A co-evolving agentic AI system for medical imaging analysis

Songhao Li[1,2], Jonathan Xu[3,*], Tiancheng Bao[2,*], Yuxuan Liu[4,*], Yuchen Liu[2,*], Yihang Liu[2,*], Lilin Wang[2], Wenhui Lei[1], Sheng Wang[1], Yinuo Xu[5], Yan Cui[1,4], Jialu Yao[1,2], Shunsuke Koga[1], Zhi Huang[1,6,✉]

[1] Department of Pathology and Laboratory Medicine, University of Pennsylvania
[2] Department of Electrical and System Engineering, University of Pennsylvania
[3] The Wharton School, University of Pennsylvania
[4] Department of Bioengineering, University of Pennsylvania
[5] Department of Computer and Information Science, University of Pennsylvania
[6] Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania

* Equal contribution
✉ Correspondence: Zhi Huang (zhi.huang@pennmedicine.upenn.edu)

Open-source code repository   TissueLab platform access

Abstract

Agentic AI is rapidly advancing in healthcare and biomedical research. However, in medical image analysis, their performance and adoption remains limited due to the lack of a robust ecosystem, insufficient toolsets, and the lack of real-time interactive expert feedback. Here we present "TissueLab", a co-evolving agentic AI system that allows human to ask direct research questions, automatically plan and generate explainable workflows, and conduct real-time analyses where experts can visualize intermediate results and refine them. TissueLab's ecosystem integrates tool factories spanning pathology, radiology, and spatial omics domains. By standardizing the inputs, outputs, capabilities, and use cases of diverse tools, TissueLab determines when and how to invoke these expert tools to address research and clinical questions. Through experiments across diverse tasks, where clinically meaningful quantifications directly inform staging, prognosis, and treatment planning, we show that TissueLab achieves state-of-the-art performance compared with end-to-end vision-language models (VLM) and other agentic AI systems such as GPT-5. Moreover, TissueLab ecosystem continuously learns from clinicians, accumulating knowledge and evolving toward improved classifiers and more effective decision strategies. With transparent model refinement through active learning, it can deliver accurate results in previously unseen disease contexts within minutes without requiring massive datasets or prolonged retraining. In colon cancer, TissueLab reached 94.9% accuracy in neoplastic cell quantification within 10-30 minutes of feedback with real-time updates, and in prostate cancer it adapted in two minutes to achieve 99.8% accuracy in tumor-to-duct ratio measurement, outperforming state-of-the-art VLM baselines. Released as a sustainable open-source ecosystem (tissuelab.org), we expect TissueLab to significantly advance and accelerate computational research and translational adoption in medical imaging, while establishing a foundation for the next generation of medical AI.

1 Introduction

Medical image analysis provides the computational foundation for understanding disease status and progression. It is essential for clinical decision-making, treatment planning, and, most importantly, advancing scientific discovery. Traditionally, the design and implementation of comprehensive medical image analysis pipelines required extensive theoretical and computational expertise. These pipelines were expensive to build and challenging to adapt to diverse scientific questions. Clinicians often propose important investigative questions, but relied on collaboration with computational scientists to establish workflows and carry out the research. Such question- and hypothesis-driven computational analyses, reliant on manually constructed pipelines and slow feedback cycles, were not scalable and could not accelerate discovery at the pace required for next-generation autonomous biomedical research.

Advances in medical foundation models have provided stronger baselines for many downstream tasks, yet cancers and other complex diseases present highly variable morphologies, imaging modalities, often requiring carefully customized workflows across diverse study objectives, which makes fully automated analysis across diverse medical imaging scenarios challenging. At present, no single model reliably addresses the full spectrum of clinical imaging needs. Clinical quantifications remain central to prognosis and treatment planning. For example, evaluating the tumor-to-normal cell ratio in histology images is a critical yet tedious task that cannot be performed manually at scale. Current vision-language models (VLMs) often fail on such questions: they are not transparent about the computational process and are prone to hallucinations, which degrades performance, undermines trust, and ultimately leads to algorithm aversion [1]. Building effective medical AI therefore requires more than off-the-shelf foundation models [2, 3]. These models must be fine-tuned or embedded within more sophisticated, adaptive pipelines that can directly answer research and clinical questions in a reliable, reproducible, and transparent manner. The central challenge is translating rapidly evolving AI capabilities into safe, trustworthy systems that adapt to unseen scenarios without manual workflow design or extensive retraining.

Recent breakthroughs in large language models (LLMs) driven agentic AI [4–8] demonstrated the potential to generalize across diverse tasks. By enabling flexible orchestration of tools through LLMs, these agents promise a path toward more universal AI applications that could be directly usable in clinical settings [9–11]. However, current VLM-based agentic AI still faces intrinsic limitations. Most agents depend on a fixed toolbox of pretrained models [12] that quickly become obsolete as the field evolves, requiring substantial maintenance and redevelopment costs. Moreover, hallucinations [13, 14], token overload, and attention dilution make it difficult for VLMs to safely coordinate multiple models or process high-dimensional data such as gigapixel whole-slide images or volumetric CT scans [15]. Lastly, clinicians and researchers are unable to intervene and refine intermediate tools dynamically, limiting adaptability, performance, and trust in the overall workflow [16]. Moreover, expert knowledge and preferences are not retained for future reuse. As a result, current agentic systems often reduce to pipelines of proprietary tools, which are costly to develop, fragile to maintain, and poorly suited for unseen clinical cases where accurate quantification can directly alter prognosis and treatment strategies. These gaps perpetuate the divide between AI researchers and clinical practitioners, limiting research and real-world translational adoption.

To address these challenges, we introduce TissueLab, a co-evolving agentic AI system designed for medical imaging analysis that continuously evolves with new tools and user feedback. TissueLab emphasizes four principles: (i) **adaptivity**, realized through a modular plugin architecture that allows any state-of-the-art model to be mounted as a task node, further enhanced by semantic function-calling, which enables the agent to interpret and operate on heterogeneous data stored in the memory layer regardless of format, and by parallel workflow execution, which leverages topological sorting to distribute inference across independent branches so that total runtime is effectively bounded by the longest critical path rather than the sum of all tasks; (ii) **co-evolution**, by incorporating clinician feedback into active learning loops and maintaining an editable memory layer that persistently records intermediate results as training examples. Feedback provided by clinicians is transformed into labeled data that supports lightweight fine-tuning of downstream modules, enabling the system to rapidly adapt to clinical needs. In this way, the editable memory layer not only ensures transparency and traceability, but also serves as a reservoir of supervision that drives continual system refinement through feedback-driven fine-tuning; (iii) **safety**, as every diagnosis is explicitly grounded in authoritative clinical standards rather than unconstrained model reasoning. To achieve this, TissueLab employs the Model Context Protocol (MCP) to dynamically retrieve external guidelines and criteria ensuring that outputs are always supported by traceable references. In addition, all intermediate states are persisted within the system and remain fully transparent and visualizable to users, so that every decision can be reviewed, verified, and reproduced. By coupling external evidence retrieval with persistent transparency, TissueLab minimizes hallucinations and guarantees that its recommendations align with evolving standards of care. and (iv) **community value**, by enabling both AI researchers and clinicians to contribute models, annotations, and workflows within a shared ecosystem, thereby enhancing collective capability while allowing users to tailor the system to their own needs. Together, these principles position TissueLab as a bridge between rapid advances in AI research and their safe translation into reproducible, guideline-aligned, and clinically meaningful applications that remain responsive to evolving standards of care.

We demonstrate the versatility of TissueLab across pathology, radiology, and spatial omics, where clinically meaningful quantification directly informs staging, prognosis, and treatment planning. Beyond immediate clinical applications, TissueLab contributes to biomedical research by providing expert-level, scalable assessments that accelerate discovery and support cancer research and other disease studies. Through relational analyses and task-oriented evaluations, TissueLab enables universal yet trustworthy AI agents that bridge the divide between research innovation and clinical adoption-laying the foundation for an open, sustainable ecosystem that evolves in tandem with both medical needs and technological progress.

TissueLab is released as an open-source software available across Windows, macOS, and Linux. The online ecosystem is also available at tissuelab.org. All clinicians and researchers are free to access to the system, perform medical imaging analysis and annotations, contribute AI models, and co-develop agentic AI workflows. This commitment to openness reflects our team's mission to advance open-source medical AI and community value, ensuring that medical AI evolves collaboratively, sustainably, and in alignment with day-to-day practice.

2 Results

2.1 Creating a co-evolving agentic AI system for medical imaging

We developed TissueLab, the first co-evolving agentic AI ecosystem for medical image analysis. Inspired by the Factory Method design pattern [17], TissueLab abstracts diverse imaging problems into a small set of fundamental operations such as tissue segmentation, classification, and local code analysis, thereby enabling clinicians to access advanced AI analysis without requiring AI research or coding expertise.

