



<https://doi.org/10.1038/s41698-025-00903-y>

A promptable CT foundation model for solid tumor evaluation

 Check for updates

**Léo Machado^{1,2}, Léo Alberge¹, Hélène Philippe^{1,2,3}, Elodie Ferreres¹, Julien Khlaut^{1,4}, Julie Dupuis¹,
Korentin Le Floch^{1,4}, Denis Habip Gatenyo⁵, Pascal Roux⁶, Jules Grégory^{2,3}, Maxime Ronot^{2,3}✉,
Corentin Dancette¹, Tom Boeken⁴, Daniel Tordjman¹, Pierre Manceron¹ & Paul Hérent^{1,6}**

Carcinogenesis is inherently complex, resulting in heterogeneous tumors with variable outcomes and frequent metastatic potential. Conventional longitudinal evaluation methods like RECIST 1.1 remain labor-intensive and prone to measurement errors, while existing AI solutions face critical limitations due to tumor heterogeneity, insufficient annotations, and lack of user interaction. We developed ONCOPILLOT, an interactive CT-based foundation model dedicated to 3D tumor segmentation, significantly refining RECIST 1.1 evaluations with active radiologist engagement. Trained on more than 8000 CT scans, ONCOPILLOT employs intuitive visual prompts, including point-click, bounding boxes, and edit-points. It attains segmentation accuracy that matches or exceeds state-of-the-art methods, provides radiologist-level precision for RECIST 1.1 measurements, reduces inter-observer variability, and enhances workflow efficiency. Integrating clinical expertise with interactive AI capabilities, ONCOPILLOT facilitates widespread access to advanced biomarkers, notably volumetric tumor analyses, thereby supporting improved clinical decision-making, patient stratification, and accelerating advancements in oncology research.

The wide variability in tumor appearance and location makes precise monitoring of oncological disease a critical challenge for both clinical care and research. Effective evaluation is essential for assessing tumor aggressiveness, predicting prognosis, and guiding treatment decisions. RECIST v1.1 has long been regarded as the gold standard for assessing solid tumors over time¹, allowing for patient stratification based on disease response or progression.

However, this method has significant limitations: low information yield from linear long-axis measurement compared to total tumor burden^{2,3}, arbitrary and non-reproducible selection of target lesions leading to misclassification of disease status⁴, and significant inaccuracies in measuring the long axis, with inter-reader variability exceeding 20%⁵, further contributing to classification errors.

Traditionally, long and short axes of the tumor are used as proxies for estimating tumor size on CT scans. However, linear measurements are increasingly considered inadequate as the field shifts toward more informative markers, such as volumetry⁶ and shape assessments, including tumor eccentricity and irregularity⁷. Volumetric analysis, more sensitive to change due to its proportionality to the cube of the radius, is advantageous for detecting tumor burden changes, especially for tumors with irregular shapes, where linear measurements fail⁸. Novel radiomics biomarkers

derived from volumetric analysis show promise in oncological evaluation, notably in colon and lung cancers^{9,10}.

Despite its promise, volumetric measurement is time-consuming¹¹ and impractical to perform manually. Early models relying on manual feature extraction to deep-learning approaches using convolutional neural networks¹² have limited success. Most are organ-specific and effective primarily in straightforward cases, such as lung nodules, but struggle with more complex lesions. Furthermore, these methods lack interactivity and adaptability, restricting their clinical integration.

The emergence of foundation models, powered by transformer architecture and self-attention mechanisms¹³, could alleviate these issues. Pre-trained on extensive unannotated datasets in a self-supervised manner, foundation models tend to outperform traditional deep-learning systems when evaluated on a wide range of downstream-tasks¹⁴. Remarkably, these models also exhibit emergent properties, where complex capabilities arise naturally from their scale and training, further boosting their adaptability. Their capacity for transfer learning and zero-shot classification allows them to tackle previously unseen challenges-or do so with minimal fine-tuning-making them a transformative force in medical imaging^{15,16}. In computer vision, for instance, they can generate reliable segmentation masks from simple visual cues like bounding boxes or point-click inputs¹⁷. This ability to

¹Raidium, Paris Biotech Santé, Paris, France. ²AP-HP. Nord, Department of Radiology, FHU MOSAIC, Beaujon Hospital, Clichy, France. ³Université Paris Cité, Paris, France. ⁴Department of Vascular and Oncological Interventional Radiology, Université Paris Cité, AP-HP, Hôpital Européen Georges Pompidou, HEKA INRIA, Paris, France. ⁵Department of Radiology, Hôpital Cochin, AP-HP, Paris, France. ⁶Centre d'Imagerie du Nord. Saint-Denis, France. ✉ e-mail: maxime.ronot@aphp.fr



THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

dynamically refine segmentations paves the way for more explainable AI, enhancing usability for radiologists.

In response, we developed ONCOPILLOT, an interactive foundation model trained on more than 8000 publicly available CT scans of general anatomy and tumors. ONCOPILLOT aims to deliver precise RECIST measurements and facilitate volumetric analysis, integrating seamlessly into the radiologist’s workflow.

Results

Foundation model

ONCOPILLOT is a foundation model adapted from SAM¹⁷, developed specifically for segmenting biomedical images. Similar methods have been concurrently explored in MedSAM¹⁸, SegVol¹⁹, and SAM-Med3D²⁰. It was pre-trained on a diverse dataset comprising normal

anatomy and oncological lesions, totaling 2374 CT scans including 104 anatomical structures (e.g., organs, bones) and 4 oncological lesions regardless of histology and malignancy (i.e., lung, liver, pancreas and colon tumors) from the MSD dataset (Fig. 1a), without distinction regarding their histological type or malignancy. To become specialized for oncology the model was subsequently fine-tuned on a comprehensive dataset of 6229 tumors from various organs (e.g., pancreas, bone, liver, kidney, lung, lymph nodes).

