	0.36 ± 0.11	0.53 ± 0.23	0.44 ± 0.32	0.24 ± 0.28	0.15 ± 0.22	0.19 ± 0.28	0.20 ± 0.25	0.29 ± 0.28	0.43 ± 0.33	0.19 ± 0.22
ResEnc-L	0.55 ± 0.28	0.84 ± 0.24	0.36 ± 0.20	0.57 ± 0.23	0.48 ± 0.31	0.30 ± 0.29	0.17 ± 0.25	0.22 ± 0.30	0.29 ± 0.31	0.42 ± 0.29	0.56 ± 0.32	0.22 ± 0.25
ResEnc-XL	0.60 ± 0.28	0.83 ± 0.24	0.44 ± 0.22	0.60 ± 0.21	0.49 ± 0.31	0.42 ± 0.32	0.28 ± 0.30	0.25 ± 0.29	0.26 ± 0.29	0.49 ± 0.30	0.61 ± 0.31	0.26 ± 0.28
MedNeXt-M	0.32 ± 0.28	0.84 ± 0.24	0.41 ± 0.15	0.54 ± 0.25	0.41 ± 0.29	0.20 ± 0.24	0.22 ± 0.27	0.20 ± 0.31	0.20 ± 0.26	0.26 ± 0.26	0.45 ± 0.33	0.12 ± 0.17
MedNeXt-L	0.36 ± 0.28	0.81 ± 0.26	0.44 ± 0.18	0.49 ± 0.25	0.38 ± 0.30	0.23 ± 0.26	0.23 ± 0.24	0.22 ± 0.32	0.23 ± 0.29	0.29 ± 0.28	0.47 ± 0.34	0.17 ± 0.20
Primus-M	0.41 ± 0.32	0.65 ± 0.36	0.48 ± 0.22	0.48 ± 0.26	0.43 ± 0.33	0.14 ± 0.23	0.12 ± 0.24	0.18 ± 0.28	0.14 ± 0.23	0.21 ± 0.27	0.38 ± 0.33	0.11 ± 0.18
Primus-L	0.39 ± 0.31	0.71 ± 0.26	0.50 ± 0.20	0.50 ± 0.25	0.39 ± 0.33	0.17 ± 0.23	0.16 ± 0.26	0.19 ± 0.30	0.12 ± 0.19	0.21 ± 0.26	0.36 ± 0.33	0.11 ± 0.18

Model	MSD Pa	Pano	HECKTOR	SegRap	HN-Cet	HN-HN1	RADCURE	HNSCC	HN-PET	AdACC	AutoPet	ULS23
nnUNet	0.16 ± 0.21	0.09 ± 0.14	0.22 ± 0.19	0.23 ± 0.10	0.33 ± 0.15	0.27 ± 0.27	0.29 ± 0.24	0.25 ± 0.24	0.39 ± 0.26	0.58 ± 0.28	0.09 ± 0.14	0.23 ± 0.31
ResEnc-M	0.29 ± 0.27	0.18 ± 0.22	0.29 ± 0.19	0.33 ± 0.11	0.36 ± 0.13	0.30 ± 0.30	0.30 ± 0.24	0.26 ± 0.23	0.41 ± 0.26	0.64 ± 0.27	0.13 ± 0.16	0.35 ± 0.33
ResEnc-L	0.44 ± 0.28	0.23 ± 0.23	0.38 ± 0.26	0.39 ± 0.11	0.37 ± 0.14	0.42 ± 0.32	0.46 ± 0.28	0.51 ± 0.30	0.52 ± 0.24	0.79 ± 0.19	0.18 ± 0.19	0.61 ± 0.27
ResEnc-XL	0.43 ± 0.31	0.30 ± 0.25	0.39 ± 0.27	0.41 ± 0.11	0.32 ± 0.17	0.36 ± 0.34	0.49 ± 0.27	0.54 ± 0.30	0.48 ± 0.27	0.80 ± 0.18	0.17 ± 0.19	0.68 ± 0.23
MedNeXt-M	0.20 ± 0.25	0.13 ± 0.16	0.24 ± 0.20	0.35 ± 0.11	0.30 ± 0.16	0.27 ± 0.27	0.28 ± 0.23	0.26 ± 0.24	0.36 ± 0.24	0.60 ± 0.28	0.12 ± 0.15	0.34 ± 0.32
MedNeXt-L	0.26 ± 0.25	0.15 ± 0.18	0.24 ± 0.19	0.38 ± 0.09	0.28 ± 0.16	0.26 ± 0.27	0.29 ± 0.24	0.28 ± 0.25	0.36 ± 0.25	0.59 ± 0.32	0.13 ± 0.15	0.36 ± 0.33
Primus-M	0.21 ± 0.26	0.15 ± 0.19	0.26 ± 0.21	0.28 ± 0.12	0.33 ± 0.12	0.35 ± 0.32	0.35 ± 0.26	0.34 ± 0.29	0.43 ± 0.26	0.52 ± 0.35	0.13 ± 0.16	0.22 ± 0.31
Primus-L	0.18 ± 0.24	0.14 ± 0.18	0.23 ± 0.21	0.28 ± 0.14	0.28 ± 0.14	0.33 ± 0.31	0.34 ± 0.27	0.33 ± 0.29	0.39 ± 0.28	0.51 ± 0.30	0.11 ± 0.15	0.24 ± 0.32

performance ranges from 0.23 and 0.27 (nnU-Net) to 0.53 and 0.45 (the best performing ResEnc-XL model). For example, Wasserthal et al. (2023) report a DSC > 0.90 on a large range of organ targets. Although lesion segmentation does not appear to be a solved problem even under large-scale multi-institutional training, pure CNN models seem to deliver higher performance when compared to transformer- and hybrid-based ones. However, no single architecture family dominates outright. ResEnc variants occupy the top three positions by global DSC: ResEnc-XL (0.53), ResEnc-L (0.48) with ResEnc-M (0.31) and MedNeXt-L (0.31) tying for third place. This trend is also reflected in the macro-averaged DSC, which shows a more balanced distribution across all models, down-weighting the impact of ULS23.

Breaking down the results by anatomical region, the model performance trends remain similar, even though locations such as kidney, H&N, lung, and liver deliver higher performance compared to colon and pancreas. These results reflect the clinical complexity of colon and pancreatic cancer identification. Adrenal scores highest of all (0.51/0.80), but with only 11 held-out cases and no out-of-distribution (OOD) comparison, we feel it is unfair to compare it alongside the other lesion regions. Furthermore, colon and pancreas are also the two organs where pooling costs the most relative to training a specialist model. Comparing the same architecture in both regimes: a ResEnc-XL specialist trained only on MSD-Colon reaches 0.43 DSC in-domain (Table 4), while the pooled ResEnc-XL reaches only 0.26 on the same colon cases (Table 2). The pattern repeats for pancreas, where the ResEnc-XL specialist reaches 0.52 DSC across its in-domain/OOD dataset (Table 4) versus 0.33 for the pooled model (Table 2). In both cases, pooling roughly halves achievable DSC relative to a model that only ever had to learn that one lesion location, while on head and neck datasets, the same model improves DSC when pooled (0.47) compared to its specialist version trained on RADCURE when averaging DSC cross its in-domain/OOD cases (0.43).

At the dataset level, models trained on the LesionBench training set achieve lower dataset-specific DSCs compared to their specialist counterparts for most of the sources (Table 2 and Appendix A.7 Table 9). As for the location analysis, the best performing ResEnc-XL model on head and neck datasets (e.g., RADCURE, Head-Neck-PET-CT, and HNSCC) shows negligible reductions. From our analysis, there is no clear pattern in terms of dataset size or cancer location to predict how the generalist model will perform compare to its specialist configuration.

Complete-miss rates vary considerably across sources (Appendix A.7 Figures 6 and 7). Colon is the most challenging location, with miss rates of 38-65% for every model, while ResEnc-L and ResEnc-XL consistently achieve the lowest miss rates across both datasets and lesion locations. On cases where a lesion was detected, median recall is comparable across architectures (0.59-0.74), whereas

Table 3: **Evaluation on LesionBench (#3,035) SRVE** ($\mu \pm \sigma$), the lower the better. SRVE is bounded in $[0, 100]\%$: 0 is a perfect volume match and 100 a complete miss. Arrows denote net over- (↑) or under-prediction (↓). We report the SRVE averaged over the entire LesionBench test set, macro-averaged per dataset, and the SRVE breakdown per lesion location and per dataset. All the evaluated models have been trained on the LesionBench training set (#12,824) for 2k epochs.

Model	Params	Epoch Speed (s)	Global	Macro	Kidney	Lung	Liver	Colon	Pancreas	Head & Neck	Adrenal	Multi-Loc
nnU-Net	88.4M	41.54	54.63 ± 32.50 ↑	49.19 ± 17.17 ↑	40.18 ± 29.42 ↑	61.67 ± 33.23 ↑	55.55 ± 32.11 ↓	53.66 ± 27.51 ↑	73.87 ± 23.06 ↑	51.36 ± 29.27 ↑	18.56 ± 21.16 ↑	54.18 ± 34.07 ↑
ResEnc-M	383.2M	38.51	49.32 ± 31.98 ↑	45.15 ± 14.66 ↑	30.73 ± 25.62 ↑	50.75 ± 34.22 ↑	51.22 ± 30.97 ↓	46.83 ± 30.68 ↑	65.28 ± 27.21 ↑	52.48 ± 28.46 ↑	19.68 ± 19.56 ↑	47.00 ± 33.30 ↑
ResEnc-L	527.4M	96.67	36.64 ± 31.00 ↑	41.35 ± 15.73 ↑	31.05 ± 28.12 ↑	51.08 ± 34.60 ↑	48.67 ± 33.66 ↑	60.77 ± 24.32 ↑	59.12 ± 31.28 ↑	34.26 ± 27.72 ↑	18.23 ± 19.29 ↑	31.61 ± 29.55 ↑
ResEnc-XL	527.4M	163.36	32.21 ± 29.08 ↑	37.82 ± 14.18 ↑	26.29 ± 26.25 ↑	40.18 ± 32.85 ↑	44.03 ± 34.31 ↓	50.06 ± 25.72 ↑	53.65 ± 31.22 ↑	33.79 ± 27.58 ↑	16.47 ± 18.79 ↑	26.56 ± 26.55 ↑
MedNeXt-M	17.55M	468.32	53.22 ± 32.15 ↑	48.22 ± 17.44 ↑	46.21 ± 30.82 ↑	57.66 ± 33.38 ↑	53.15 ± 33.09 ↓	56.10 ± 28.17 ↑	74.42 ± 23.95 ↑	54.02 ± 28.39 ↑	23.13 ± 20.32 ↑	50.18 ± 33.68 ↑
MedNeXt-L	61.78M	481.12	52.28 ± 32.24 ↑	48.31 ± 15.96 ↑	47.16 ± 30.43 ↑	57.95 ± 33.11 ↑	50.42 ± 33.15 ↓	55.50 ± 26.71 ↑	71.53 ± 24.76 ↑	53.51 ± 28.90 ↑	23.64 ± 22.18 ↑	48.90 ± 33.71 ↑
Primus-M	244.51M	53.92	51.02 ± 33.06 ↑	44.67 ± 12.66 ↓	29.37 ± 26.56 ↑	51.37 ± 33.29 ↑	48.91 ± 33.25 ↓	50.54 ± 24.70 ↑	63.51 ± 28.28 ↑	44.05 ± 29.06 ↑	26.99 ± 26.20 ↑	55.57 ± 34.62 ↑
Primus-L	428.29M	81.94	49.98 ± 33.02 ↑	45.32 ± 12.87 ↑	32.39 ± 26.64 ↑	51.54 ± 32.97 ↑	51.43 ± 32.58 ↓	53.70 ± 28.01 ↑	66.32 ± 28.19 ↑	42.52 ± 28.61 ↑	35.49 ± 22.43 ↑	53.41 ± 34.96 ↑

Model	KiTS23	KiPA22	MSD Lung	NSCLC Radgen	NSCLC Rad	LIDC-IDRI	LNDb	MSD Liver	TRIALS	C_Liver Met	HCC-TACE-Seg	MSD Colon
nnU-Net	43.37 ± 28.67 ↑	17.86 ± 24.46 ↓	38.05 ± 31.37 ↑	31.15 ± 26.87 ↓	42.66 ± 30.42 ↑	72.97 ± 28.48 ↑	77.21 ± 26.60 ↑	61.99 ± 33.98 ↑	69.34 ± 23.67 ↓	49.65 ± 32.79 ↑	39.82 ± 27.99 ↓	53.66 ± 27.51 ↑
ResEnc-M	33.25 ± 25.31 ↑	13.09 ± 20.35 ↓	38.64 ± 23.68 ↑	30.69 ± 25.70 ↓	36.25 ± 28.29 ↑	59.11 ± 34.40 ↑	59.38 ± 35.47 ↑	61.94 ± 34.12 ↑	59.10 ± 28.74 ↓	46.38 ± 26.56 ↑	35.69 ± 29.80 ↓	46.83 ± 30.68 ↑
ResEnc-L	33.69 ± 27.52 ↑	12.57 ± 25.17 ↑	40.39 ± 20.47 ↑	24.54 ± 21.28 ↓	36.16 ± 27.91 ↑	58.52 ± 34.95 ↑	69.45 ± 33.33 ↑	63.56 ± 33.76 ↑	56.71 ± 33.71 ↓	43.80 ± 29.59 ↑	27.17 ± 27.79 ↓	60.77 ± 24.32 ↑
ResEnc-XL	28.43 ± 26.13 ↑	11.31 ± 21.83 ↑	26.91 ± 25.67 ↑	25.78 ± 17.08 ↓	36.00 ± 29.51 ↑	42.75 ± 34.47 ↑	51.88 ± 36.11 ↑	58.42 ± 33.79 ↑	59.31 ± 32.24 ↑	34.70 ± 31.17 ↑	22.69 ± 24.64 ↓	50.06 ± 25.72 ↑
MedNeXt-M	51.77 ± 28.78 ↑	7.33 ± 8.47 ↓	36.65 ± 22.58 ↑	28.38 ± 22.18 ↓	47.14 ± 27.99 ↑	67.97 ± 32.44 ↑	59.65 ± 34.59 ↑	63.70 ± 34.99 ↑	61.21 ± 28.36 ↓	52.79 ± 30.31 ↑	27.65 ± 28.73 ↓	56.10 ± 28.17 ↑
MedNeXt-L	51.67 ± 28.16 ↑	15.61 ± 26.84 ↓	32.03 ± 27.60 ↑	31.57 ± 27.02 ↓	48.41 ± 28.18 ↑	67.14 ± 32.10 ↑	62.59 ± 31.98 ↑	62.07 ± 37.03 ↑	54.07 ± 31.07 ↓	50.56 ± 28.50 ↑	28.64 ± 29.46 ↓	55.50 ± 26.71 ↑
Primus-M	29.51 ± 25.02 ↑	28.39 ± 35.47 ↓	26.51 ± 18.20 ↓	26.11 ± 22.13 ↓	37.87 ± 28.87 ↑	60.84 ± 31.74 ↑	61.13 ± 35.73 ↑	56.80 ± 36.34 ↑	58.00 ± 30.51 ↓	40.81 ± 31.61 ↑	42.83 ± 30.11 ↓	50.54 ± 24.70 ↑
Primus-L	34.02 ± 27.20 ↑	20.96 ± 18.62 ↓	28.35 ± 17.58 ↓	29.08 ± 24.29 ↓	39.58 ± 29.68 ↑	61.19 ± 31.43 ↑	54.75 ± 35.98 ↑	54.71 ± 36.21 ↑	60.64 ± 29.37 ↓	49.12 ± 30.67 ↑	38.70 ± 30.95 ↓	53.70 ± 28.01 ↑

Model	MSD Pancreas	PANORAMA	HECKTOR	SegRap2023	H&N Cetuximab	H&N-Rad-HN1	RADCURE	HNSCC	H&N-PET-CT	Adrenal-ACC	AUTO-PET	ULS23
nnU-Net	59.11 ± 27.44 ↑	78.47 ± 19.34 ↑	49.16 ± 28.38 ↑	45.34 ± 19.39 ↓	23.53 ± 12.40 ↓	59.39 ± 33.46 ↑	52.59 ± 29.37 ↑	56.48 ± 27.78 ↑	24.58 ± 21.16 ↑	18.56 ± 21.16 ↑	62.71 ± 32.54 ↑	52.90 ± 34.11 ↑
ResEnc-M	50.43 ± 28.51 ↑	69.90 ± 25.06 ↑	50.46 ± 25.72 ↑	36.90 ± 19.49 ↓	30.58 ± 18.15 ↓	61.07 ± 32.68 ↑	53.89 ± 28.95 ↑	57.98 ± 25.87 ↑	26.43 ± 20.76 ↑	19.68 ± 19.56 ↑	60.93 ± 33.24 ↑	44.91 ± 32.80 ↑
ResEnc-L	38.05 ± 29.70 ↑	65.68 ± 28.76 ↑	43.23 ± 29.38 ↑	49.57 ± 12.96 ↓	39.03 ± 20.49 ↓	44.47 ± 33.96 ↑	33.00 ± 28.08 ↑	25.47 ± 19.92 ↑	22.93 ± 18.67 ↑	18.23 ± 19.29 ↑	57.81 ± 34.35 ↑	27.69 ± 26.62 ↑
ResEnc-XL	39.09 ± 30.39 ↑	58.19 ± 30.06 ↑	42.96 ± 30.61 ↑	47.19 ± 15.13 ↓	41.18 ± 28.27 ↓	52.70 ± 36.37 ↑	32.12 ± 26.98 ↑	23.82 ± 19.28 ↑	25.68 ± 20.78 ↓	16.47 ± 18.79 ↑	58.28 ± 33.79 ↑	21.81 ± 21.59 ↓
MedNeXt-M	60.49 ± 26.30 ↑	78.76 ± 21.39 ↑	58.14 ± 25.60 ↑	28.11 ± 15.48 ↓	26.05 ± 8.98 ↓	65.39 ± 31.57 ↑	55.12 ± 28.69 ↑	53.77 ± 26.97 ↑	33.43 ± 22.27 ↑	23.13 ± 20.32 ↑	66.99 ± 31.45 ↑	47.67 ± 33.27 ↑
MedNeXt-L	58.58 ± 26.11 ↑	75.57 ± 22.88 ↑	59.41 ± 25.44 ↑	27.33 ± 14.19 ↓	36.34 ± 16.90 ↑	66.95 ± 30.20 ↑	54.23 ± 29.36 ↑	52.25 ± 27.58 ↑	30.94 ± 22.82 ↑	23.64 ± 22.18 ↑	68.24 ± 31.20 ↑	46.01 ± 33.11 ↑
Primus-M	51.57 ± 29.16 ↑	67.23 ± 26.94 ↑	42.11 ± 28.61 ↑	49.16 ± 18.25 ↓	41.53 ± 15.10 ↓	49.12 ± 36.75 ↑	45.49 ± 29.62 ↑	42.26 ± 26.51 ↑	23.06 ± 19.04 ↑	26.99 ± 26.20 ↑	59.29 ± 33.43 ↑	55.02 ± 34.76 ↑
Primus-L	50.35 ± 30.71 ↑	71.30 ± 25.38 ↑	44.43 ± 28.98 ↑	47.65 ± 17.75 ↓	47.01 ± 16.75 ↓	49.58 ± 37.02 ↑	43.20 ± 28.87 ↑	39.22 ± 25.81 ↑	19.66 ± 17.86 ↑	35.49 ± 22.43 ↑	62.98 ± 32.66 ↑	51.98 ± 35.06 ↑

median precision varies roughly three times as widely (0.23-0.68; Appendix A.7 Figure 8). This suggests that models locate lesions at similar rates, but differ in how accurately they delineate them.

Table 3 reports the evaluation in terms of volume variations compared to gold standard (boxplots in Appendix A.7 Figures 9 and 10). SRVE shows a broadly consistent pattern of over-segmentation with seven of the eight organ groups showing a net positive volume bias, meaning models tend to predict more lesion volume than is annotated. Liver is the exception, with a net negative bias, driven primarily by HCC-TACE-Seg and TRIALS. We suspect this reflects under-detection rather than systematic under-sizing of correctly located lesions. Qualitative review of TRIALS suggests that a subset of lesions are genuinely difficult to localize within the liver. At the organ level, no organ shows a catastrophic SRVE score in the way some datasets do at the DSC level (or NSD in Appendix A.7 Table 8). On average, Adrenal (22.8%) and kidney (35.4%) are the most volumetrically accurate, while the pancreas is the hardest (66.0%). KiPA22 and Adrenal-ACC are the strongest performers, consistent with their tight crops, while PANORAMA is the single worst dataset for volume accuracy (70.6%). ULS23 is a partial exception to the crop hypothesis. Despite ranking second-best for NSD (Appendix A.7 Table 8), it ranks only eleventh of 24 datasets for SRVE (43.5%). This suggests that a lesion-centered crop helps a model locate the correct boundary but does not by itself guarantee that the predicted lesion volume matches the annotated burden. This hypothesis is also directly supported by the NSD (Appendix A.7 Table 8 Figure 11 and 12), where KiPA22 and ULS23, two datasets with crops around the target lesion, achieve the highest NSD of any source in LesionBench (0.57 and 0.41, respectively), while AutoPET and PANORAMA, both whole-body or thoracic-abdomen acquisitions, are the worst (0.06 and 0.09). Within the lung datasets, NSD is not uniform. NSCLC-Radiogenomics and MSD-Lung, which contain larger lesions, achieve markedly higher NSD than the small-lesion cohorts LIDC-IDRI and LNDb, mirroring the same primary-lesion-versus-small-lesion split observed in the specialist transfer results. Together, these results are consistent with FoV, rather than organ identity alone, being a key factor in boundary-level segmentation quality.

Finally, when compared to DSC, NSD is much lower, which may be caused by false-positive regions in the considered scan that disproportionately penalize NSD, since even a small spurious prediction far from the true lesion boundary pulls the surface-distance score down sharply. Colon and pancreas remain the hardest on both metrics, consistent with the DSC results already reported. This further supports the FoV hypothesis, where lesion-centered crops leave little room for false-positive predictions elsewhere in the volume that damage NSD on wider-FoV scans.

3.2 Per Dataset Evaluation

In the specialist setting, we report the DSC ResEnc-XL results in Table 4, being the best-performing solution in the cancer-agnostic settings. All the other configurations are reported in Appendix A.7

Table 4: **ResEnc-XL per dataset evaluation DSC** ($\mu \pm \sigma$). For each ResEnc-XL trained model we report the in distribution evaluation (test sets from the same training dataset) and out of distribution evaluation (same cancer type, but test set from a different dataset), when available. Results reported only for same-organ targets, empty otherwise as not meaningful. All models trained for 1k epochs.