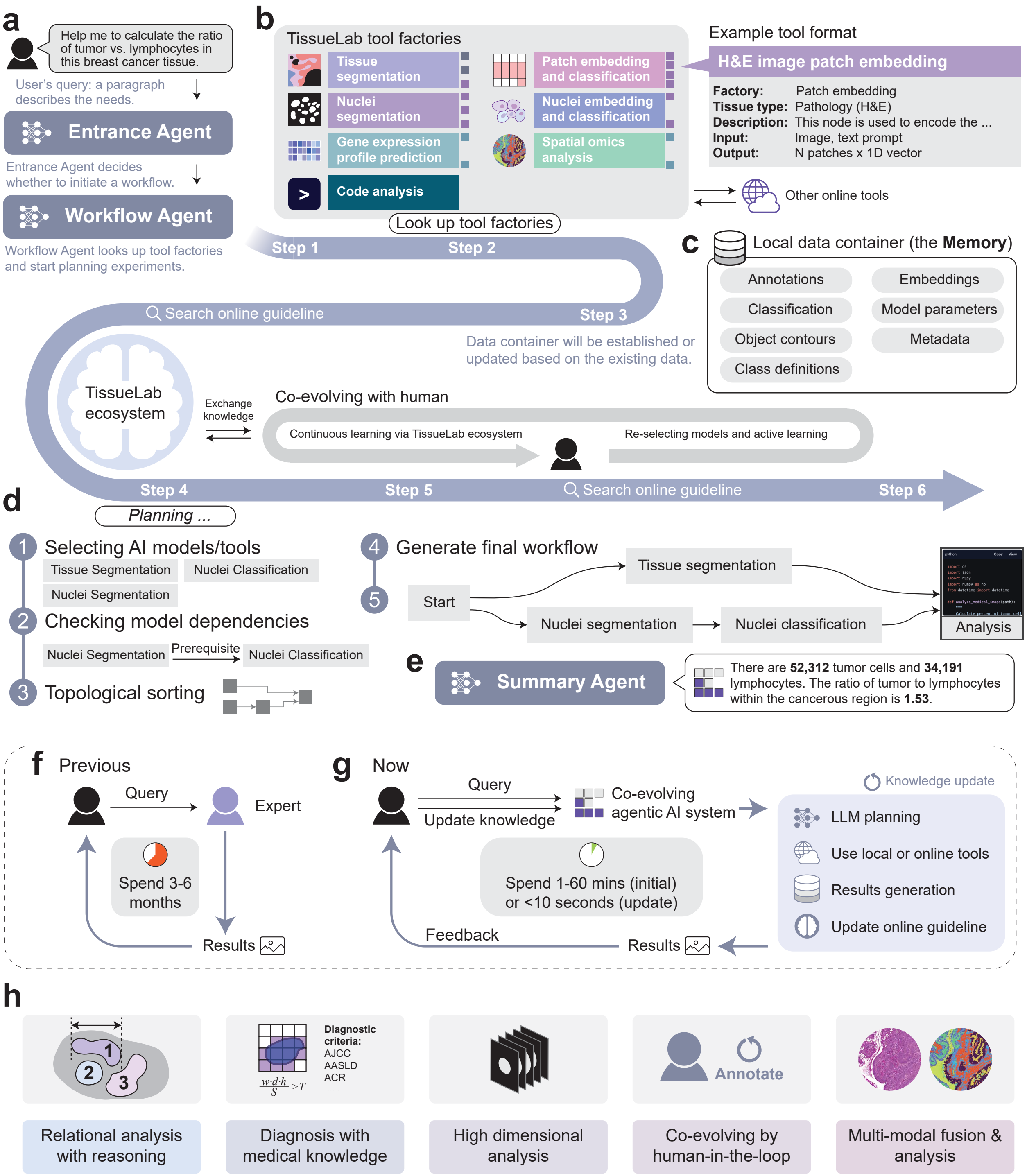
The TissueLab agentic AI system starts with users' query, e.g., "*Calculate the ratio of tumor cells vs. lymphocytes in this breast cancer tissue.*" Based on the user query, our **entrance agent** will ask **workflow agent** to initiate a workflow (**Figure 1a**). The available local tool factories, along with the capability of integrating new tools (**Figure 1b**) without modifying the current codebase, as well as storing the intermediate results into a local data container (**Figure 1c**), offers a practical solution for LLM as an orchestrator to invoke appropriate AI tools and avoid token overload[1]. When selecting AI tools, our agentic AI system will check model denpendencies according to the model card, and perform topolocial sorting to enable distributed inference[2] (**Figure 1d**). Once the workflow has been generated (**Supplementary Figure 1a**), TissueLab will directly analyze the results stored at the local data container, and the **summary agent** will produce a direct answer in natural language (**Figure 1e**).

With TissueLab platform, user can further provide additional feedback through chat box therefore adjust/improve the analysis module (**Supplementary Figure 1b**). During the workflow generation and execution in **Figure 1d**, all intermediate results can be visualized and assessed by human experts through our software platform (**Supplementary Figure 1c**). This interactive component is the key to enable the "coevloving" function, which allows human experts to (i) provide expert annotations (**Supplementary Figure 1c**); (ii) improve AI tools in real-time through iterative feedback (**Supplementary Figure 2**); (iii) adjusting or reselecting appropriate models; and (iv) contribute new knowledge or reuse others knowledge in our TissueLab ecosystem.

Building on top of the fixed foundation models, TissueLab AI system enables users to contribute an-

[1] Token overload is avoided by storing all intermediate model results (e.g., NumPy arrays, CSV files) in an HDF5 file, allowing the analysis code to access and reuse them later.

[2] Ditributed inference allows TissueLab agentic AI system to handle high-throughput requests while efficiently organizing tasks.


a
Help me to calculate the ratio of tumor vs. lymphocytes in this breast cancer tissue.
User's query: a paragraph describes the needs.
Entrance Agent
Entrance Agent decides whether to initiate a workflow.
Workflow Agent
Workflow Agent looks up tool factories and start planning experiments.
b
TissueLab tool factories
Tissue segmentation
Nuclei segmentation
Gene expression profile prediction
Code analysis
Patch embedding and classification
Nuclei embedding and classification
Spatial omics analysis
Example tool format
H&E image patch embedding
Factory: Patch embedding
Tissue type: Pathology (H&E)
Description: This node is used to encode the ...
Input: Image, text prompt
Output: N patches x 1D vector
Other online tools
Look up tool factories
Step 1
Step 2
c
Local data container (the Memory)
Annotations
Embeddings
Classification
Model parameters
Object contours
Metadata
Class definitions
Search online guideline
Step 3
Data container will be established or updated based on the existing data.
TissueLab ecosystem
Exchange knowledge
Co-evolving with human
Continuous learning via TissueLab ecosystem
Re-selecting models and active learning
Step 4
Step 5
Search online guideline
Step 6
d
Planning ...
1 Selecting AI models/tools
Tissue Segmentation
Nuclei Classification
Nuclei Segmentation
2 Checking model dependencies
Nuclei Segmentation
Prerequisite
Nuclei Classification
3 Topological sorting
4 Generate final workflow
5
Start
Tissue segmentation
Nuclei segmentation
Nuclei classification
Analysis
python Copy View
e
Summary Agent
f
Previous
Query
Expert
Spend 3-6 months
Results
g
Now
Query
Update knowledge
Co-evolving agentic AI system
Knowledge update
LLM planning
Use local or online tools
Results generation
Update online guideline
Spend 1-60 mins (initial) or <10 seconds (update)
Feedback
Results
h
Diagnostic criteria:
AJCC
AASLD
ACR
......
$\frac{w \cdot d \cdot h}{S} > T$
Annotate
Relational analysis with reasoning
Diagnosis with medical knowledge
High dimensional analysis
Co-evolving by human-in-the-loop
Multi-modal fusion & analysis

```python
import os
import json
import h5py
import numpy as np
from datetime import datetime

def analyze_medical_image(path):
    """
    Calculate percent of tumor cell
```

There are **52,312** tumor cells and **34,191** lymphocytes. The ratio of tumor to lymphocytes within the cancerous region is **1.53**.

Figure 1: Overview of the TissueLab agentic AI ecosystem for accelerating biomedical research and clinical decision support. **a**. Orchestration phase in which the large language model plans workflows. **b**. Modular plugin architecture enabling seamless integration of new models without maintaining a fixed toolbox. **c**. Editable memory layer that supports co-evolution, allowing clinician feedback to align outputs with expert expectations and mitigate attention dilution. **d**. Execution phase where modular task nodes (e.g., segmentation, classification, spatial analysis) are scheduled via topological sorting, supporting both serial and parallel execution. **e**. Natural language summary agent for user-facing interpretation of results. **f**. Traditional clinical workflow, in which queries often required 3-6 months from initial question to expert-derived results. **g**. TissueLab workflow, where the same queries can be addressed within 1-60 minutes for initial setup and updated in less than 10 seconds after feedback. **h**. Representative applications, including relational analysis with reasoning, diagnosis with medical knowledge, high dimensional analysis, and multi-modal fusion and analysis.

notations to build lightweight classifiers that can be shared within the community and be reused immediately through active learning [18]. Compared with previous and traditional AI research workflow, where clinicians asks expert AI scientist to design and implement a computational pipeline, which is often a time-consuming process (**Figure 1f**), TissueLab can now autonomously generate and execute expert-level workflows within 60 minutes (**Figure 1g**). Moreover, the process can be iteratively refined with expert feedback in just 10 seconds per iteration. By incorporating human feedback, clinicians can actively participate in workflow design rather than remain passive users within the ecosystem, enabling the agentic AI system to continuously improve and become smarter over time.

Through comprehensive experimental evaluations in pathology, radiology, and spatial omics, we demonstrated the advancements of the TissueLab system, as detailed in later sections: (i) performing relational analysis with reasoning, (ii) enabling trustworthy diagnosis informed by medical knowledge, (iii) addressing high-dimensional data challenges, (iv) co-evolving with human experts, and (v) understanding and solving multi-modal analysis tasks (**Figure 1h**).