ONCOPILLOT is designed to interactively segment oncological lesions in 3D, utilizing visual prompts such as a bounding box (referred to as bbox) around the lesion of interest or a point-click (referred to as point) inside it (Fig. 1b). To simulate the dynamic refinement of the predicted segmentation masks by radiologists we developed an editing mechanism (referred to as point-edit, see Methods) which performs 4 successive prompting of the

Fig. 1 | ONCOPILLOT Foundation Model Training and Evaluation. **A** Overview of the datasets used for training the ONCOPILLOT segmentation model, including the distribution across train, test, and validation sets. **B** Diagram illustrating the ONCOPILLOT segmentation model’s workflow. The model accepts visual prompts (either point-clicks or bounding boxes) of 3D tumor volumes and outputs corresponding 3D segmentation masks. Optional editing allows for real or simulated radiologist interaction, where positive and negative edit-points can be set manually in a viewer environment or automatically during evaluation.

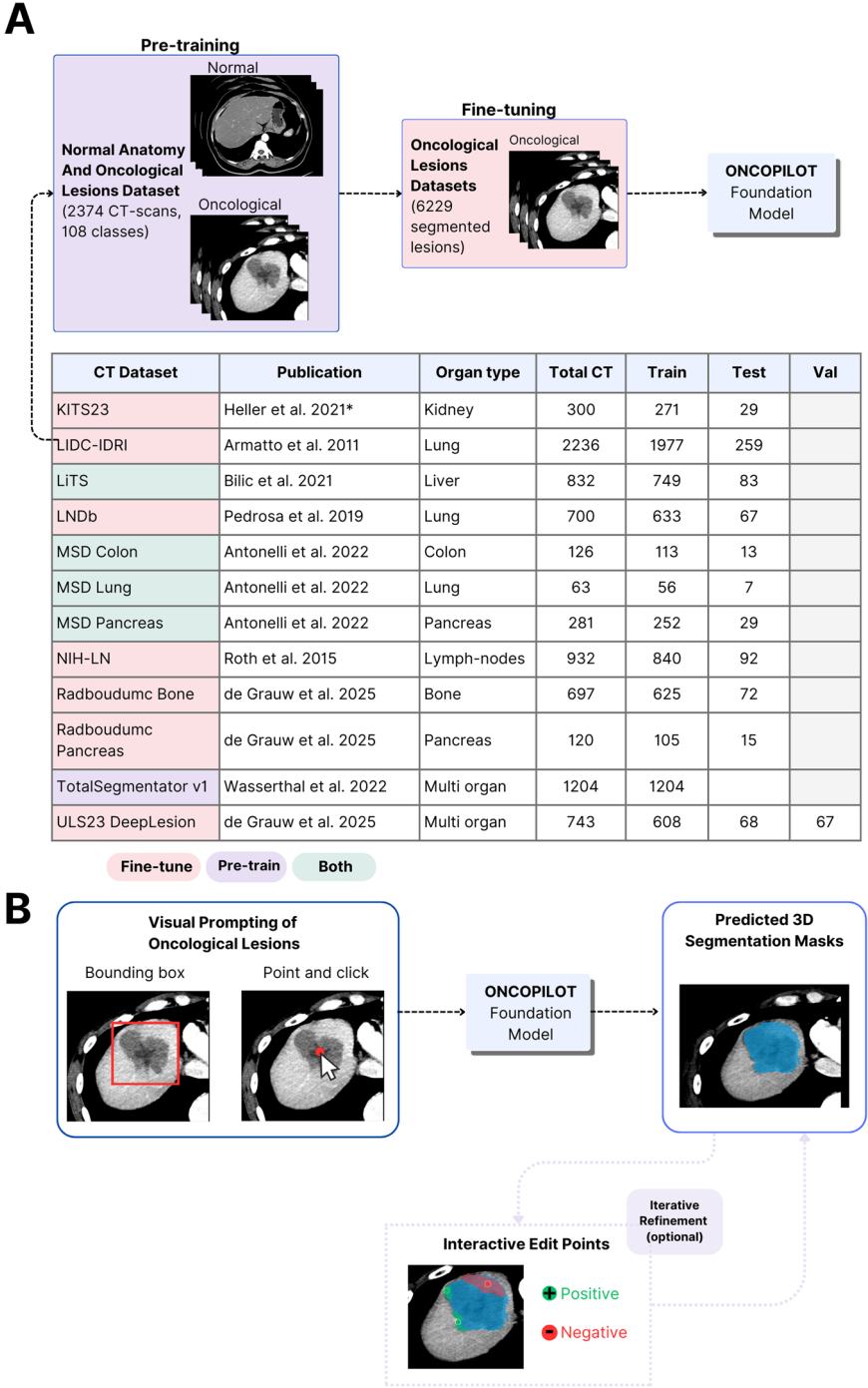
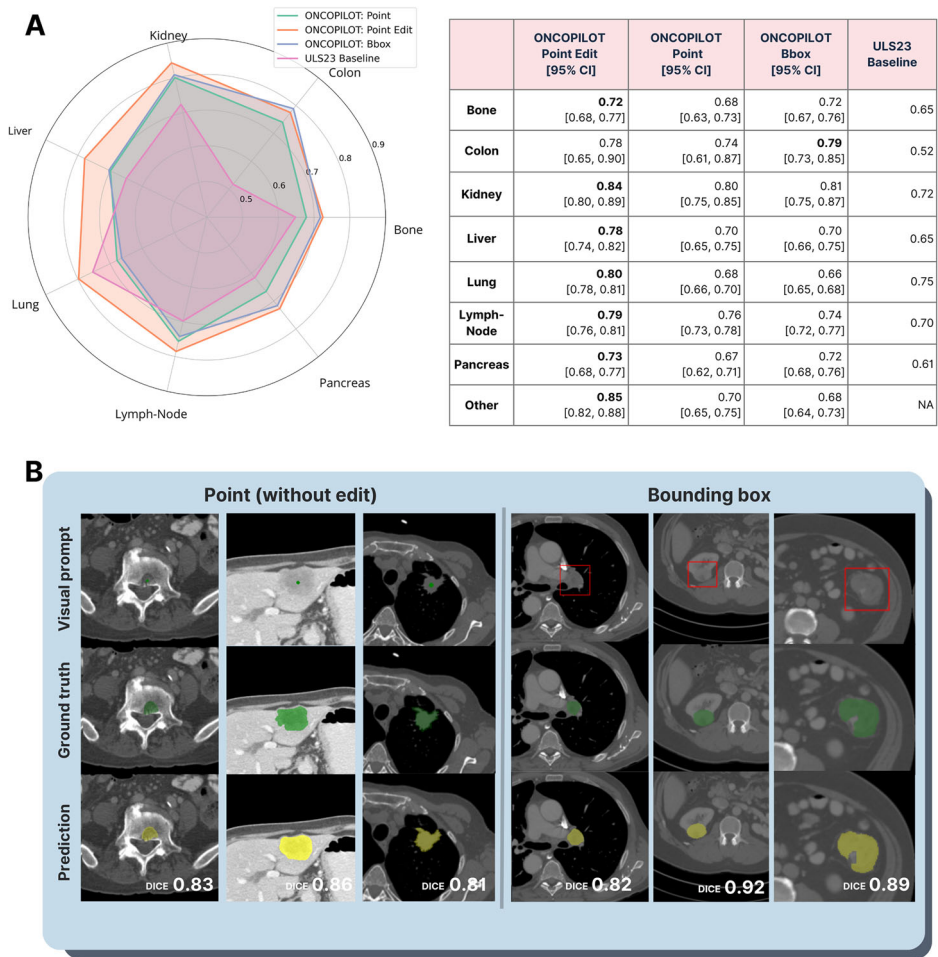