Model	Params	Kidney −− Test		Lung −− Test					Liver −− Test				Colon −− Test
		KiTS23	KiPA22	MSD-Lu	NSCLC-RG	NSCLC-RM	LIDC	LNDb	MSD-Lv	TRIALS	CLiMet	HCC-TS	MSD-Co
ResEnc-XL-KiTS23	527.70M	0.85 ± 0.17	0.00 ± 0.00										
ResEnc-XL-KiPA22	383.47M	0.00 ± 0.00	0.91 ± 0.08										
ResEnc-XL-MSD-Lu	527.70M			0.58 ± 0.34	0.69 ± 0.26	0.39 ± 0.34	0.24 ± 0.34	0.12 ± 0.27					
ResEnc-XL-NSCLC-RG	527.70M			0.71 ± 0.29	0.74 ± 0.22	0.37 ± 0.31	0.26 ± 0.33	0.12 ± 0.29					
ResEnc-XL-NSCLC-RM	527.47M			0.13 ± 0.21	0.16 ± 0.24	0.58 ± 0.33	0.06 ± 0.16	0.00 ± 0.01					
ResEnc-XL-LIDC	527.70M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
ResEnc-XL-LNDb	527.70M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
ResEnc-XL-MSD-Lv	383.47M								0.51 ± 0.34	0.39 ± 0.31	0.63 ± 0.29	0.47 ± 0.35	
ResEnc-XL-TRIALS	383.47M								0.35 ± 0.34	0.67 ± 0.27	0.57 ± 0.29	0.59 ± 0.34	
ResEnc-XL-CLiMet	527.70M								0.46 ± 0.32	0.34 ± 0.32	0.66 ± 0.28	0.48 ± 0.34	
ResEnc-XL-HCC-TS	527.47M								0.25 ± 0.35	0.27 ± 0.34	0.41 ± 0.34	0.69 ± 0.30	
ResEnc-XL-MSD-Co	527.47M												0.43 ± 0.39

Model	Params	Pancreas −− Test		Head & Neck −− Test							Adrenal −− Test	Multi-Loc −− Test	
		MSD-Pa	PANO	HECKTOR	SegRap	HN-Cet	HN-HN1	RADCURE	HNSCC	HN-PET	AdACC	AutoPET	ULS23
ResEnc-XL-MSD-Pa	527.47M	0.52 ± 0.34	0.43 ± 0.32										
ResEnc-XL-PANO	527.88M	0.50 ± 0.34	0.63 ± 0.31										
ResEnc-XL-HECKTOR	527.70M			0.65 ± 0.20	0.06 ± 0.07	0.41 ± 0.17	0.30 ± 0.29	0.21 ± 0.23	0.27 ± 0.25	0.55 ± 0.19			
ResEnc-XL-SegRap	525.52M			0.07 ± 0.08	0.78 ± 0.08	0.19 ± 0.11	0.02 ± 0.06	0.05 ± 0.13	0.01 ± 0.06	0.11 ± 0.12			
ResEnc-XL-HN-Cet	527.47M			0.45 ± 0.28	0.04 ± 0.06	0.40 ± 0.25	0.19 ± 0.28	0.18 ± 0.25	0.26 ± 0.26	0.45 ± 0.25			
ResEnc-XL-HN-HN1	527.47M			0.23 ± 0.25	0.01 ± 0.02	0.20 ± 0.14	0.48 ± 0.34	0.32 ± 0.31	0.33 ± 0.31	0.28 ± 0.31			
ResEnc-XL-RADCURE	527.47M			0.25 ± 0.20	0.22 ± 0.12	0.34 ± 0.19	0.40 ± 0.34	0.54 ± 0.28	0.39 ± 0.27	0.44 ± 0.31			
ResEnc-XL-HNSCC	527.88M			0.38 ± 0.25	0.00 ± 0.01	0.29 ± 0.17	0.37 ± 0.34	0.28 ± 0.29	0.53 ± 0.32	0.49 ± 0.24			
ResEnc-XL-HN-PET	527.70M			0.37 ± 0.25	0.07 ± 0.06	0.49 ± 0.23	0.29 ± 0.32	0.30 ± 0.29	0.27 ± 0.27	0.54 ± 0.26			
ResEnc-XL-AdACC	527.70M										0.91 ± 0.05		
ResEnc-XL-AutoPET	383.47M											0.33 ± 0.28	
ResEnc-XL-ULS23	527.70M												0.60 ± 0.29

Table 9 for DSC and Tables 10 and 11 for NSD and SRVE. Specialist configurations reach overall fair results on their own held-out split. However, performance often does not carry over to their respective OOD evaluation sets. We find that OOD performance is particularly poor in the kidney, where all evaluated models collapse to ≤ 0.004 (Appendix A.7 Table 9) in both cross-domain directions. We attribute this primarily to a FoV mismatch, as KiTS volumes are thoraco-abdominal scans, whereas KiPA volumes are tightly cropped around the kidney. Additionally, not all datasets transfer equally. The six non-SegRap H&N datasets experience a moderate drop in performance consistent with domain shift, while SegRap sits outside this pattern. Transferring any other H&N specialist to SegRap and vice-versa yields a DSC ≤ 0.25 (Appendix A.7 Table 9). This is despite the SegRap-trained ResEnc-XL reaches 0.78 in-domain performance, the highest in-domain score of any H&N specialist. This pattern is observed to a smaller extent within the lungs, where MSD-Lu and NSCLC-RG transfer well to each-other but poorly to LIDC and LNDb. It is also worth noting the training collapse of ResEnc-XL (LIDC, LNDb), resulting in 0.0 DSC across in/out-of-domain evaluation (Table 4). We found the liver and pancreas to be the most stable organ groups. Across the liver datasets, in-domain DSC averages 0.62 and cross-dataset DSC averages 0.43 (Appendix A.7 Table 9). In certain cases (MSD-Lv on CLiMet) cross-dataset performance exceeds the target's own in-domain score. Pancreas lesion segmentation shows a similar profile in stability. MSD-Pa and PANO specialists score 0.50 and 0.56 in-domain and 0.39/0.46 across datasets (Appendix A.7 Table 9).

NSD largely mirrors DSC trends (Appendix A.7 Table 10), but does diverge certain areas. In H&N, cross-dataset transfer is markedly more forgiving for DSC than for boundary accuracy. Excluding SegRap, models retain 65% of their in-domain DSC when applied OOD, but only 53% of their in-domain NSD (Appendix A.7 Table 10). This could indicate that a transferred model still finds lesions in roughly the right location, but the delineated boundary is comparatively worse than the overlap alone would suggest. SegRap remains the clear outlier for every model regardless of architecture. In the lung, this pattern reverses. DSC for models transferred onto LIDC is poor (mean cross-domain DSC 0.17, or 0.19 after excluding the three collapsed LNDb specialists), yet the corresponding NSD is considerably less damaged at 0.32/0.37 with/without collapses (Appendix A.7 Table 10). This suggests that even when a transferred model over- or under-segments a nodule, it still locates the lesion boundary reasonably well; the failure is closer to a sizing error than a localization error.

Volumetric agreement follows closely with DSC in the kidney failures (Appendix A.7 Table 11). Cross-dataset transfer between KiTS23 and KiPA22 produces SRVE of 85–100%, which we attribute to the two datasets' very different FoV as opposed to difficulty with lesion segmentation in the kidney location. A model trained on one crop size has no mechanism to constrain predictions when applied to images with a substantially different anatomical context. Lung shows a

more structured failure mode tied to lesion size: models evaluated on the larger lesion datasets (NSCLC-Radiogenomics, NSCLC-Radiomics) tend to under-predict volume, while the same models evaluated on the small-lesion datasets (LIDC, LNDb) over-predict (Appendix Table A.7 11). This asymmetry is not absolute, as a small number of cross-dataset pairs still under-predict on the small-lesion targets, but the dominant direction is consistent across architectures.

Overall, single-dataset training is not a reliable proxy for "lesion segmentation" even when the anatomical target and modality are held fixed: the transfer penalty ranges from mild (liver, pancreas) to large (lung, head and neck, kidney). Increasing model capacity does not rescue cross-dataset performance; it only widens or preserves the in-domain/cross-domain gap.

4 Discussion and Limitations

Discussion: We present LesionBench, a cancer-agnostic benchmark for lesion segmentation spanning multiple lesion locations and pathologies. Using LesionBench, we evaluate popular segmentation architectures under both specialist (single-dataset) and generalist (pooled-dataset) training regimes. Our results point to a gap rather than a solution: lesion segmentation remains substantially more difficult than the organ segmentation tasks that have traditionally driven progress in the field. Performance is highly heterogeneous across datasets, transfer between lesion types and imaging protocols is inconsistent, and even the strongest models exhibit systematic failure modes that organ-segmentation benchmarks rarely surface. Specialists achieve reasonable in-domain DSCs but transfer poorly to other datasets of the same organ, in some cases collapsing entirely. Pooling all 24 datasets into a single ResEnc-XL model largely resolves this transfer failure: the complete-miss rate falls from 34.2% (29.8% excluding collapsed training runs) for the ResEnc-XL specialist to 7.1% for the pooled counterpart, making the datasets on which the specialist failed to train (LIDC, LNDb) usable. However, this improvement comes at a cost: pooled training reduces in-domain DSC relative to the matched specialist on the large majority of datasets, with reductions exceeding 0.25 DSC on several sources (e.g., TRIALS, SegRap2023, PANORAMA). A single shared model is therefore not simply "the best specialist plus generalization"; it is a genuine trade-off between per-dataset sharpness and cross-dataset robustness. This suggests that neither extreme (specialists vs. generalists) evaluated here is sufficient for lesion segmentation at scale. Specialists lack robustness to acquisition and protocol shifts pervasive across oncology data, while a naively pooled generalist lacks the per-source precision a specialist can reach when training and test distributions match. We do not view this as a limitation of any single architecture in our comparison, but as evidence of a capability gap that current training strategies do not address: no model we evaluated combines specialist-level in-domain performance with specialist-free transfer. We hope LesionBench provides a useful foundation for the community to develop and evaluate segmentation models that are capable of robustly handling lesion segmentation alongside, and not merely as an extension of, organ segmentation.

Limitations: Evaluation Protocol: Since datasets within LesionBench lack consistent instance identifiers, each scan contributes a single overlap score regardless of the number of lesions it contains. Thus, strong overlap on a dominant lesion can mask missed or poorly segmented additional lesions, while false-positive regions can depress DSC, precision, and volume error, even if the annotated lesion is well delineated. Future work could address this by adding instance identifiers, at least to the test set, and reporting per-lesion detection, instance DSC, and false-positive counts. **Model Selection:** We chose models that can be trained within nnU-Net's automated configuration pipeline, which allows for fair comparison of architectures under a shared framework. However, this also means that we did not optimize each model individually. So the reported scores should be read as performance under a common, largely default recipe, rather than as an upper bound for each network. Even so, lesion segmentation at scale remains challenging for several configurations due to variations in FoV, acquisition protocol, pathology, and lesion size and location across LesionBench. Further work should focus on developing novel solutions that can deliver promising results in cancer-agnostic segmentation. **Label Noise:** We treated the lesion masks supplied with each source dataset as the reference standard, without performing a systematic manual audit. Therefore, LesionBench inherits any inconsistencies present in the public annotations. Where multiple rater masks were available, we fused them using STAPLE to obtain a consensus reference. Future work should prioritize label integrity, including a structured review of the LesionBench masks, correction of clear errors, and additional multi-rater annotation to quantify residual label noise.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with programming tasks and grammar/spell checking. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

LesionBench is assembled entirely from previously published, de-identified CT collections; no new patient data were collected and no imaging data are redistributed. We release only conversion code, label mappings, consensus-mask generation code, and case-identifier split lists. Of the 24 constituent collections, 19 are openly licensed and 5 require completion of a data-use agreement with the originating provider; LesionBench is therefore fully reproducible only for users who independently obtain access to those five. Per-dataset licenses are listed in Table 5. Benchmark rank is not evidence of clinical readiness. All work that is not our own (in particular the datasets used within LesionBench) have been appropriately referenced as the respective authors work, not ours. Ethics considerations: patients are approximately 2:1 male between 50 and 80 years of age, and mostly White; ethnicity is unreported for the majority of cases, and scanner manufacturers are concentrated in two vendors. Therefore, results should not be assumed to transfer to under-represented groups.

REPRODUCIBILITY STATEMENT

We have released the full source code for LesionBench at `https://tinyurl.com/LesionBenchmark`. LesionBench is assembled entirely from publicly available collections; we redistribute no imaging data ourselves. The repository provides the complete list of constituent datasets together with their sources, licences, and access conditions. We release the code that converts each source collection into the common binary segmentation format used throughout the benchmark, including the label mapping. Train and test partitions are distributed as explicit case-identifier lists rather than being regenerated at runtime, so no result depends on a random split. We list the Docker images required to reproduce training for every method evaluated. These images are public and may be pulled freely; we pin them by digest so that the environment, framework versions, and CUDA stack are fixed. The evaluation code computing Dice Similarity Coefficient, Normalised Surface Distance, and Symmetric Relative Volume Error is released in full.

ACKNOWLEDGMENTS

Removed for double-blind revision

REFERENCES

Foundation models for pan-cancer segmentation in ct images. `https://www.codabench.org/competitions/7149/`, 2025. CodaBench competition.

Triphasic-aided liver lesion segmentation in non-contrast ct (trials) challenge, 2025. URL `https://zenodo.org/records/15087646`.

Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):39, 2018.

HJWL Aerts, L Wee, E Rios Velazquez, RTH Leijenaar, C Parmar, P Grossmann, S Carvalho, J Bussink, R Monshouwer, B Haibe-Kains, et al. Data from nsclc-radiomics (version 4)[data set]. *The Cancer Imaging Archive*, 10:K9, 2014.

Natalia Alves, Megan Schuurmans, Dawid Rutkowski, Anindo Saha, Pierpaolo Vendittelli, Nancy Obuchowski, Marjolein H Liedenbaum, Ingfrid S Haldorsen, Anders Molven, Derya Yakar, et al. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care ct scans (panorama): an international, paired, non-inferiority, confirmatory, observational study. *The Lancet Oncology*, 27(1):116–124, 2026.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. Data from lidc-idri. *(The Cancer Imageing Archive)*, 2015.

S Bakr, O Gevaert, S Echegaray, K Ayers, M Zhou, M Shafiq, H Zheng, W Zhang, A Leung, M Kadoch, et al. Data for nsclc radiogenomics (version 4)[data set]. *The Cancer Imaging Archive*, 2017.

Pedro R Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Saikat Roy, Yannick Kirchhoff, Maximilian Rokuss, et al. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Advances in Neural Information Processing Systems*, 37:15184–15201, 2024.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

WR Bosch, WL Straube, JW Matthews, and JA Purdy. Head-neck cetuximab. *The Cancer Imaging Archive*, 2015.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

MJJ de Grauw, E Th Scholten, Ewoud J Smit, Matthieu JCM Rutten, Mathias Prokop, Bram van Ginneken, and Alessa Hering. The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *Medical image analysis*, 102:103525, 2025.

Aaron Grossberg, Hesham Elhalawani, Abdallah Mohamed, Sam Mulder, Bowman Williams, Aubrey L White, James Zafereo, Andrew J Wong, Joel E Berends, Shady AboHashem, et al. Hnscc. *(No Title)*, 2020.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Yuting He, Guanyu Yang, Jian Yang, Yang Chen, Youyong Kong, Jiasong Wu, Lijun Tang, Xiaomei Zhu, Jean-Louis Dillenseger, Pengfei Shao, et al. Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical image analysis*, 63:101722, 2020.

Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

Ziyan Huang, Zhongying Deng, Jin Ye, Haoyu Wang, Yanzhou Su, Tianbin Li, Hui Sun, Junlong Cheng, Jianpin Chen, Junjun He, et al. A-eval: A benchmark for cross-dataset and cross-modality evaluation of abdominal multi-organ segmentation. *Medical Image Analysis*, 101:103499, 2025.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 488–498. Springer, 2024.

Burak Kocak, Michail E Klontzas, Arnaldo Stanzione, Aymen Meddeb, Aydın Demircioğlu, Christian Bluethgen, Keno Bressem, Lorenzo Ugga, Nathaniel Mercaldo, Oliver Díaz, et al. Evaluation metrics in medical imaging ai: fundamentals, pitfalls, misapplications, and recommendations. *European journal of radiology artificial intelligence*, pp. 100030, 2025.

Thomas Küstner, Sergios Gatidis, Ornela Megne, Michael Ingrisch, Matthias Fabritius, Jakob Dexl, Katharina Jeblick, Andreas Mittermeier, Balthasar Schachtner, Theresa Stüber, Clemens Cyran, Zdravko Marinov, Felix Peisen, Andreas Wagner, Ahmed Othman, Antoine Sanner, Tanja Loßau, Jan Hendrik Moltz, Temke Kohlbrandt, and Alessa Hering. Automated lesion segmentation in whole-body pet/ct and longitudinal (autopet/ct iv), March 2025. URL `https://doi.org/10.5281/zenodo.15045096`.

Yun Liu, Timo Kohlberger, Mohammad Norouzi, George E Dahl, Jenny L Smith, Arash Mohtashamian, Niels Olson, Lily H Peng, Jason D Hipp, and Martin C Stumpe. Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Archives of pathology & laboratory medicine*, 143(7):859–868, 2019.

Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical image analysis*, 101:103447, 2025.

A Moawad, D Fuentes, A Morshid, A Khalaf, M Elmohr, A Abusaif, JD Hazle, AO Kaseb, M Hassan, A Mahvash, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation [data set]. *The Cancer Imaging Archive*, 5, 2021.

Ahmed W Moawad, Ayahallah A Ahmed, Mohab ElMohr, Mohamed Eltaher, Mouhammed Amir Habra, Sarah Fisher, Nancy Perrier, Miao Zhang, David Fuentes, and Khaled Elsayes. Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with ki-67 expression (adrenal-acc-ki67-seg)[data set]. *The Cancer Imaging Archive*, 8, 2023.

João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al. Lndb: a lung nodule database on computed tomography. *arXiv preprint arXiv:1911.08434*, 2019.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36:36620–36636, 2023.

Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, 21(2):182–194, 2024.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 405–415. Springer, 2023.

Numan Saeed, Salma Hassan, Shahad Hardan, Ahmed Aly, Darya Taratynova, Umair Nawaz, Ufaq Khan, Muhammad Ridzuan, Vincent Andrearczyk, Adrien Depeursinge, et al. A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. *arXiv preprint arXiv:2509.00367*, 2025.

Pengfei Shao, Chao Qin, Changjun Yin, Xiaoxin Meng, Xiaobing Ju, Jie Li, Qiang Lv, Wei Zhang, and Zhengquan Xu. Laparoscopic partial nephrectomy with segmental renal artery clamping: technique and clinical outcomes. *European urology*, 59(5):849–855, 2011.

Pengfei Shao, Lijun Tang, Pu Li, Yi Xu, Chao Qin, Qiang Cao, Xiaobing Ju, Xiaoxin Meng, Qiang Lv, Jie Li, et al. Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy. *European urology*, 62(6):1001–1008, 2012.

Amber L Simpson, Jacob Peoples, John M Creasy, Gabor Fichtinger, Natalie Gangai, Krishna N Keshavamurthy, Andras Lasso, Jinru Shia, Michael I D'Angelica, and Richard KG Do. Preoperative ct and survival data for patients undergoing resection of colorectal liver metastases. *Scientific data*, 11(1):172, 2024.

Lalith Kumar Shiyam Sundar, Josef Yu, Otto Muzik, Oana C Kulterer, Barbara Fueger, Daria Kifjak, Thomas Nakuz, Hyung Min Shin, Annika Katharina Sima, Daniela Kitzmantl, et al. Fully automated, semantic segmentation of whole-body 18f-fdg pet/ct images based on data-centric artificial intelligence. *Journal of Nuclear Medicine*, 63(12):1941–1948, 2022.

Michele Svanera, Mattia Savardi, Alberto Signoroni, Sergio Benini, and Lars Muckli. Fighting the scanner effect in brain mri segmentation with a progressive level-of-detail network trained on multi-site data. *Medical Image Analysis*, 93:103090, 2024.

Martin Vallieres, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. Data from head-neck-pet-ct the cancer imaging archive, 2017.

Tassilo Wald, Saikat Roy, Fabian Isensee, Constantin Ulrich, Sebastian Ziegler, Dasha Trofimova, Raphael Stock, Michael Baumgartner, Gregor Köhler, and Klaus Maier-Hein. Primus: Enforcing attention usage for 3d medical image segmentation. *arXiv preprint arXiv:2503.01835*, 2025.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Leonard Wee and Andre Dekker. Data from head-neck-radiomics-hn1. *(The Cancer Imaging Archive)*, 2019.

ML Welch, S Kim, A Hope, SH Huang, Z Lu, J Marsilla, M Kazmierski, K Rey-McIntyre, T Patel, B O'Sullivan, et al. Computed tomography images from large head and neck cohort (radcure). *The Cancer Imaging Archive*, 4, 2023.

Zhongchen Zhao, Huai Chen, Jiang Li, and Lisheng Wang. Boundary attention u-net for kidney and kidney tumor segmentation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1540–1543. IEEE, 2022.

Yuming Zhong and Yi Wang. Simple: Similarity-aware propagation learning for weakly-supervised breast cancer segmentation in dce-mri. In *International conference on medical image computing and computer-assisted intervention*, pp. 567–577. Springer, 2023.

A APPENDIX

A.1 DATASET SOURCES

To facilitate data access we have reported in Table 5 the links where we have originally accessed the 24 datasets included in LesionBench.

Table 5: Shorthand names, full dataset names, licenses, and links for all datasets included in the LesionBench benchmark. While all constituent datasets are publicly accessible, the degree of accessibility varies across sources. Some datasets are freely available without restriction, whereas others require completion of data access agreements, which may include commitments to restrict commercial use or prohibit redistribution of the data. NIH stands for NIH Controlled Data Access Policy.