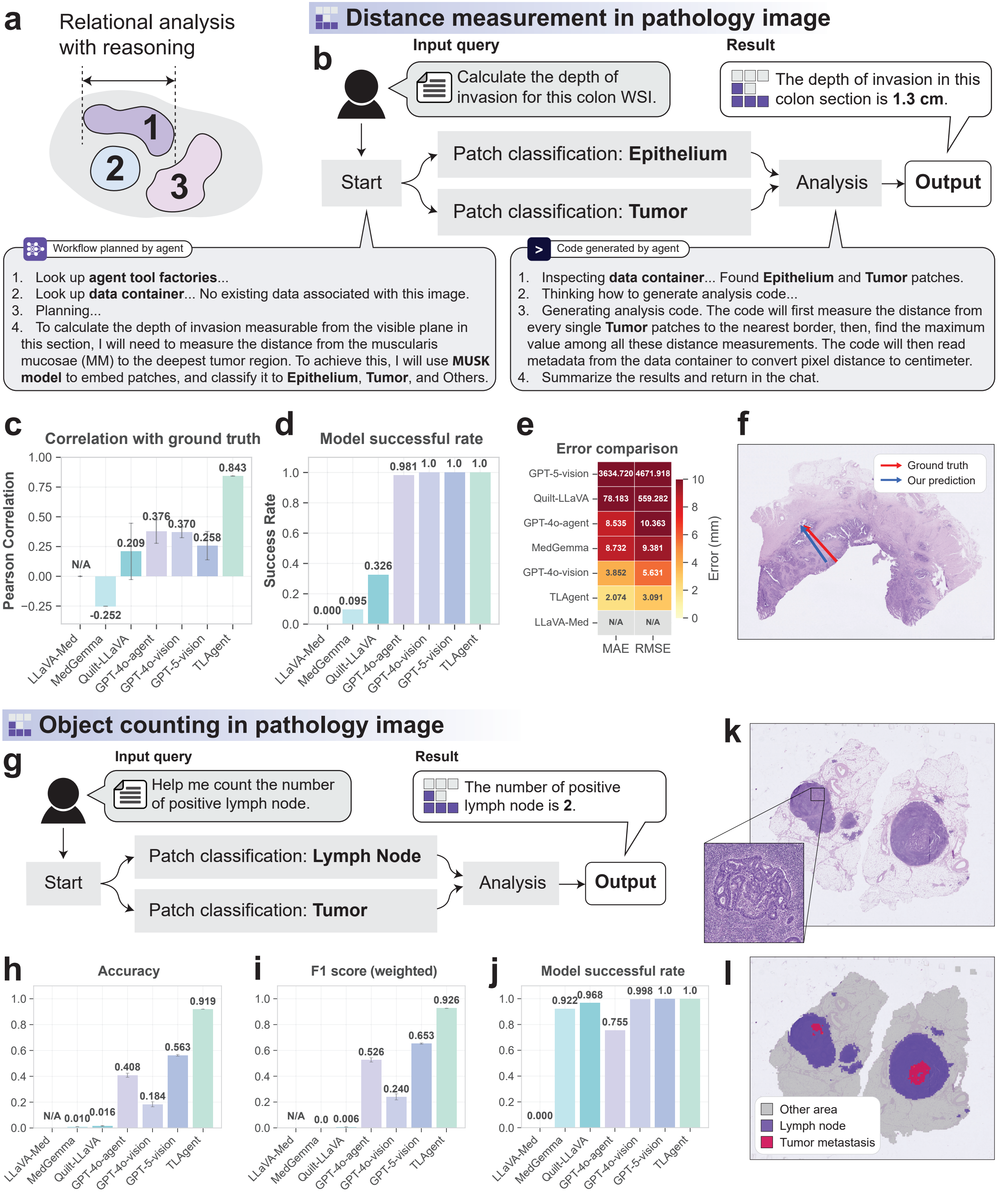
2.2 Co-evolving agentic AI accelerates expert-level assessment of tissue measurement

Many clinically important measurements in pathology remain labor-intensive and lack suitable foundation models that can generalize across disease contexts. Tasks such as quantifying tumor invasion depth typically require expert pathologists to manually delineate tumor boundaries on whole-slide images, a process that is both time-consuming and difficult to scale (**Figure 2a**). No current foundation model is specifically trained to address these fine-grained measurements. Moreover, training a dedicated vision-language model for each fine-grained measurement task would demand substantial data and computational resources, making it impractical in real-world clinical settings. Using TissueLab to automatically generates functional workflows and incorporates clinician annotations in a co-evolving loop (**Supplementary Figure 2**), lightweight modules can be rapidly updated and refined, sometimes within minutes, thereby ensuring that the system continuously aligns with clinical requirements. This adaptive strategy allows TissueLab to transform tasks traditionally dependent on extensive human labor into efficient, reproducible, and expert-level analyses.

We first evaluated TissueLab on predicting tumor invasion depth from whole-slide pathology images. Depth of invasion (DoI) is a critical histopathological feature that informs tumor staging and prognosis, and accurate and interpretable quantification is essential for clinical decision-making. Automatic DoI evaluation enables manual tasks to be completed autonomously and can support a wide range of clinical and research applications, such as Breslow thickness assessment in skin cancer. In this work, we study the DoI in regional lymph node metastasis in colon adenocarcinoma (LNCO2) dataset [19] across 105 primary tumor resections from 49 patients.

As illustrated in **Figure 2b**, TissueLab agent (TLAgent) designs a structured workflow for measuring DoI in colorectal cancer. After the user query, the LLM orchestrator will look up agent tool factories, checking available data outputs, and come up with a workflow by invoking expert AI tools. After the process has finished, TLAgent will then perform automatic code analysis by inspecting data, generating and executing Python source code, and summarize the output in natural language.

As a result, the predicted invasion depths were strongly correlated with pathologist annotated ground truth (Pearson $\rho = 0.843$), demonstrating robust consistency and superior performance across the full range of invasion severities compared to the second best model GPT-4o-agent (Pearson $\rho = 0.376$) (**Figure 2c**). To ensure fair comparisons, we provided large thumbnail inputs for baseline models that could not process high-resolution whole-slide images and repeated each experiment five times. Nevertheless, some baseline models, such as Quilt-LLaVA, still failed under these conditions, either producing outputs unrelated to the query or explicitly returning no valid response. TLAgent consistently delivered the answer with 100% in task completion (success) rate (**Figure 2d**). In terms of numerical accuracy, TLAgent achieved a mean absolute error (MAE) of 2.047 mm and a root mean square error (RMSE) of 3.091 mm relative to expert annotations, reflecting expert-level accuracy (**Figure 2e**). While GPT-4o-vision reported competitive MAE values, its correlations with ground truth were poor (Pearson $\rho = 0.37$), indicating that the predictions lacked directional consistency and often failed to reflect the true progression of invasion depth. For other



Workflow planned by agent

1. Look up **agent tool factories**...
2. Look up **data container**... No existing data associated with this image.
3. Planning...
4. To calculate the depth of invasion measurable from the visible plane in this section, I will need to measure the distance from the muscularis mucosae (MM) to the deepest tumor region. To achieve this, I will use **MUSK model** to embed patches, and classify it to **Epithelium**, **Tumor**, and Others.

Code generated by agent

1. Inspecting **data container**... Found **Epithelium** and **Tumor** patches.
2. Thinking how to generate analysis code...
3. Generating analysis code. The code will first measure the distance from every single **Tumor** patches to the nearest border, then, find the maximum value among all these distance measurements. The code will then read metadata from the data container to convert pixel distance to centimeter.
4. Summarize the results and return in the chat.

Figure 2: Agentic AI performs relational analysis with reasoning. a, Conceptual schematic of relational analysis with reasoning. **b**, Workflow generated by TLAgent for measuring depth of invasion (DoI) in colorectal cancer. **c**, Correlation with ground truth across methods for DoI. **d**, Task-completion (success) rate for DoI among baseline VLM agents versus TLAgent. **e**, Error metrics for DoI (MAE and RMSE, mm). **f**, An example of WSI showing expert annotation and TissueLab prediction overlaid, illustrating explainable alignment. **g**, Workflow generated by TLAgent for positive lymph-node counting in H&E pathology. **h**, Accuracy by method for lymph-node counting. **i**, Weighted F1-score (to account for class imbalance) for lymph-node counting. **j**, Task-completion (success) rate for lymph-node counting. **k**, Original WSI for the lymph-node task, serving as the ground truth reference. **l**, TissueLab segmentation overlaid on the whole slide images, providing explainable correspondence between predicted metastasis regions and expert labels.

approaches, MedGemma and GPT-4o-agent showed higher errors, with MAEs in the range of 8-9 mm; MedGemma also displayed a negative correlation with ground truth (Pearson $\rho = -0.22$). Both Quilt-LLaVA (MAE = 78.183 mm, RMSE $>$ 550 mm) and GPT-5-vision (MAE = 3634 mm, RMSE = 4671 mm) exhibited excessively large errors, indicating that these predictions were highly unreliable, and LLaVA-Med often did not generate valid outputs. Note for all baseline models, experiments were repeated 5 times and the average performance was reported.

By taking a closer inspection on other approaches' result through TissueLab software, we found that these outputs were either random or lacking consistent pattern between predicted and true invasion depth. In contrast, TissueLab autonomously constructed a valid computational workflow: segmenting the relevant tumor tissue using domain-specific models, extracting the contours, and computing invasion depth as the maximum of the shortest distances between each tumor boundary point and the epithelial surface contour as detailed in Algorithm 1 [20]. **Figure 2f** shows that this approach reflects the clinical definition of the measurement and yields results that are not only numerically accurate but also directionally consistent with ground truth. These findings highlight the robustness of TissueLabâĂŹs co-evolving relational reasoning design in capturing clinically relevant spatial relationships.

