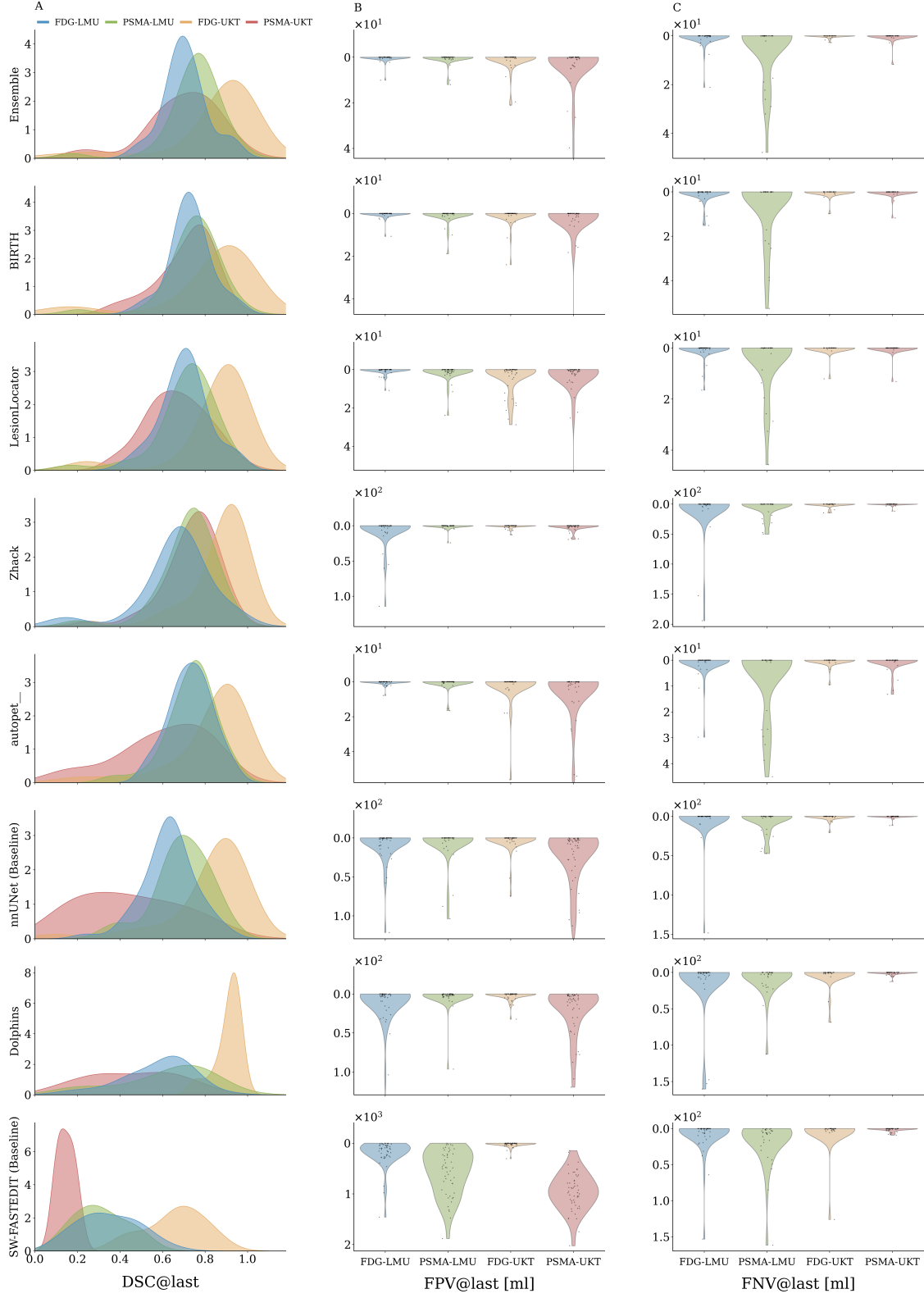


Supplementary Materials



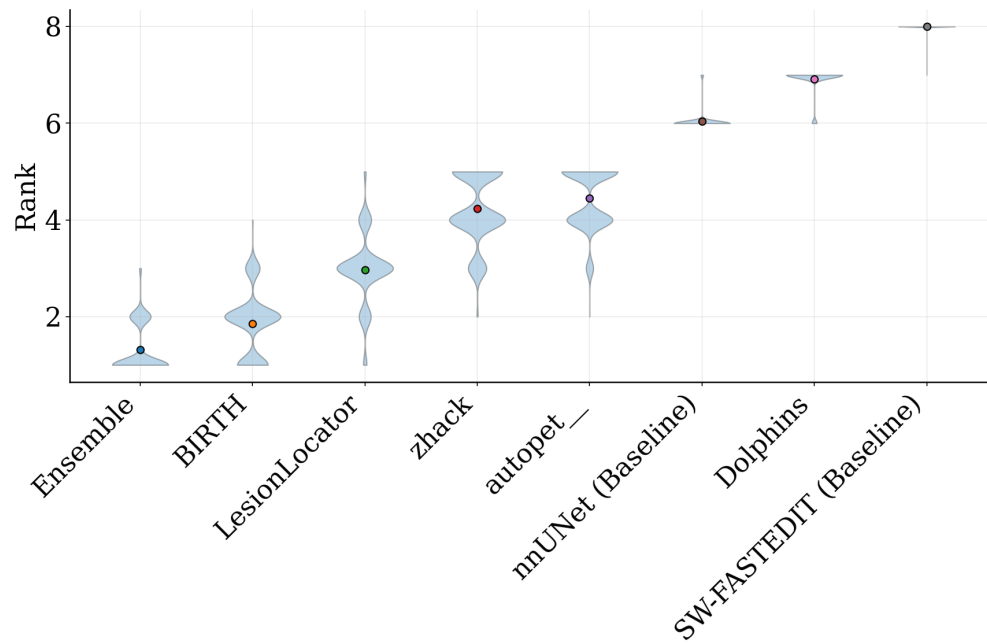
Supplementary Figure S1: Task 1. Stratified evaluation of challenge algorithms across tracers (FDG, PSMA) and centers (UKT, LMU) for (A) DSC@last, (B) FNV@last, and (C) FPV@last.