Shorthand	Full Name	License	Dataset Link
Kidney			
KiTS23	KiTS23	CC BY-NC-SA 4.0	https://kits-challenge.org/kits23/
KiPA22	KiPA22	CC BY-NC-ND	https://kipa22.grand-challenge.org/
Lung			
MSD-Lu	MSD Lung	CC BY-SA 4.0	http://medicaldecathlon.com/
NSCLC-RG	NSCLC Radiogenomics	CC BY 3.0	https://www.cancerimagingarchive.net/collection/nsclc-radiogenomics/
NSCLC-RM	NSCLC Radiomics	CC BY-NC 3.0	https://www.cancerimagingarchive.net/collection/nsclc-radiomics/
LIDC	LIDC-IDRI	CC BY 3.0	https://www.cancerimagingarchive.net/collection/lidc-idri/
LNDb	LNDb	CC BY-NC-ND 4.0	https://lndb.grand-challenge.org/
Liver			
MSD-Lv	MSD Liver	CC BY-SA 4.0	http://medicaldecathlon.com/
TRIALS	TRIALS	Not Explicitly Stated	https://synapse.org/Synapse:syn53285416/wiki/
CLiMet	Colorectal Liver Metastasis	CC BY 4.0	https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases/
HCC-TS	HCC-TACE-Seg	CC BY 4.0	https://www.cancerimagingarchive.net/collection/hcc-tace-seg/
Colon & Pancreas			
MSD-Co	MSD Colon	CC BY-SA 4.0	http://medicaldecathlon.com/
MSD-Pa	MSD Pancreas	CC BY-SA 4.0	http://medicaldecathlon.com/
PANO	PANORAMA	CC BY-NC 4.0	https://panorama.grand-challenge.org/datasets-imaging-labels/
Head & Neck			
HECKTOR	HECKTOR	CC BY-NC-SA	https://hecktor25.grand-challenge.org/
SegRap	SegRap2023	CC BY-NC 4.0	https://segrap2023.grand-challenge.org/
HN-Cet	Head-Neck Cetuximab	NIH	https://www.cancerimagingarchive.net/collection/head-neck-cetuximab/
HN-HN1	Head-Neck-Radiomics-HN1	NIH	https://www.cancerimagingarchive.net/collection/head-neck-radiomics-hn1/
RADCURE	RADCURE	NIH	https://www.cancerimagingarchive.net/collection/radcure/
HNSCC	HNSCC	NIH	https://www.cancerimagingarchive.net/collection/hnscc/
HN-PET	Head-Neck-PET-CT	NIH	https://www.cancerimagingarchive.net/collection/head-neck-pet-ct/
Adrenal			
AdACC	Adrenal-ACC-Ki67-Seg	CC BY 4.0	https://www.cancerimagingarchive.net/collection/adrenal-acc-ki67-seg/
Multi-Location			
AutoPET	AUTO-PET	CC BY-NC 4.0	https://www.autopet.org/autopetiv.html
ULS23	ULS23	CC BY-NC-SA 4.0	https://uls23.grand-challenge.org/

A.2 DEDUPLICATION

As described in the main paper, we performed manual deduplication of LesionBench to prevent data leakage from the training set into the test set. We report the number of images flagged at each similarity threshold in Table 6, with visual representations in Figures 2 and 3. Analysis of the similarity score distribution revealed an inflection point at a threshold of 0.998, beyond which the number of flagged images increased rapidly (148 to 372 images between 0.998 and 0.997; Table 6), suggesting a transition from genuine duplicates to borderline cases requiring manual review. To ensure comprehensive duplicate detection while remaining tractable, all image pairs with a similarity score ≥ 0.995 were subject to manual inspection, yielding 963 candidates in total. Manual review indicated that all pairs with a similarity score ≥ 0.997 were true positives; below this, the false-positive rate rose sharply, and no additional images below the 0.995 threshold were examined. Of the 963 candidate pairs, 154 were confirmed as true positives and were subsequently removed from the test set and reassigned to the training set to prevent data leakage.

Table 6: Number of images at or above each cosine similarity threshold for near-duplicate detection in the LesionBench benchmark.

Threshold	Images $\geq$ Threshold
1.000	16
0.999	130
0.998	148
0.997	372
0.996	762
0.995	963
0.994	1,090
0.993	1,192
0.992	1,284
0.991	1,416
0.990	1,567
0.989	1,702
0.988	1,866
0.987	2,085
0.986	2,279
0.985	2,459
0.984	2,669
0.983	2,852
0.982	3,048
0.981	3,219
0.980	3,398

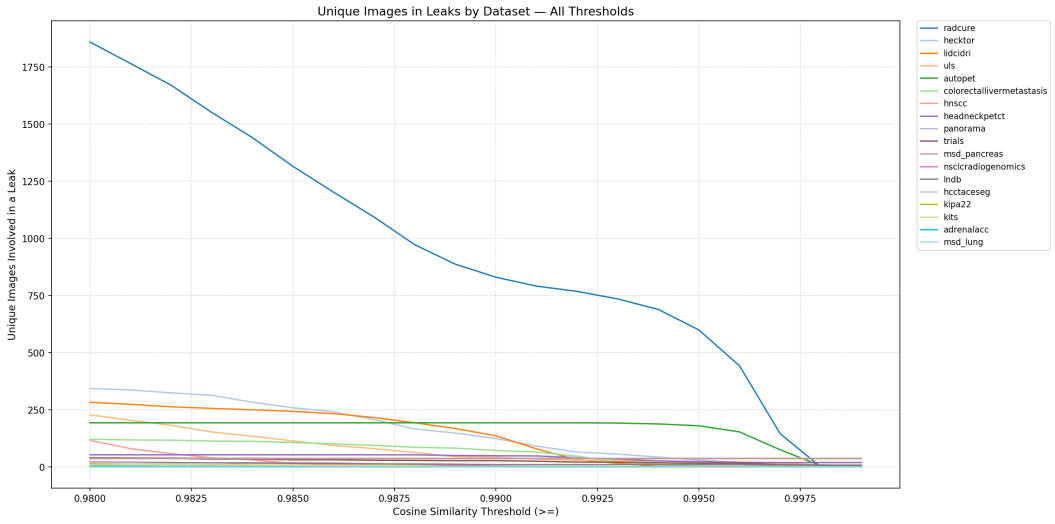


Figure 2: Number of images flagged as near-duplicates across the full range of cosine similarity thresholds (0.10–1.00). Each point represents the cumulative count of images with at least one duplicate match at or above the given threshold. We observed that when the number of flagged images remained stable across the 1.00–0.98 range, these pairs were highly likely to be true positives, whereas a sharp increase in flagged images as the threshold decreased was strongly indicative of a higher false-positive rate.

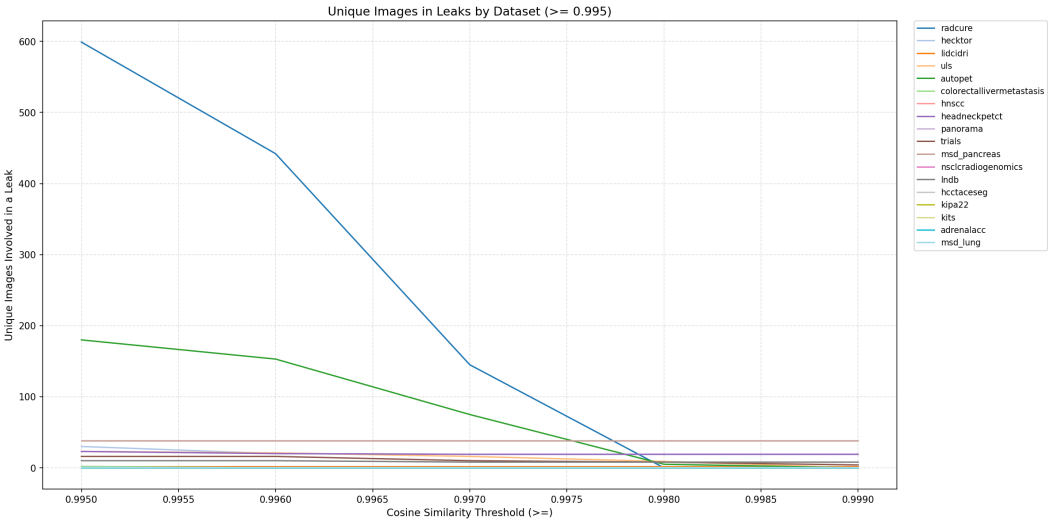


Figure 3: Zoomed-in view of the high-similarity regime (0.997–1.000), highlighting the marked increase in flagged images at 0.998. This range is most relevant for identifying true near-duplicates and exact matches across the dataset. Consistent with our broader observation, pairs flagged within this stable high-similarity regime were highly likely to be true positives, whereas larger increases in flagged images at lower thresholds were associated with a higher false-positive rate.

A.3 LESION RETENTION STRATEGY

Table 7 report for each datasets included the available labels and which have been retained as lesion in LesionBench.

Table 7: **Per-dataset lesion label retention policy.** Original label names are shown next to the labels retained in LesionBench. A comma separates distinct labels that are kept and merged into one foreground class (label 1) for our binary segmentation task.

Dataset	Original Labels	Label Kept
Kidney		
KiTS23	Kidney, Tumor, Masses	Tumor, Masses
KiPA22	Renal vein, Kidney, Renal artery, Tumor	Tumor
Lung		
MSD-Lu	Cancer	Cancer
NSCLC-RG	Lesion	Lesion
NSCLC-RM	Esophagus, Mass, Heart, Lung (left), Lung (right), Spinal cord	Mass
LIDC	Lesion	Lesion
LNDb	Lesion	Lesion
Liver		
MSD-Lv	Liver, Cancer	Cancer
TRIALS	Liver Tumor	Liver Tumor
CLiMet	Liver, Liver remnant, Hepatic, Portal, Tumor	Tumor
HCC-TS	Liver, Mass, Portal vein, Abdominal aorta	Mass
Colon & Pancreas		
MSD-Co	Colon Cancer	Colon Cancer
MSD-Pa	Pancreas, Cancer	Cancer
PANO	PDAC lesion, Veins, Arteries, Pancreas parenchyma, Pancreatic duct, Common bile duct	PDAC Lesion
Head & Neck		
HECKTOR	GTVp, GTVn	GTVp, GTVn
SegRap2023	Brain, Brainstem, Chiasm, Cochlea left, Cochlea right, Esophagus, Eustachian tube left, Eustachian tube right, Eye left, Eye right, Hippocampus left, Hippocampus right, Internal auditory canal left, Internal auditory canal right, Larynx, Larynx glottic, Larynx supraglottic, Lens left, Lens right, Mandible left, Mandible right, Mastoid left, Mastoid right, Middle Ear left, Middle ear right, Optic nerve left, Optic nerve right, Oral cavity, Parotid left, Parotid right, Pharynx, Pituitary, Spinal cord, Submandibular left, Submandibular right, Temporal lobe left, Temporal lobe right, Thyroid, Temporomandibular joint left, Temporomandibular joint right, Trachea, Tympanic cavity left, Tympanic cavity right, Vestibular semicircular canal left, Vestibular semicircular canal right, GTVp, GTVnd	GTVp, GTVnd
HN-Cet	GTV, Spinal cord, Trachea, Esophagus, Larynx, Hypothalamus, Brain, Brain stem, Pituitary, Thyroid	GTV
HN-HN1	GTV	GTV
RADCURE	GTVp, Brachial plexus (left), Brachial plexus (right), Brainstem, Cochlea (left), Cochlea (right), Esophagus, Eye (left), Eye (right), Larynx, Lens (left), Lens (right), Lips, Mandible, Optic nerve (left), Optic nerve (right), Optic chiasm, Parotid (left), Parotid (right), Spinal cord, Brain, Submandibular gland (right), Submandibular gland (left)	GTVp
HNSCC	GTVp, Brain, GTV, Esophagus	GTVp, GTV
HN-PET	GTV, Brain, Esophagus, Spinal cord, Trachea	GTV
Adrenal		
AdACC	Adrenal Tumor	Adrenal Tumor
Multi-Location		
AutoPET	Lesion	Lesion
ULS23	Lesion	Lesion

A.4 METRICS

Below we report the mathematical formulation employed when computing the DSC and NSD.

Dice Similarity Coefficient which quantifies volumetric overlap between the predicted and ground-truth masks:

$$\mathrm{DSC}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|} \tag{1}$$

where: A, B are the predicted and ground-truth segmentation masks, $|\cdot|$ indicates the number of voxels within the 3D mask, and $|A \cap B|$ is the number of elements common to both A and B.

Normalized Surface Distance:

$$\mathrm{NSD}^{(\tau)}(A, B) = \frac{\left|\partial A \cap B^{(\tau)}\right| + \left|\partial B \cap A^{(\tau)}\right|}{|\partial A| + |\partial B|}, \tag{2}$$

where: A, B are again the predicted and ground-truth segmentation masks, $\partial A, \partial B$ are the surfaces of A and B, and τ is the tolerance threshold, set to $1\,\mathrm{mm}$ in this work.

A.5 Data Split Strategy

It is common within medical segmentation challenges for organizers to withhold their test sets while only the training set is publicly available in many cases. For each dataset within LesionBench, we adopted the following strategy: where a public test set was available, we used it directly; otherwise, we split the publicly available data 80/20 at the patient level into training and test sets to avoid any data leakage. For TCIA data, candidate images were first reviewed for suitability for lesion annotation, and the resulting set was likewise split 80/20 at the patient level. To construct LesionBench, we combined all training and test images and applied the de-duplication procedure described in Section 2.1.2 to identify and remove any overlap between splits.

A.6 Network Configuration

Each model was planned separately, so nnU-Net could select a configuration suited to that dataset. Under this per-dataset planning, two size presets (for example Large and XL) can arrive at the same parameter count, and a single preset can arrive at different counts on different datasets. The ResEnc presets are GPU memory budgets (8 GB for M, 24 GB for L, 40 GB for XL). For each dataset the planner starts from the spacing and median shape, sets the number of stages and the channel counts (32, doubled up to 320), and reduces the patch until the estimated activation memory fits the budget. Any memory left over is spent on batch size. Patch size and batch size change the epoch time and leave the weights unchanged. On the pooled generalist, ResEnc-L and ResEnc-XL are both the 7-stage network (527.4M); XL's patch ($112 \times 320 \times 256$) contains $1.75\times$ the voxels of L's ($80 \times 256 \times 256$), which lengthens its epoch. Specialists follow the same rule. On KiPA22 both are the 6-stage network (383.47M) with the same patch ($224 \times 160 \times 160$), and XL only raises the batch size from 3 to 5. Where the XL budget can support it, XL is given one extra stage of 320 channels, and the counts differ. In contrast MedNeXt-M uses one fixed published configuration, so it is 17.55M on every dataset.

A.7 Additional Results

Due to space limitations in the main manuscript we have only reported DSC and SRVE for the pooled cancer-agnostic evaluation and the best performing ResEnc-XL on single-cancer datasets. We here reports results in terms of Normalized Surface Distance (NSD) in Table 8 for the cancer-agnostic configuration, and DSC, NSD, and SRVE for single-cancer datasets in Tables 9, 10, and 11, respectively.

Table 8: **Evaluation on LesionBench (#3,035) NSD:** NSD score ($\mu \pm \sigma$), the higher the better. We report the overall NSD averaged over the entire LesionBench test set, macro-averaged per dataset, and the NSD score breakdown per lesion location and per dataset. All the evaluated models have been trained on the LesionBench training set (#12,824) for 2k epochs.

Model	Params	Epoch Speed (sec)	Global	Macro	Kidney	Lung	Liver	Colon	Pancreas	Head & Neck	Adrenal	Multi-Loc
nnU-Net	88.4M	41.54	0.19 ± 0.26	0.15 ± 0.11	0.23 ± 0.20	0.17 ± 0.20	0.09 ± 0.12	0.07 ± 0.10	0.05 ± 0.08	0.15 ± 0.13	0.22 ± 0.11	0.24 ± 0.34
ResEnc-M	383.2M	38.51	0.26 ± 0.28	0.21 ± 0.12	0.35 ± 0.23	0.28 ± 0.24	0.14 ± 0.15	0.11 ± 0.13	0.11 ± 0.14	0.16 ± 0.12	0.27 ± 0.14	0.34 ± 0.35
ResEnc-L	527.4M	96.67	0.43 ± 0.32	0.28 ± 0.15	0.41 ± 0.22	0.33 ± 0.25	0.21 ± 0.18	0.14 ± 0.16	0.18 ± 0.18	0.29 ± 0.20	0.37 ± 0.11	0.59 ± 0.34
ResEnc-XL	527.4M	163.36	0.47 ± 0.33	0.31 ± 0.16	0.45 ± 0.21	0.44 ± 0.31	0.23 ± 0.20	0.18 ± 0.18	0.20 ± 0.19	0.31 ± 0.20	0.38 ± 0.13	0.63 ± 0.34
MedNeXt-M	17.55M	468.32	0.24 ± 0.27	0.20 ± 0.13	0.25 ± 0.23	0.26 ± 0.23	0.14 ± 0.15	0.07 ± 0.10	0.08 ± 0.12	0.15 ± 0.12	0.25 ± 0.12	0.31 ± 0.33
MedNeXt-L	61.78M	481.12	0.25 ± 0.28	0.20 ± 0.12	0.27 ± 0.22	0.27 ± 0.23	0.15 ± 0.15	0.11 ± 0.12	0.10 ± 0.13	0.16 ± 0.13	0.24 ± 0.14	0.33 ± 0.34
Primus-M	244.51M	53.92	0.19 ± 0.24	0.17 ± 0.10	0.25 ± 0.21	0.18 ± 0.22	0.10 ± 0.13	0.06 ± 0.10	0.09 ± 0.12	0.18 ± 0.15	0.23 ± 0.16	0.22 ± 0.31
Primus-L	428.29M	81.94	0.20 ± 0.26	0.17 ± 0.10	0.24 ± 0.19	0.21 ± 0.22	0.10 ± 0.13	0.06 ± 0.10	0.08 ± 0.10	0.17 ± 0.15	0.22 ± 0.12	0.23 ± 0.32

Model	KiTS23	KiPA22	MSD Lung	NSCLC Radgen	NSCLC Rad	LIDC-IDRI	LNDb	MSD Liver	TRIALS	C_Liver Met	HCC-TACE-Seg	MSD Colon
nnU-Net	0.18 ± 0.16	0.57 ± 0.17	0.26 ± 0.20	0.26 ± 0.16	0.26 ± 0.20	0.13 ± 0.19	0.07 ± 0.12	0.08 ± 0.11	0.04 ± 0.06	0.11 ± 0.13	0.14 ± 0.12	0.07 ± 0.10
ResEnc-M	0.32 ± 0.20	0.58 ± 0.26	0.30 ± 0.11	0.36 ± 0.20	0.30 ± 0.21	0.27 ± 0.26	0.19 ± 0.24	0.09 ± 0.13	0.08 ± 0.08	0.20 ± 0.18	0.18 ± 0.15	0.11 ± 0.13
ResEnc-L	0.37 ± 0.19	0.66 ± 0.23	0.30 ± 0.18	0.38 ± 0.19	0.32 ± 0.21	0.34 ± 0.27	0.24 ± 0.27	0.12 ± 0.14	0.11 ± 0.09	0.31 ± 0.20	0.24 ± 0.15	0.14 ± 0.16
ResEnc-XL	0.43 ± 0.20	0.64 ± 0.21	0.31 ± 0.20	0.40 ± 0.22	0.34 ± 0.22	0.50 ± 0.33	0.40 ± 0.33	0.12 ± 0.13	0.10 ± 0.08	0.37 ± 0.21	0.27 ± 0.15	0.18 ± 0.18
MedNeXt-M	0.19 ± 0.16	0.66 ± 0.21	0.36 ± 0.17	0.39 ± 0.20	0.27 ± 0.19	0.23 ± 0.23	0.26 ± 0.28	0.11 ± 0.16	0.08 ± 0.09	0.18 ± 0.16	0.18 ± 0.14	0.07 ± 0.10
MedNeXt-L	0.22 ± 0.17	0.60 ± 0.28	0.36 ± 0.17	0.35 ± 0.20	0.25 ± 0.19	0.26 ± 0.24	0.26 ± 0.24	0.11 ± 0.16	0.08 ± 0.08	0.20 ± 0.16	0.18 ± 0.13	0.11 ± 0.12
Primus-M	0.23 ± 0.18	0.46 ± 0.27	0.35 ± 0.19	0.31 ± 0.22	0.27 ± 0.22	0.13 ± 0.19	0.12 ± 0.22	0.08 ± 0.14	0.06 ± 0.10	0.12 ± 0.14	0.14 ± 0.14	0.06 ± 0.10
Primus-L	0.22 ± 0.17	0.40 ± 0.24	0.38 ± 0.17	0.33 ± 0.20	0.25 ± 0.22	0.17 ± 0.21	0.17 ± 0.26	0.09 ± 0.15	0.05 ± 0.09	0.13 ± 0.14	0.12 ± 0.13	0.06 ± 0.10

Model	MSD Pancreas	PANORAMA	HECKTOR	SegRap2023	H&N Cetuximab	H&N-Rad-HN1	RADCURE	HNSCC	H&N-PET-CT	Adrenal-ACC	AUTO-PET	ULS23
nnU-Net	0.09 ± 0.13	0.04 ± 0.06	0.10 ± 0.09	0.04 ± 0.03	0.11 ± 0.07	0.17 ± 0.17	0.17 ± 0.13	0.13 ± 0.13	0.17 ± 0.15	0.22 ± 0.11	0.04 ± 0.04	0.27 ± 0.35
ResEnc-M	0.18 ± 0.18	0.09 ± 0.11	0.15 ± 0.10	0.07 ± 0.04	0.11 ± 0.04	0.18 ± 0.18	0.17 ± 0.12	0.13 ± 0.12	0.18 ± 0.15	0.27 ± 0.14	0.06 ± 0.06	0.38 ± 0.35
ResEnc-L	0.32 ± 0.22	0.13 ± 0.13	0.24 ± 0.19	0.06 ± 0.04	0.12 ± 0.07	0.31 ± 0.23	0.31 ± 0.19	0.35 ± 0.22	0.24 ± 0.17	0.37 ± 0.11	0.08 ± 0.08	0.66 ± 0.30
ResEnc-XL	0.32 ± 0.25	0.17 ± 0.16	0.25 ± 0.20	0.08 ± 0.04	0.11 ± 0.07	0.26 ± 0.24	0.32 ± 0.18	0.39 ± 0.24	0.22 ± 0.18	0.38 ± 0.13	0.08 ± 0.08	0.71 ± 0.28
MedNeXt-M	0.13 ± 0.17	0.07 ± 0.09	0.12 ± 0.11	0.08 ± 0.03	0.08 ± 0.03	0.16 ± 0.16	0.17 ± 0.12	0.15 ± 0.14	0.16 ± 0.14	0.25 ± 0.12	0.05 ± 0.06	0.35 ± 0.34
MedNeXt-L	0.17 ± 0.18	0.08 ± 0.10	0.13 ± 0.10	0.09 ± 0.03	0.08 ± 0.02	0.16 ± 0.16	0.17 ± 0.13	0.17 ± 0.15	0.16 ± 0.14	0.24 ± 0.14	0.06 ± 0.06	0.37 ± 0.35
Primus-M	0.13 ± 0.17	0.08 ± 0.10	0.12 ± 0.10	0.05 ± 0.03	0.12 ± 0.06	0.23 ± 0.23	0.20 ± 0.15	0.19 ± 0.17	0.18 ± 0.15	0.23 ± 0.16	0.05 ± 0.06	0.24 ± 0.32
Primus-L	0.10 ± 0.13	0.07 ± 0.09	0.11 ± 0.10	0.06 ± 0.04	0.10 ± 0.06	0.21 ± 0.22	0.19 ± 0.15	0.19 ± 0.17	0.16 ± 0.16	0.22 ± 0.12	0.05 ± 0.05	0.27 ± 0.34

Table 9: **Per dataset evaluation DSC:** performances are reported as DSC score ($\mu \pm \sigma$). For each trained model we report the in distribution evaluation (test set corresponding to training set), out of distribution evaluation (same cancer type, but test set from a different dataset, excluding colon and adrenal gland), and cancer-agnostic performances (test set from different cancer types). Test sets cardinality is reported in Table 1. All models trained for 1k epochs.