Algorithm 1 Pseudocode for DOI measurement generated by TissueLab

```
Input: Tumor boundary vertices $C_{\text{tumor}} = \{p_i\}_{i=1}^{N}$, Epithelial contour vertices $C_{\text{epi}} = \{e_j\}_{j=1}^{M}$, Pixel size pixel_size$_{\mu m}$
Initialize: $D_{\max} = 0,\ p^\star = \text{None},\ q^\star = \text{None}$
# Build spatial index (e.g., KD-tree) for $C_{\text{epi}}$
for each $p \in C_{\text{tumor}}$ do
    $q \leftarrow \text{NearestPoint}(C_{\text{epi}}, p)$
    $d \leftarrow \|p - q\|$ # Euclidean distance in pixels
    if $d > D_{\max}$ then
        $D_{\max} \leftarrow d$
        $p^\star \leftarrow p$
        $q^\star \leftarrow q$
    end if
end for
$\text{DOI}_{\mu m} \leftarrow D_{\max} \times \text{pixel\_size}_{\mu m}$
Output: $\text{DOI}_{\mu m},\ p^\star,\ q^\star$
# If epithelium has multiple contours, take $C_{\text{epi}}$ as the union of all vertices
```

We next evaluated the TissueLab agentic AI system on predicting the number of metastatic lymph nodes per slide in the regional lymph node metastasis in colon adenocarcinoma (LNCO2) dataset across 321 adjacent lymph node slides from 22 patients. This task is clinically significant for staging as the number of metastatic lymph nodes directly influences treatment decisions and the level of therapeutic aggressiveness [21–23]. Since many slides do not contain metastatic lymph nodes, the dataset is imbalanced across categories. Therefore, the weighted F1-score is reported alongside overall accuracy.

As illustrated in **Figure 2g**, TLAgent accomplishes this task by first segmenting lymph nodes and tumor regions, then determining nodal positivity through code-based computation of spatial overlap between the two. This structured pipeline underlies the performance gains shown in **Figure 2h–i**. To be specific, TLAgent achieved accuracy = 0.919 and weighted F1 = 0.926, substantially outperforming baseline models Quilt-LLaVA (accuracy = 0.016, F1 = 0.006), MedGemma (accuracy = 0.010, F1 = 0.000), GPT-4o-vision (accuracy = 0.184, F1 = 0.240), GPT-5-vision (accuracy = 0.563, F1 = 0.653), and GPT-4o-agent (accuracy = 0.408, F1 = 0.526). On another note, LLaVA-Med frequently failed to complete the task. Compared to the best-performing baseline (GPT-5-vision), TLAgent improved accuracy by +0.356 absolute points ($\approx$ 63% relative gain) and weighted F1 by +0.273 ($\approx$ 41% relative gain). In addition to its superior performance, TLAgent consistently achieved a 100% task success rate (**Figure 2j**). Note for all baseline models, experiments were repeated 5 times and the average performance was reported.

These comparisons underscore a key distinction between TLAgent and existing agent-based or VLM-based models. Whereas single-model prompting often yields arbitrary or clinically meaningless outputs, TissueLab relies on co-evolving agentic orchestration to assemble domain-appropriate workflows. By integrating segmentation models, contour extraction, and geometric computation into a structured pipeline, the system ensures that its reasoning is clinically grounded and can be validated through visual inspection through the software. This design explains why TLAgent not only achieves lower error metrics but also produces results that align directionally with ground truth, highlighting the necessity of agentic orchestration for clinically meaningful relational reasoning. **Figure 2k–l** further illustrates both the original whole-slide image and intermediate results generated by TLAgent that can be visualized from TissueLab, showing that the system not only correctly counts positive lymph nodes but also identifies tissue types within each region and presents them through an interpretable visualization. Such explainable outputs position TissueLab as a clinically trustworthy assistant, fostering greater safety and confidence in the use of AI by pathologists.

2.3 TissueLab enables guideline-aligned diagnosis using an effectively infinite medical knowledge database

Leveraging online searching for diagnosis aligned with up-to-date clinical guidelines and expansive knowledge database

Through the Model Context Protocol (MCP) [24], TissueLab agents can access authoritative medical resources, such as AJCC [20], CAP [25], WHO [26], or AASLD [27] and other guidelines in real time. By incorporating these continuously updated knowledge sources and leveraging retrieval-augmented generation [28] in diagnostic reasoning, the system ensures that outputs remain aligned with current clinical standards. In contrast to models that rely on static internal knowledge, TissueLab dynamically incorporates external criteria, enabling guideline-consistent decisions for tasks such as metastasis classification or disease staging. In this way, the ecosystem effectively functions as an infinite and continuously updated knowledge base without any maintenance, while remaining anchored to authoritative clinical sources. This approach ensures that diagnostic outputs are not only accurate but also clinically safe, reproducible, and aligned with medical practice [29].

Assessment of lymph node metastasis using online diagnostic criteria

To evaluate this capability, we further evaluated lymph node metastasis classification on the LNCO2 dataset across 321 lymph node slides, a staging-critical task that requires distinguishing macrometastasis ($>$ 2.0 mm), micrometastasis ($\geq$ 0.2 mm), and isolated tumor cells ($<$ 0.2 mm) [20]. These categories determine whether a lymph node is considered positive and thereby directly influence N stage assignment in the AJCC system and are clinically decisive for prognosis and the use of adjuvant therapy in colorectal and other cancers. For consistency across methods, baseline models that could not handle full whole-slide inputs were evaluated using large thumbnail representations, and each experiment was repeated five times. As illustrated in **Figure 3a**, TLAgent retrieved the relevant thresholds directly from these latest authoritative guidelines and applied them consistently to case measurements. Based on such criteria, the system achieved a weighted F1-score of 0.939 and an accuracy of 0.931, closely matching expert-level diagnosis (**Figure 3b–c**). Finally, **Figure 3d** summarizes recall, precision, F1-score, accuracy, and correlation in a radar plot, highlighting that TissueLab agentic AI system outperforms all baselines. Complete numerical results are provided in **Table 1**.

Table 1: Comparison of metastasis classification performance across models. Best performed model results are highlighted with bold font.

Method	Accuracy	F1-score	Precision	Recall	Correlation
TLAgent (ours)	**0.931 ± 0.000**	**0.939 ± 0.000**	**0.951 ± 0.000**	**0.931 ± 0.000**	**0.849 ± 0.000**
GPT-5-vision	0.388 ± 0.012	0.468 ± 0.013	0.802 ± 0.021	0.388 ± 0.012	0.070 ± 0.018
Quilt-LLaVA	0.162 ± 0.023	0.173 ± 0.036	0.780 ± 0.042	0.162 ± 0.023	−0.013 ± 0.012
GPT-4o-vision	0.117 ± 0.006	0.063 ± 0.011	0.783 ± 0.025	0.117 ± 0.006	−0.001 ± 0.001
LLaVA-Med	0.104 ± 0.010	0.036 ± 0.011	0.733 ± 0.178	0.104 ± 0.010	0.015 ± 0.006
GPT-4o-agent	0.096 ± 0.006	0.028 ± 0.010	0.710 ± 0.442	0.096 ± 0.006	0.048 ± 0.004
MedGemma	0.090 ± 0.000	0.033 ± 0.000	0.887 ± 0.000	0.090 ± 0.000	0.125 ± 0.000

Note: We report Cohen's κ as the correlation metric because it measures agreement on categorical labels while correcting for chance. Pearson targets continuous variables; MCC is valid but less commonly interpreted in multi-class contexts.