#	Team	Mean rank	DSC@last (rank)					FNV@last (rank)					FPV@last (rank)				
			FDG-LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean DSC	FDG-LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean FNV	FDG-LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean FPV
1	BIRTH	1.25	0.7010 (3)	0.7437 (1)	0.8242 (1)	0.6959 (2)	0.7412±0.0514	1.1150 (2)	6.6527 (3)	0.4740 (2)	0.4245 (3)	2.1665±2.6044	0.3896 (1)	0.8895 (2)	0.9871 (2)	2.8930 (2)	1.2898±0.9530
2	LesionLocator	2.25	0.7040 (2)	0.7178 (4)	0.8029 (4)	0.6697 (3)	0.7236±0.0490	0.8397 (1)	5.3520 (1)	0.4407 (1)	0.3720 (1)	1.7511±2.0867	0.6165 (3)	1.4394 (4)	3.8858 (6)	3.9582 (3)	2.4750±1.4762
3	Zhack	3.25	0.6490 (4)	0.7298 (3)	0.8089 (2)	0.7356 (1)	0.7308±0.0566	8.4465 (5)	6.6464 (2)	1.3300 (5)	0.4630 (4)	4.2215±3.3992	6.7693 (4)	0.9080 (3)	0.7990 (1)	2.1481 (1)	2.6561±2.4332
4	autopet_	3.375	0.7320 (1)	0.7349 (2)	0.8076 (3)	0.5895 (4)	0.7160±0.0791	1.1491 (3)	6.6830 (4)	0.5029 (3)	0.9349 (5)	2.3175±2.5312	0.4183 (2)	0.8549 (1)	2.2152 (3)	6.4935 (4)	2.4955±2.4015
5	naUNet [†]	5.25	0.6306 (5)	0.6925 (5)	0.7998 (5)	0.4168 (5)	0.6349±0.1397	3.9837 (4)	6.6863 (5)	1.1791 (4)	0.4174 (2)	3.0666±2.4761	8.7541 (5)	6.9473 (6)	3.8308 (5)	23.8253 (6)	10.8394±7.7015
6	Dolphin	5.75	0.5455 (6)	0.5394 (6)	0.7327 (6)	0.3897 (6)	0.5518±0.1216	11.2429 (7)	9.9620 (6)	4.0891 (6)	0.9378 (6)	6.5580±4.2195	12.0569 (6)	3.6937 (5)	2.2252 (4)	20.8322 (5)	9.7020±7.4403
7	SW-FASTEDIT [†]	6.875	0.2507 (7)	0.1184 (7)	0.5324 (7)	0.0228 (7)	0.2311±0.1919	8.7727 (6)	19.0827 (7)	5.1807 (7)	1.4831 (7)	8.6298±6.562	190.0509 (7)	576.1275 (7)	23.4361 (7)	959.5820 (7)	437.2991±362.1004

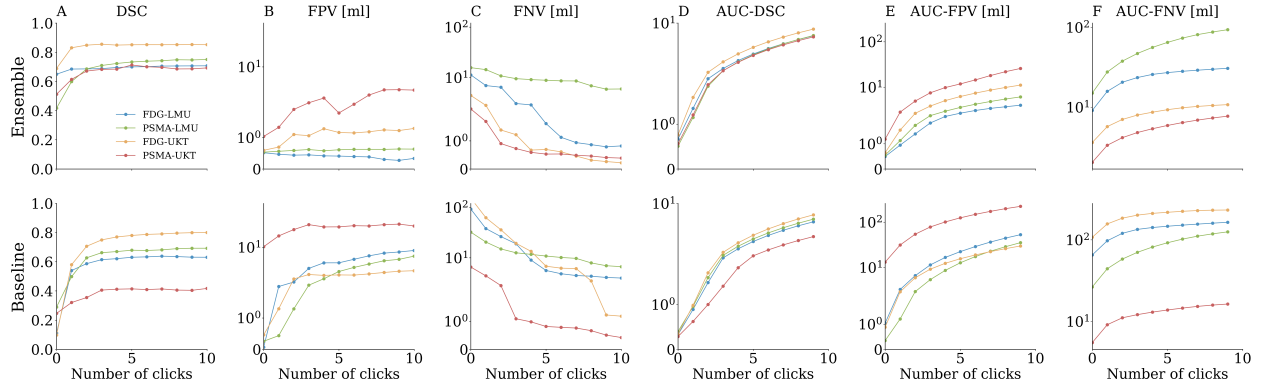
Supplementary Table S1: Task 1. Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV) for all submitted algorithms across centers and tracers. [†] Out of competition.

#	Team	AUC-DSC (rank)					AUC-FNV (rank)					AUC-FPV (rank)				
		FDG-LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean AUC-DSC	FDG_LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean AUC-FNV	FDG-LMU	PSMA-LMU	FDG-UKT	PSMA-UKT	Mean AUC-FPV
1	BIRTH	6-9196 (3)	7-0838 (1)	8-1705 (1)	6-7937 (2)	7.2419±0.5459	23-3049 (2)	96-3897 (2)	15-3223 (2)	6-6894 (2)	35-4266±35.6842	4-3815 (2)	8-7053 (3)	10-5639 (2)	31-3134 (2)	13-7411±10.3904
2	LesionLocator	6-9646 (2)	6-8726 (2)	7-9455 (3)	6-4661 (3)	7.0622±0.5434	14-5625 (1)	79-0628 (1)	5-5547 (1)	3-4653 (1)	25-6613±31.1121	7-6836 (3)	15-5263 (4)	39-5529 (6)	44-1859 (3)	26-7372±15.4711
3	Zhack	6-3488 (4)	6-7943 (3)	7-9480 (2)	6-8603 (1)	6.9878±0.5882	115-5239 (6)	98-9267 (3)	31-6819 (3)	10-0667 (4)	64-0498±44.2375	69-4458 (5)	7-9248 (2)	6-0634 (1)	16-1855 (1)	24-9049±25.9964
4	autopet_...	6-9967 (1)	6-7771 (4)	7-7660 (4)	5-3709 (4)	6.7277±0.8651	39-5739 (3)	116-8194 (5)	66-4700 (6)	21-7453 (6)	61-1521±35.8666	3-8145 (1)	6-0607 (1)	16-2253 (3)	64-0434 (4)	22-5360±24.4163
5	naUNet [†]	5-9005 (5)	6-3661 (5)	7-2066 (6)	3-8690 (6)	5.8355±1.2281	163-9669 (7)	125-1135 (6)	231-9554 (7)	16-2937 (5)	134-3324±78.1480	52-8512 (4)	35-1703 (5)	29-7207 (5)	223-2989 (6)	85-2603±80.1540
6	Dolphin	5-4550 (6)	5-3937 (6)	7-3272 (5)	3-8974 (5)	5.5183±1.2164	112-4295 (5)	99-6204 (4)	40-8908 (4)	9-3782 (3)	65-5797±42.1954	120-5687 (6)	36-9369 (6)	22-2520 (4)	208-3224 (5)	97-0200±74.4026
7	SW-FASTEDIT [†]	2-4818 (7)	1-1707 (7)	5-2685 (7)	0-2192 (7)	2.2851±1.9006	89-8120 (4)	196-6383 (7)	56-6229 (5)	22-8706 (7)	91-4860±65.1600	1913-3372 (7)	5949-7750 (7)	230-1135 (7)	10077-7511 (7)	4542-744±3812.1058