Model	Params	Kidney −− Test		Lung −− Test					Liver −− Test				Colon −− Test
		KiTS23	KiPA22	MSD-Lu	NSCLC-RG	NSCLC-RM	LIDC	LNDb	MSD-Lv	TRIALS	CLiMet	HCC-TS	MSD-Co
nnU-Net-KiTS23	88.62M	0.84 ± 0.16	0.03 ± 0.06										
nnU-Net-KiPA22	88.21M	0.00 ± 0.00	0.85 ± 0.25										
ResEnc-L-KiTS23	383.47M	0.85 ± 0.17	0.00 ± 0.00										
ResEnc-L-KiPA22	383.47M	0.00 ± 0.00	0.91 ± 0.07										
ResEnc-XL-KiTS23	527.70M	0.85 ± 0.17	0.00 ± 0.00										
ResEnc-XL-KiPA22	383.47M	0.00 ± 0.00	0.91 ± 0.08										
MedNeXt-M-KiTS23	17.55M	0.86 ± 0.15	0.00 ± 0.00										
MedNeXt-M-KiPA22	17.55M	0.01 ± 0.02	0.86 ± 0.24										
Primus-M-KiTS23	244.73M	0.87 ± 0.14	0.00 ± 0.00										
Primus-M-KiPA22	245.06M	0.00 ± 0.01	0.86 ± 0.24										
Primus-L-KiTS23	428.56M	0.85 ± 0.17	0.00 ± 0.00										
Primus-L-KiPA22	428.96M	0.01 ± 0.01	0.88 ± 0.15										
nnU-Net-MSD-Lu	88.21M			0.58 ± 0.34	0.65 ± 0.28	0.36 ± 0.32	0.20 ± 0.32	0.08 ± 0.23					
nnU-Net-NSCLC-RG	88.21M			0.73 ± 0.24	0.71 ± 0.24	0.34 ± 0.32	0.21 ± 0.32	0.11 ± 0.25					
nnU-Net-NSCLC-RM	88.02M			0.13 ± 0.17	0.23 ± 0.27	0.57 ± 0.34	0.03 ± 0.12	0.0 ± 0.0					
nnU-Net-LIDC	88.21M			0.52 ± 0.26	0.48 ± 0.24	0.06 ± 0.17	0.48 ± 0.32	0.29 ± 0.30					
nnU-Net-LNDb	88.21M			0.54 ± 0.27	0.45 ± 0.27	0.07 ± 0.20	0.45 ± 0.33	0.30 ± 0.35					
ResEnc-L-MSD-Lu	527.29M			0.53 ± 0.37	0.57 ± 0.37	0.40 ± 0.34	0.20 ± 0.32	0.04 ± 0.15					
ResEnc-L-NSCLC-RG	527.29M			0.74 ± 0.24	0.70 ± 0.27	0.34 ± 0.32	0.20 ± 0.31	0.09 ± 0.23					
ResEnc-L-NSCLC-RM	527.47M			0.16 ± 0.21	0.19 ± 0.24	0.62 ± 0.31	0.05 ± 0.15	0.00 ± 0.00					
ResEnc-L-LIDC	383.47M			0.52 ± 0.23	0.53 ± 0.23	0.10 ± 0.23	0.59 ± 0.30	0.32 ± 0.33					
ResEnc-L-LNDb	383.47M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
ResEnc-XL-MSD-Lu	527.70M			0.58 ± 0.34	0.69 ± 0.26	0.39 ± 0.34	0.24 ± 0.34	0.12 ± 0.27					
ResEnc-XL-NSCLC-RG	527.70M			0.71 ± 0.29	0.74 ± 0.22	0.37 ± 0.31	0.26 ± 0.33	0.12 ± 0.29					
ResEnc-XL-NSCLC-RM	527.47M			0.13 ± 0.21	0.16 ± 0.24	0.58 ± 0.33	0.06 ± 0.16	0.00 ± 0.01					
ResEnc-XL-LIDC	527.70M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
ResEnc-XL-LNDb	527.70M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
MedNeXt-M-MSD-Lu	17.55M			0.50 ± 0.36	0.63 ± 0.32	0.36 ± 0.31	0.21 ± 0.32	0.04 ± 0.15					
MedNeXt-M-NSCLC-RG	17.55M			0.73 ± 0.27	0.69 ± 0.28	0.39 ± 0.32	0.21 ± 0.32	0.10 ± 0.26					
MedNeXt-M-NSCLC-RM	17.55M			0.17 ± 0.24	0.17 ± 0.25	0.61 ± 0.30	0.04 ± 0.14	0.00 ± 0.00					
MedNeXt-M-LIDC	17.55M			0.50 ± 0.26	0.44 ± 0.29	0.10 ± 0.22	0.56 ± 0.31	0.36 ± 0.33					
MedNeXt-M-LNDb	17.55M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
Primus-M-MSD-Lu	245.34M			0.57 ± 0.34	0.62 ± 0.35	0.37 ± 0.32	0.20 ± 0.32	0.05 ± 0.19					
Primus-M-NSCLC-RG	245.34M			0.75 ± 0.27	0.69 ± 0.30	0.40 ± 0.31	0.22 ± 0.33	0.12 ± 0.28					
Primus-M-NSCLC-RM	244.67M			0.13 ± 0.21	0.18 ± 0.25	0.55 ± 0.35	0.03 ± 0.11	0.00 ± 0.00					
Primus-M-LIDC	245.34M			0.53 ± 0.27	0.52 ± 0.27	0.10 ± 0.22	0.57 ± 0.31	0.34 ± 0.35					
Primus-M-LNDb	245.34M			0.51 ± 0.32	0.49 ± 0.27	0.07 ± 0.20	0.39 ± 0.37	0.22 ± 0.32					
Primus-L-MSD-Lu	429.30M			0.57 ± 0.34	0.65 ± 0.32	0.40 ± 0.32	0.21 ± 0.33	0.04 ± 0.15					
Primus-L-NSCLC-RG	429.30M			0.76 ± 0.26	0.69 ± 0.28	0.40 ± 0.31	0.21 ± 0.32	0.11 ± 0.26					
Primus-L-NSCLC-RM	428.49M			0.13 ± 0.20	0.19 ± 0.25	0.58 ± 0.33	0.03 ± 0.11	0.00 ± 0.00					
Primus-L-LIDC	429.30M			0.55 ± 0.24	0.52 ± 0.25	0.09 ± 0.22	0.55 ± 0.32	0.33 ± 0.34					
Primus-L-LNDb	429.30M			0.54 ± 0.29	0.49 ± 0.27	0.07 ± 0.20	0.44 ± 0.36	0.28 ± 0.36					
nnU-Net-MSD-Lv	88.62M								0.48 ± 0.31	0.36 ± 0.31	0.62 ± 0.27	0.45 ± 0.34	
nnU-Net-TRIALS	88.62M								0.41 ± 0.33	0.60 ± 0.27	0.60 ± 0.25	0.57 ± 0.34	
nnU-Net-CLiMet	88.21M								0.40 ± 0.34	0.31 ± 0.29	0.62 ± 0.29	0.50 ± 0.33	
nnU-Net-HCC-TS	88.02M								0.22 ± 0.35	0.24 ± 0.36	0.41 ± 0.33	0.67 ± 0.31	
ResEnc-L-MSD-Lv	383.47M								0.51 ± 0.33	0.41 ± 0.31	0.63 ± 0.29	0.50 ± 0.34	
ResEnc-L-TRIALS	383.47M								0.35 ± 0.36	0.64 ± 0.30	0.57 ± 0.26	0.60 ± 0.32	
ResEnc-L-CLiMet	527.29M								0.42 ± 0.33	0.32 ± 0.28	0.66 ± 0.27	0.48 ± 0.33	
ResEnc-L-HCC-TS	527.06M								0.25 ± 0.36	0.32 ± 0.38	0.39 ± 0.36	0.72 ± 0.27	
ResEnc-XL-MSD-Lv	383.47M								0.51 ± 0.34	0.39 ± 0.31	0.63 ± 0.29	0.47 ± 0.35	
ResEnc-XL-TRIALS	383.47M								0.35 ± 0.34	0.67 ± 0.27	0.57 ± 0.29	0.59 ± 0.34	
ResEnc-XL-CLiMet	527.70M								0.46 ± 0.32	0.34 ± 0.32	0.66 ± 0.28	0.48 ± 0.34	
ResEnc-XL-HCC-TS	527.47M								0.25 ± 0.35	0.27 ± 0.34	0.41 ± 0.34	0.69 ± 0.30	
MedNeXt-M-MSD-Lv	17.55M								0.60 ± 0.28	0.38 ± 0.31	0.64 ± 0.28	0.46 ± 0.36	
MedNeXt-M-TRIALS	17.55M								0.42 ± 0.34	0.64 ± 0.29	0.62 ± 0.25	0.53 ± 0.36	
MedNeXt-M-CLiMet	17.55M								0.46 ± 0.32	0.37 ± 0.32	0.69 ± 0.25	0.48 ± 0.35	
MedNeXt-M-HCC-TS	17.55M								0.25 ± 0.35	0.28 ± 0.34	0.37 ± 0.36	0.73 ± 0.27	
Primus-M-MSD-Lv	244.73M								0.48 ± 0.35	0.40 ± 0.30	0.63 ± 0.29	0.47 ± 0.35	
Primus-M-TRIALS	244.73M								0.37 ± 0.35	0.59 ± 0.31	0.62 ± 0.27	0.59 ± 0.32	
Primus-M-CLiMet	245.17M								0.45 ± 0.33	0.32 ± 0.33	0.66 ± 0.27	0.48 ± 0.31	
Primus-M-HCC-TS	244.58M								0.26 ± 0.36	0.23 ± 0.33	0.37 ± 0.35	0.67 ± 0.32	
Primus-L-MSD-Lv	428.56M								0.51 ± 0.34	0.38 ± 0.31	0.64 ± 0.28	0.48 ± 0.33	
Primus-L-TRIALS	428.56M								0.37 ± 0.35	0.58 ± 0.30	0.59 ± 0.28	0.63 ± 0.30	
Primus-L-CLiMet	429.10M								0.45 ± 0.32	0.32 ± 0.32	0.65 ± 0.29	0.49 ± 0.31	
Primus-L-HCC-TS	428.37M								0.25 ± 0.35	0.27 ± 0.36	0.35 ± 0.35	0.68 ± 0.32	
nnU-Net-MSD-Co	88.02M												0.42 ± 0.36
ResEnc-L-MSD-Co	527.47M												0.42 ± 0.40
ResEnc-XL-MSD-Co	527.47M												0.43 ± 0.39
MedNeXt-M-MSD-Co	17.55M												0.47 ± 0.38
Primus-M-MSD-Co	244.67M												0.37 ± 0.38
Primus-L-MSD-Co	428.49M												0.38 ± 0.39

Table 9: **Per dataset evaluation DSC (continued).**

Model	Params	Pancreas −− Test		Head & Neck −− Test							Adrenal −− Test	Multi-Loc −− Test	
		MSD-Pa	PANO	HECKTOR	SegRap	HN-Cet	HN-HN1	RADCURE	HNSCC	HN-PET	AdACC	AutoPET	ULS23
nnU-Net-MSD-Pa	88.02M	0.47 ± 0.30	0.38 ± 0.33										
nnU-Net-PANO	88.43M	0.51 ± 0.35	0.43 ± 0.32										
ResEnc-L-MSD-Pa	527.06M	0.59 ± 0.31	0.43 ± 0.32										
ResEnc-L-PANO	383.24M	0.47 ± 0.36	0.61 ± 0.32										
ResEnc-XL-MSD-Pa	527.47M	0.52 ± 0.34	0.43 ± 0.32										
ResEnc-XL-PANO	527.88M	0.50 ± 0.34	0.63 ± 0.31										
MedNeXt-M-MSD-Pa	17.55M	0.49 ± 0.35	0.39 ± 0.32										
MedNeXt-M-PANO	17.55M	0.46 ± 0.34	0.53 ± 0.36										
Primus-M-MSD-Pa	244.58M	0.50 ± 0.33	0.37 ± 0.33										
Primus-M-PANO	244.64M	0.46 ± 0.34	0.55 ± 0.34										
Primus-L-MSD-Pa	428.37M	0.43 ± 0.34	0.37 ± 0.33										
Primus-L-PANO	428.46M	0.44 ± 0.33	0.50 ± 0.34										
nnU-Net-HECKTOR	88.21M			0.63 ± 0.22	0.06 ± 0.07	0.43 ± 0.16	0.28 ± 0.30	0.21 ± 0.24	0.26 ± 0.27	0.52 ± 0.20			
nnU-Net-SegRap	126.23M			0.07 ± 0.08	0.78 ± 0.08	0.21 ± 0.10	0.02 ± 0.06	0.05 ± 0.13	0.01 ± 0.05	0.11 ± 0.12			
nnU-Net-HN-Cet	88.02M			0.46 ± 0.27	0.05 ± 0.07	0.42 ± 0.23	0.20 ± 0.27	0.18 ± 0.26	0.23 ± 0.26	0.45 ± 0.27			
nnU-Net-HN-HN1	88.02M			0.21 ± 0.24	0.0 ± 0.02	0.30 ± 0.14	0.42 ± 0.37	0.32 ± 0.32	0.36 ± 0.31	0.33 ± 0.33			
nnU-Net-RADCURE	88.43M			0.20 ± 0.19	0.18 ± 0.15	0.36 ± 0.18	0.35 ± 0.32	0.41 ± 0.30	0.33 ± 0.26	0.41 ± 0.32			
nnU-Net-HNSCC	88.43M			0.33 ± 0.22	0.02 ± 0.02	0.34 ± 0.17	0.34 ± 0.32	0.29 ± 0.26	0.48 ± 0.30	0.44 ± 0.21			
nnU-Net-HN-PET	88.21M			0.38 ± 0.23	0.08 ± 0.07	0.56 ± 0.16	0.30 ± 0.30	0.31 ± 0.27	0.29 ± 0.25	0.56 ± 0.23			
ResEnc-L-HECKTOR	383.47M			0.64 ± 0.22	0.07 ± 0.08	0.42 ± 0.19	0.28 ± 0.30	0.20 ± 0.23	0.28 ± 0.26	0.54 ± 0.20			
ResEnc-L-SegRap	525.52M			0.05 ± 0.07	0.78 ± 0.07	0.19 ± 0.12	0.02 ± 0.06	0.05 ± 0.13	0.01 ± 0.06	0.09 ± 0.12			
ResEnc-L-HN-Cet	527.47M			0.48 ± 0.25	0.05 ± 0.07	0.42 ± 0.25	0.26 ± 0.30	0.20 ± 0.26	0.26 ± 0.27	0.50 ± 0.24			
ResEnc-L-HN-HN1	527.47M			0.22 ± 0.24	0.01 ± 0.02	0.21 ± 0.14	0.45 ± 0.35	0.32 ± 0.31	0.36 ± 0.32	0.30 ± 0.30			
ResEnc-L-RADCURE	527.47M			0.25 ± 0.20	0.22 ± 0.13	0.35 ± 0.20	0.40 ± 0.34	0.52 ± 0.28	0.41 ± 0.27	0.46 ± 0.31			
ResEnc-L-HNSCC	383.24M			0.42 ± 0.23	0.00 ± 0.01	0.29 ± 0.17	0.37 ± 0.34	0.28 ± 0.29	0.53 ± 0.33	0.48 ± 0.24			
ResEnc-L-HN-PET	527.29M			0.37 ± 0.25	0.10 ± 0.09	0.52 ± 0.23	0.26 ± 0.32	0.29 ± 0.29	0.26 ± 0.27	0.56 ± 0.22			
ResEnc-XL-HECKTOR	527.70M			0.65 ± 0.20	0.06 ± 0.07	0.41 ± 0.17	0.30 ± 0.29	0.21 ± 0.23	0.27 ± 0.25	0.55 ± 0.19			
ResEnc-XL-SegRap	525.52M			0.07 ± 0.08	0.78 ± 0.08	0.19 ± 0.11	0.02 ± 0.06	0.05 ± 0.13	0.01 ± 0.06	0.11 ± 0.12			
ResEnc-XL-HN-Cet	527.47M			0.45 ± 0.28	0.04 ± 0.06	0.40 ± 0.25	0.19 ± 0.28	0.18 ± 0.25	0.26 ± 0.26	0.45 ± 0.25			
ResEnc-XL-HN-HN1	527.47M			0.23 ± 0.25	0.01 ± 0.02	0.20 ± 0.14	0.48 ± 0.34	0.32 ± 0.31	0.33 ± 0.31	0.28 ± 0.31			
ResEnc-XL-RADCURE	527.47M			0.25 ± 0.20	0.22 ± 0.12	0.34 ± 0.19	0.40 ± 0.34	0.54 ± 0.28	0.39 ± 0.27	0.44 ± 0.31			
ResEnc-XL-HNSCC	527.88M			0.38 ± 0.25	0.00 ± 0.01	0.29 ± 0.17	0.37 ± 0.34	0.28 ± 0.29	0.53 ± 0.32	0.49 ± 0.24			
ResEnc-XL-HN-PET	527.70M			0.37 ± 0.25	0.07 ± 0.06	0.49 ± 0.23	0.29 ± 0.32	0.30 ± 0.29	0.27 ± 0.27	0.54 ± 0.26			
MedNeXt-M-HECKTOR	17.55M			0.64 ± 0.21	0.07 ± 0.08	0.45 ± 0.19	0.25 ± 0.29	0.22 ± 0.24	0.26 ± 0.27	0.55 ± 0.19			
MedNeXt-M-SegRap	17.55M			0.09 ± 0.10	0.78 ± 0.08	0.19 ± 0.11	0.02 ± 0.06	0.06 ± 0.13	0.01 ± 0.06	0.12 ± 0.13			
MedNeXt-M-HN-Cet	17.55M			0.47 ± 0.27	0.05 ± 0.07	0.44 ± 0.21	0.32 ± 0.32	0.24 ± 0.28	0.25 ± 0.26	0.52 ± 0.26			
MedNeXt-M-HN-HN1	17.55M			0.23 ± 0.25	0.02 ± 0.03	0.34 ± 0.20	0.49 ± 0.35	0.35 ± 0.31	0.37 ± 0.32	0.37 ± 0.33			
MedNeXt-M-RADCURE	17.55M			0.19 ± 0.18	0.23 ± 0.18	0.37 ± 0.19	0.37 ± 0.34	0.43 ± 0.29	0.31 ± 0.26	0.39 ± 0.33			
MedNeXt-M-HNSCC	17.55M			0.17 ± 0.20	0.00 ± 0.00	0.09 ± 0.07	0.32 ± 0.31	0.08 ± 0.12	0.47 ± 0.35	0.15 ± 0.15			
MedNeXt-M-HN-PET	17.55M			0.41 ± 0.24	0.08 ± 0.08	0.52 ± 0.24	0.30 ± 0.31	0.32 ± 0.28	0.28 ± 0.25	0.55 ± 0.26			
Primus-M-HECKTOR	245.34M			0.64 ± 0.21	0.05 ± 0.07	0.43 ± 0.18	0.27 ± 0.29	0.22 ± 0.24	0.26 ± 0.27	0.54 ± 0.19			
Primus-M-SegRap	243.84M			0.05 ± 0.07	0.76 ± 0.07	0.16 ± 0.09	0.03 ± 0.09	0.06 ± 0.13	0.02 ± 0.06	0.10 ± 0.12			
Primus-M-HN-Cet	244.67M			0.42 ± 0.30	0.04 ± 0.07	0.39 ± 0.25	0.21 ± 0.30	0.16 ± 0.25	0.22 ± 0.28	0.42 ± 0.29			
Primus-M-HN-HN1	244.67M			0.18 ± 0.23	0.01 ± 0.01	0.23 ± 0.20	0.44 ± 0.35	0.29 ± 0.31	0.34 ± 0.30	0.25 ± 0.32			
Primus-M-RADCURE	244.51M			0.19 ± 0.20	0.20 ± 0.15	0.40 ± 0.18	0.36 ± 0.33	0.45 ± 0.31	0.36 ± 0.26	0.38 ± 0.35			
Primus-M-HNSCC	244.64M			0.29 ± 0.23	0.01 ± 0.02	0.36 ± 0.17	0.40 ± 0.34	0.32 ± 0.29	0.51 ± 0.31	0.45 ± 0.25			
Primus-M-HN-PET	245.17M			0.35 ± 0.25	0.09 ± 0.10	0.55 ± 0.26	0.33 ± 0.30	0.32 ± 0.28	0.29 ± 0.27	0.53 ± 0.27			
Primus-L-HECKTOR	429.30M			0.63 ± 0.22	0.05 ± 0.07	0.43 ± 0.19	0.26 ± 0.30	0.22 ± 0.24	0.26 ± 0.28	0.55 ± 0.19			
Primus-L-SegRap	427.48M			0.04 ± 0.05	0.76 ± 0.07	0.17 ± 0.09	0.04 ± 0.09	0.06 ± 0.13	0.01 ± 0.06	0.09 ± 0.12			
Primus-L-HN-Cet	428.49M			0.43 ± 0.29	0.05 ± 0.08	0.39 ± 0.23	0.24 ± 0.29	0.17 ± 0.25	0.23 ± 0.27	0.44 ± 0.26			
Primus-L-HN-HN1	428.49M			0.16 ± 0.23	0.01 ± 0.03	0.22 ± 0.22	0.43 ± 0.35	0.30 ± 0.31	0.33 ± 0.30	0.24 ± 0.32			
Primus-L-RADCURE	428.29M			0.18 ± 0.20	0.18 ± 0.16	0.37 ± 0.17	0.36 ± 0.35	0.44 ± 0.31	0.36 ± 0.28	0.38 ± 0.35			
Primus-L-HNSCC	428.46M			0.26 ± 0.21	0.01 ± 0.02	0.34 ± 0.16	0.41 ± 0.34	0.32 ± 0.28	0.50 ± 0.31	0.39 ± 0.24			
Primus-L-HN-PET	429.10M			0.37 ± 0.24	0.08 ± 0.09	0.55 ± 0.24	0.32 ± 0.33	0.34 ± 0.28	0.30 ± 0.26	0.50 ± 0.30			
nnU-Net-AdACC	88.21M										0.84 ± 0.18		
ResEnc-L-AdACC	383.47M										0.85 ± 0.16		
ResEnc-XL-AdACC	527.70M										0.91 ± 0.05		
MedNeXt-M-AdACC	17.55M										0.86 ± 0.17		
Primus-M-AdACC	245.34M										0.85 ± 0.18		
Primus-L-AdACC	429.30M										0.84 ± 0.19		
nnU-Net-AutoPET	88.62M											0.24 ± 0.26	
nnU-Net-ULS23	88.21M												0.28 ± 0.30
ResEnc-L-AutoPET	383.47M											0.33 ± 0.27	
ResEnc-L-ULS23	383.47M												0.54 ± 0.32
ResEnc-XL-AutoPET	383.47M											0.33 ± 0.28	
ResEnc-XL-ULS23	527.70M												0.60 ± 0.29
MedNeXt-M-AutoPET	17.55M											0.08 ± 0.11	
MedNeXt-M-ULS23	17.55M												0.31 ± 0.30
Primus-M-AutoPET	244.73M											0.30 ± 0.28	
Primus-M-ULS23	245.34M												0.35 ± 0.31
Primus-L-AutoPET	428.56M											0.30 ± 0.27	
Primus-L-ULS23	429.30M												0.33 ± 0.30

Table 10: **Per dataset evaluation NSD:** Performances are reported as NSD score ($\mu \pm \sigma$), the higher the better. For each trained model we report the in distribution evaluation (test set corresponding to training set), out of distribution evaluation (same cancer type, but test set from a different dataset, excluding colon and adrenal gland), and cancer-agnostic performances (test set from different cancer types). Test sets cardinality is reported in Table 1. All models trained for 1k epochs.