Fig. 2 | ONCOPILOT performance against baseline. **A** Radar plot (left) and table (right) displaying segmentation mean DICE scores across 7 lesion types for 3 different ONCOPILOT models (point, point-edit, bbox) compared to the best-performing baseline from the ULS23 segmentation challenge on the 10% held-out test set. **B** Examples of successful segmentations from the test set, comparing point mode (left columns) and bbox mode (right columns). The top row shows the visual prompt provided to the model, the middle row displays the ground truth mask for that slice, and the bottom row presents the ONCOPILOT model's predicted segmentation.



model with negative and positive edit-points targeting over-segmented and under-segmented areas respectively.

Segmentation performance

The state-of-the-art baseline for comparing ONCOPILOT’s performance was the model used in the ULS23 oncological lesion segmentation challenge²¹, based on nnUnet²². ONCOPILOT matched or surpassed the ULS23 model²¹ in all evaluation metrics-point, point-edit, and bbox-across all lesion types, with the exception of lung tumors, where only the point-edit model demonstrated superior performance (Fig. 2a, with examples of successful segmentations in Fig. 2b). ONCOPILOT achieved mean DICE scores of 0.70 (95% CI: 0.69–0.71) for point mode, 0.70 (95% CI: 0.68–0.71) for bbox mode, and 0.79 (95% CI: 0.78–0.80) for point-edit mode, compared to 0.70 for the baseline.

The distribution of lesion sizes by organs is provided in Supplementary Fig. 1b. To be noted, the test dataset was imbalanced, with over 40% of the lesions being lung tumors (further addressed in the Discussion section).

Morphology analysis

The segmentation masks produced by the model in point mode were influenced by the lesion morphology and size. Indeed, ONCOPILOT exhibited lower DICE scores for lesions with irregular, non-spherical shapes, with a mean DICE of 0.66 (95% CI: 0.63–0.69) for tumors with a sphericity index below 0.6, compared to 0.71 (95% CI: 0.7–0.73) for more spherical tumors in point mode ($p < 0.001$, Fig. 3a, Supplementary Fig. 2a).

Similarly, smaller lesions yielded lower DICE scores, with a mean of 0.67 (95% CI: 0.65–0.69) for lesions with a long axis <15 mm versus 0.73 (95% CI: 0.72–0.75) for larger lesions ($p < 0.001$, Fig. 3b, Supplementary