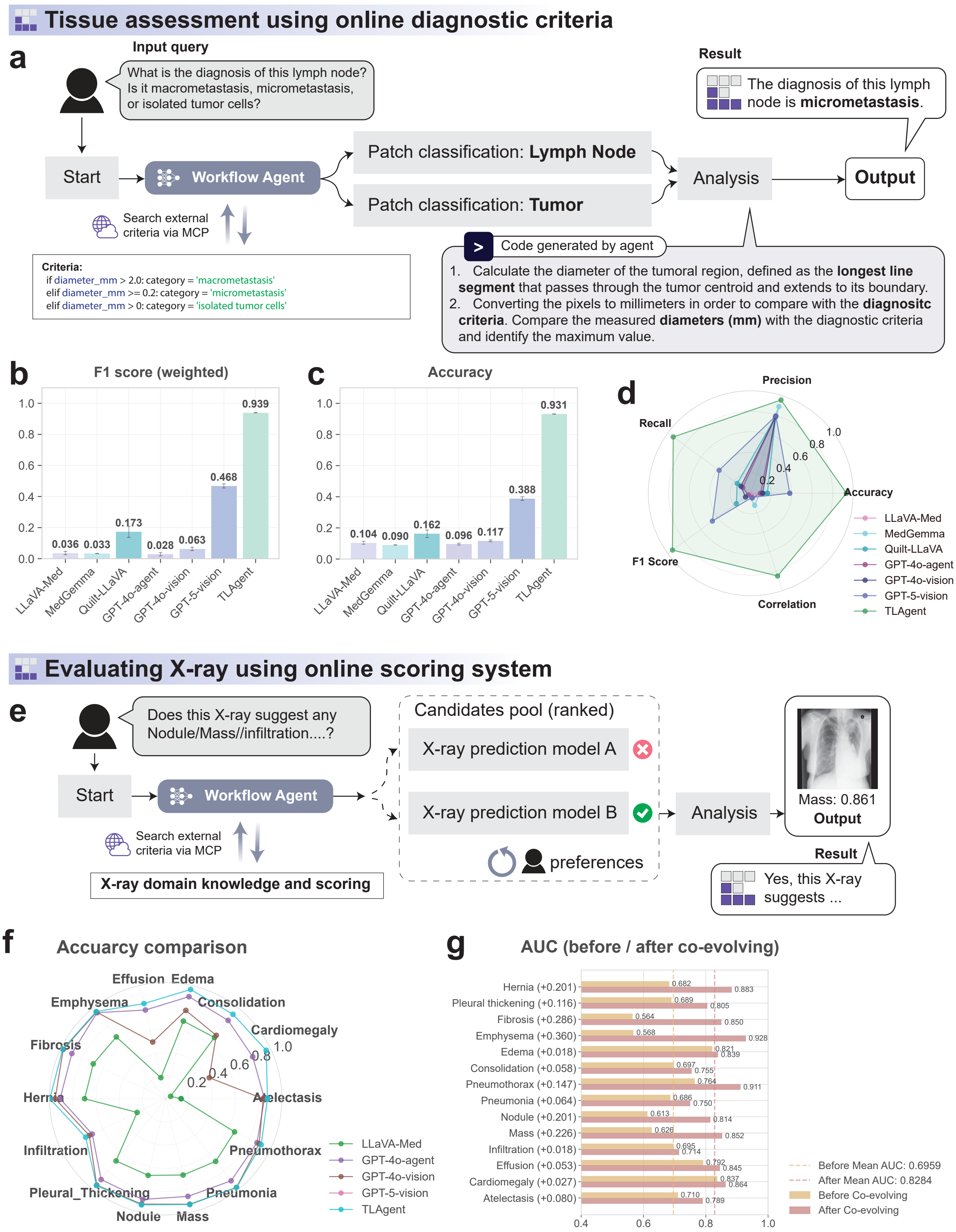
In contrast, baseline VLMs lack access and interpretation to external, up-to-date criteria therefore performed poorly across all metrics. For example, GPT-5-vision reached only moderate performance (F1 = 0.468, Accuracy = 0.388), while other widely used VLMs such as Quilt-LLaVA (F1 = 0.173, Accuracy = 0.162), LLaVA-Med (F1 = 0.036, Accuracy = 0.104), MedGemma (F1 = 0.033, Accuracy = 0.090), GPT-4o-agent (F1 = 0.028, Accuracy = 0.096), and GPT-4o-vision (F1 = 0.063, Accuracy = 0.117). From a clinical standpoint, such misclassifications are critical. These results demonstrate that TissueLab, by leveraging online search restricted to authoritative medical guidelines, produces clinically valid, guideline-aligned diagnostic outputs, instead of unguided single-model output.

Assessment of chest X-ray diagnosis using online guideline and human feedback

Besides pathology image, TLAgent can also perform accurate and reliable chest X-ray diagnosis through online guideline retrieval and human feedback. Chest X-ray diagnosis task is clinically important but faces unique challenges: latest vision-language medical AI models are rapidly evolving, often surpassing older ones, yet their performance varies substantially across datasets and disease categories. To handle this variability, TissueLab incorporates a co-evolving mechanism that integrates clinician-in-the-loop oversight. As illustrated in **Figure 3e**, the system first generates a candidate pool of models for each query and selects the one with the best score. When clinicians provide negative feedback, TLAgent reselects from the same pool and updates model scores, allowing future rankings to better reflect clinical needs. In this way, the system continuously improves its clinical intelligence while ensuring interpretability.

We evaluated TLAgent on chest X-ray classification using the NIH Chest X-ray Dataset [30], which comprises 112,120 frontal-view radiographs from 30,805 unique patients. Disease labels were derived by text-mining the original radiology reports using natural language processing.

While the dataset suitable for our experiment, accuracy alone can be misleading given severe class imbalance. Many thoracic conditions have low prevalence, so baseline VLMs appear to achieve high accuracy simply by defaulting to the negative class, without capturing true disease signals. As shown in the per-disease accuracy summary (**Table 2**) and the accuracy radar plot (**Figure 3f**), this will inflate overall accuracy but fail to capture clinically meaningful predictions.



What is the diagnosis of this lymph node? Is it macrometastasis, micrometastasis, or isolated tumor cells?

The diagnosis of this lymph node is **micrometastasis**.

```
Criteria:
  if diameter_mm > 2.0: category = 'macrometastasis'
  elif diameter_mm >= 0.2: category = 'micrometastasis'
  elif diameter_mm > 0: category = 'isolated tumor cells'
```

1. Calculate the diameter of the tumoral region, defined as the **longest line segment** that passes through the tumor centroid and extends to its boundary.
2. Converting the pixels to millimeters in order to compare with the **diagnositc criteria**. Compare the measured **diameters (mm)** with the diagnostic criteria and identify the maximum value.

Does this X-ray suggest any Nodule/Mass//infiltration....?

Yes, this X-ray suggests ...

Figure 3: Guideline-aligned diagnosis using agentic AI. **a**, Workflow for lymph-node metastasis classification in pathology. **b**, Weighted F1-score comparison across models. **c**, Accuracy comparison across models. **d**, Radar plot summarizing recall, precision, F1-score, accuracy, and correlation for metastasis classification. **e**, Workflow for chest X-ray diagnosis by clinicians-in-the-loop co-evolution. **f**, Per-disease accuracy radar across 14 conditions for each model. **g**, AUC before and after co-evolution with clinician feedback, showing systematic improvement.

Table 2: Comparison of disease classification performance across different agentic AI systems. Given the large size of the evaluation dataset, each experiment was conducted once, and mean or standard deviation values were not reported. Best performed model results are highlighted with bold font.

Disease	LLaVA-Med	GPT-4o-agent	GPT-4o-vision	TLAgent (ours)
Atelectasis	0.133	0.850	0.839	**0.876**
Cardiomegaly	0.047	0.831	0.416	**0.960**
Consolidation	0.674	0.862	0.697	**0.929**
Edema	0.685	0.901	0.779	**0.964**
Effusion	0.186	0.783	0.500	**0.840**
Emphysema	0.681	0.947	0.954	**0.959**
Fibrosis	0.691	0.896	0.978	**0.983**
Hernia	0.697	0.909	0.951	**0.997**
Infiltration	0.272	0.701	0.729	**0.764**
Mass	0.674	0.862	0.932	**0.934**
Nodule	0.677	0.889	0.931	**0.937**
Pleural Thickening	0.685	0.890	0.951	**0.955**
Pneumonia	0.690	0.903	0.977	**0.978**
Pneumothorax	0.658	0.872	0.893	**0.912**

For this reason, the area under the ROC curve (AUC) provides a more reliable evaluation metric (**Figure 3g**). Unlike other VLM baselines which return categorical outputs without calibrated scores, our agentic AI framework generates step-wise prediction scores through adaptive orchestration (candidate pooling, best-score selection, and clinician-gated reselection[3] with preference updates). Such calibrated scores provide clinicians with interpretable confidence levels that can be directly referenced in decision-making, improving diagnostic safety and transparency. In turn, clinicians use these scores to provide feedback that guides the system's preference updates. This bidirectional loop establishes a genuine form of co-evolution: clinicians benefit from interpretable outputs, while their feedback continuously improves the agent's intelligence. With this design, TLAgent raised the mean AUC from 0.6959 [31] to 0.8284 [32] across 14 thoracic conditions, leveraging a candidate pool mechanism to dynamically adapt and evolve with clinician's feedback. Other baseline models do not produce well-defined confidence scores, and therefore AUC cannot be computed.

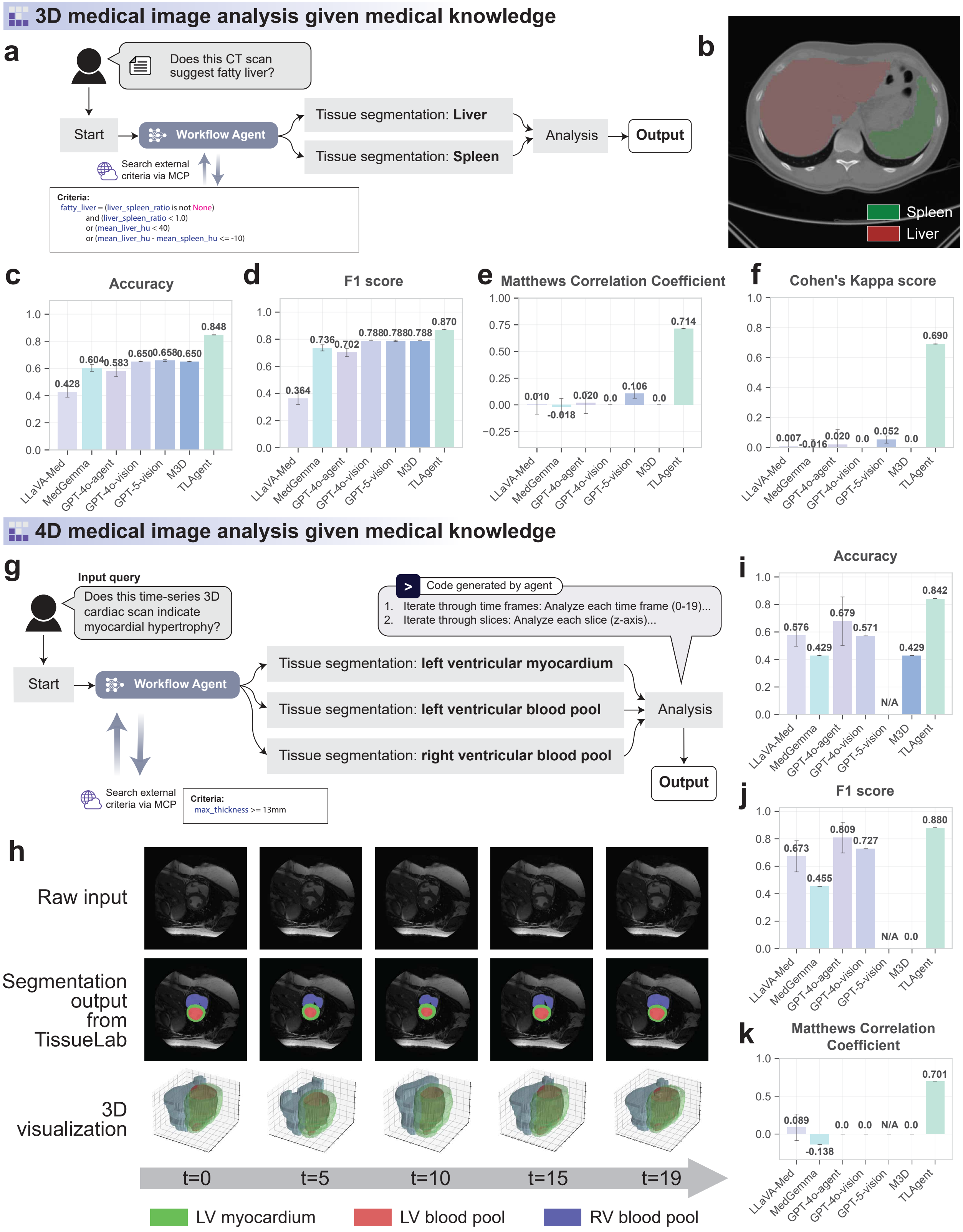
2.4 TissueLab enables high-dimensional radiology image analysis