Model	Params	Kidney – Test		Lung – Test					Liver – Test				Colon – Test
		KiTS23	KiPA22	MSD-Lu	NSCLC-RG	NSCLC-RM	LIDC	LNDb	MSD-Lv	TRIALS	CLiMet	HCC-TS	MSD-Co
nnU-Net-KiTS23	88.62M	0.71 ± 0.16	0.02 ± 0.01										
nnU-Net-KiPA22	88.21M	0.00 ± 0.00	0.78 ± 0.13										
ResEnc-L-KiTS23	383.47M	0.74 ± 0.16	0.00 ± 0.00										
ResEnc-L-KiPA22	383.47M	0.00 ± 0.00	0.76 ± 0.14										
ResEnc-XL-KiTS23	527.70M	0.74 ± 0.16	0.00 ± 0.00										
ResEnc-XL-KiPA22	383.47M	0.00 ± 0.00	0.76 ± 0.13										
MedNeXt-M-KiTS23	17.55M	0.75 ± 0.16	0.00 ± 0.00										
MedNeXt-M-KiPA22	17.55M	0.00 ± 0.00	0.72 ± 0.24										
Primus-M-KiTS23	244.73M	0.75 ± 0.15	0.00 ± 0.00										
Primus-M-KiPA22	245.06M	0.00 ± 0.00	0.74 ± 0.23										
Primus-L-KiTS23	428.56M	0.73 ± 0.17	0.00 ± 0.00										
Primus-L-KiPA22	428.96M	0.00 ± 0.00	0.75 ± 0.17										
nnU-Net-MSD-Lu	88.21M			0.56 ± 0.34	0.56 ± 0.29	0.19 ± 0.17	0.45 ± 0.31	0.53 ± 0.39					
nnU-Net-NSCLC-RG	88.21M			0.69 ± 0.28	0.60 ± 0.24	0.14 ± 0.15	0.42 ± 0.32	0.42 ± 0.37					
nnU-Net-NSCLC-RM	88.02M			0.05 ± 0.07	0.12 ± 0.13	0.46 ± 0.27	0.04 ± 0.12	0.00 ± 0.00					
nnU-Net-LIDC	88.21M			0.41 ± 0.23	0.30 ± 0.26	0.06 ± 0.14	0.57 ± 0.31	0.43 ± 0.36					
nnU-Net-LNDb	88.21M			0.45 ± 0.25	0.29 ± 0.28	0.07 ± 0.19	0.52 ± 0.32	0.43 ± 0.40					
ResEnc-L-MSD-Lu	527.29M			0.50 ± 0.36	0.55 ± 0.33	0.29 ± 0.18	0.53 ± 0.29	0.40 ± 0.37					
ResEnc-L-NSCLC-RG	527.29M			0.75 ± 0.22	0.64 ± 0.22	0.20 ± 0.15	0.47 ± 0.30	0.42 ± 0.38					
ResEnc-L-NSCLC-RM	527.47M			0.10 ± 0.13	0.11 ± 0.13	0.51 ± 0.25	0.08 ± 0.16	0.00 ± 0.00					
ResEnc-L-LIDC	383.47M			0.37 ± 0.23	0.34 ± 0.24	0.09 ± 0.19	0.68 ± 0.29	0.51 ± 0.39					
ResEnc-L-LNDb	383.47M			0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0					
ResEnc-XL-MSD-Lu	527.70M			0.55 ± 0.34	0.57 ± 0.29	0.27 ± 0.19	0.51 ± 0.30	0.47 ± 0.39					
ResEnc-XL-NSCLC-RG	527.70M			0.68 ± 0.33	0.62 ± 0.25	0.16 ± 0.15	0.43 ± 0.31	0.33 ± 0.40					
ResEnc-XL-NSCLC-RM	527.47M			0.15 ± 0.16	0.11 ± 0.14	0.49 ± 0.26	0.13 ± 0.19	0.02 ± 0.04					
ResEnc-XL-LIDC	527.70M			0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0					
ResEnc-XL-LNDb	527.70M			0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0					
MedNeXt-M-MSD-Lu	17.55M			0.50 ± 0.34	0.54 ± 0.32	0.19 ± 0.14	0.39 ± 0.32	0.17 ± 0.31					
MedNeXt-M-NSCLC-RG	17.55M			0.74 ± 0.27	0.64 ± 0.22	0.19 ± 0.15	0.45 ± 0.32	0.49 ± 0.40					
MedNeXt-M-NSCLC-RM	17.55M			0.09 ± 0.16	0.08 ± 0.14	0.49 ± 0.24	0.06 ± 0.14	0.00 ± 0.00					
MedNeXt-M-LIDC	17.55M			0.39 ± 0.25	0.30 ± 0.27	0.09 ± 0.19	0.64 ± 0.31	0.49 ± 0.37					
MedNeXt-M-LNDb	17.55M			0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00					
Primus-M-MSD-Lu	245.34M			0.52 ± 0.35	0.56 ± 0.34	0.23 ± 0.16	0.51 ± 0.29	0.39 ± 0.41					
Primus-M-NSCLC-RG	245.34M			0.72 ± 0.34	0.62 ± 0.28	0.18 ± 0.13	0.49 ± 0.32	0.60 ± 0.38					
Primus-M-NSCLC-RM	244.67M			0.07 ± 0.11	0.10 ± 0.14	0.48 ± 0.27	0.08 ± 0.14	0.00 ± 0.01					
Primus-M-LIDC	245.34M			0.45 ± 0.25	0.34 ± 0.26	0.09 ± 0.19	0.66 ± 0.30	0.48 ± 0.39					
Primus-M-LNDb	245.34M			0.48 ± 0.29	0.33 ± 0.28	0.13 ± 0.23	0.57 ± 0.33	0.57 ± 0.32					
Primus-L-MSD-Lu	429.30M			0.54 ± 0.34	0.57 ± 0.31	0.24 ± 0.15	0.51 ± 0.31	0.35 ± 0.34					
Primus-L-NSCLC-RG	429.30M			0.73 ± 0.32	0.60 ± 0.27	0.16 ± 0.15	0.45 ± 0.32	0.54 ± 0.36					
Primus-L-NSCLC-RM	428.49M			0.07 ± 0.09	0.10 ± 0.13	0.52 ± 0.22	0.07 ± 0.13	0.00 ± 0.00					
Primus-L-LIDC	429.30M			0.41 ± 0.27	0.32 ± 0.26	0.08 ± 0.18	0.66 ± 0.30	0.48 ± 0.38					
Primus-L-LNDb	429.30M			0.43 ± 0.31	0.32 ± 0.29	0.10 ± 0.21	0.59 ± 0.33	0.62 ± 0.34					
nnU-Net-MSD-Lv	88.62M								0.40 ± 0.25	0.23 ± 0.28	0.50 ± 0.23	0.21 ± 0.22	
nnU-Net-TRIALS	88.62M								0.29 ± 0.23	0.34 ± 0.24	0.46 ± 0.21	0.23 ± 0.18	
nnU-Net-CLiMet	88.21M								0.28 ± 0.22	0.20 ± 0.23	0.50 ± 0.24	0.22 ± 0.21	
nnU-Net-HCC-TS	88.02M								0.18 ± 0.20	0.16 ± 0.21	0.35 ± 0.22	0.36 ± 0.19	
ResEnc-L-MSD-Lv	383.47M								0.48 ± 0.20	0.26 ± 0.29	0.54 ± 0.23	0.22 ± 0.22	
ResEnc-L-TRIALS	383.47M								0.28 ± 0.28	0.38 ± 0.27	0.44 ± 0.21	0.26 ± 0.22	
ResEnc-L-CLiMet	527.29M								0.31 ± 0.22	0.17 ± 0.21	0.55 ± 0.22	0.21 ± 0.20	
ResEnc-L-HCC-TS	527.06M								0.20 ± 0.24	0.19 ± 0.21	0.39 ± 0.22	0.41 ± 0.14	
ResEnc-XL-MSD-Lv	383.47M								0.50 ± 0.23	0.26 ± 0.28	0.51 ± 0.23	0.22 ± 0.22	
ResEnc-XL-TRIALS	383.47M								0.30 ± 0.24	0.43 ± 0.25	0.46 ± 0.22	0.27 ± 0.21	
ResEnc-XL-CLiMet	527.70M								0.39 ± 0.24	0.24 ± 0.25	0.54 ± 0.25	0.21 ± 0.21	
ResEnc-XL-HCC-TS	527.47M								0.21 ± 0.21	0.13 ± 0.16	0.35 ± 0.21	0.41 ± 0.17	
MedNeXt-M-MSD-Lv	17.55M								0.52 ± 0.25	0.24 ± 0.29	0.53 ± 0.23	0.21 ± 0.22	
MedNeXt-M-TRIALS	17.55M								0.31 ± 0.24	0.39 ± 0.26	0.49 ± 0.22	0.24 ± 0.22	
MedNeXt-M-CLiMet	17.55M								0.35 ± 0.19	0.23 ± 0.25	0.61 ± 0.18	0.23 ± 0.22	
MedNeXt-M-HCC-TS	17.55M								0.24 ± 0.24	0.15 ± 0.14	0.36 ± 0.24	0.41 ± 0.19	
Primus-M-MSD-Lv	244.73M								0.44 ± 0.26	0.25 ± 0.28	0.53 ± 0.24	0.26 ± 0.23	
Primus-M-TRIALS	244.73M								0.29 ± 0.24	0.35 ± 0.25	0.48 ± 0.23	0.27 ± 0.21	
Primus-M-CLiMet	245.17M								0.35 ± 0.24	0.22 ± 0.25	0.54 ± 0.24	0.20 ± 0.21	
Primus-M-HCC-TS	244.58M								0.22 ± 0.23	0.14 ± 0.17	0.35 ± 0.21	0.42 ± 0.17	
Primus-L-MSD-Lv	428.56M								0.47 ± 0.25	0.25 ± 0.29	0.53 ± 0.23	0.25 ± 0.23	
Primus-L-TRIALS	428.56M								0.32 ± 0.23	0.35 ± 0.25	0.45 ± 0.22	0.29 ± 0.22	
Primus-L-CLiMet	429.10M								0.35 ± 0.21	0.22 ± 0.26	0.53 ± 0.26	0.21 ± 0.21	
Primus-L-HCC-TS	428.37M								0.24 ± 0.21	0.13 ± 0.17	0.30 ± 0.24	0.40 ± 0.18	
nnU-Net-MSD-Co	88.02M												0.53 ± 0.28
ResEnc-L-MSD-Co	527.47M												0.50 ± 0.34
ResEnc-XL-MSD-Co	527.47M												0.57 ± 0.30
MedNeXt-M-MSD-Co	17.55M												0.52 ± 0.32
Primus-M-MSD-Co	244.67M												0.51 ± 0.31
Primus-L-MSD-Co	428.49M												0.54 ± 0.29

Table 10: **Single-dataset NSD (continued).**

Model	Params	Pancreas −− Test		Head & Neck −− Test							Adrenal −− Test	Multi-Loc −− Test	
		MSD-Pa	PANO	HECKTOR	SegRap	HN-Cet	HN-HN1	RADCURE	HNSCC	HN-PET	AdACC	AutoPET	ULS23
nnU-Net-MSD-Pa	88.02M	0.39 ± 0.28	0.29 ± 0.25										
nnU-Net-PANO	88.43M	0.32 ± 0.25	0.39 ± 0.29										
ResEnc-L-MSD-Pa	527.06M	0.51 ± 0.24	0.30 ± 0.24										
ResEnc-L-PANO	383.24M	0.45 ± 0.29	0.53 ± 0.28										
ResEnc-XL-MSD-Pa	527.47M	0.43 ± 0.27	0.29 ± 0.25										
ResEnc-XL-PANO	527.88M	0.41 ± 0.29	0.51 ± 0.30										
MedNeXt-M-MSD-Pa	17.55M	0.42 ± 0.28	0.27 ± 0.23										
MedNeXt-M-PANO	17.55M	0.37 ± 0.27	0.43 ± 0.30										
Primus-M-MSD-Pa	244.58M	0.39 ± 0.27	0.26 ± 0.24										
Primus-M-PANO	244.64M	0.40 ± 0.27	0.43 ± 0.29										
Primus-L-MSD-Pa	428.37M	0.37 ± 0.26	0.26 ± 0.24										
Primus-L-PANO	428.46M	0.34 ± 0.26	0.38 ± 0.28										
nnU-Net-HECKTOR	88.21M			0.47 ± 0.19	0.03 ± 0.03	0.14 ± 0.08	0.18 ± 0.19	0.10 ± 0.12	0.17 ± 0.17	0.20 ± 0.15			
nnU-Net-SegRap	126.23M			0.05 ± 0.05	0.51 ± 0.11	0.13 ± 0.07	0.01 ± 0.03	0.02 ± 0.03	0.01 ± 0.02	0.07 ± 0.07			
nnU-Net-HN-Cet	88.02M			0.31 ± 0.17	0.03 ± 0.04	0.21 ± 0.14	0.16 ± 0.19	0.13 ± 0.16	0.14 ± 0.15	0.21 ± 0.19			
nnU-Net-HN-HN1	88.02M			0.13 ± 0.13	0.00 ± 0.01	0.13 ± 0.05	0.34 ± 0.27	0.24 ± 0.20	0.25 ± 0.18	0.18 ± 0.19			
nnU-Net-RADCURE	88.43M			0.09 ± 0.08	0.02 ± 0.03	0.14 ± 0.08	0.23 ± 0.21	0.26 ± 0.18	0.15 ± 0.14	0.20 ± 0.18			
nnU-Net-HNSCC	88.43M			0.18 ± 0.13	0.01 ± 0.01	0.11 ± 0.06	0.23 ± 0.22	0.17 ± 0.17	0.37 ± 0.24	0.20 ± 0.15			
nnU-Net-HN-PET	88.21M			0.19 ± 0.14	0.03 ± 0.03	0.22 ± 0.14	0.21 ± 0.21	0.18 ± 0.16	0.15 ± 0.13	0.26 ± 0.15			
ResEnc-L-HECKTOR	383.47M			0.49 ± 0.18	0.03 ± 0.04	0.15 ± 0.09	0.18 ± 0.18	0.11 ± 0.13	0.19 ± 0.17	0.22 ± 0.16			
ResEnc-L-SegRap	525.52M			0.05 ± 0.05	0.51 ± 0.12	0.13 ± 0.07	0.01 ± 0.03	0.02 ± 0.04	0.01 ± 0.02	0.06 ± 0.07			
ResEnc-L-HN-Cet	527.47M			0.31 ± 0.17	0.03 ± 0.04	0.21 ± 0.15	0.21 ± 0.22	0.14 ± 0.16	0.17 ± 0.16	0.24 ± 0.19			
ResEnc-L-HN-HN1	527.47M			0.14 ± 0.14	0.01 ± 0.01	0.08 ± 0.08	0.38 ± 0.26	0.22 ± 0.20	0.24 ± 0.20	0.15 ± 0.18			
ResEnc-L-RADCURE	527.47M			0.11 ± 0.09	0.02 ± 0.03	0.14 ± 0.09	0.28 ± 0.23	0.36 ± 0.18	0.20 ± 0.17	0.22 ± 0.20			
ResEnc-L-HNSCC	383.24M			0.24 ± 0.14	0.00 ± 0.01	0.09 ± 0.07	0.24 ± 0.23	0.18 ± 0.19	0.42 ± 0.28	0.23 ± 0.20			
ResEnc-L-HN-PET	527.29M			0.19 ± 0.15	0.04 ± 0.06	0.24 ± 0.17	0.19 ± 0.22	0.20 ± 0.18	0.14 ± 0.16	0.27 ± 0.15			
ResEnc-XL-HECKTOR	527.70M			0.48 ± 0.18	0.03 ± 0.03	0.14 ± 0.08	0.19 ± 0.19	0.10 ± 0.12	0.17 ± 0.16	0.22 ± 0.16			
ResEnc-XL-SegRap	525.52M			0.06 ± 0.05	0.51 ± 0.11	0.13 ± 0.07	0.01 ± 0.02	0.02 ± 0.04	0.01 ± 0.02	0.07 ± 0.08			
ResEnc-XL-HN-Cet	527.47M			0.31 ± 0.18	0.03 ± 0.03	0.19 ± 0.15	0.14 ± 0.19	0.13 ± 0.15	0.17 ± 0.16	0.19 ± 0.19			
ResEnc-XL-HN-HN1	527.47M			0.14 ± 0.14	0.01 ± 0.01	0.08 ± 0.08	0.39 ± 0.25	0.23 ± 0.20	0.22 ± 0.19	0.14 ± 0.19			
ResEnc-XL-RADCURE	527.47M			0.10 ± 0.09	0.02 ± 0.03	0.12 ± 0.08	0.28 ± 0.24	0.38 ± 0.20	0.19 ± 0.17	0.22 ± 0.19			
ResEnc-XL-HNSCC	527.88M			0.22 ± 0.15	0.01 ± 0.01	0.08 ± 0.07	0.25 ± 0.24	0.16 ± 0.19	0.43 ± 0.28	0.22 ± 0.18			
ResEnc-XL-HN-PET	527.70M			0.20 ± 0.15	0.02 ± 0.02	0.23 ± 0.15	0.26 ± 0.23	0.20 ± 0.18	0.15 ± 0.15	0.26 ± 0.18			
MedNeXt-M-HECKTOR	17.55M			0.47 ± 0.19	0.04 ± 0.04	0.17 ± 0.10	0.15 ± 0.18	0.11 ± 0.13	0.16 ± 0.17	0.23 ± 0.16			
MedNeXt-M-SegRap	17.55M			0.06 ± 0.05	0.51 ± 0.12	0.12 ± 0.07	0.01 ± 0.02	0.02 ± 0.03	0.01 ± 0.02	0.07 ± 0.08			
MedNeXt-M-HN-Cet	17.55M			0.31 ± 0.18	0.04 ± 0.04	0.22 ± 0.15	0.22 ± 0.22	0.16 ± 0.17	0.15 ± 0.15	0.26 ± 0.19			
MedNeXt-M-HN-HN1	17.55M			0.14 ± 0.14	0.01 ± 0.01	0.12 ± 0.09	0.39 ± 0.27	0.24 ± 0.21	0.23 ± 0.20	0.20 ± 0.20			
MedNeXt-M-RADCURE	17.55M			0.07 ± 0.08	0.04 ± 0.05	0.13 ± 0.07	0.24 ± 0.24	0.28 ± 0.17	0.14 ± 0.15	0.20 ± 0.20			
MedNeXt-M-HNSCC	17.55M			0.10 ± 0.10	0.01 ± 0.01	0.03 ± 0.02	0.21 ± 0.22	0.05 ± 0.07	0.39 ± 0.30	0.08 ± 0.10			
MedNeXt-M-HN-PET	17.55M			0.20 ± 0.15	0.03 ± 0.03	0.25 ± 0.14	0.22 ± 0.23	0.19 ± 0.17	0.15 ± 0.15	0.30 ± 0.18			
Primus-M-HECKTOR	245.34M			0.47 ± 0.18	0.02 ± 0.02	0.14 ± 0.08	0.17 ± 0.17	0.11 ± 0.12	0.17 ± 0.16	0.22 ± 0.15			
Primus-M-SegRap	243.84M			0.05 ± 0.04	0.47 ± 0.10	0.10 ± 0.06	0.02 ± 0.03	0.02 ± 0.04	0.01 ± 0.02	0.06 ± 0.06			
Primus-M-HN-Cet	244.67M			0.31 ± 0.18	0.04 ± 0.04	0.18 ± 0.13	0.22 ± 0.21	0.14 ± 0.16	0.14 ± 0.16	0.21 ± 0.20			
Primus-M-HN-HN1	244.67M			0.13 ± 0.13	0.01 ± 0.01	0.11 ± 0.08	0.36 ± 0.26	0.22 ± 0.20	0.22 ± 0.17	0.16 ± 0.18			
Primus-M-RADCURE	244.51M			0.08 ± 0.08	0.03 ± 0.04	0.16 ± 0.07	0.21 ± 0.22	0.28 ± 0.20	0.16 ± 0.15	0.19 ± 0.20			
Primus-M-HNSCC	244.64M			0.16 ± 0.12	0.01 ± 0.01	0.12 ± 0.07	0.29 ± 0.25	0.20 ± 0.19	0.40 ± 0.27	0.20 ± 0.16			
Primus-M-HN-PET	245.17M			0.19 ± 0.15	0.03 ± 0.05	0.24 ± 0.17	0.22 ± 0.20	0.21 ± 0.16	0.16 ± 0.16	0.27 ± 0.16			
Primus-L-HECKTOR	429.30M			0.47 ± 0.19	0.02 ± 0.03	0.14 ± 0.08	0.15 ± 0.18	0.11 ± 0.13	0.17 ± 0.18	0.22 ± 0.15			
Primus-L-SegRap	427.48M			0.04 ± 0.04	0.45 ± 0.10	0.11 ± 0.06	0.02 ± 0.04	0.02 ± 0.04	0.01 ± 0.02	0.05 ± 0.06			
Primus-L-HN-Cet	428.49M			0.30 ± 0.17	0.06 ± 0.05	0.18 ± 0.13	0.22 ± 0.20	0.14 ± 0.16	0.14 ± 0.15	0.22 ± 0.19			
Primus-L-HN-HN1	428.49M			0.11 ± 0.13	0.01 ± 0.01	0.09 ± 0.09	0.38 ± 0.25	0.22 ± 0.20	0.21 ± 0.17	0.16 ± 0.18			
Primus-L-RADCURE	428.29M			0.07 ± 0.08	0.03 ± 0.04	0.15 ± 0.07	0.22 ± 0.23	0.28 ± 0.20	0.17 ± 0.16	0.19 ± 0.20			
Primus-L-HNSCC	428.46M			0.14 ± 0.11	0.01 ± 0.01	0.11 ± 0.06	0.30 ± 0.24	0.19 ± 0.17	0.39 ± 0.26	0.17 ± 0.15			
Primus-L-HN-PET	429.10M			0.19 ± 0.14	0.02 ± 0.03	0.24 ± 0.16	0.25 ± 0.24	0.21 ± 0.17	0.16 ± 0.15	0.24 ± 0.16			
nnU-Net-AdACC	88.21M										0.62 ± 0.20		
ResEnc-L-AdACC	383.47M										0.64 ± 0.17		
ResEnc-XL-AdACC	527.70M										0.67 ± 0.13		
MedNeXt-M-AdACC	17.55M										0.66 ± 0.18		
Primus-M-AdACC	245.34M										0.63 ± 0.17		
Primus-L-AdACC	429.30M										0.63 ± 0.18		
nnU-Net-AutoPET	88.62M											0.11 ± 0.12	
nnU-Net-ULS23	88.21M												0.30 ± 0.32
ResEnc-L-AutoPET	383.47M											0.19 ± 0.15	
ResEnc-L-ULS23	383.47M												0.58 ± 0.36
ResEnc-XL-AutoPET	383.47M											0.20 ± 0.15	
ResEnc-XL-ULS23	527.70M												0.65 ± 0.32
MedNeXt-M-AutoPET	17.55M											0.03 ± 0.04	
MedNeXt-M-ULS23	17.55M												0.32 ± 0.33
Primus-M-AutoPET	244.73M											0.15 ± 0.14	
Primus-M-ULS23	245.34M												0.35 ± 0.32
Primus-L-AutoPET	428.56M											0.15 ± 0.14	
Primus-L-ULS23	429.30M												0.33 ± 0.32

Table 11: **Per Dataset Evaluation SRVE:** Performances are reported as symmetric relative volume error (SRVE, $\mu \pm \sigma$), the lower the better. SRVE is bounded in $[0, 100]\%$: 0 is a perfect volume match and 100 a complete miss. Arrows denote net over- ($\uparrow$) or under-prediction ($\downarrow$). For each trained model we report the in distribution evaluation (test set corresponding to training set), out of distribution evaluation (same cancer type, but test set from a different dataset, excluding colon and adrenal gland), and cancer-agnostic performances (test set from different cancer types). Test sets cardinality is reported in Table 1. All models trained for 1k epochs.