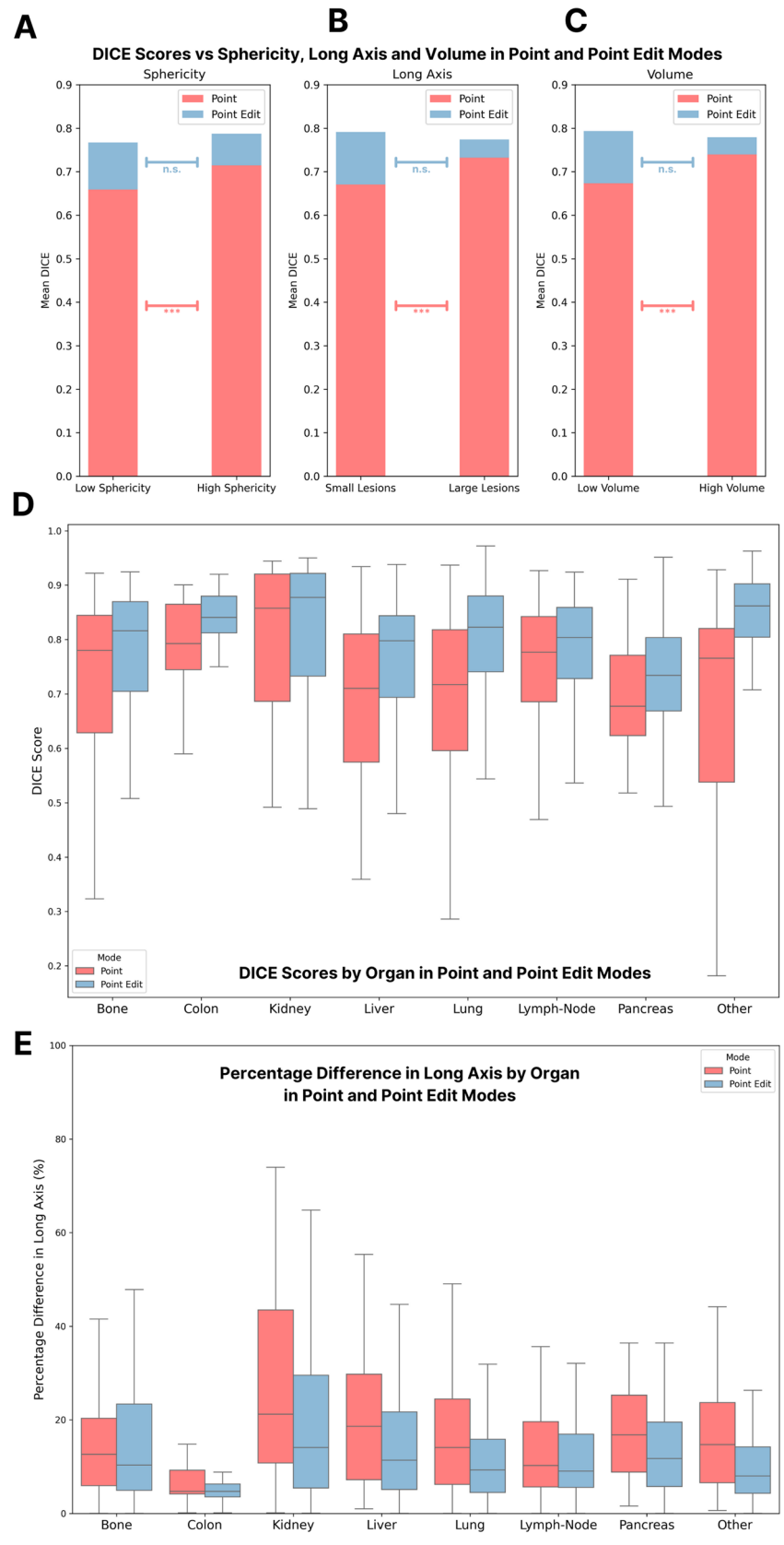
Fig. 2b). This trend persisted when using volume as a metric: lesions under 1 mL had a mean DICE of 0.67 (95% CI: 0.66–0.69), compared to 0.74 (95% CI: 0.72–0.76) for larger lesions ($p < 0.001$, Fig. 3c, Supplementary Fig. 2c). Crucially, interactive editing mitigated these biases, eliminating significant differences ($p > 0.05$) in DICE scores between lesions of varying sphericity, long axis, or volume in point-edit mode. This approach also reduced disparities in DICE between lesion types (Fig. 3d). Additionally, when using RECIST measurements for the long axis instead of DICE scores, interactive editing significantly reduced measurement errors, with the median error decreasing from 14.1% in point mode to 9.6% in point-edit mode ($p < 0.001$). This level of accuracy is consistent with the reported inter-reader variability among radiologists for single-lesion measurements⁵ (Fig. 3e).

ONCOPILOT evaluation against radiologists

To evaluate ONCOPILOT’s accuracy in clinical setting its long axis measurements were compared to those of radiologists. A validation set of 67 tumors from the ULS23 DeepLesion dataset was used, selected according to RECIST v1.1 guidelines (long axis ≥ 10 mm for solid lesions, short axis ≥ 15 mm for lymph nodes) and segmentation quality.

ONCOPILOT demonstrated radiologist-level performance in point, point-edit, and bbox modes (Fig. 4a, b). There was no statistically significant difference ($p > 0.05$) between the different ONCOPILOT models when evaluated against radiologists, with a median absolute error in long axis measurement of 1.3 mm (95% CI: 1–2.1) for radiologists (8.6% of the median lesion size) versus 1.1 mm (95% CI: 0.9–1.3) for ONCOPILOT in point-edit mode (7.4%), 1.6 mm (95% CI: 1–3) in point mode (10.8%), and 1.5 mm (95% CI: 0.9–2.3) in bbox mode (10.4%).

Fig. 3 | ONCOPILLOT Performance on Different Lesion Types. **A** Bar plot showing the mean DICE scores from ONCOPILLOT segmentation masks in point mode (red) and point-edit mode (blue) for spherical lesions (sphericity > 0.6) versus irregular lesions (see Methods for the sphericity formula). **B** Bar plot showing the mean DICE scores from ONCOPILLOT segmentation masks in point mode (red) and point-edit mode (blue) for large lesions (long axis > 15 mm) versus smaller lesions. **C** Bar plot showing the mean DICE scores from ONCOPILLOT segmentation masks in point mode (red) and point-edit mode (blue) for voluminous lesions (volume > 1 mL) versus smaller lesions. **D** Boxplot displaying the distribution of DICE scores produced by ONCOPILLOT in point mode (red) and point-edit mode (blue) across various lesion types in the 10% held-out test set, with median values and inter-quartile ranges highlighted. **E** Boxplot showing RECIST measurements error against the ground truth, derived from ONCOPILLOT's predicted masks in point mode (red) and point-edit mode (blue) across different lesion types in the 10% held-out test set, highlighting median values and interquartile ranges. The long axis is defined as the longest possible line in the axial plane across the predicted 3D mask. ***: $p < 0.001$; n.s: non-significant.



ONCOPILLOT integration into radiologist's workflow

To assess whether ONCOPILLOT could serve as an AI companion, its integration into the workflow was evaluated. ONCOPILLOT enhanced the reproducibility and efficiency of radiologist measurements, with an inter-

reader deviation of 1.7 mm when assisted by ONCOPILLOT versus 2.4 mm manually (Fig. 4c, d, $p < 0.05$), leading to an increase of ICC(2,1) from 0.88 to 0.93. Additionally, radiologists demonstrated a faster measurement speed using ONCOPILLOT, with an average time of 17.2 s per measurement