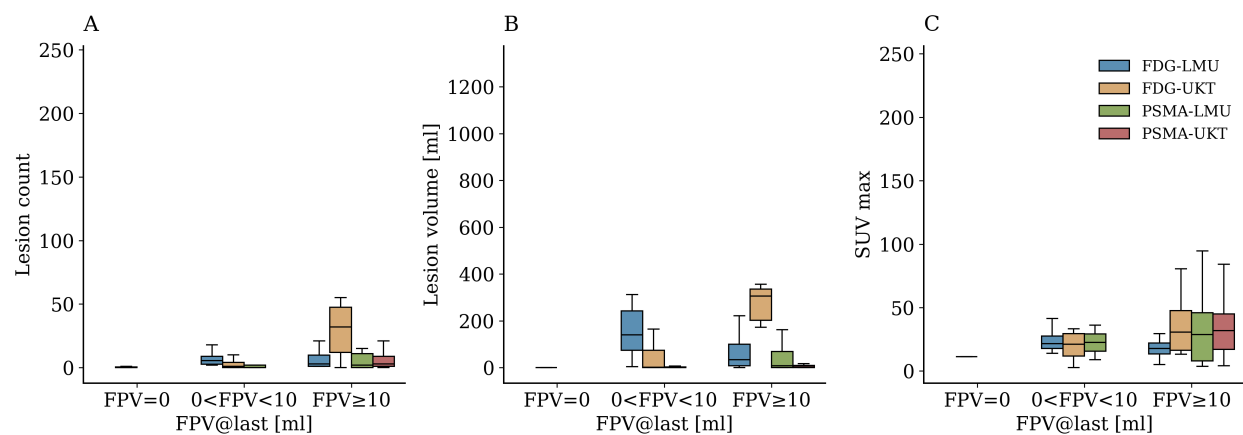
Supplementary Table S2: Task 1. Area-under-the-curve (AUC)-based metrics for Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV): AUC-DSC, AUC-FNV, AUC-FPV; evaluated in all submitted algorithms across centers and tracers. [†] Out of competition.



Supplementary Figure S2: Task 1. Ranking stability of participating teams based on 1,000 bootstrap iterations. Violin plots show the distribution of ranks per team. The mean rank (dot) corresponds with the final challenge ranking, the robustness of the results to sampling from the test dataset.



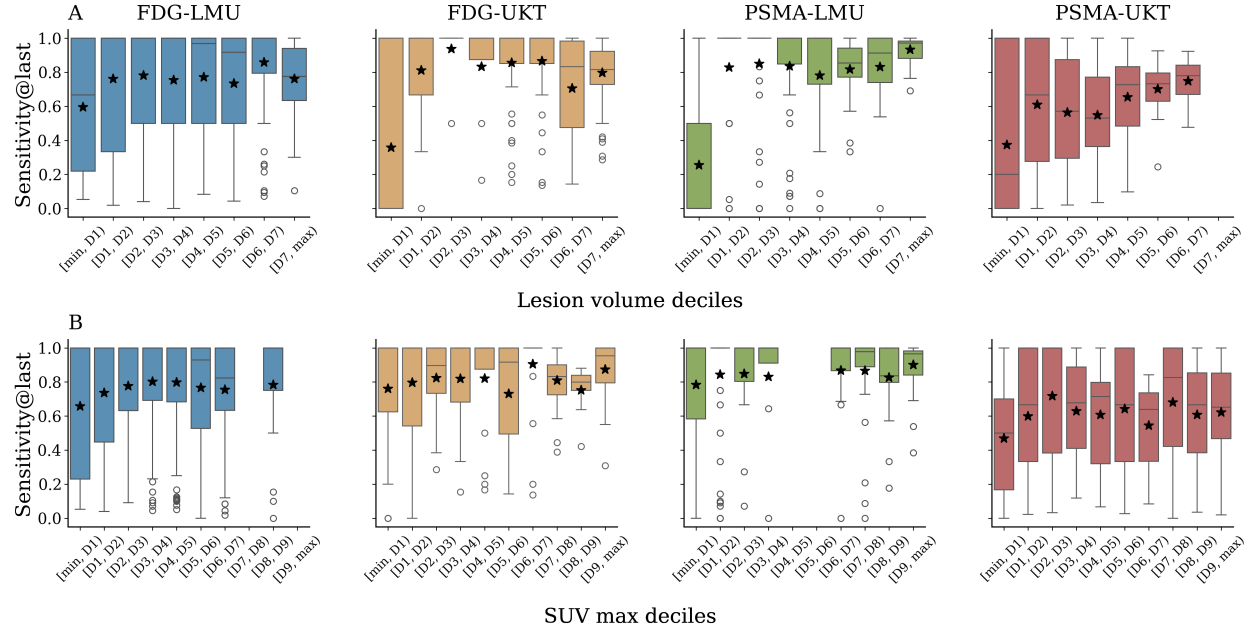
Supplementary Figure S3: Task 1. (A) DSC, (B) FPV, (C) FNV, (D) AUC-DSC, AUC-FPV (E), (F) AUC-FNV as a function of the number of clicks (NoC) across tracers and centers for the ensemble and baseline models, comparing initial performance with best achievable results.



Supplementary Figure S4: Task 1. FPV behaviour across centers and tracers. Case-level distributions stratified by FPV at the final interaction (FPV@last), where FPV was averaged across all evaluated algorithms per case. (A) Lesion count, (B) lesion volume, and (C) uptake values (SUV max) are shown across centers and tracers. Increasing FPV is associated with greater lesion burden and higher uptake values.

Model	Lesion count	Lesion volume [ml]	SUV max
Ensemble	0.61 (p<0.0001)	0.45 (p<0.0001)	0.44 (p<0.0001)
BIRTH	0.60 (p<0.0001)	0.46 (p<0.0001)	0.37 (p<0.0001)
LesionLocator	0.18 (p=0.19)	0.05 (p=1)	0.28 (p=0.006)
Zhack	0.26 (p=0.002)	0.20 (p=0.076)	0.28 (p=0.008)
Baseline	0.60 (p<0.0001)	0.41 (p<0.0001)	0.29 (p=0.004)

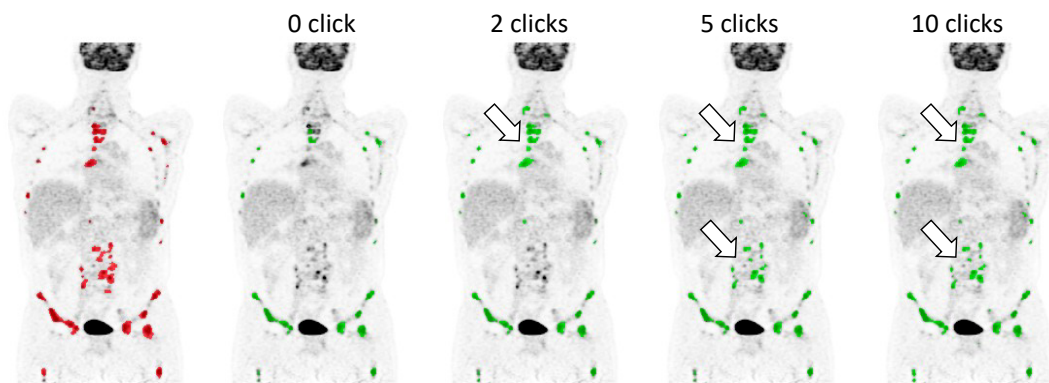
Supplementary Table S3: Task 1. Spearman rank correlations (ρ) between FPV@last and lesion characteristics for the ensemble, baseline model, and the three top-performing algorithms. Correlations were computed separately per model across the 200 test cases using two-sided tests. P-values (p) were Bonferroni corrected for multiple comparisons (15 tests).