TLAgent predict fatty liver in 3D chest CT

The TissueLab platform and TLAgent operate not only on 2D medical images but also on high-dimensional radiological data, such as volumetric and time-resolved imaging like CT and MRI scans. One example is fatty liver in 3D chet CT. Fatty liver disease is a prevalent and clinically important condition associated with metabolic syndrome, cardiovascular risk, and long-term liver outcomes. In radiology practice, non-invasive diagnosis is based on guideline-defined Hounsfield unit (HU) values [27]. Here we evaluated TLAgent on the fatty liver diagnosis across 38 chest CT patients from held-out test set of the UNIFESP Chest CT Fatty Liver Competition Dataset [33]. As shown in **Figure 4a**, TLAgent autonomously retrieved the diagnostic criteria from AASLD and constructed a workflow that first segmented the liver and spleen volumes (**Figure 4b**). HU values were then computed from the segmented regions and evaluated against the guideline thresholds to determine the presence or absence of steatosis.

TLAgent demonstrated consistently high performance, with accuracy of 0.848 and F1-score of 0.870,

[3] Effective clinician-gated reselection requires an online, open access ecosystem that can both retain user preferences and enable community-wide sharing. Achieving this was nontrivial, and at present this capability is uniquely implemented by TissueLab platform (tissuelab.org).



```
Criteria:
fatty_liver = (liver_spleen_ratio is not None)
        and (liver_spleen_ratio < 1.0)
        or (mean_liver_hu < 40)
        or (mean_liver_hu - mean_spleen_hu <= -10)
```

```
1.  Iterate through time frames: Analyze each time frame (0-19)...
2.  Iterate through slices: Analyze each slice (z-axis)...
```

```
Criteria:
max_thickness >= 13mm
```

Figure 4: High-dimensional radiological analysis and diagnosis with medical knowledge. a, Workflow diagram for fatty liver diagnosis. **b**, Representative CT slice showing intermediate segmentation data during the TissueLab workflow. **c**, Accuracy comparison across methods for fatty liver diagnosis. **d**, Binary F1-score comparison for fatty liver diagnosis. **e**, Matthews correlation coefficient for fatty liver diagnosis. **f**, CohenâĂŹs kappa score for fatty liver diagnosis. **g**, Workflow diagram for myocardial hypertrophy assessment. **h**, Representative sample showing intermediate segmentation outputs from time-series 3D cardiac image. **i**, Accuracy comparison of hypertrophy diagnosis. **j**, Binary F1-score for hypertrophy diagnosis. **k**, Correlation between predictions and ground truth.

supported by strong agreement-based metrics (MCC[4] = 0.714, Cohen's kappa = 0.690) (**Figure 4c–f**). In contrast, most baseline models approached fatty liver diagnosis through either 2D slice-based recognition or simplistic code-based analysis, rather than leveraging true 3D medical imaging as TLAgent did, with the exception of the latest M3D agentic AI model. As a result, they showed weaker and often misleading performance. To ensure fair comparison, we manually provided 2D slices as inputs to these baselines and repeated the entire experiment five times. Accuracy values were 0.428 (LLaVA-Med), 0.604 (MedGemma), 0.583 (GPT-4o-agent), 0.650 (GPT-4o-vision), 0.658 (GPT-5-vision), and 0.650 (M3D). While some baselines produced comparable F1-scores (up to 0.788 for GPT-5-vision), their agreement-based metrics were nearly zero (**Figure 4e–f**). In particular, M3D and GPT-4o-vision defaulted to always predicting the positive class, which inflated recall and F1-scores but yielded no discriminative power (see **Supplementary Figure 3** for confusion matrices). These results highlight the advantage of agentic orchestration in the high-dimensional imaging tasks where these baselines may hallucinate plausible-seeming outputs that fail to align with clinical reliability.

TLAgent improves detection of intracranial hemorrhage from 3D CT scans

In addition, TLAgent can also perform 3D brain MRI analysis. Here we demostrate the use case of identifying Intracranial hemorrhage (ICH). ICH is a life-critical neurological emergency in which rapid and accurate diagnosis is essential, as delays in detection can directly affect patient outcomes. However, reliable detection requires experts to examine hundreds of slices per scan, making the process time-consuming and prone to oversight. Automated systems capable of accurate and interpretable hemorrhage detection therefore have high clinical value. For benchmarking, we used the PhysioNet ICH dataset [34–36], which provides 82 3D CT scans with voxel-level annotations, enabling quantitative evaluation of hemorrhage segmentation and classification.

We ensured fair comparisons by evaluating baseline models on appropriate inputs and repeated each experiment five times. As illustrated in **Supplementary Figure 4a**, TLAgent searched authoritative diagnostic criteria for ICH and assembled a workflow that applied TotalSegmentator for volumetric hemorrhage segmentation and anatomical brain structure delineation, followed by code-based spatial overlap analysis to determine the presence of bleeding. This decomposition reflects the typical radiological reasoning process, beginning with hemorrhage localization and subsequently contextualizing the findings within neuroanatomical structures. Representative outputs of volumetric segmentation and slice-wise overlays are shown in **Supplementary Figure 4b** through TissueLab platform, providing transparent evidence for each diagnostic decision.

As quantified in **Supplementary Figure 4c–d**, TLAgent achieved an accuracy of 78.7% with an F1-score of 0.750, closely matching expert-level performance. In contrast, general-purpose vision-language models struggled to provide clinically reliable outputs. LLaVA-Med effectively failed (F1 = 0.000) despite superficial accuracy of 52%, while MedGemma (Acc = 45.3%, F1 = 0.226) and GPT-4o-vision (Acc = 57.3%, F1 = 0.273) produced unstable and anatomically implausible predictions across slices. GPT-4o-agent performed somewhat better (Acc = 51.0%, F1 = 0.645) but still lagged behind TLAgent, reflecting frequent mislocalizations. GPT-5-vision and M3D were unable to complete the task. These comparisons emphasize that TLAgent was the only system able to combine high numerical accuracy with clinically coherent predictions.

For the more fine-grained sub-question of intraventricular hemorrhage detection in **Supplementary**

[4] MCC: Matthews correlation coefficient.

Figure 4e–f, TLAgent maintained meaningful discriminative ability, achieving an F1-score of 0.500 and an accuracy of 0.947. However, LLaVA-Med and GPT-4o-vision both failed completely (F1 = 0.000), while MedGemma (F1 = 0.091, Acc = 0.733) and GPT-4o-agent (F1 = 0.084, Acc = 0.441) produced unstable or near-random predictions. GPT-5-vision and M3D were unable to provide evaluable outputs. Notably, although GPT-4o-vision appeared competitive in accuracy (0.896), its zero F1-score indicated predictions that were inconsistent and clinically meaningless.

TLAgent enables dynamic detection of myocardial hypertrophy from 4D cardiac imaging

Lastly, in addition to 3D image, TLAgent can also handle 4D imaging data, where volumetric data are captured across sequential time points. For example, myocardial hypertrophy is a clinically critical condition associated with heart failure, arrhythmia, and sudden cardiac death. Accurate diagnosis depends on precise measurement of left ventricular wall thickness across spatial slices and temporal frames of the cardiac cycle. In routine practice, this requires experts to manually delineate endocardial and epicardial boundaries frame by frame, a process that is prohibitively time-consuming and unsuitable for large-scale or real-time deployment. For benchmarking, we used the Sunnybrook Cardiac MRI dataset [37], which consist of 45 cine-MRI images from a mixed of patients and pathologies: healthy, hypertrophy, heart failure with infarction and heart failure without infarction.