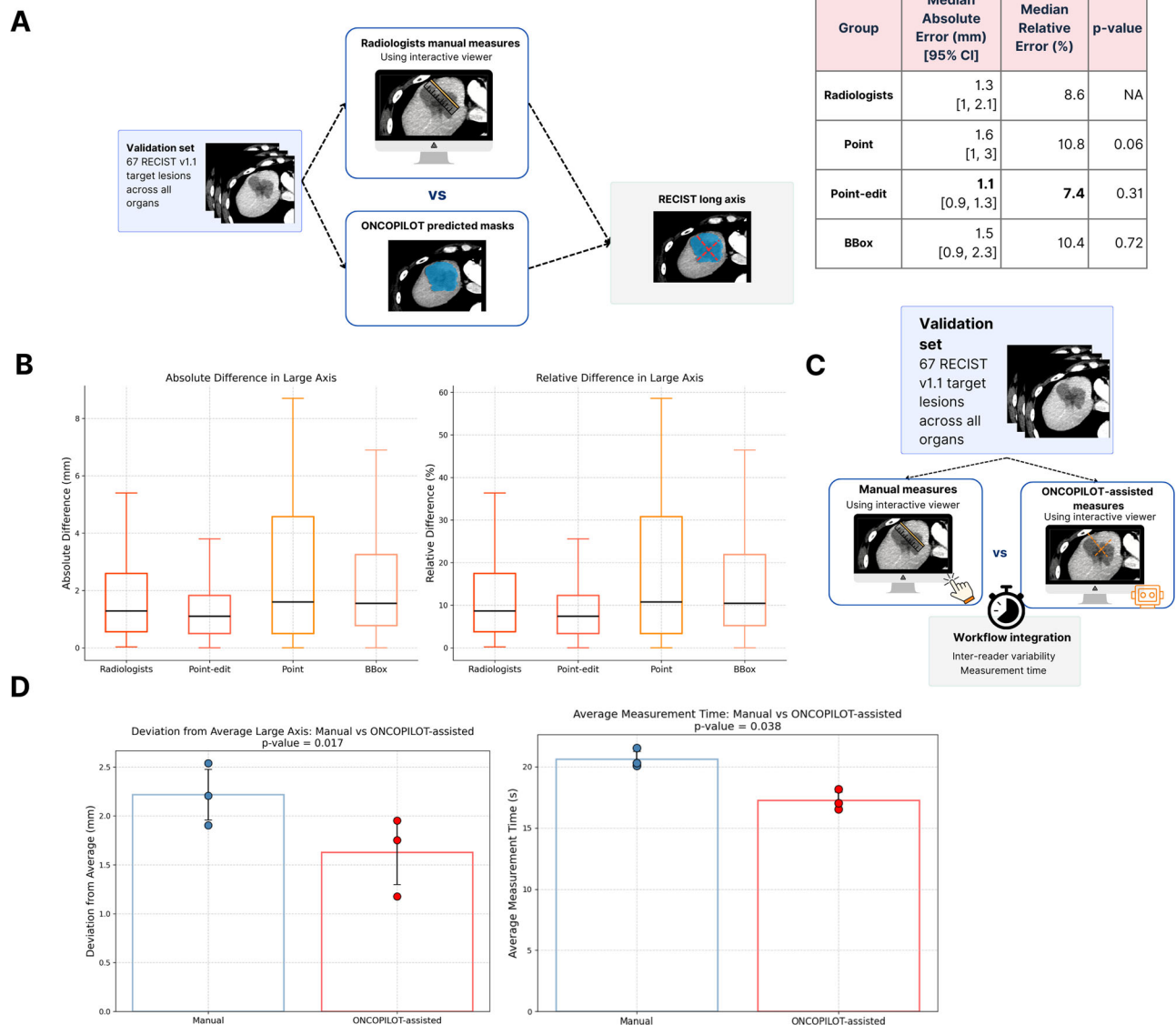


Fig. 4 | ONCOPILLOT Integration Into Radiologist's Workflow. **A** Diagram and results comparing ONCOPILLOT in point, point-edit, and bbox modes against three radiologists for the long-axis measurement of diverse oncological lesions. Median absolute error (mm) and median relative error (% of lesion size) are shown. *p* values from *t*-tests compare ONCOPILLOT models to radiologists for long-axis measurement error, without statistical significance $p \geq 0.05$. The long axis is the longest line in the axial plane across the predicted 3D mask. **B** Boxplot (bottom) of ONCOPILLOT's tumors long-axis measurement performance against radiologists. Left:

median absolute error (mm) vs. ground truth. Right: median relative error (% of lesion size). Median and interquartile ranges are shown. **C** Diagram of an experiment evaluating radiologists' inter-operator variability and measurement time while measuring tumors' long-axis using a digital viewer for manual vs. ONCOPILLOT-assisted (bbox mode) long-axis assessments. **D** Boxplots show radiologists' inter-operator variability in measurement error (left) and measurement time (right) using manual vs. ONCOPILLOT-assisted annotations across diverse tumors, with *t*-test *p* values; $n = 3$.

compared to 20.6 s with manual annotations ($p < 0.05$). Notably, this improvement in speed was achieved without focusing on speed optimization, as it operated on a non-optimized web-based platform (showcased in Supplementary Fig. 3a–e). Most of the measurement time was spent locating the lesion within the exam, suggesting that ONCOPILLOT could be further accelerated with targeted improvements.

Segmentation performances on a external dataset

To demonstrate out-of-distribution generalizability, we evaluated ONCOPILLOT on an external multicentric cohort of real-world patients undergoing longitudinal oncological evaluation. Moreover, we compared ONCOPILLOT to the ULS23 baseline and to non-specialized models: the state-of-the-art promptable foundation model SAM-Med3D-turbo²⁰ and ONCOPILLOT-zero the pre-trained version of ONCOPILLOT (see Methods).

On this dataset, ONCOPILLOT achieved a mean DICE score of 0.79 (95% CI: 0.75–0.84), 0.68 (95% CI: 0.63–0.74), and 0.70 (95% CI: 0.65–0.76) for point-edit, point, and bbox modes, respectively. These results match or slightly exceed the ULS23 baseline, which exhibited a mean DICE of 0.66 (95% CI: 0.61–0.71) (see Supplementary Fig. 4a).