Model	Params	Kidney – Test		Lung – Test					Liver – Test				Colon – Test
		KiTS23	KiPA22	MSD-Lu	NSCLC-RG	NSCLC-RM	LIDC	LNDb	MSD-Lv	TRIALS	CLiMet	HCC-TS	MSD-Co
nnU-Net-KiTS23	88.62M	$9.85 \pm 15.04 \uparrow$	$86.22 \pm 24.77 \uparrow$										
nnU-Net-KiPA22	88.21M	$95.24 \pm 9.52 \uparrow$	$10.98 \pm 25.53 \downarrow$										
ResEnc-L-KiTS23	383.47M	$9.58 \pm 17.18 \downarrow$	$100.00 \pm 0.00 \downarrow$										
ResEnc-L-KiPA22	383.47M	$54.08 \pm 27.54 \downarrow$	$4.10 \pm 4.13 \downarrow$										
ResEnc-XL-KiTS23	527.70M	$9.80 \pm 17.07 \downarrow$	$100.00 \pm 0.00 \downarrow$										
ResEnc-XL-KiPA22	383.47M	$85.40 \pm 22.62 \uparrow$	$3.89 \pm 2.68 \downarrow$										
MedNeXt-M-KiTS23	17.55M	$8.01 \pm 13.30 \uparrow$	$100.00 \pm 0.00 \downarrow$										
MedNeXt-M-KiPA22	17.55M	$97.00 \pm 6.73 \uparrow$	$5.77 \pm 7.27 \downarrow$										
Primus-M-KiTS23	244.73M	$7.67 \pm 13.67 \downarrow$	$100.00 \pm 0.00 \downarrow$										
Primus-M-KiPA22	245.06M	$99.15 \pm 1.82 \uparrow$	$9.23 \pm 21.47 \uparrow$										
Primus-L-KiTS23	428.56M	$9.41 \pm 17.80 \uparrow$	$100.00 \pm 0.00 \downarrow$										
Primus-L-KiPA22	428.96M	$99.35 \pm 1.47 \uparrow$	$8.52 \pm 13.97 \uparrow$										
nnU-Net-MSD-Lu	88.21M			$29.44 \pm 30.69 \downarrow$	$26.84 \pm 25.20 \downarrow$	$50.00 \pm 36.07 \downarrow$	$74.48 \pm 37.38 \uparrow$	$90.69 \pm 24.27 \downarrow$					
nnU-Net-NSCLC-RG	88.21M			$11.41 \pm 10.76 \uparrow$	$18.19 \pm 19.07 \downarrow$	$50.29 \pm 31.36 \uparrow$	$74.43 \pm 36.78 \uparrow$	$87.35 \pm 26.42 \uparrow$					
nnU-Net-NSCLC-RM	88.02M			$65.61 \pm 22.24 \uparrow$	$71.43 \pm 27.28 \downarrow$	$30.70 \pm 33.39 \uparrow$	$80.64 \pm 28.67 \uparrow$	$89.13 \pm 21.44 \uparrow$					
nnU-Net-LIDC	88.21M			$36.62 \pm 28.84 \downarrow$	$38.37 \pm 27.70 \downarrow$	$80.28 \pm 24.80 \downarrow$	$34.89 \pm 32.27 \downarrow$	$47.22 \pm 33.89 \uparrow$					
nnU-Net-LNDb	88.21M			$36.60 \pm 28.77 \downarrow$	$44.25 \pm 29.12 \downarrow$	$87.72 \pm 25.85 \downarrow$	$39.11 \pm 33.06 \downarrow$	$53.23 \pm 38.49 \uparrow$					
ResEnc-L-MSD-Lu	527.29M			$36.57 \pm 34.81 \downarrow$	$35.89 \pm 35.42 \downarrow$	$51.09 \pm 38.26 \downarrow$	$75.58 \pm 38.03 \uparrow$	$95.82 \pm 16.75 \downarrow$					
ResEnc-L-NSCLC-RG	527.29M			$16.38 \pm 25.75 \downarrow$	$23.77 \pm 27.79 \downarrow$	$57.73 \pm 38.06 \downarrow$	$76.58 \pm 36.39 \uparrow$	$89.97 \pm 25.60 \uparrow$					
ResEnc-L-NSCLC-RM	527.47M			$77.36 \pm 25.51 \downarrow$	$75.80 \pm 26.27 \downarrow$	$26.68 \pm 29.43 \uparrow$	$84.66 \pm 24.09 \uparrow$	$86.74 \pm 24.61 \uparrow$					
ResEnc-L-LIDC	383.47M			$35.98 \pm 23.74 \downarrow$	$37.56 \pm 24.27 \downarrow$	$79.78 \pm 30.78 \downarrow$	$26.52 \pm 26.97 \downarrow$	$47.80 \pm 36.54 \uparrow$					
ResEnc-L-LNDb	383.47M			$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$					
ResEnc-XL-MSD-Lu	527.70M			$29.74 \pm 30.11 \downarrow$	$22.81 \pm 21.95 \downarrow$	$51.15 \pm 38.84 \downarrow$	$69.13 \pm 40.13 \uparrow$	$83.71 \pm 31.27 \uparrow$					
ResEnc-XL-NSCLC-RG	527.70M			$13.12 \pm 20.97 \uparrow$	$16.86 \pm 15.03 \downarrow$	$50.23 \pm 34.56 \uparrow$	$64.25 \pm 39.47 \uparrow$	$82.79 \pm 31.79 \uparrow$					
ResEnc-XL-NSCLC-RM	527.47M			$85.85 \pm 23.36 \downarrow$	$82.85 \pm 24.77 \downarrow$	$29.53 \pm 32.44 \downarrow$	$89.22 \pm 21.82 \uparrow$	$96.66 \pm 12.45 \uparrow$					
ResEnc-XL-LIDC	527.70M			$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$					
ResEnc-XL-LNDb	527.70M			$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$					
MedNeXt-M-MSD-Lu	17.55M			$36.52 \pm 34.77 \downarrow$	$25.70 \pm 25.67 \downarrow$	$53.47 \pm 36.30 \downarrow$	$71.33 \pm 38.01 \uparrow$	$93.91 \pm 16.39 \uparrow$					
MedNeXt-M-NSCLC-RG	17.55M			$16.38 \pm 29.64 \uparrow$	$23.55 \pm 29.10 \downarrow$	$50.13 \pm 36.25 \uparrow$	$74.75 \pm 37.34 \uparrow$	$88.10 \pm 27.50 \uparrow$					
MedNeXt-M-NSCLC-RM	17.55M			$61.42 \pm 30.78 \downarrow$	$72.42 \pm 27.59 \downarrow$	$24.08 \pm 28.06 \uparrow$	$77.01 \pm 30.45 \uparrow$	$81.20 \pm 25.28 \uparrow$					
MedNeXt-M-LIDC	17.55M			$30.07 \pm 24.54 \downarrow$	$44.37 \pm 32.92 \downarrow$	$75.85 \pm 32.10 \downarrow$	$28.09 \pm 29.88 \uparrow$	$40.20 \pm 38.55 \uparrow$					
MedNeXt-M-LNDb	17.55M			$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$	$100.00 \pm 0.00 \downarrow$					
Primus-M-MSD-Lu	245.34M			$22.61 \pm 26.40 \downarrow$	$25.91 \pm 32.22 \downarrow$	$51.91 \pm 36.93 \downarrow$	$75.35 \pm 38.24 \uparrow$	$91.59 \pm 23.39 \downarrow$					
Primus-M-NSCLC-RG	245.34M			$8.58 \pm 15.61 \uparrow$	$20.66 \pm 27.57 \downarrow$	$50.34 \pm 36.06 \uparrow$	$74.78 \pm 37.23 \uparrow$	$85.81 \pm 30.94 \uparrow$					
Primus-M-NSCLC-RM	244.67M			$85.83 \pm 22.64 \downarrow$	$73.46 \pm 31.39 \downarrow$	$35.47 \pm 38.15 \downarrow$	$90.93 \pm 21.13 \uparrow$	$91.48 \pm 23.77 \downarrow$					
Primus-M-LIDC	245.34M			$34.87 \pm 27.53 \downarrow$	$39.52 \pm 28.97 \downarrow$	$76.73 \pm 28.88 \downarrow$	$30.48 \pm 30.71 \downarrow$	$47.85 \pm 39.34 \uparrow$					
Primus-M-LNDb	245.34M			$40.68 \pm 35.34 \downarrow$	$44.58 \pm 28.20 \downarrow$	$86.52 \pm 28.39 \downarrow$	$51.72 \pm 39.91 \downarrow$	$70.91 \pm 36.95 \downarrow$					
Primus-L-MSD-Lu	429.30M			$29.13 \pm 31.26 \downarrow$	$24.00 \pm 26.99 \downarrow$	$50.88 \pm 37.71 \downarrow$	$74.07 \pm 38.84 \uparrow$	$95.71 \pm 15.99 \downarrow$					
Primus-L-NSCLC-RG	429.30M			$7.15 \pm 11.93 \uparrow$	$20.85 \pm 25.42 \downarrow$	$47.61 \pm 35.34 \uparrow$	$74.80 \pm 37.13 \uparrow$	$86.50 \pm 27.80 \uparrow$					
Primus-L-NSCLC-RM	428.49M			$84.31 \pm 22.10 \downarrow$	$72.28 \pm 31.24 \downarrow$	$30.88 \pm 35.03 \uparrow$	$89.74 \pm 22.63 \uparrow$	$95.20 \pm 12.80 \downarrow$					
Primus-L-LIDC	429.30M			$33.20 \pm 24.39 \downarrow$	$39.09 \pm 26.12 \downarrow$	$71.20 \pm 30.55 \downarrow$	$31.60 \pm 32.04 \downarrow$	$50.53 \pm 37.58 \uparrow$					
Primus-L-LNDb	429.30M			$36.13 \pm 30.92 \downarrow$	$44.24 \pm 28.55 \downarrow$	$87.00 \pm 26.90 \downarrow$	$45.50 \pm 38.67 \downarrow$	$63.39 \pm 42.14 \uparrow$					
nnU-Net-MSD-Lv	88.62M								$41.98 \pm 33.31 \downarrow$	$48.46 \pm 33.85 \downarrow$	$22.12 \pm 23.30 \downarrow$	$49.75 \pm 36.44 \downarrow$	
nnU-Net-TRIALS	88.62M								$35.46 \pm 30.57 \downarrow$	$20.88 \pm 19.92 \downarrow$	$23.08 \pm 18.99 \downarrow$	$35.01 \pm 33.70 \downarrow$	
nnU-Net-CLiMet	88.21M								$46.55 \pm 36.65 \downarrow$	$60.49 \pm 32.43 \downarrow$	$19.53 \pm 21.77 \downarrow$	$41.57 \pm 33.13 \downarrow$	
nnU-Net-HCC-TS	88.02M								$66.84 \pm 38.94 \downarrow$	$62.04 \pm 39.75 \downarrow$	$46.43 \pm 36.27 \downarrow$	$24.93 \pm 33.22 \downarrow$	
ResEnc-L-MSD-Lv	383.47M								$41.59 \pm 35.66 \downarrow$	$52.35 \pm 31.66 \downarrow$	$22.74 \pm 27.49 \downarrow$	$45.43 \pm 35.21 \downarrow$	
ResEnc-L-TRIALS	383.47M								$53.79 \pm 40.50 \downarrow$	$20.12 \pm 24.89 \downarrow$	$29.19 \pm 24.49 \downarrow$	$32.05 \pm 32.31 \downarrow$	
ResEnc-L-CLiMet	527.29M								$44.37 \pm 37.87 \downarrow$	$50.04 \pm 33.70 \downarrow$	$17.84 \pm 22.09 \downarrow$	$40.50 \pm 34.20 \downarrow$	
ResEnc-L-HCC-TS	527.06M								$66.10 \pm 39.62 \downarrow$	$53.24 \pm 40.39 \downarrow$	$50.06 \pm 39.78 \downarrow$	$18.69 \pm 29.55 \downarrow$	
ResEnc-XL-MSD-Lv	383.47M								$41.96 \pm 36.81 \downarrow$	$56.81 \pm 32.96 \downarrow$	$23.34 \pm 26.76 \downarrow$	$47.39 \pm 34.74 \downarrow$	
ResEnc-XL-TRIALS	383.47M								$55.54 \pm 39.58 \downarrow$	$18.92 \pm 25.96 \downarrow$	$32.64 \pm 28.67 \downarrow$	$33.51 \pm 34.33 \downarrow$	
ResEnc-XL-CLiMet	527.70M								$43.48 \pm 36.42 \downarrow$	$58.75 \pm 34.95 \downarrow$	$20.68 \pm 23.04 \downarrow$	$47.49 \pm 36.31 \downarrow$	
ResEnc-XL-HCC-TS	527.47M								$65.16 \pm 38.58 \downarrow$	$60.25 \pm 34.53 \downarrow$	$46.68 \pm 39.40 \downarrow$	$23.20 \pm 31.31 \downarrow$	
MedNeXt-M-MSD-Lv	17.55M								$28.12 \pm 30.72 \downarrow$	$48.22 \pm 35.33 \downarrow$	$20.96 \pm 26.96 \downarrow$	$48.90 \pm 37.17 \downarrow$	
MedNeXt-M-TRIALS	17.55M								$48.04 \pm 37.39 \downarrow$	$20.03 \pm 23.20 \downarrow$	$21.99 \pm 16.50 \downarrow$	$35.91 \pm 33.12 \downarrow$	
MedNeXt-M-CLiMet	17.55M								$40.48 \pm 38.01 \downarrow$	$48.06 \pm 36.42 \downarrow$	$15.97 \pm 21.46 \downarrow$	$42.33 \pm 35.85 \downarrow$	
MedNeXt-M-HCC-TS	17.55M								$60.08 \pm 42.31 \uparrow$	$61.08 \pm 35.29 \downarrow$	$51.74 \pm 39.83 \downarrow$	$16.43 \pm 27.36 \uparrow$	
Primus-M-MSD-Lv	244.73M								$40.65 \pm 36.59 \downarrow$	$49.45 \pm 31.81 \downarrow$	$21.64 \pm 25.75 \downarrow$	$47.77 \pm 38.19 \downarrow$	
Primus-M-TRIALS	244.73M								$47.58 \pm 37.55 \uparrow$	$23.94 \pm 25.69 \downarrow$	$25.22 \pm 23.75 \downarrow$	$33.85 \pm 32.27 \downarrow$	
Primus-M-CLiMet	245.17M								$43.12 \pm 39.30 \downarrow$	$53.34 \pm 36.91 \downarrow$	$19.84 \pm 23.90 \downarrow$	$42.11 \pm 31.98 \downarrow$	
Primus-M-HCC-TS	244.58M								$65.54 \pm 38.66 \downarrow$	$61.73 \pm 35.98 \downarrow$	$53.62 \pm 40.60 \downarrow$	$24.66 \pm 34.79 \uparrow$	
Primus-L-MSD-Lv	428.56M								$37.45 \pm 35.50 \downarrow$	$52.64 \pm 32.35 \downarrow$	$21.61 \pm 25.54 \downarrow$	$45.42 \pm 35.71 \downarrow$	
Primus-L-TRIALS	428.56M								$44.48 \pm 39.70 \downarrow$	$23.52 \pm 26.22 \downarrow$	$27.37 \pm 23.77 \downarrow$	$27.11 \pm 27.69 \downarrow$	
Primus-L-CLiMet	429.10M								$45.02 \pm 36.39 \downarrow$	$54.98 \pm 36.24 \downarrow$	$22.09 \pm 26.62 \downarrow$	$39.91 \pm 30.10 \downarrow$	
Primus-L-HCC-TS	428.37M								$69.58 \pm 38.24 \downarrow$	$56.46 \pm 35.15 \downarrow$	$54.60 \pm 40.08 \downarrow$	$22.88 \pm 34.46 \uparrow$	
nnU-Net-MSD-Co	88.02M												$46.60 \pm 38.76 \downarrow$
ResEnc-L-MSD-Co	527.47M												$43.19 \pm 37.87 \downarrow$
ResEnc-XL-MSD-Co	527.47M												$45.49 \pm 42.06 \downarrow$
MedNeXt-M-MSD-Co	17.55M												$39.92 \pm 36.39 \downarrow$
Primus-M-MSD-Co	244.67M												$51.78 \pm 39.18 \downarrow$
Primus-L-MSD-Co	428.49M												$50.06 \pm 43.39 \downarrow$

Table 11: **Per Dataset Evaluation SRVE** (continued).