In our experiment, TLAgent searched the authoritative clinical definition of hypertrophy, specified as a maximal left ventricular wall thickness $\geq$ 15 mm at end-diastole in the absence of secondary causes [38, 39]. As illustrated in **Figure 4g**, the system assembled a workflow that segmented the left ventricular myocardium (LVMYO), left ventricular blood pool (LVBP), and right ventricular blood pool (RVBP). From these segmentations, endocardial contours were derived from the LVBP boundary, epicardial contours from the outer boundary of LVBP and LVMYO regions, and RV insertion points from RVBP-LVMYO junctions. Slice-wise wall thickness was then computed in 2D by measuring the distance from each endocardial point to its intersection with the epicardial contour along the outward normal direction. These measurements were aggregated across frames to reconstruct dynamic wall-thickness curves. Representative outputs are shown in **Figure 4h**, which display 3D segmentations across multiple time frames together with short-axis 2D overlays on raw cine MRI slices, highlighting the practical utility of TissueLab by providing clinicians with transparent and directly interpretable visual results. The full computational procedure for wall-thickness measurement generated by TLAgent is summarized in Algorithm 2.

With TLAgent, the automatic design and implementation of this workflow reached an accuracy of 84.2%, F1 = 0.880, and a strong correlation with ground truth thickness (MCC = 0.701). In contrast, baseline models lack direct 3D data support. To ensure fair comparisons, we manually slice volumetric data into 2D images, specifying the temporal assignment of each slice, and repeating each experiment five times to ensure consistency. The best-performing baseline, GPT-4o-vision, achieved only 57.1% accuracy and F1 = 0.727, with near-zero correlation to the ground truth, indicating predictions that failed to capture clinically defined patterns. GPT-4o-agent reached F1 = 0.809 but collapsed to predicting nearly all cases as positive, eliminating specificity. LLaVA-MedâĂŹs performance was essentially random (accuracy = 57.6%, F1 = 67.3%, MCC = 0.089), and MedGemma also showed limited performance (accuracy = 42.9%, F1 = 0.455). These comparisons emphasize that only TLAgent was exhibited high accuracy with non-trivial recall, sustaining robust performance in a challenging sub-task where vision-language baselines degraded to chance-level behavior.

2.5 TissueLab interactive ecosystem enables customized tool development

Among above experiments, it is worth noting that TLAgent performance is bottlenecked by underlying tools, such as segmentation model, rather than in the agent orchestration strategy itself. Although the system can seamlessly integrate improved segmentation tools in the future to further boost accuracy while preserving interpretability and guideline alignment, in reality there are limited tools available for medical imaging analysis, and many research and clinical questions fall into edge cases for which no existing tool

Algorithm 2 Pseudocode for LV hypertrophy detection generated by TLAgent

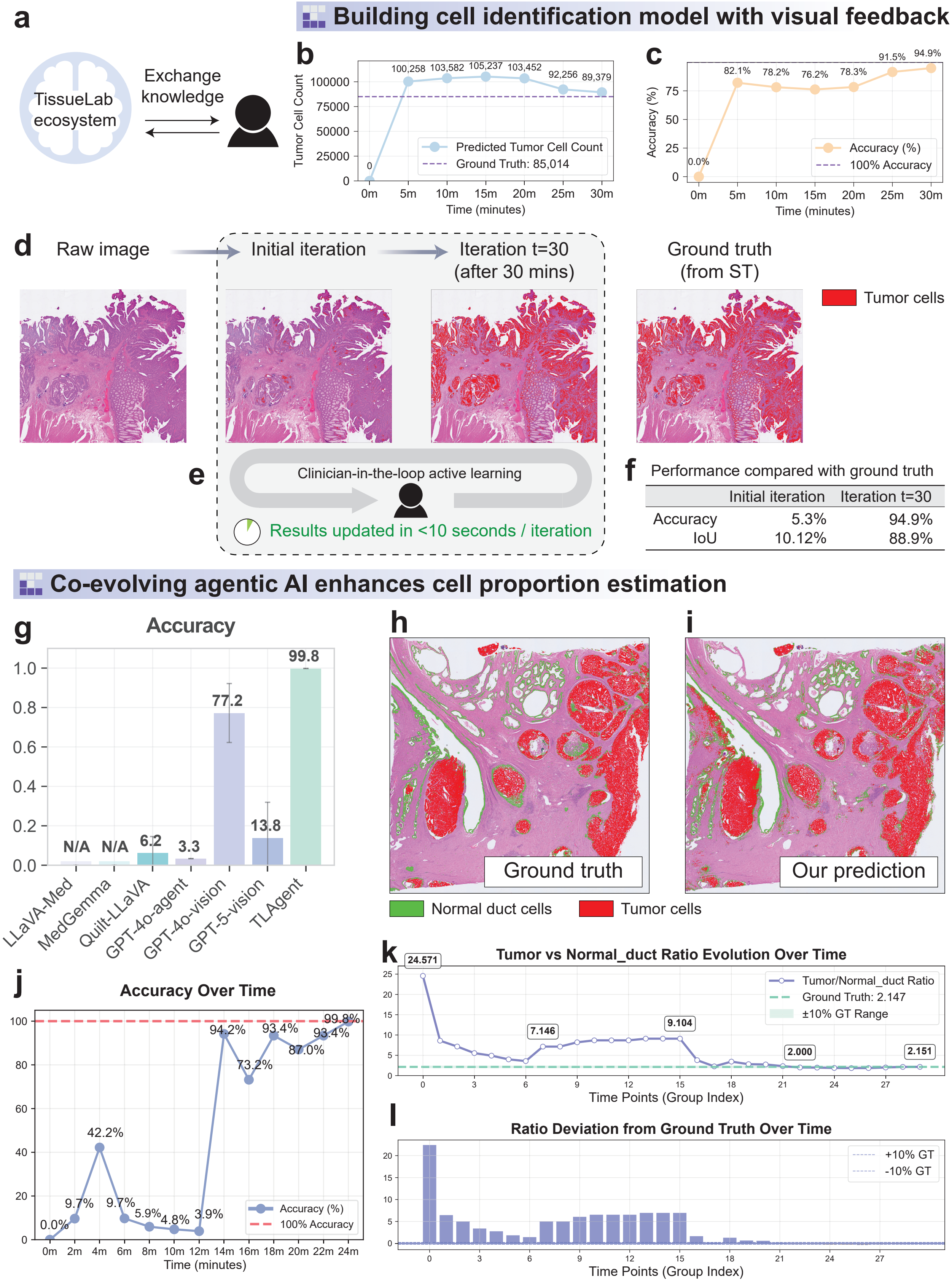
```
**Input:** LV blood pool mask $M_{\text{LVBP}}$, LV myocardium mask $M_{\text{LVMYO}}$, RV blood pool mask $M_{\text{RVBP}}$, Pixel spacing $(\Delta x, \Delta y)$, Threshold $\tau = 15$ mm
**Initialize:** $WT^{\text{ED}}_{\max} = 0$
Identify ED frame $t_{\text{ED}}$ # Frame with maximal LVBP area
for each slice $z$ in $t_{\text{ED}}$ do
    $C_{\text{endo}} \leftarrow \partial(M_{\text{LVBP}}[t_{\text{ED}}, z])$ # Endocardial contour
    $C_{\text{epi}} \leftarrow \partial(M_{\text{LVBP}}[t_{\text{ED}}, z] \cup M_{\text{LVMYO}}[t_{\text{ED}}, z])$ # Epicardial contour
    for each $p \in C_{\text{endo}}$ do
        $\mathbf{n}(p) \leftarrow$ outward normal at $p$
        $q \leftarrow$ first intersection of ray $(p, \mathbf{n}(p))$ with $C_{\text{epi}}$
        $d \leftarrow \|p - q\| \times \text{PixelScale}(\Delta x, \Delta y)$ # Local thickness (mm)
        if $d > WT^{\text{ED}}_{\max}$ then
            $WT^{\text{ED}}_{\max} \leftarrow d$
        end if
    end for
end for
if $WT^{\text{ED}}_{\max} \geq \tau$ then
    Hypertrophy $\leftarrow$ Yes # Guideline-aligned diagnosis
else
    Hypertrophy $\leftarrow$ No
end if
**Output:** Hypertrophy label, $WT^{\text{ED}}_{\max}$
```

can provide solutions.

To address this limitation, TLAgent is deeply integrated into our TissueLab software and ecosystem (tissuelab.org). The TissueLab platform enables users to create customized tools by (i) developing Python pipelines that can be uploaded to the ecosystem; and (ii) fine-tuning existing foundation models for diverse downstream tasks. Once built, these models can be shared through the TissueLab ecosystem, fostering collaboration and reuse. This capability allows the TissueLab agentic AI system to substantially accelerate medical imaging research and scientific discovery (**Figure 5a**). We evaluated this capability of cell identification on two different cancer pathology slides (Colon and Prostate) from the Visium HD dataset [40, 41], which includes high-resolution spatial transcriptomics to serve as ground-truth labels.

In this experiment, the objective was to determine how many cells are tumor. Despite our best efforts to ensure fair comparisons, baseline models consistently failed to produce valid outputs for this task, as they may not have had access to the same high-resolution data as TLAgent. However, guided by clinician's active learning annotation, TissueLab was able to evolve from unseen cases toward refined clinical alignment. By allowing pathologist to finetune our in-house H&E cell classification foundation model "NuClass", the system rapidly improved tumor cell identification performance from near-zero accuracy to 82.1% within five minutes of incremental training, and ultimately reached 94.9% accuracy after 30 minutes (**Figure 5b–c**). At the same time, **Figure 5d–e** show that predictions not only matched overall counts but also reproduced the spatial distribution of neoplastic versus non-neoplastic cells, closely aligning with ground truth labels. The intersection-over-union (IoU) between predicted and ground-truth tumor regions improved dramatically from 10.1% at the initial iteration to 88.9% in 30 minutes (**Figure 5f**), highlighting TissueLab's capability to enable users to efficiently build a cell-counting model from scratch.