For comparison, ONCOPILLOT-zero and SAM-Med3D (turbo version, 5 edits) achieved a mean DICE of 0.58 (95% CI: 0.53–0.64) and 0.44 (95% CI: 0.38–0.5) respectively, demonstrating that specialized model outperform generic ones.

To be noted, ONCOPILLOT's performance on lung lesions was superior in this validation cohort compared to previously reported results, with a mean DICE of 0.71 (95% CI: 0.60–0.80) vs 0.68 (95% CI: 0.66–0.70) for point, and 0.80 (95% CI: 0.73–0.86) vs 0.66 (95% CI: 0.65–0.69) for bbox. Consistently, the lung tumors in the external validation cohort were far

bigger with a median large-axis of 23 mm vs 9 mm (Supplementary Fig. 4b). Some successful and failed segmentation cases from the external cohort are displayed in Supplementary Fig. 5.

Discussion

In summary, ONCOPILOT matched state-of-the-art performance in tumor segmentation across a diverse set of oncological lesions, achieving radiologist-level accuracy in RECIST 1.1 measurements. The model's flexibility, enabled by interactive visual prompts and refinement capabilities in a radiological viewer, marks a significant advancement in integrating an explainable AI copilot into the imaging workflow while keeping the radiologist in the loop. This strategy not only minimizes inter-reader variability and reduces measurement time, but also offers greater flexibility than rigid, task-specific segmentation models.

Indeed, while recent advances have improved segmentation models considerably, they still sometimes generate errors ranging from significant failures reminiscent of generative model hallucinations to more subtle inaccuracies along tumor boundaries. For example, delineating certain tumors poses inherent challenges that demand nuanced expert intervention and manual refinement, especially when dealing with heterogeneous tumors, necrotic lymph node clusters, confluent lesions, or diffuse infiltrative masses. Consequently, retaining radiologist oversight is crucial, underscoring the value of an interactive editing mode. Without the ability to correct initial segmentation shortcomings, the clinical adoption of these tools could be severely limited.

Clinical evaluation of this segmentation model is needed to establish how these gains in efficiency and precision translate into meaningful improvements in real-world longitudinal oncological evaluation and influence disease status assessment. Addressing the management of multiple lesions per patient across various time points, in line with RECIST 1.1 guidelines, is a critical prerequisite for future studies.

Additionally, it will be important to examine the learning curve and workflow implications for radiologists employing this segmentation tool, especially for the editing mechanism, as well as to identify potential pitfalls in its use, given that human-AI collaborations do not always yield optimal outcomes²³. However, due to the straightforward nature of our visual prompts, relying solely on bounding boxes and click points, we are confident that these segmentation tools can be easily mastered by users across a wide range of skill levels.

Regarding ONCOPILOT suboptimal performance on lung tumors, lesion size appears to be the main factor involved, highlighting a limitation of our model that should be addressed in future versions. Indeed, lung tumors in the test set were predominantly small nodules of uncertain oncological relevance, in part not acceptable under the RECIST 1.1 guidelines with a median size of 9 mm compared to 20 mm for non-lung tumors. This disproportionate representation (more than 40%) of lung lesions in the test dataset skewed the overall results, disadvantageous to our model's performance. This hypothesis is supported by the more convincing performance of ONCOPILOT on the lung lesions from the external validation cohort, which were larger.

ONCOPILOT not only enhances the precision and consistency of RECIST-based oncological assessments but also goes beyond traditional RECIST measurements by enabling volumetric analysis and uncovering previously unexplored radiomic features. Volumetric biomarkers, such as tumor growth rate and total tumor burden, combined with morphology-based markers, offer more comprehensive and accurate indicators of tumor mass and aggressiveness compared to conventional long and short axis measurements. These novel radiomic biomarkers will better accommodate the variability in tumor presentations, providing a more precise characterization of oncological disease.

This study leverages publicly available baseline and data²¹ as well as open-source model architecture¹⁷, demonstrating that foundation models are already capable of delivering impactful results in the biomedical field without significant technical hurdles. ONCOPILOT showcases the promising potential of this technology, with future iterations expected to be

significantly more advanced and effective, which will require extensive clinical validations. These results reinforce our belief that foundation models are a pivotal step toward the next generation of AI-assisted radiology.

Through this work, we aim to demonstrate oncological evaluation as the first use case for the native integration of foundation-model-based AI assistants into the radiologist's workflow, paving the way for improved patient stratification, optimized clinical trial monitoring, more informed treatment decisions, and ultimately enhanced patient care.

Methods

Foundation model

ONCOPILOT is trained to perform segmentation tasks on 2D images using prompts such as a bounding box, a point, or a mask. It generates a 3D prediction of an anatomical structure from the input image and visual prompts. The segmentation is propagated sequentially along the z-axis, starting from the initial 2D slice, until the object's boundaries are reached, resulting in a segmentation mask. Alternatively, propagation can stop based on predefined criteria.

Our foundation model was initialized using SAM model weights¹⁷ and underwent supervised training on diverse anatomical segmentation tasks to provide an intermediary model, ONCOPILOT-zero. It is then fine-tuned specifically on tumors to deliver the final model, ONCOPILOT. The model's objective is a combination of DICE and cross-entropy loss. The model is trained to segment various anatomical structures on all axes and is prompted using either a point that is sampled randomly on the mask, a bounding box that is sampled around the mask with a random pixel offset, or a mask that is rotated from -10 to 10 degrees, scale up or down by up to 10%, move up, down, right or left by up to 10% and finally eroded and dilated. Random windowing augmentation is performed with a shift of up to 10%.

Finally, the model is fine-tuned on tumors following the same procedure. The pre-training took 40 h using 32 V100 GPUs (1280 GPU h) with a constant learning rate of 10^{-5} . The fine-tuning required 10 h on a Nvidia 4090 GPU and uses the same configuration.

Baseline

The baseline model (nnUnet-ResEnc+SS) was evaluated on the 10% held-out test set from their labeled dataset of 38,693 lesions. It takes as input volume-of-interest images of size $128 \times 256 \times 256$ padded to the minimum intensity value when necessary. To simulate a radiologist's click on the lesion, the volume-of-interest is centered on the lesion.