Model	Params	Pancreas – Test		Head & Neck – Test							Adrenal – Test	Multi-Loc – Test	
		MSD-Pa	PANO	HECKTOR	SegRap	HN-Cet	HN-HN1	RADCURE	HNSCC	HN-PET	AdACC	AutoPET	ULS23
nnU-Net-MSD-Pa	88.02M	34.79 ± 33.83 ↓	44.60 ± 34.28 ↓										
nnU-Net-PANO	88.43M	37.86 ± 31.28 ↓	29.53 ± 28.62 ↑										
ResEnc-L-MSD-Pa	527.06M	27.93 ± 31.73 ↓	37.89 ± 30.30 ↓										
ResEnc-L-PANO	383.24M	41.51 ± 38.68 ↓	29.66 ± 32.00 ↓										
ResEnc-XL-MSD-Pa	527.47M	33.14 ± 35.05 ↓	39.16 ± 30.52 ↓										
ResEnc-XL-PANO	527.88M	35.85 ± 33.31 ↓	24.40 ± 26.72 ↓										
MedNeXt-M-MSD-Pa	17.55M	32.69 ± 32.52 ↓	44.48 ± 33.01 ↑										
MedNeXt-M-PANO	17.55M	37.80 ± 30.97 ↓	31.06 ± 31.58 ↓										
Primus-M-MSD-Pa	244.58M	30.75 ± 28.69 ↓	45.45 ± 32.34 ↓										
Primus-M-PANO	244.64M	38.88 ± 34.12 ↓	28.62 ± 27.49 ↓										
Primus-L-MSD-Pa	428.37M	36.21 ± 35.06 ↓	43.16 ± 32.37 ↓										
Primus-L-PANO	428.46M	40.46 ± 33.09 ↓	32.11 ± 29.61 ↑										
nnU-Net-HECKTOR	88.21M			18.87 ± 18.73 ↓	92.85 ± 7.72 ↓	53.02 ± 18.81 ↓	44.34 ± 33.38 ↓	45.53 ± 29.54 ↓	42.00 ± 27.79 ↑	31.31 ± 17.72 ↓			
nnU-Net-SegRap	126.23M			52.49 ± 24.21 ↑	11.83 ± 8.24 ↓	22.54 ± 17.17 ↑	70.11 ± 29.78 ↑	64.06 ± 26.82 ↑	74.08 ± 21.37 ↑	40.08 ± 20.45 ↑			
nnU-Net-HN-Cet	88.02M			32.82 ± 29.91 ↓	93.04 ± 7.85 ↓	53.31 ± 26.51 ↓	53.03 ± 36.84 ↓	54.67 ± 35.06 ↓	49.21 ± 27.81 ↑	33.91 ± 28.65 ↓			
nnU-Net-HN-HN1	88.02M			55.65 ± 34.48 ↓	96.72 ± 5.96 ↓	64.62 ± 18.06 ↓	40.05 ± 35.81 ↑	50.82 ± 36.70 ↓	47.56 ± 33.53 ↓	54.20 ± 36.93 ↓			
nnU-Net-RADCURE	88.43M			40.25 ± 26.94 ↑	68.30 ± 15.66 ↓	35.54 ± 27.39 ↓	51.36 ± 35.14 ↑	38.33 ± 28.81 ↑	46.65 ± 23.85 ↑	22.16 ± 20.30 ↓			
nnU-Net-HNSCC	88.43M			47.77 ± 27.67 ↑	89.97 ± 8.88 ↓	42.12 ± 29.76 ↓	41.27 ± 34.23 ↑	41.63 ± 28.62 ↑	28.99 ± 22.25 ↑	25.34 ± 17.94 ↑			
nnU-Net-HN-PET	88.21M			38.44 ± 29.16 ↑	87.87 ± 8.96 ↓	33.32 ± 23.40 ↓	53.07 ± 34.26 ↑	43.59 ± 30.86 ↑	46.86 ± 26.01 ↑	28.89 ± 21.72 ↓			
ResEnc-L-HECKTOR	383.47M			20.68 ± 23.28 ↓	91.46 ± 8.81 ↓	54.20 ± 21.25 ↓	45.24 ± 33.45 ↓	52.63 ± 32.52 ↓	43.63 ± 28.64 ↑	33.05 ± 17.63 ↓			
ResEnc-L-SegRap	525.52M			49.93 ± 25.15 ↑	12.03 ± 8.26 ↓	19.87 ± 17.00 ↑	69.97 ± 29.57 ↑	62.94 ± 26.89 ↑	73.69 ± 21.05 ↑	37.33 ± 20.50 ↑			
ResEnc-L-HN-Cet	527.47M			30.50 ± 26.94 ↑	91.51 ± 9.26 ↓	53.66 ± 29.03 ↓	56.74 ± 33.57 ↓	54.88 ± 35.01 ↓	48.17 ± 28.24 ↑	31.28 ± 26.08 ↓			
ResEnc-L-HN-HN1	527.47M			57.78 ± 35.12 ↓	96.21 ± 5.21 ↓	77.06 ± 16.87 ↓	42.04 ± 34.50 ↓	52.94 ± 34.12 ↓	45.37 ± 32.24 ↓	60.21 ± 33.89 ↓			
ResEnc-L-RADCURE	527.47M			37.00 ± 26.03 ↑	75.14 ± 13.20 ↓	49.71 ± 22.48 ↓	44.37 ± 34.85 ↑	29.29 ± 26.64 ↑	42.78 ± 23.59 ↑	32.11 ± 26.84 ↓			
ResEnc-L-HNSCC	383.24M			32.06 ± 24.09 ↑	97.26 ± 2.70 ↓	65.28 ± 25.17 ↓	42.92 ± 29.67 ↓	52.09 ± 30.91 ↓	24.83 ± 23.41 ↓	25.42 ± 19.38 ↓			
ResEnc-L-HN-PET	527.29M			37.77 ± 27.56 ↑	87.05 ± 11.00 ↓	37.81 ± 28.33 ↓	54.47 ± 35.28 ↑	50.36 ± 34.99 ↓	53.47 ± 27.59 ↑	27.52 ± 21.86 ↓			
ResEnc-XL-HECKTOR	527.70M			18.92 ± 19.04 ↓	90.01 ± 7.05 ↓	55.44 ± 20.15 ↓	44.44 ± 32.61 ↓	46.98 ± 29.19 ↓	41.13 ± 27.79 ↑	32.29 ± 17.63 ↓			
ResEnc-XL-SegRap	525.52M			50.46 ± 25.19 ↑	11.84 ± 8.14 ↓	22.44 ± 16.62 ↑	70.09 ± 28.52 ↑	62.67 ± 26.86 ↑	74.02 ± 21.22 ↑	37.88 ± 20.36 ↑			
ResEnc-XL-HN-Cet	527.47M			34.44 ± 31.33 ↓	93.36 ± 7.19 ↓	56.57 ± 28.44 ↓	59.03 ± 35.82 ↓	58.65 ± 35.03 ↓	48.58 ± 28.70 ↑	37.34 ± 27.72 ↓			
ResEnc-XL-HN-HN1	527.47M			59.40 ± 35.21 ↓	94.76 ± 7.52 ↓	76.48 ± 17.80 ↓	41.07 ± 34.54 ↓	52.27 ± 34.54 ↓	49.57 ± 31.67 ↓	60.97 ± 33.87 ↓			
ResEnc-XL-RADCURE	527.47M			45.44 ± 30.00 ↑	75.32 ± 12.54 ↓	43.22 ± 32.20 ↓	46.08 ± 34.83 ↑	27.53 ± 25.26 ↑	43.16 ± 24.36 ↑	30.14 ± 25.62 ↓			
ResEnc-XL-HNSCC	527.88M			35.10 ± 27.35 ↑	97.12 ± 2.45 ↓	59.50 ± 27.38 ↓	47.70 ± 34.00 ↓	48.11 ± 28.89 ↓	24.74 ± 21.22 ↓	26.56 ± 18.01 ↓			
ResEnc-XL-HN-PET	527.70M			37.53 ± 29.12 ↑	90.92 ± 7.13 ↓	44.68 ± 26.33 ↓	54.99 ± 37.35 ↓	49.86 ± 34.35 ↓	49.20 ± 25.08 ↑	26.46 ± 27.67 ↓			
MedNeXt-M-HECKTOR	17.55M			16.85 ± 17.30 ↓	89.01 ± 9.07 ↓	47.87 ± 22.66 ↓	45.19 ± 30.78 ↓	41.57 ± 27.39 ↓	40.98 ± 27.55 ↑	28.63 ± 14.96 ↓			
MedNeXt-M-SegRap	17.55M			52.20 ± 25.00 ↑	11.48 ± 8.41 ↓	21.23 ± 16.18 ↑	69.97 ± 28.68 ↑	64.39 ± 26.29 ↑	74.49 ± 21.10 ↑	41.27 ± 19.27 ↑			
MedNeXt-M-HN-Cet	17.55M			33.82 ± 27.20 ↑	88.84 ± 10.43 ↓	45.76 ± 28.83 ↓	47.97 ± 34.79 ↑	46.71 ± 33.44 ↑	50.66 ± 25.34 ↑	27.57 ± 21.22 ↓			
MedNeXt-M-HN-HN1	17.55M			48.14 ± 33.27 ↓	94.84 ± 5.56 ↓	55.99 ± 27.92 ↓	35.22 ± 34.01 ↑	42.53 ± 32.28 ↓	45.87 ± 31.36 ↑	53.85 ± 38.31 ↓			
MedNeXt-M-RADCURE	17.55M			45.15 ± 28.20 ↑	62.40 ± 16.77 ↓	26.07 ± 20.24 ↓	54.96 ± 35.22 ↑	40.92 ± 29.13 ↑	51.47 ± 26.38 ↑	24.07 ± 23.60 ↑			
MedNeXt-M-HNSCC	17.55M			51.34 ± 30.00 ↑	89.93 ± 15.40 ↓	35.79 ± 24.76 ↑	54.25 ± 35.34 ↑	68.65 ± 26.88 ↑	35.01 ± 29.98 ↑	48.53 ± 24.32 ↑			
MedNeXt-M-HN-PET	17.55M			34.81 ± 24.63 ↑	69.38 ± 30.26 ↓	32.75 ± 25.54 ↓	49.49 ± 32.98 ↑	42.80 ± 30.97 ↑	49.08 ± 24.24 ↑	19.73 ± 15.79 ↓			
Primus-M-HECKTOR	245.34M			19.67 ± 19.92 ↓	93.68 ± 7.35 ↓	52.54 ± 20.78 ↓	50.98 ± 31.58 ↓	45.08 ± 28.81 ↓	43.34 ± 28.76 ↓	30.23 ± 15.02 ↓			
Primus-M-SegRap	243.84M			55.85 ± 25.00 ↑	12.16 ± 9.23 ↓	24.53 ± 19.00 ↑	71.54 ± 28.89 ↑	66.28 ± 25.41 ↑	75.78 ± 20.85 ↑	42.76 ± 20.86 ↑			
Primus-M-HN-Cet	244.67M			39.98 ± 35.01 ↓	93.73 ± 7.98 ↓	58.07 ± 26.75 ↓	66.12 ± 38.76 ↓	60.60 ± 36.11 ↓	52.00 ± 30.97 ↑	43.67 ± 33.05 ↓			
Primus-M-HN-HN1	244.67M			63.71 ± 35.22 ↓	90.66 ± 14.89 ↓	72.39 ± 24.57 ↓	45.01 ± 38.82 ↑	55.81 ± 36.42 ↓	48.05 ± 30.27 ↓	63.99 ± 37.32 ↓			
Primus-M-RADCURE	244.51M			46.31 ± 28.81 ↑	69.15 ± 18.45 ↓	36.00 ± 21.45 ↓	51.25 ± 31.47 ↑	33.48 ± 27.64 ↑	46.67 ± 24.61 ↑	36.02 ± 31.94 ↓			
Primus-M-HNSCC	244.64M			37.17 ± 27.24 ↑	94.39 ± 3.96 ↓	56.81 ± 24.10 ↓	43.45 ± 35.59 ↑	40.60 ± 26.31 ↓	26.56 ± 23.27 ↑	21.61 ± 13.89 ↓			
Primus-M-HN-PET	245.17M			41.23 ± 31.62 ↑	88.65 ± 11.89 ↓	31.69 ± 24.43 ↓	43.70 ± 31.96 ↑	45.99 ± 33.70 ↓	51.47 ± 28.82 ↑	26.84 ± 23.49 ↓			
Primus-L-HECKTOR	429.30M			19.59 ± 20.04 ↓	93.25 ± 7.92 ↓	53.91 ± 20.91 ↓	48.47 ± 31.66 ↓	49.49 ± 30.99 ↓	43.91 ± 31.23 ↓	30.43 ± 15.29 ↓			
Primus-L-SegRap	427.48M			53.48 ± 25.57 ↑	12.59 ± 8.54 ↓	23.22 ± 20.27 ↑	71.68 ± 28.29 ↑	65.74 ± 25.63 ↑	74.61 ± 21.61 ↑	40.91 ± 21.58 ↑			
Primus-L-HN-Cet	428.49M			39.28 ± 34.03 ↓	93.29 ± 8.85 ↓	58.30 ± 25.00 ↓	59.64 ± 38.73 ↓	59.03 ± 36.09 ↓	50.28 ± 31.33 ↑	42.18 ± 31.53 ↓			
Primus-L-HN-HN1	428.49M			63.31 ± 34.01 ↓	90.73 ± 19.71 ↓	73.29 ± 25.48 ↓	36.13 ± 36.59 ↑	52.44 ± 36.05 ↓	50.32 ± 32.23 ↓	63.10 ± 37.38 ↓			
Primus-L-RADCURE	428.29M			47.26 ± 29.46 ↑	71.89 ± 17.81 ↓	42.12 ± 21.77 ↓	47.70 ± 33.03 ↑	34.05 ± 28.43 ↑	43.01 ± 25.04 ↑	34.31 ± 30.40 ↓			
Primus-L-HNSCC	428.46M			46.79 ± 27.08 ↑	95.35 ± 3.83 ↓	54.99 ± 26.35 ↓	44.85 ± 32.87 ↓	40.18 ± 27.14 ↓	26.47 ± 22.25 ↑	28.23 ± 19.57 ↑			
Primus-L-HN-PET	429.10M			40.21 ± 29.42 ↑	88.27 ± 12.47 ↓	30.19 ± 22.63 ↓	48.24 ± 37.62 ↑	41.26 ± 31.73 ↓	49.87 ± 26.59 ↑	25.24 ± 19.69 ↓			
nnU-Net-AdACC	88.21M										11.57 ± 19.13 ↓		
ResEnc-L-AdACC	383.47M										10.46 ± 16.74 ↓		
ResEnc-XL-AdACC	527.70M										4.98 ± 6.13 ↓		
MedNeXt-M-AdACC	17.55M										9.28 ± 18.06 ↓		
Primus-M-AdACC	245.34M										11.18 ± 18.78 ↓		
Primus-L-AdACC	429.30M										12.06 ± 20.36 ↓		
nnU-Net-AutoPET	88.62M											41.45 ± 28.34 ↓	
nnU-Net-ULS23	88.21M												56.54 ± 32.79 ↑
ResEnc-L-AutoPET	383.47M											36.51 ± 28.30 ↓	
ResEnc-L-ULS23	383.47M												30.81 ± 28.01 ↑
ResEnc-XL-AutoPET	383.47M											42.19 ± 32.89 ↓	
ResEnc-XL-ULS23	527.70M												26.88 ± 25.83 ↑
MedNeXt-M-AutoPET	17.55M											74.39 ± 26.21 ↑	
MedNeXt-M-ULS23	17.55M												57.38 ± 33.39 ↑
Primus-M-AutoPET	244.73M											34.94 ± 27.51 ↓	
Primus-M-ULS23	245.34M												48.08 ± 32.67 ↑
Primus-L-AutoPET	428.56M											35.93 ± 27.21 ↓	
Primus-L-ULS23	429.30M												49.01 ± 32.57 ↑



Figure 4: **Boxplots showing DSC of generalist model performance on all cases in LesionBench split by dataset:** Performance is highly heterogeneous across datasets and varies by model. KiPA22 is the easiest dataset to segment by a wide margin (mean DSC 0.79 across all eight models, versus 0.63 for the next-best, Adrenal-ACC). AutoPET, MSD-Colon, LNDb, PANORAMA, MSD-Liver, and TRIALS are the most difficult, with every model scoring below 0.20 mean DSC.

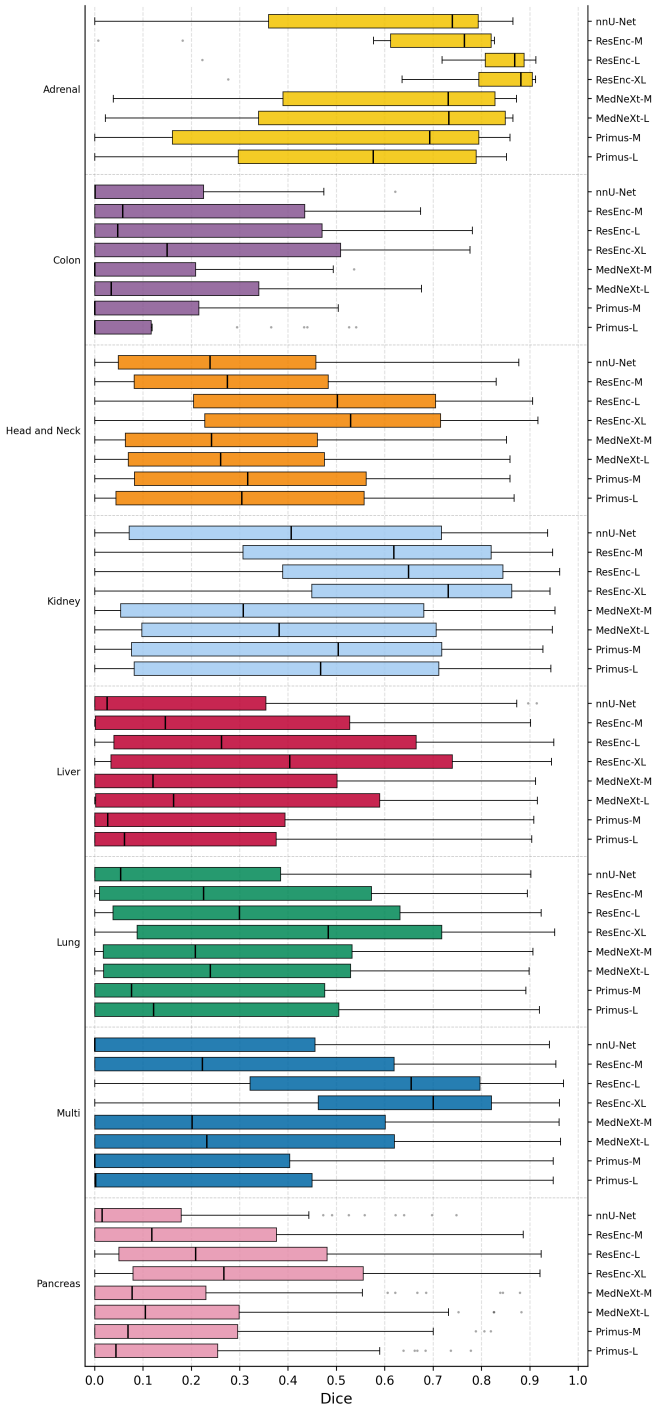


Figure 5: **Boxplots showing DSC of generalist model performance on all cases in LesionBench split by organ:** Performance is highly heterogeneous across organs and varies by model. Adrenal is the easiest lesions to segment, followed by kidney. Colon and pancreas are the most difficult with every model scoring below 0.33 on either.

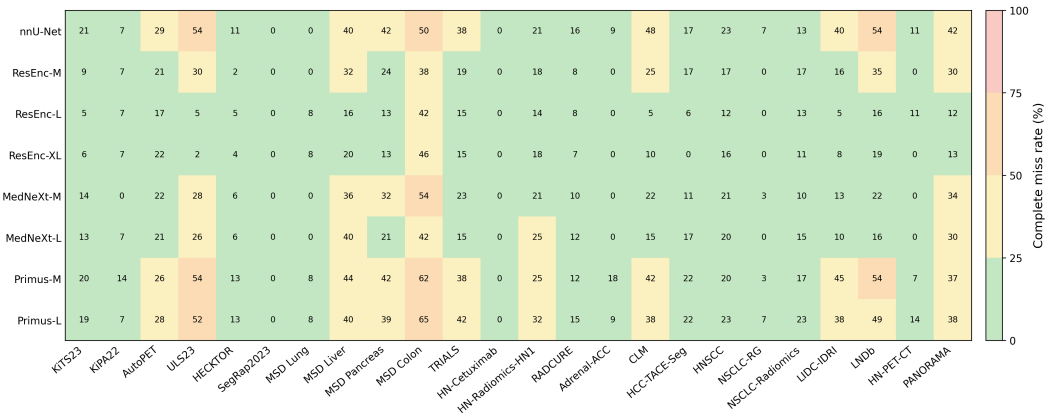


Figure 6: **Complete-miss rate by dataset in % per model:** Missed lesions are not equally distributed across datasets. Missed lesions are not equally distributed across datasets. MSD-Colon, MSD-Liver, LNDb, and ULS23 have the highest mean miss rates across all eight models, while SegRap2023 and Head-Neck-Cetuximab have no misses for any model. Nearly all datasets follow the same per-model ordering: nnU-Net has the highest miss rate and ResEnc-L or ResEnc-XL the lowest, with the gap between them often exceeding 30 percentage points.

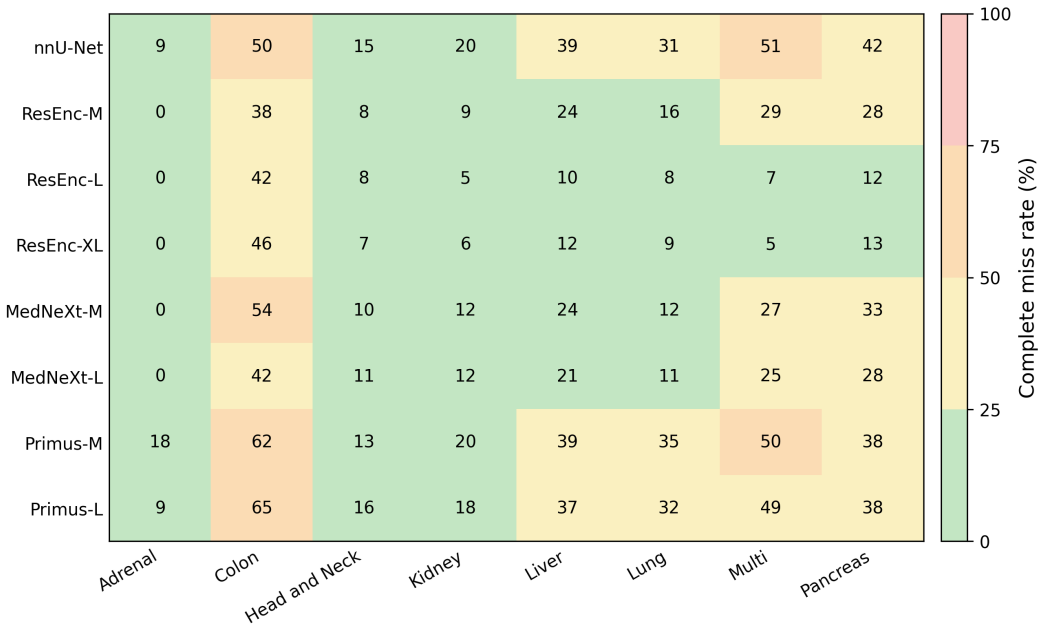


Figure 7: **Complete-miss rate by organ in % per model:** Missed lesions are not equally distributed across organs. Colon has the highest mean miss rate and remains difficult for every model regardless of capacity. Multi-location and pancreas are the next most difficult, while adrenal and H&N have the lowest miss rates.

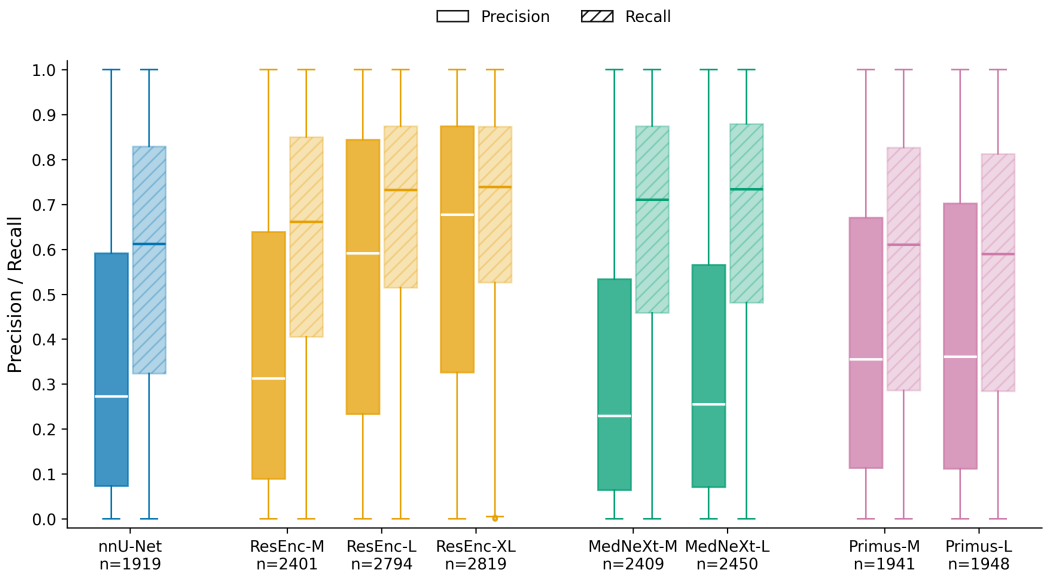


Figure 8: **Precision and recall of generalist models on all cases in LesionBench where a prediction was not missed (DSC $>$ 0):** Once a model successfully detects a lesion, median recall is comparatively stable across architectures (0.59–0.74), while median precision varies far more widely (0.23–0.68, roughly three times the spread of recall). This indicates that the primary axis of disagreement between models is the rate of false-positive voxels, whether from over-segmenting around a correctly located lesion or from spurious predictions elsewhere in the volume rather than how much of the true lesion extent they manage to cover.