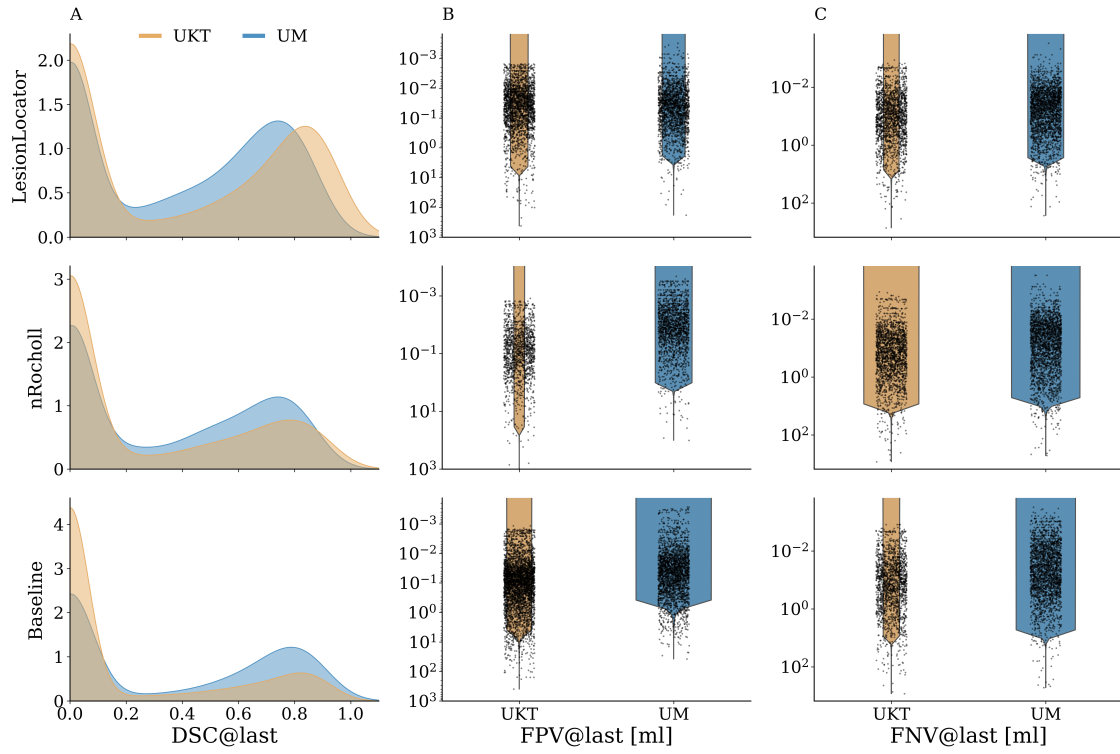
Supplementary Figure S5: Task 1. Distribution of lesion sensitivity by lesion volume and tracer uptake (SUV max). Lesions were stratified by volume (A) or maximum SUV (B) into interdecile ranges, and sensitivity was calculated at the last click (sensitivity@last) across all final leaderboard algorithms, including the ensemble model. Detection sensitivity is lower for smaller lesions and for lesions with very low tracer uptake compared with larger lesions and those with higher uptake.

Team	DSC@1 $\uparrow$	DSC@5 $\uparrow$	DSC@10 $\uparrow$	FPV@1 $\downarrow$	FPV@5 $\downarrow$	FPV@10 $\downarrow$	FNV@1 $\downarrow$	FNV@5 $\downarrow$	FNV@10 $\downarrow$	NoC@85 $\downarrow$	NoC@90 $\downarrow$
Ensemble	0.677 $\pm$ 0.205	0.740 $\pm$ 0.154	0.741 $\pm$ 0.164	0.765 $\pm$ 2.780	0.994 $\pm$ 3.145	1.536 $\pm$ 6.149	6.277 $\pm$ 19.310	2.671 $\pm$ 9.381	1.668 $\pm$ 6.338	8.930 $\pm$ 3.857	9.542 $\pm$ 3.506
BIRTH	0.681 $\pm$ 0.213	0.730 $\pm$ 0.184	0.733 $\pm$ 0.182	1.30 $\pm$ 6.618	1.262 $\pm$ 5.691	1.290 $\pm$ 5.545	5.163 $\pm$ 13.822	3.019 $\pm$ 9.816	1.973 $\pm$ 7.199	8.980 $\pm$ 3.864	9.535 $\pm$ 3.429
LesionLocator	0.679 $\pm$ 0.198	0.706 $\pm$ 0.178	0.717 $\pm$ 0.171	2.891 $\pm$ 7.435	2.648 $\pm$ 7.281	2.475 $\pm$ 7.063	2.979 $\pm$ 10.519	2.366 $\pm$ 8.576	1.589 $\pm$ 6.020	9.155 $\pm$ 3.797	9.750 $\pm$ 3.221
Zhack	0.630 $\pm$ 0.237	0.710 $\pm$ 0.197	0.721 $\pm$ 0.191	2.453 $\pm$ 11.432	2.593 $\pm$ 10.377	2.656 $\pm$ 10.687	9.683 $\pm$ 30.786	5.846 $\pm$ 22.399	4.502 $\pm$ 20.917	9.185 $\pm$ 3.548	9.875 $\pm$ 2.946
autopet_	0.584 $\pm$ 0.238	0.693 $\pm$ 0.194	0.709 $\pm$ 0.187	1.505 $\pm$ 7.491	2.085 $\pm$ 6.865	2.495 $\pm$ 9.394	13.774 $\pm$ 45.818	3.871 $\pm$ 10.914	2.134 $\pm$ 7.215	9.390 $\pm$ 3.298	9.990 $\pm$ 2.766
nnU-Net (Baseline) [†]	0.480 $\pm$ 0.259	0.612 $\pm$ 0.229	0.620 $\pm$ 0.230	4.902 $\pm$ 15.804	8.802 $\pm$ 23.114	10.839 $\pm$ 25.567	30.111 $\pm$ 89.716	5.673 $\pm$ 18.186	3.038 $\pm$ 13.814	9.900 $\pm$ 2.782	10.285 $\pm$ 2.255
Dolphins	0.539 $\pm$ 0.292	0.539 $\pm$ 0.292	0.539 $\pm$ 0.292	9.702 $\pm$ 23.394	9.702 $\pm$ 23.394	9.702 $\pm$ 23.394	6.776 $\pm$ 23.900	6.776 $\pm$ 23.900	6.776 $\pm$ 23.900	9.800 $\pm$ 3.250	9.950 $\pm$ 3.066
SW-FASTEDIT (Baseline) [†]	0.211 $\pm$ 0.253	0.215 $\pm$ 0.254	0.217 $\pm$ 0.256	473.425 $\pm$ 525.913	449.446 $\pm$ 502.787	437.299 $\pm$ 483.338	9.511 $\pm$ 24.109	8.600 $\pm$ 23.592	8.278 $\pm$ 23.079	10.915 $\pm$ 0.859	11.000 $\pm$ 0.000