Datasets

ONCOPILOT's pre-training was performed using publicly available datasets with medical images and segmentation masks for general anatomy and oncological lesions:

- 1204 CT scans from TotalSegmentator v1²⁴, with 104 labeled anatomical structures (27 organs, 59 bones, 10 muscles, 8 vessels).
- 743 diverse tumors from the DeepLesion dataset²⁵, curated and segmented for the ULS23 challenge²¹, referred to as ULS23 DeepLesion.
- 697 bone oncological lesions and 120 pancreatic tumors from the Radboudumc hospital, available through the ULS23 dataset²¹.
- 470 volumes from the multimodal MSD challenge²⁶, using only the Lung, Colon, Pancreas datasets.
- 700 lung nodules from the LNDdb dataset²⁷.
- 300 kidney tumors from the KITS23 dataset²⁸.
- 832 liver tumors from the LiTS dataset²⁹, also part of the MSD challenge.
- 932 mediastinal and abdominal lymph nodes from the NIH-LN dataset³⁰.
- 2236 lung oncological lesions from the LIDC-IDRI dataset³¹.

A 90% training set was selected randomly, leaving 10% as a held-out test dataset, following the ULS23 challenge methodology. A validation set

comprising 67 tumors ≥ 10 mm (≥ 15 mm for lymph nodes) was reserved for comparison against radiologists, selected from the ULS23 DeepLesion training set.

Segmentation process

The model had access to the entire volume and visual prompts. The volume was thresholded using a wide windowing range of $[-500, 1000]$ HU, which yielded optimal results on the diverse ULS23 DeepLesion dataset (Supplementary Fig. 1a).

The model initially outputs a 2D segmentation mask for the middle slice. Then, autoregressive propagation of segmentation masks occurs across adjacent slices, using the prior mask as a prompt for the next slice, thus producing a 3D segmentation mask.

ONCOPILLOT was evaluated using three experimental settings simulating real-life visual prompts:

- **Bounding box:** A 2D bounding box outlines the lesion from the middle slice of the ground-truth mask, expanded by a 15-pixel offset.
- **Point-click:** A single point is placed at the barycenter of the ground-truth mask or the nearest point within the mask.
- **Point-edit:** To simulate radiologist interactions, the 3D segmentation mask from point-click mode is refined by up to 4 edits, targeting areas with prediction errors (either reducing over-segmentation or expanding under-segmentation).

On the ULS23 DeepLesion dataset, each segmentation inference from a visual prompt takes approximately 50 ms allowing real-time edits. The initial exam preprocessing step takes around 7 seconds for the entire CT scan on a Nvidia 4090 GPU, can be performed asynchronously, and is only performed once per exam.

Segmentation performances

DICE scores were reported to compare segmentation models' performances on the test sets. 95% confidence intervals (CI) were computed using the standard normal distribution approach, assuming the normality of the sampling distribution.

Morphology analysis

A sphericity index is used to evaluate shape regularity, calculated as the ratio of the surface area of a perfect sphere to the surface area of the ground-truth segmentation mask for objects of equal volume. A perfect sphere has a sphericity index of 1, while irregular structures have values closer to 0. The formula is

$$S = \frac{\pi^{1/3} \cdot (6V)^{2/3}}{A} \quad (1)$$

where S represents sphericity, V the object volume, and A the surface area. Independent Student's t tests were performed to compare mean DICE/RECIST measurements across predefined size, volume and sphericity groups.

RECIST measurement

ONCOPILLOT's RECIST measurements were based on the segmentation masks in bounding box, point, and point-edit modes. The primary measurement was the long axis of the lesion, simplified and restricted to the axial plane, even for lymph nodes. The axial plane is generally recommended as the default plane for RECIST 1.1 measures on CT scans, even when provided with isotropic sequences, to ensure comparison with non-isotropic follow-ups or previous exams¹.

Measurements from ONCOPILLOT and three radiologists with ≥ 18 months of experience were compared against measurements inferred from ground-truth segmentation masks to calculate measurement errors. Wilcoxon Signed-Rank tests were performed to compare median RECIST measurements across different visual prompts versus radiologists measurements. 95% confidence intervals (CI) for median measurement errors

were computed using the standard normal distribution approach, assuming the normality of the sampling distribution.

Radiologists used a web-based viewer for both manual and ONCOPILLOT-assisted measurements. They could freely adjust the windowing and navigate the volume but without multi-planar reconstruction.

ONCOPILLOT integration into radiologist's workflow

Inter-operator variability was calculated as the absolute deviation of each radiologist's measurement from the overall average for each lesion, using manual measures or ONCOPILLOT-assisted with bounding box prompts. A paired Student's t test was performed to compare interoperator deviation with and without ONCOPILLOT assistance. ICC(2,1) (two-way random effects model for absolute agreement) coefficients were reported with and without ONCOPILLOT. Measurement duration was defined as the time from the initial display of the CT to the final measurement. A paired Student's t -test was performed to compare mean measurement durations.

Segmentation performance on a external multicentric dataset

A private anonymized external dataset was curated from Gradient Health data platform. 37 patients from multiple centers in North and South America who underwent CT examination for RECIST 1.1 longitudinal analysis were included, totaling 87 lesions. The lesions were segmented in 3D by one radiologist followed by a review from another. This dataset was used uniquely for testing the model, without any overlap with the training datasets. DICE scores were reported, as well as their stratification by lesion type.

Data availability

No datasets were generated or analysed during the current study.

Code availability

Due to intellectual property constraints, the code used to train and evaluate ONCOPILLOT is not publicly available at the time of publication.