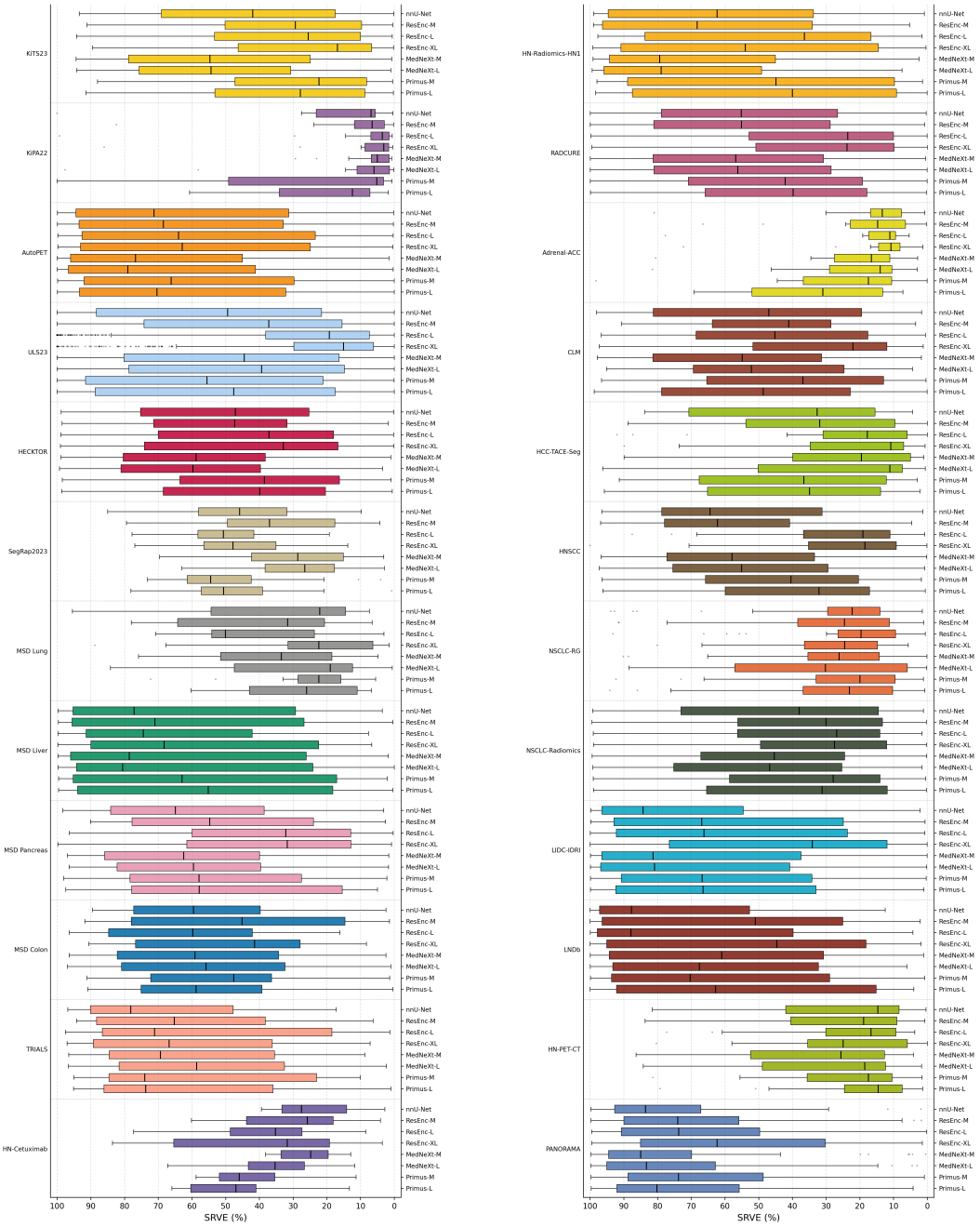


Figure 9: **Boxplots showing SRVE of generalist model performance on all cases in LesionBench, split by dataset:** SRVE is highly heterogeneous by dataset, and no single factor cleanly explains which datasets perform best. KiPA22 and Adrenal-ACC are the two strongest datasets, consistent with a tight-crop hypothesis. Both are small, single-organ regions of interest but this does not hold universally, since ULS23 performs only moderately despite being similarly curated. Within Head & Neck, HN-PET-CT is a clear outlier in the right direction (third-best overall), while the remaining six Head & Neck datasets are all considerably worse.

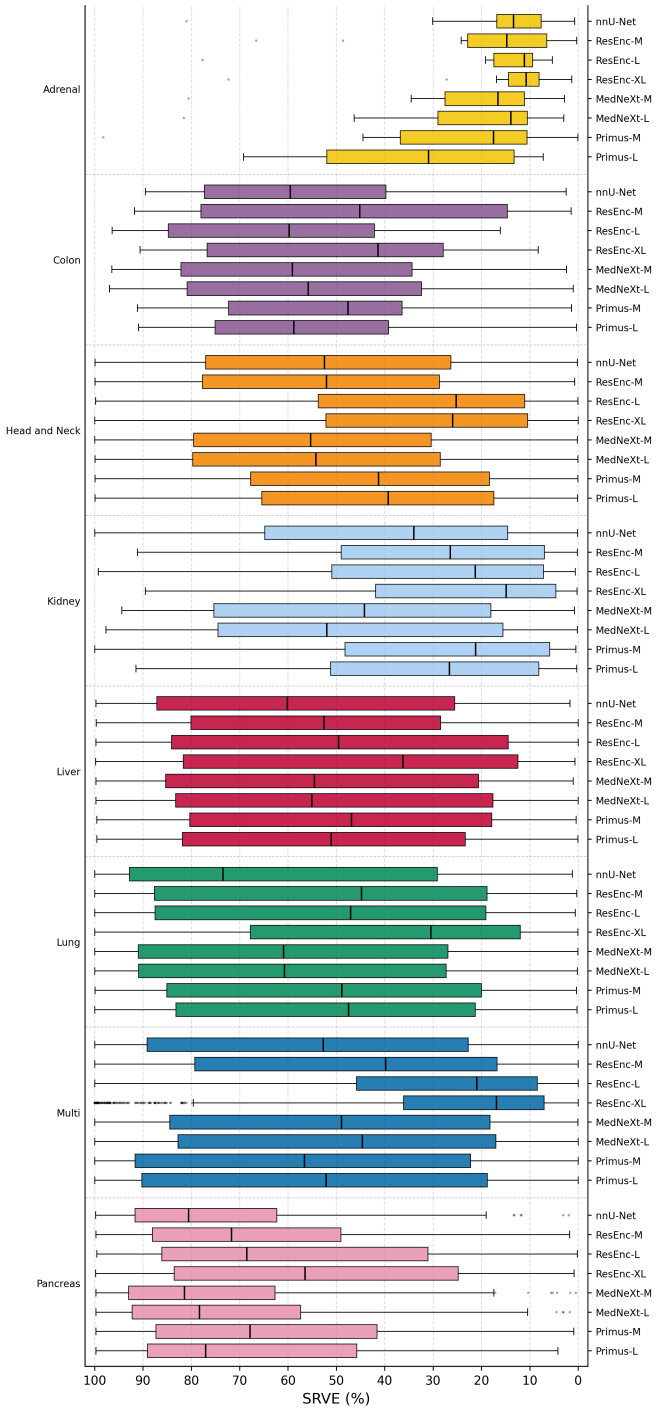


Figure 10: **Boxplots showing SRVE of generalist model performance on all cases in Lesion-Bench, split by organ:** SRVE remains heterogeneous when aggregated by organ. Adrenal and Kidney lesions are the two best-performing. Pancreas is the weakest organ by a clear margin, followed by Colon and Lung, indicating that volumetric agreement is substantially harder to achieve for diffuse, variably sized lesions than for the more spatially constrained organs at the top of the ranking.

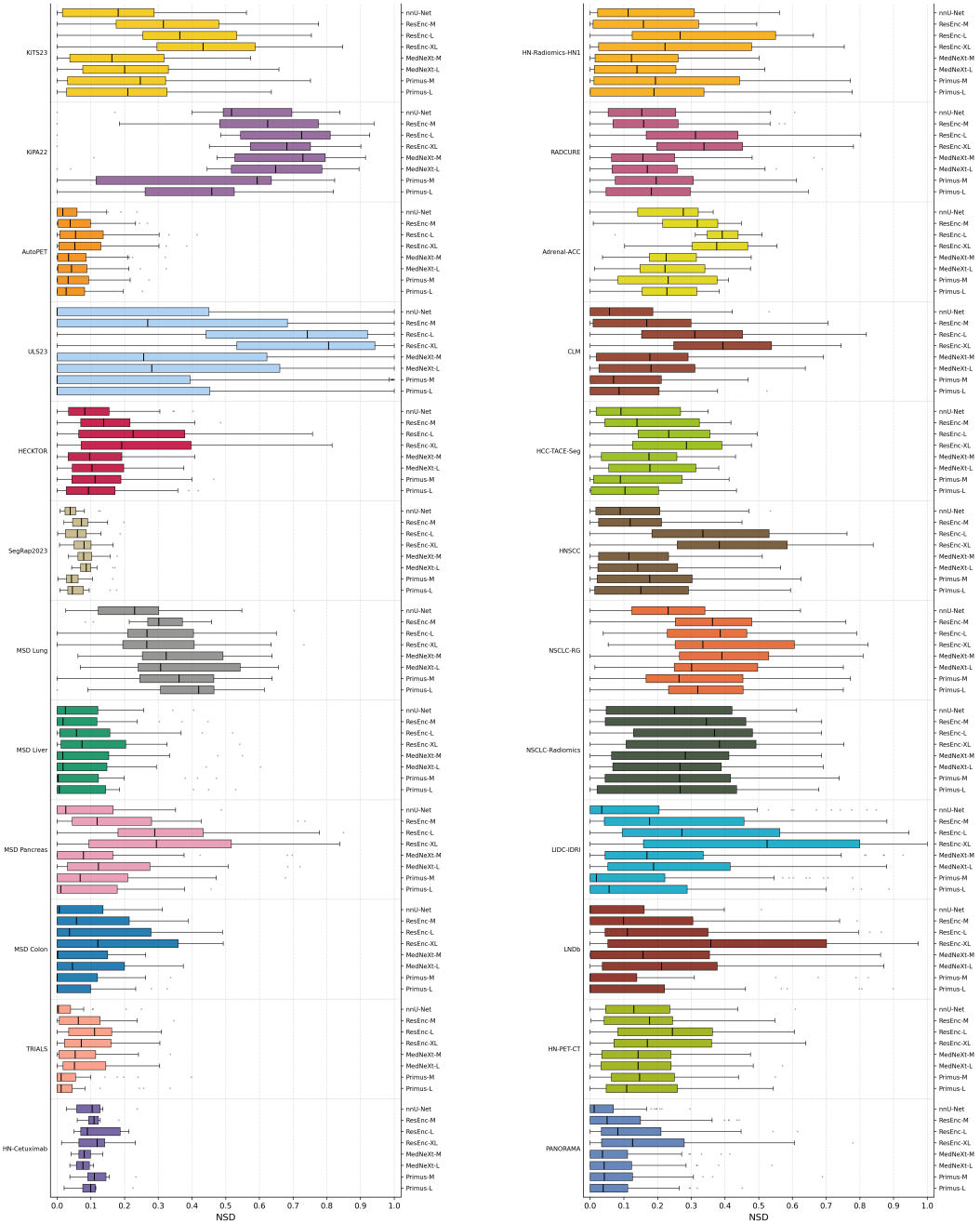


Figure 11: **Boxplots showing NSD of generalist model performance on all cases in LesionBench, split by dataset:** NSD is highly heterogeneous both across and within datasets. Some datasets show large disagreement between models. ULS23 has the widest spread in mean NSD across the eight models, followed by CLiMet, LIDC-IDRI, and LNDb. Others are uniformly difficult regardless of model choice, with AutoPET, SegRap2023, and HN-Cetuximab all scoring consistently low NSD.

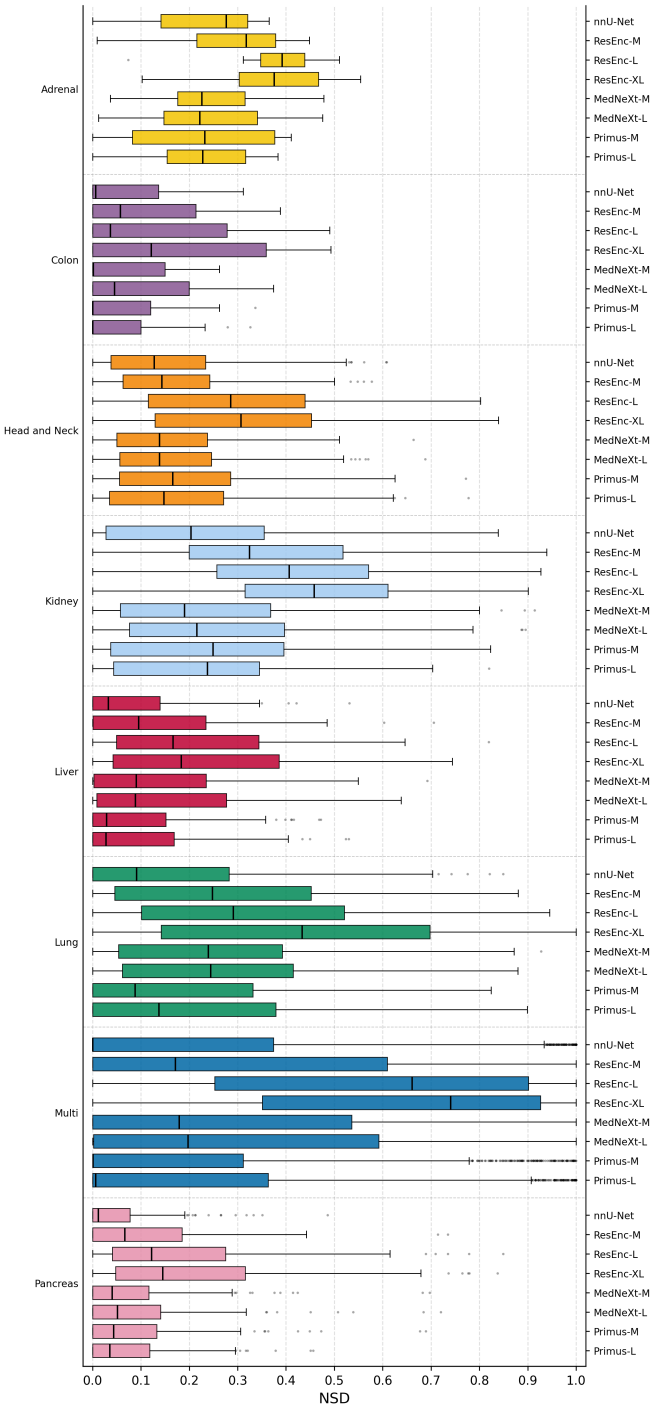


Figure 12: **Boxplots showing NSD of generalist model performance on all cases in Lesion-Bench, split by organ:** NSD remains highly heterogeneous when aggregated by organ, mirroring the dataset-level heterogeneity above. Kidney and Multi-Location yield the highest NSD overall, though the latter is driven almost entirely by ULS23, which accounts for a large proportion of Multi-Location cases. Colon, Pancreas, and Liver are the most difficult locations, with Colon showing the lowest mean NSD of any organ.

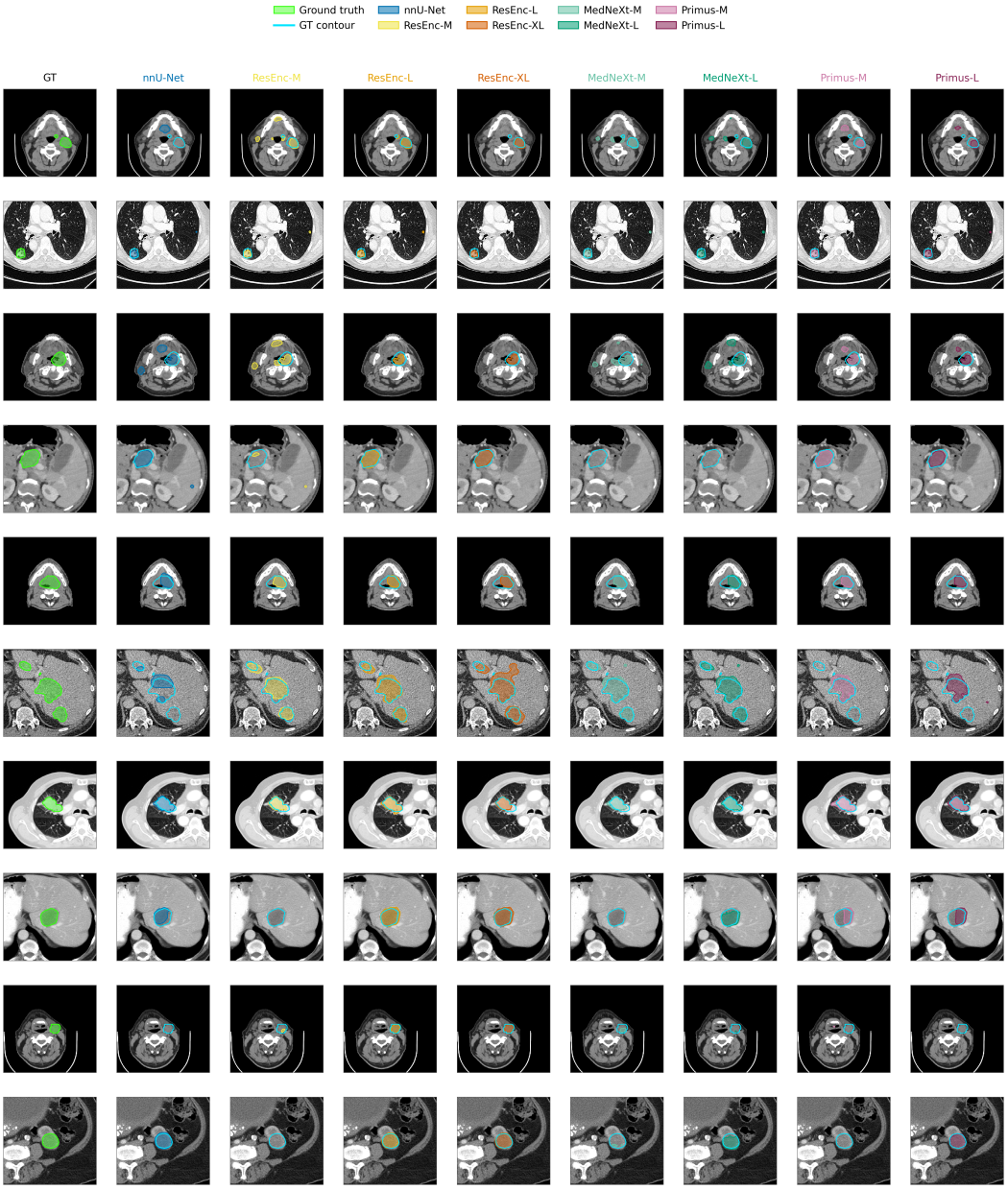


Figure 13: **Qualitative examples of high-DSC cases on LesionBench:** Axial CT slices for ten cases with high ResEnc-XL Dice, with the ground-truth lesion overlaid in green (leftmost column) and each model's prediction overlaid in its corresponding colour, both shown with the ground-truth contour for reference. These examples illustrate cases where model predictions closely match the annotated lesion extent across most or all architectures.

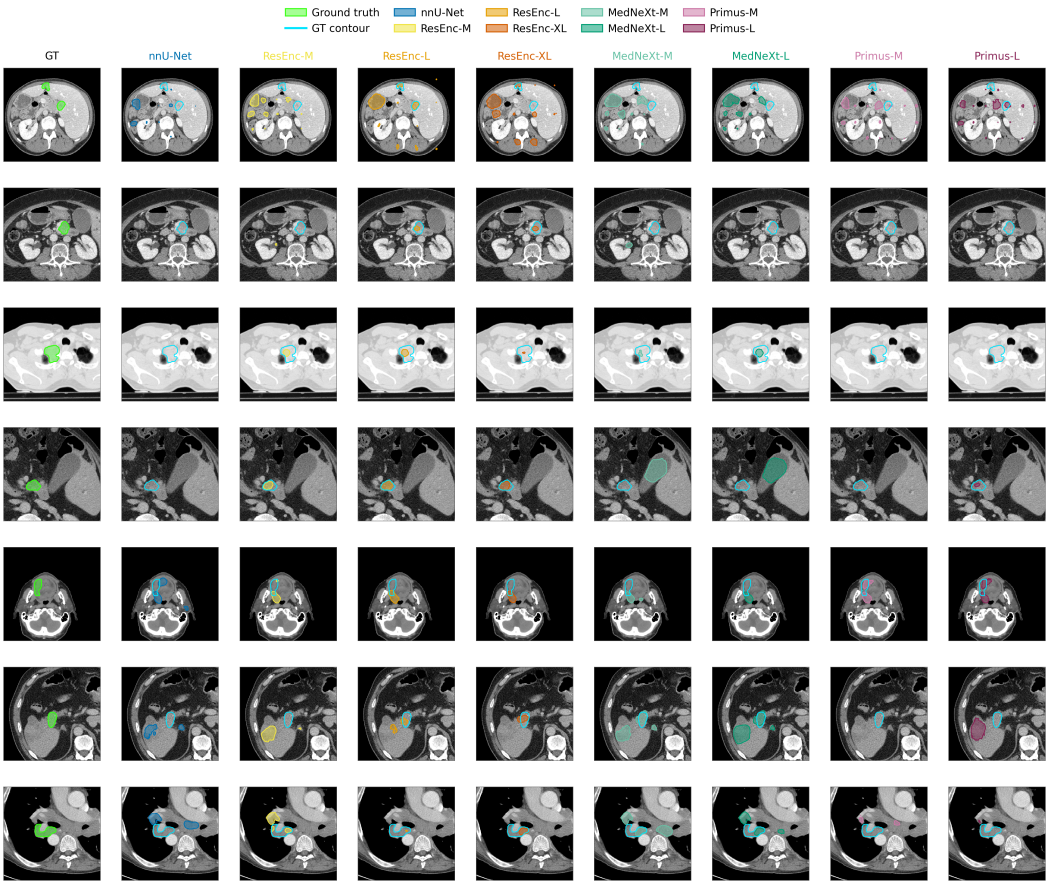


Figure 14: **Qualitative examples of low-DSC cases on LesionBench:** Axial CT slices for seven cases with low ResEnc-XL DSC, with the ground-truth lesion overlaid in green (leftmost column) and each model's prediction overlaid in its corresponding colour, both shown with the ground-truth contour for reference. These examples illustrate typical failure modes across models, including missed lesions, boundary disagreement, and false-positive predictions elsewhere in the slice.