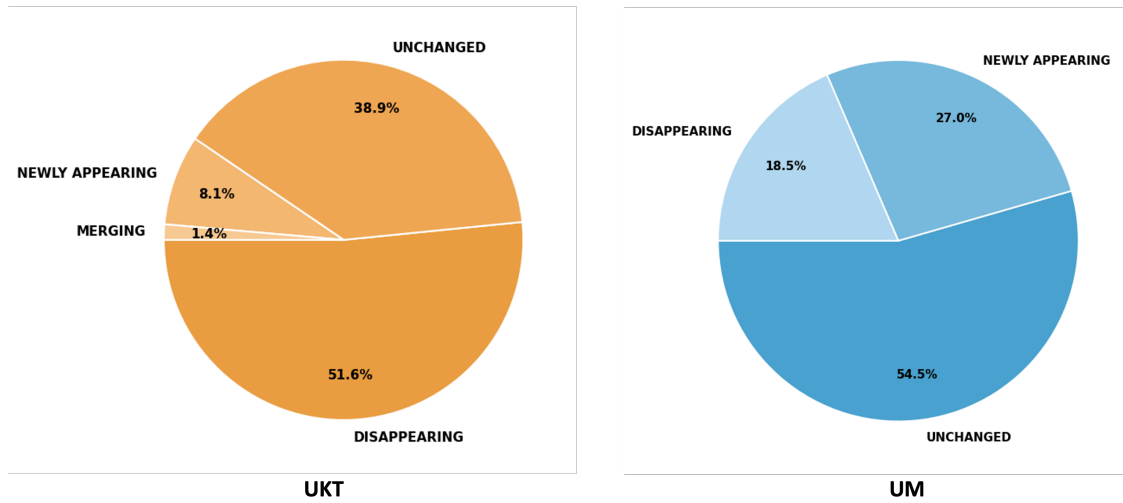
Supplementary Table S4: Task 1. Segmentation performance (mean $\pm$ standard deviation) at 1, 5, and 10 clicks for Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV) (DSC@1/5/10, FNV@1/5/10, FPV@1/5/10), and number of clicks (NoC) required to reach DSC thresholds of 0.85 and 0.9 (NoC@85, NoC@90). $\uparrow$ indicates higher is better, $\downarrow$ lower is better, [†] Out of competition.



Supplementary Figure S6: Task 1. Example PET image with manual segmentation (red) and corresponding nnU-Net baseline predicted masks (green) at different interaction levels. The arrows show the refinement of the prediction with increasing user clicks.



Supplementary Figure S7: Task 2. Center-specific (UKT, UM) evaluation of lesion-level performance for the submitted algorithms. Despite domain differences and unequal lesion burden between centers, the leading method maintains stable performance, whereas the others show performance degradation on the high lesion-burden UKT cohort.



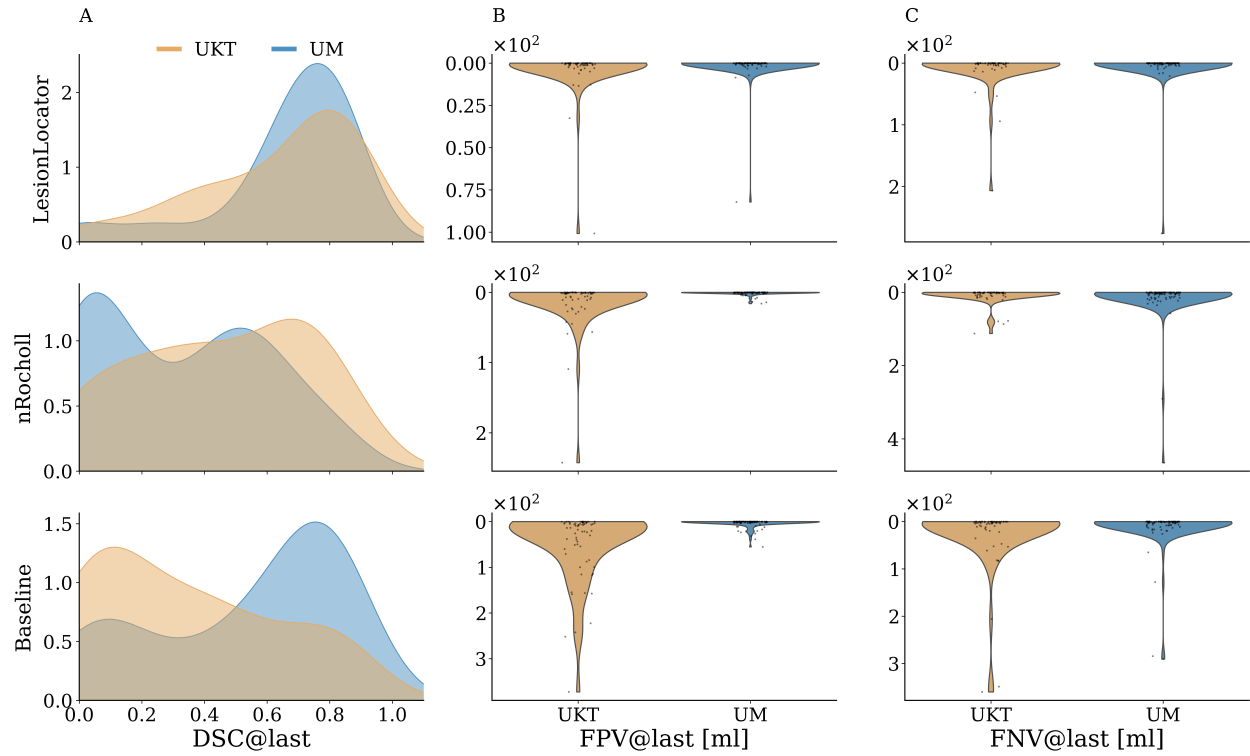
Supplementary Figure S8: Task 2. Distribution of lesion topology from baseline to follow-up in test cases across centers. The topology class describes temporal evolution: DISAPPEARING (lesions present at baseline but absent at follow-up), UNCHANGED (stable), NEWLY APPEARING (lesions absent at baseline but present at follow-up), and MERGING (multiple lesions at baseline merging into one at follow-up).