Received: 26 December 2024; Accepted: 5 April 2025;

Published online: 25 April 2025

References

1. Eisenhauer, E. A., Therasse, P. & Bogaerts, J. et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
2. Meignan, M. et al. Baseline metabolic tumor volume predicts outcome in high-tumor-burden follicular lymphoma: a pooled analysis of three multicenter studies. *J. Clin. Oncol.* **34**, 3618–3626 (2016).
3. Chung, H. W., Lee, K. Y. & Kim, H. J. et al. Fdg pet/ct metabolic tumor volume and total lesion glycolysis predict prognosis in patients with advanced lung adenocarcinoma. *J. Cancer Res. Clin. Oncol.* **140**, 89–98 (2014).
4. Kuhl, C. K. et al. Validity of recist version 1.1 for response assessment in metastatic cancer: a prospective, multireader study. *Radiology* **290**, 349–356 (2019).
5. Yoon, S. H., Kim, K. W. & Goo, J. M. et al. Observer variability in recist-based tumour burden measurements: a meta-analysis. *Eur. J. Cancer* **53**, 5–15 (2016).
6. Planz, V. B., Lubner, M. G. & Pickhardt, P. J. Volumetric analysis at abdominal ct: Oncologic and non-oncologic applications. *Br. J. Radiol.* **92**, 20180631 (2019).
7. Wang, S., Chen, A. & Yang, L. et al. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Sci. Rep.* **8**, 10393 (2018).
8. Hayes, S. A., Pietanza, M. C. & O'Driscoll, D. et al. Comparison of ct volumetric measurement with recist response in patients with lung cancer. *Eur. J. Radiol.* **85**, 524–533 (2016).
9. Dercle, L., Zhao, B. & Gönen, M. et al. An imaging signature to predict outcome in metastatic colorectal cancer using routine computed tomography scans. *Eur. J. Cancer* **161**, 138–147 (2022).

10. Dercle, L., Fronheiser, M. & Rizvi, N. A. et al. Baseline radiomic signature to estimate overall survival in patients with nscl. *J. Thorac. Oncol.* **18**, 587–598 (2023).
11. Zimmermann, M., Kuhl, C. K., Engelke, H., Bettermann, G. & Keil, S. Ct-based whole-body tumor volumetry versus recist 1.1: Feasibility and implications for inter-reader variability. *Eur. J. Radiol.* **135**, 109514 (2021).
12. Tandon, R., Agrawal, S., Rathore, N. P. S., Mishra, A. K. & Jain, S. K. A systematic review on deep learning-based automated cancer diagnosis models. *J. Cell. Mol. Med.* **28**, e18144 (2024).
13. Vaswani, A., Shazeer, N. M. & Parmar, N. et al. Attention is all you need. *Adv. Neural Inform. Process. Syst.* **30**, 5998–6008 (2017).
14. Bommasani, R., Hudson, D. A. & Adeli, E. et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
15. Azad, B. et al. Foundational models in medical imaging: a comprehensive survey and future vision. arXiv preprint arXiv:2310.18689 (2023).
16. Paschali, M. et al. Foundation models in radiology: What, how, why, and why not. *Radiology* **314**, e240597 (2025).
17. Kirillov, A., Mintun, E. & Ravi, N. et al. Segment anything. arXiv preprint arXiv:2304.02643 (2023).
18. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
19. Du, Y., Bai, F. & Huang, T. et al. Segvol: Universal and interactive volumetric medical image segmentation. arXiv preprint arXiv:2311.13385 (2023).
20. Wang, H. et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023).
21. de Grauw, M. J. J. et al. The ULS23 challenge: A baseline model and benchmark dataset for 3D universal lesion segmentation in computed tomography. *Med. Image Anal.* **102**, 103525 (2025).
22. Isensee, F., Jaeger, P. F. & Kohl, S. A. A. et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Med. Methods* **18**, 203–211 (2021).
23. Yu, F. et al. Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
24. Wasserthal, J. et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.* **5**, e220024 (2023).
25. Yan, K., Wang, X., Lu, L. & Summers, R. M. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J. Med. Imaging* **5**, 036501 (2018).
26. Antonelli, M., Reinke, A. & Bakas, S. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
27. Pedrosa, J. et al. Lndb challenge on automatic lung cancer patient management. *Med. Image Anal.* **70**, 102027 (2021).
28. Heller, N. et al. The KITS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. Preprint at <https://arxiv.org/abs/2307.01984> (2023).
29. Bilic, P., Christ, P. & Li, H. B. et al. The liver tumor segmentation benchmark (lits). *Med. Image Anal.* **84**, 102680 (2023).
30. Roth, H., Farag, A. & Lu, L. et al. A new 2.5 d representation for lymph node detection in ct (ct lymph nodes). *Cancer Imaging Archive*. <https://doi.org/10.7937/k9/tcia.2015.aqiidcnm> (2015).
31. Armato, S. G., McLennan, G. & Bidaut, L. et al. The lung image database consortium (lidc) and image database resource initiative

(idri): A completed reference database of lung nodules on ct scans. *Med. Phys.* **38**, 915–931 (2011).

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011013489R2 made by GENCI. We are grateful to the Hering laboratory at Radboud University Medical Center, Nijmegen (The Netherlands) for the ULS23 Challenge initiative.

Author contributions

Conceptualization: L.M., L.A., E.F., J.K., C.D., D.T., P.M. & P.H. Experiment design: L.M., L.A., E.F., J.K., C.D., D.T., P.M. Experiment implementation: L.M., L.A., E.F., J.K., C.D., K.L.F., D.H.G. & D.T. Result investigation: L.M., L.A., D.T. & P.M. Funding acquisition: M.R., J.G., P.M. & P.H. Project administration: P.R., T.B., M.R., J.G., P.M. & P.H. Supervision: M.R., J.G., P.R., T.B., P.M. & P.H. Writing original draft: L.M., H.P., J.K., E.F., J.D., P.R., J.G., M.R., D.T., P.M. Writing review and editing: L.M., L.A., D.T., M.R. & P.M.

Competing interests

L.M., L.A., H.P., E.F., J.K., J.D., C.D., D.T., and K.L.F. receive financial compensation from the company Raidium. P.R., T.B., P.M., and P.H. hold partial ownership (stocks) in the company Raidium. All other authors declare no financial or non-financial competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41698-025-00903-y>.

Correspondence and requests for materials should be addressed to Maxime Ronot.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025