#	Team	DSC@last			FNV@last			FPV@last		
		UKT	UM	Mean DSC	UKT	UM	Mean FNV	UKT	UM	Mean FPV
1	LesionLocator	0.3812	0.3833	0.3823±0.001	0.8599	0.3656	0.6128±0.247	1.3529	0.7805	1.0667±0.2862
2	nRocholl [†]	0.2547	0.3387	0.2967±0.042	2.7338	0.2296	1.4817±1.252	2.9994	2.0950	2.5472±0.452
2	Baseline [†]	0.1835	0.3437	0.2636±0.008	1.4269	0.2977	0.8623±0.564	1.5910	1.6026	1.5968±0.006

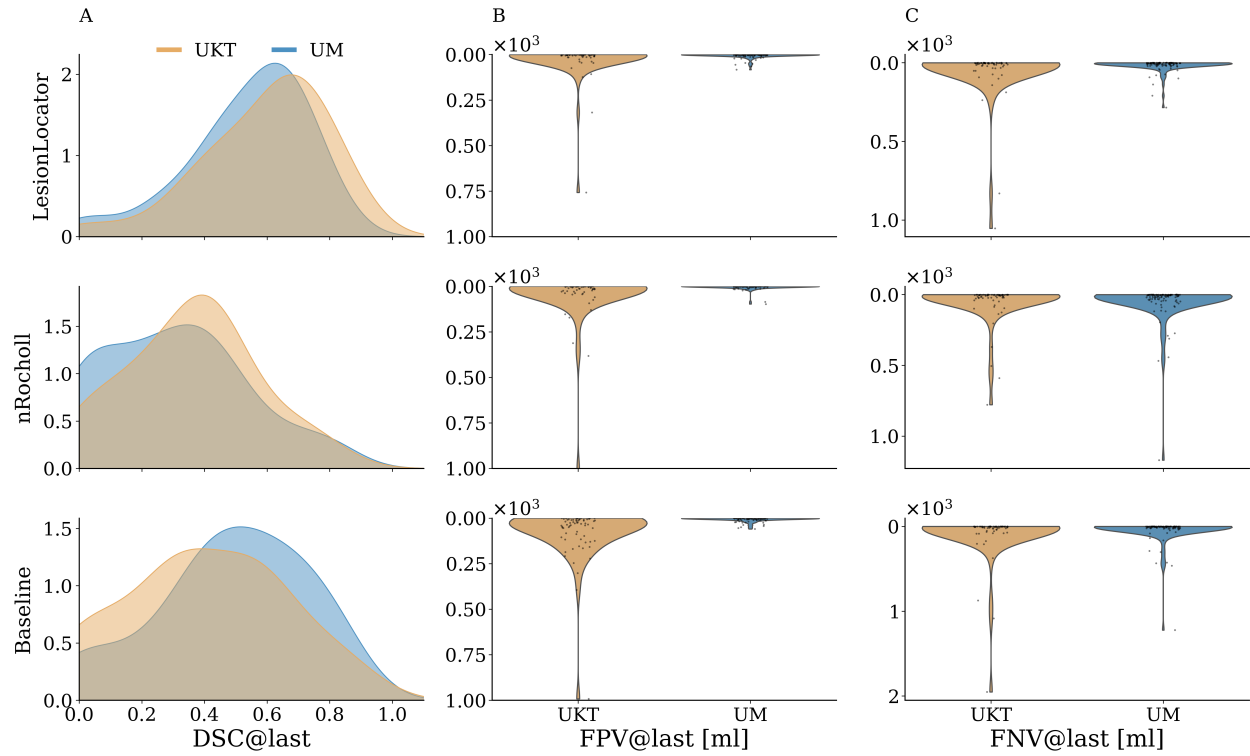
Supplementary Table S5: Task 2. Performance across institutions (UKT, UM). Mean values of Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV) are evaluated in the follow-up image, reported as DSC@last, FNV@last, and FPV@last for each team. [†] Out of competition.

Further analysis

To complement the lesion level challenge metrics, we additionally evaluated performance at the patient level using both connected-component-based (appendix p 16) and label-based (no connected-component) (appendix p 17) formulations (appendix p 18) while also reporting summed volume metrics (FPV@last, FNV@last), as mean values alone may not capture the total volumetric load of errors across cases (appendix p 19). The ranking of methods remained consistent with the lesion-based evaluation across both patient-level definitions, although absolute scores differed, with the concentration of near-zero DSC values substantially reduced, as patient-level aggregation mitigates the effect of isolated missed lesions that would otherwise yield a DSC of zero in lesion-wise evaluation. For the leading solution, using the connected-component-based definition, mean DSC@last = 0.65, with mean FNV@last = 5.50 ml and FPV@last = 2.07 ml; using label-based definition, mean DSC@last = 0.56, with mean FNV@last = 29.37 ml and FPV@last = 14.26 ml (appendix p 19).



Supplementary Figure S9: Task 2. Patient-level (cc-based) segmentation performance across the test centers. Although relative ranking of algorithms is conserved compared with lesion-based scoring (Figure S7), absolute performance differs, reflecting connected-component-based aggregation of predictions and reference regions and reduced sensitivity to isolated lesion-level errors.



Supplementary Figure S10: Task 2. Patient-level (label-based) segmentation performance across the test centers. Compared with the connected-component-based formulation (appendix p 16), rankings are preserved and absolute performance slightly differs, reflecting direct voxel-wise aggregation of multi-label predictions and increased sensitivity to total disease burden.

Patient-level formulation

Let G and P denote the ground-truth and predicted segmentation volumes for a given patient.

Connected-component-based evaluation (CC-based)

Binary masks are first derived as $G_b = 1(G > 0)$ and $P_b = 1(P > 0)$. Connected components are then identified in G_b and P_b using a predefined connectivity (connectivity=18). False Positive Volume is defined as the total volume of predicted connected components in P_b that do not overlap with any ground-truth component in G_b . False Negative Volume is defined analogously as the total volume of ground-truth connected components in G_b that do not overlap with any predicted component in P_b . Dice similarity coefficient is computed voxel-wise on the binary masks G_b and P_b . This formulation is similar to Task 1 evaluation.

Label-based patient-level evaluation (no connected components)

Evaluation is performed directly on the multi-label segmentation without connected-component decomposition. For each ground-truth label l , a corresponding prediction mask is constructed by aggregating all predicted labels overlapping with that ground-truth region. DSC is computed per label and averaged across labels for each patient. False Positive Volume is defined as the total volume of voxels where $P \neq 0$ and $G = 0$, and False Negative Volume as the total volume of voxels where $G \neq 0$ and $P = 0$, without penalizing label mismatches between overlapping regions.

Team	Lesion-level					Patient-level (cc-based)					Patient-level (label-based)				
	DSC@last [†]	FNV@last [ml] _↓		FPV@last [ml] _↓		DSC@last [†]	FNV@last [ml] _↓		FPV@last [ml] _↓		DSC@last [†]	FNV@last [ml] _↓		FPV@last [ml] _↓	
	Mean	Mean	Sum	Mean	Sum	Mean	Mean	Sum	Mean	Sum	Mean	Mean	Sum	Mean	Sum
LesionLocator	0.382±0.365	1.088±12.869	5980.736	0.632±7.960	3470.202	0.652±0.236	5.497±27.803	923.418	2.069±10.312	351.692	0.559±0.198	29.371±108.959	4934.324	14.258±64.440	2423.791
nRocholl [†]	0.297±0.340	2.547±25.095	12552.606	1.482±50.096	7301.774	0.373±0.281	11.615±44.131	1939.629	5.818±22.036	983.287	0.326±0.220	51.502±138.575	8600.757	19.822±87.894	3349.926
Baseline [†]	0.248±0.348	1.596±20.611	11139.478	0.974±7.984	6796.253	0.462±0.3105	16.052±53.233	2696.682	20.811±51.258	3537.891	0.455±0.246	60.566±215.343	10175.092	34.305±93.637	5831.867

Supplementary Table S6: Task 2. Summary of lesion-based and patient-level segmentation performance for Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV) in follow-up image across all test cases. [†]Out of competition, cc denotes connected-component.

Challenge organization and infrastructure

The autoPET/CT-IV challenge concept and protocol were defined a priori and submitted to MICCAI in January 2025, where they underwent peer review before approval. The organizing committee consisted of clinicians and researchers with complementary expertise from University Hospital Tübingen (UKT), Ludwig-Maximilians-University in Munich (LMU), University Hospital Mainz (UM), Fraunhofer MEVIS, Karlsruhe Institute of Technology, and Radboud University Medical Center.

The final challenge description, including task definitions, evaluation methodology, training data, baseline resources, and timelines, was made publicly available on the challenge website on April 1st, 2025. Algorithm submission proceeded in staged phases: a preliminary submission phase allowed participants to validate container execution and output format on a set of five representative preliminary test samples; a final submission phase was subsequently opened with activation of the full hidden test dataset of 200 cases for Task 1 and 257 cases for Task 2. To mitigate technical failures, a limited number of one final submission per task was permitted. The challenge outcomes, including algorithm performance, ranking, and future directions, were disclosed only after the challenge at a dedicated satellite event at MICCAI 2025 on 28 September 2025.

All submissions were initially evaluated on the Grand Challenge (GC) platform (Diagnostic Image Analysis Group, Radboud University Medical Center, The Netherlands) as a type-II challenge, in which participants submitted executable algorithms that were run by the organizers on private test data. A final evaluation for both tasks was subsequently performed offline on internal servers at the organizing institutions. For Task 1, an extended evaluation was required to emulate interactive segmentation settings; accordingly, ten predictions were generated per test case, and this expanded evaluation was conducted on servers at UKT. For Task 2, test cases originated from two clinical centers (UKT, UM), and due to institutional data protection requirements, test data from the UM cohort could not be accessed by organizers outside the originating institution; therefore, evaluation for these cases was performed locally at the respective site. A maximum time limit of 15 minutes per case was set for algorithm execution on the available GC computation resources (NVIDIA T4 GPU 16 GB VRAM with 8 CPUs and a maximum memory of 30 GB).

To facilitate participation and reproducibility, the organizers provided detailed documentation, reference implementations, and example submission pipelines via the challenge website (<https://autopet-iv.grand-challenge.org/>) and public code repositories (<https://github.com/lab-midas/autoPETCTIV>).

Participation policies

The autoPET/CT-IV challenge was open to academic and industrial research teams. Participants were allowed to incorporate any publicly accessible data sources and pretrained models, including foundation models, provided that their use was transparent and documented. Participation was conditional on compliance with predefined reproducibility and reporting requirements. Specifically, teams were required to submit their final method as a containerized executable, release the corresponding source code under a permissive open-source licence, and provide a publicly available technical manuscript or preprint describing model design, training data, and inference procedures in accordance with a standardized template.

To safeguard against conflicts of interest, members of the organizing institutions were excluded from competitive participation and award eligibility. Each task was evaluated independently, with separate leaderboards and rankings. Recognition was awarded to the three highest-ranking teams in each task, contingent on adherence to all participation requirements. Monetary prizes were granted as follows (in euros): 1,500, 1,000, and 500 for first, second, and third place in Task 1, and 3,000, 2,000, and 1,000 for first, second, and third place in Task 2, respectively. Eligible teams were invited to present their methods during the MICCAI satellite session and to contribute to the post-challenge summary manuscript paper.

Submitted algorithms

This section summarizes the submitted algorithms for both tasks in the final leaderboard order and provides a table of key methodological details.

Task 1: Single-staging whole-body PET/CT

Team BIRTH. The first team’s segmentation pipeline was built upon the robust nnU-Net Residual Encoder (ResEncL) framework¹. It integrates tracer classification, dual-headed segmentation with organ supervision, and post-processing based on PET SUV thresholds to address multi-tracer and multi-center variability. Human guidance is incorporated through simulated foreground and background clicks, with two strategies: full-guidance training using all clicks and stochastic click sampling to expose the model to sparse guidance scenarios. For PSMA cases, tracer-specific model leveraging pretraining and stochastic sampling was used to better capture small and sparsely distributed lesions, whereas FDG cases were handled using a unified multi-tracer model with organ supervision and anatomical priors in a hybrid strategy.

Team LesionLocator. The team’s solution is also built on the nnU-Net framework with the ResEncL architecture¹, with large-scale multimodal pretraining, organ supervision, and tailored data augmentation to improve cross-tracer and cross-center generalization (autoPET III winning solution). Beyond this foundation, they conducted a systematic evaluation of prompt representations, showing that Euclidean distance transform (EDT)-based encodings outperform unit-normalized Gaussian kernels for interactive segmentation, while Gaussian kernels with value normalized to 1 at the click location behave similarly to an EDT. To better approximate real user interactions, they introduced a probabilistic click-sampling strategy during training, encouraging sparse guidance with prompts placed arbitrarily within the target region. Model robustness is further supported by training on multiple external datasets (DEEP-PSMA dataset², TCGA-LUAD collection³, NSCLC Radiogenomics dataset⁴).

Team Zhack. The method relies on the winning autoPET III challenge algorithm (as described above) and proposes an alternative heatmap representation approach of user interactions. Instead of using two separate heatmaps, foreground and background clicks are encoded into a single signed heatmap using Gaussian smoothing and PET SUV-based masking, enabling the network to explicitly distinguish lesion-supporting and lesion-suppressing regions.

Team autopet_. The method was based on the ResEncL architecture and used an interaction encoding strategy that integrates foreground and background clicks into a Gaussian heatmap representation. Model outputs were subsequently post-processed by removing small connected components.

Team Dolphins. The approach did not incorporate interactive click guidance and instead proposed a two-stage PET/CT lesion segmentation framework that combines self-supervised representation learning with supervised fine-tuning. In the first stage, a masked autoencoder⁵ was trained on unlabeled PET/CT scans from Task 1 and longitudinal CT scans from Task 2 to learn robust modality-specific and anatomical representations without annotations. In the second stage, the pretrained encoder was fine-tuned for voxel-wise lesion segmentation using a BiXLSTM⁶-ResNet encoder with a convolutional decoder, jointly fusing PET and CT inputs. This approach highlights the potential of self-supervised learning for PET/CT lesion segmentation.

Baselines (out of competition). Two baseline methods were provided. The first was the widely used nnU-Net⁷ framework, in which simulated foreground and background clicks were encoded as Gaussian heatmaps and supplied as an additional input channel alongside the PET and CT images, enabling the integration of sparse user feedback into the segmentation process. The second baseline was the sliding-window Fastedit framework⁸ (SW-FASTEDIT), an interactive 3D approach trained exclusively on PET imaging using a MONAI⁹ DynUNet backbone. During training, PET volumes were partitioned into overlapping patches that were segmented at full resolution and subsequently reassembled. The framework identifies regions with the poorest segmentation performance, samples corrective foreground and background clicks within these error regions, and iteratively refines the segmentation outputs.

Ensemble model (out of competition). On the basis of predictions from all algorithms described above, including the baseline models, an ensemble output was generated using pixel-wise majority voting.

Task 2: Longitudinal whole-body CT

Team LesionLocator. The winning approach proposed a longitudinal, prompt-based lesion segmentation approach that directly exploits the availability of lesion prompts in both BL and FU CT scans. Building on the LongiSeg framework¹⁰, the method processes aligned BL-FU image patches centered on the provided lesion locations and integrates both point prompts and prior segmentation masks as additional input channels. The nnU-Net Residual Encoder¹ backbone enables joint modeling of longitudinal image context, avoiding explicit prompt propagation across timepoints. Additionally, the method utilizes large-scale synthetic longitudinal¹¹ pretraining to enhance the model’s

ability to leverage temporal information compared with single-timepoint or non-longitudinal pretraining strategies and potentially improves the segmentation quality over time.

Team nRocholl (out of competition). The second team implements a dual-encoder 3D nnU-Net for longitudinal lesion segmentation, with separate encoders for BL and FU CT images. The BL encoder receives both the BL image and its segmentation mask, while the FU encoder processes the FU image and the prior Gaussian volume constructed from the provided centroids; encoder weights are shared from a pretrained nnU-Net (USL23 baseline¹²). Features from the two streams are fused before decoding, and an auxiliary baseline-mask reconstruction task is incorporated during training to improve stability. A post-processing step was used to resample label predictions based on clickpoints.

Baseline (out of competition). The baseline is based on the original ULS23 baseline¹² model for 3D universal lesion segmentation in CT. Built upon the nnU-Net framework, the model was pretrained on a large, diverse dataset to perform universal lesion segmentation without task-specific adaptations, serving as a reference for assessing the added value of longitudinal training or prompt-based approaches in this challenge.

Team	Pre-processing	Augmentation	Framework	Post-processing	Architecture	Loss	Patch size
Task 1: Interactive PET/CT lesion segmentation							
BIRTH	nnU-Net preprocessing	nnU-Net augmentations	nnU-Net (3D)	PET SUV thresholding	ResEncL U-Net (3D)	DSC + CE	192x192x192
LesionLocator	Normalization, Resampling	nnUNet augmentations, misalignment augmentations	nnU-Net (3D)	-	ResEncL U-Net (3D)	DSC + CE	192x192x192
zhack	nnU-Net preprocessing	nnU-Net augmentations, misalignment augmentations	nnU-Net (3D)	PET SUV thresholding, small connected components removal	ResEncL U-Net (3D)	DSC + CE	192x192x192
autopet__	Normalization	Flipping, Random Rotation	nnU-Net (3D)	Small connected components removal	ResEncL U-Net (3D)	DSC + CE	Dx400x400
nnU-Net (Baseline)	nnU-Net preprocessing	nnU-Net augmentations	nnU-Net (3D)	-	U-Net (3D)	DSC + CE	112x160x128
Dolphins	Normalization, Resampling	nnU-Net augmentations	nnU-Net (3D)	-	BiXLSTM + ResNet, CNN decoder (3D)	DSC + CE	128x192x160
SW-FASEDIT (Baseline)	Normalization, Cropping, Flipping, Rotation	-	MONAI	-	DynUNet (3D)	DSC	224x224x224
Task 2: Longitudinal CT lesion segmentation							
LesionLocator	Normalization, Resampling	nnU-Net augmentations, longitudinal misalignments, center jiggle	LongiSeg	-	ResEncL U-Net (3D)	DSC + CE	112x224x224
nRocholl	Clamping, Rescaling	Small affine, noise, Scale, Shift, Contrast, Gaussian smooth	MONAI + nnU-Net (3D) + Lightning	Resampling, Labelling	Dual-encoder U-Net (3D)	DSC + CE	128x128x64
Baseline	nnU-Net preprocessing without resampling	nnU-Net augmentations	nnU-Net (3D)	-	ResEncL U-Net (3D) + Semi supervision	DSC + CE	256x256x128

Supplementary Table S7: Summary of submitted methods for autoPET/CT-IV Task 1 (interactive PET/CT lesion segmentation) and Task 2 (longitudinal CT lesion segmentation). DSC: Dice Similarity Coefficient, CE: Cross Entropy.

References

- 1 Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 488-98.
- 2 Ferdinandus J, Violet J, Sandhu S, Hicks RJ, Ravi Kumar AS, Iravani A, et al. Prognostic biomarkers in men with metastatic castration-resistant prostate cancer receiving [177Lu]-PSMA-617. *European journal of nuclear medicine and molecular imaging*. 2020;47(10):2322-7.
- 3 Albertina B, Watson M, Holback C, Jarosz R, Kirk S, Lee Y, et al. The cancer genome atlas lung adenocarcinoma collection (tcga-luad). (No Title). 2016.
- 4 Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*. 2018;5(1):1-9.
- 5 Munk A, Ambsdorf J, Llambias S, Nielsen M. Amaes: Augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. *arXiv preprint arXiv:240800640*. 2024.
- 6 Beck M, Pöppel K, Spanring M, Auer A, Prudnikova O, Kopp M, et al. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*. 2024;37:107547-603.
- 7 Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*. 2021;18(2):203-11.
- 8 Hadlich M, Marinov Z, Kim M, Nasca E, Kleesiek J, Stiefelhagen R. Sliding window FastEdit: a framework for lesion annotation in whole-body pet images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2024. p. 1-5.
- 9 Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:221102701*. 2022.
- 10 Rokuss MR, Kirchhoff Y, Roy S, Kovacs B, Ulrich C, Wald T, et al. Longitudinal segmentation of MS lesions via temporal difference weighting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 64-74.
- 11 Rokuss M, Kirchhoff Y, Akbal S, Kovacs B, Roy S, Ulrich C, et al. LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-Body Imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025. p. 30872-85.
- 12 de Grauw M, Scholten ET, Smit EJ, Rutten MJ, Prokop M, van Ginneken B, et al. The ULS23 challenge: A baseline model and benchmark dataset for 3D universal lesion segmentation in computed tomography. *Medical Image Analysis*. 2025;102:103525.