We next evaluated the ratio of tumor cells to normal duct cells in prostate tissue, a measure with clinical relevance for quantifying malignant transformation, informing prognosis, and supporting cancer staging. By repeating this experiment five times, GPT-4-vision reached 77.2% accuracy but with excessively wide prediction distributions, reflecting random guesses rather than consistent reasoning; and all other baselines fell below 12% accuracy or failed outright (**Figure 5g**). In contrast, the TissueLab platform enables clinicians to observe workflow results in real time, track how models learn from their annotations, and co-evolve



Performance compared with ground truth

	Initial iteration	Iteration t=30
Accuracy	5.3%	94.9%
IoU	10.12%	88.9%

Figure 5: Clinician-in-the-loop co-evolution improves tissue and cell quantification in cancer research. **a**, Schematic illustration of the clinician-in-the-loop mechanism within the TissueLab ecosystem. **b**, Tumor cell counts over time compared with ground truth (GT). **c**, Tumor cell detection accuracy as a function of feedback iterations. **d**, Representative raw image. **e**, Image series showing progressive improvement in tumor cell detection after iterative feedback. **f**, Performance of detection compared with GT. **g**, Comparison of model accuracy in quantifying tumor-to-normal duct cell ratios across methods. **h**, Cell overlay from ground truth. **i**, Cell Overlay of TissueLab predictions. **j**, Accuracy of tumor-to-normal duct cell ratio estimation over time. **k**, Evolution of predicted tumor-to-normal duct cell ratios across feedback iterations. **l**, Deviation from GT over time.

with the system. Through this process, lightweight classifiers are trained on top of our in-house "NuClass" model and shared within the ecosystem. By applying the clinician-guided classifiers, TLAgent achieved 99.8% accuracy, and as shown in **Figure 5h–i**, the predicted tumor and duct distributions closely matched the ground truth spatial patterns.

Additionally, improvements in accuracy, cell ratio, and ratio deviation over time are presented in **Figure 5j–l**, respectively. Because expert annotations involved labeling multiple cell types, the ratio initially showed little improvement. However, once the relevant cells were incorporated, TissueLab rapidly adapted: accuracy rose from 3.9% to above 90% within two minutes of feedback, and continued to improve toward near-perfect performance, ultimately reaching 99.8%. At the same time, the predicted ratios evolved smoothly toward the ground truth with minimal deviation across iterations.

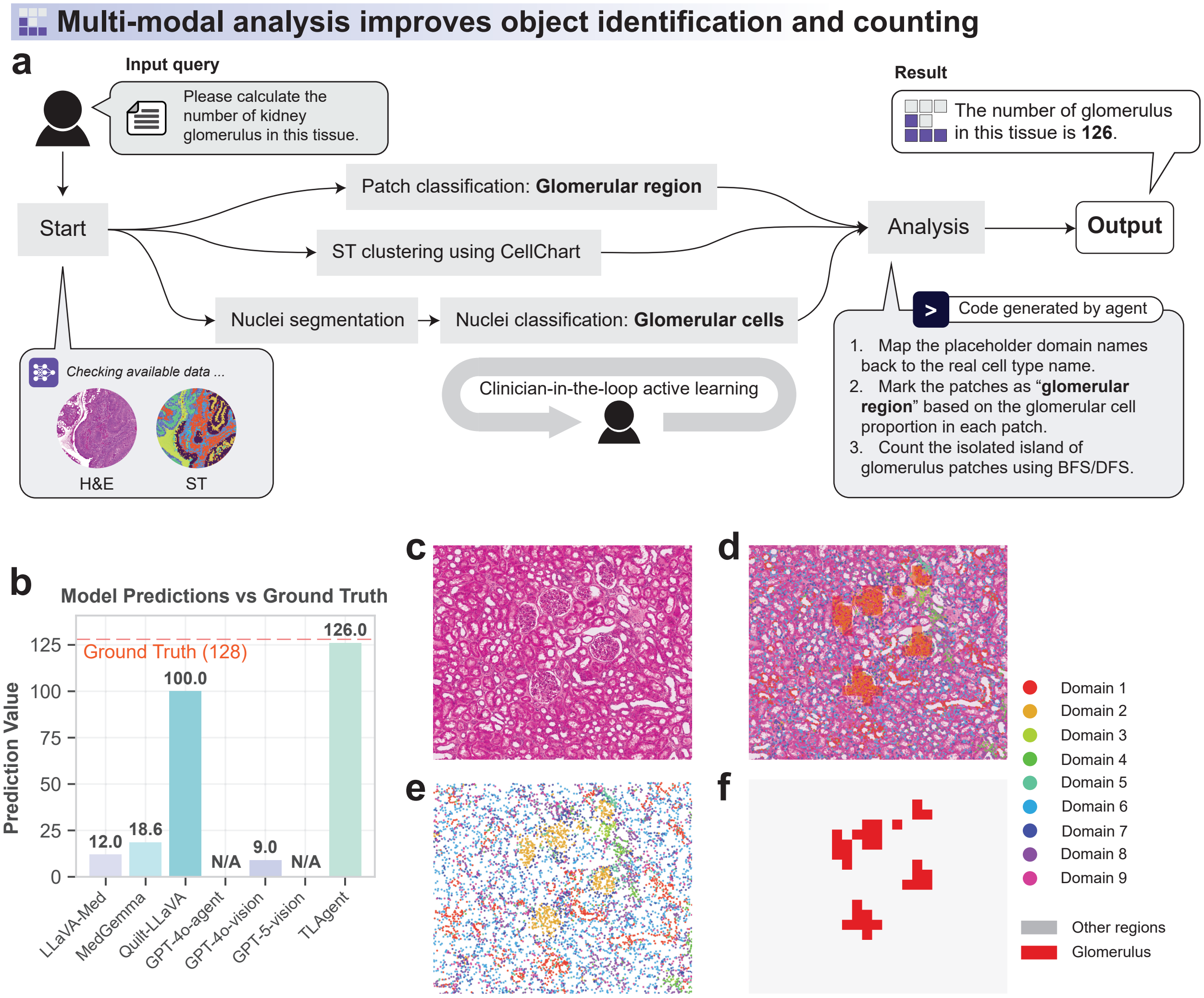
These findings confirm that co-evolving with human-in-the-loop component through TissueLab platform enables user to iteratively refine and customize existing models to achieve expert-level performances. With human expert supervision, the system uses active learning to adapt from novel cases to near-perfect concordance within minutes, offering immediate research and clinical utility. Such capability requires an interactive platform that enables timely, bidirectional feedback between human experts and the AI system, a capability not attainable with current VLM-based models or other agent systems.

2.6 TissueLab enables multi-modal integration of spatial omics and histology to improve accuracy in pathology analysis

Beyond achieving superior performances over baseline models on both pathology and radiology tasks, the TissueLab platform and TLAgent also demonstrate the ability to perform multi-modal integrative analysis. Here we demostrate a use case where TLAgent can recognize the presense of both spatial omics and histopathology modalities thereby generating more precise and clinically meaningful tissue analysis workflow [42]. Specifically, we evaluate this capability in the context of glomerulus quantification on a kidney whole slide image from the Visium HD dataset, which provides high-resolution spatial transcriptomics data. As a common quantification task, accurate glomerulus counting is critical for nephrology research and clinical care, particularly for advanced measures such as the fraction of globally sclerotic glomeruli. However, these assessments have traditionally relied on labor-intensive manual evaluation by expert pathologists.

TLAgent accomplished this task through a structured multi-modal workflow (**Figure 6a**). First, the system will check available data in the data container. If both the H&E and spatial transcriptomics data are available, it will performed cell segmentation on H&E images and extracted the centroids of individual cells. These spatial coordinates were then linked to gene expression profiles, and unsupervised clustering was performed to assign each cell to a transcriptomic domain. Because clustering yields only numerical domain labels without inherent biological meaning, TLAgent established multi-modal correspondence by mapping these domains back to annotated H&E regions. Through this alignment, the cluster corresponding to glomerular cells was identified. Once identified, TLAgent computed the local proportions of glomerular versus non-glomerular cells, aggregated glomerulus patches, and applied graph-based traversal (depth-first and breadth-first search, DFS/BFS) to obtain discrete glomerulus counts. This end-to-end workflow autonomously generated by TLAgent, highlights both the reasoning capabilities of LLMs and the local data access features embedded in our design.

With an appropriate domain configuration, TLAgent achieved 98.4% accuracy, nearly indistinguishable



Please calculate the number of kidney glomerulus in this tissue.

The number of glomerulus in this tissue is **126**.

1. Map the placeholder domain names back to the real cell type name.
2. Mark the patches as "**glomerular region**" based on the glomerular cell proportion in each patch.
3. Count the isolated island of glomerulus patches using BFS/DFS.

Figure 6: Multi-modal analysis integrating spatial omics and histopathology for glomerulus quantification. **a**, Schematic diagram of the TissueLab multi-modal analysis workflow. **b**, Comparison of model-predicted versus ground-truth (GT) glomerulus counts. **c**, Representative kidney H&E image. **d**, Patch-level and cell-level result overlays. **e**, Spatial domain clustering result obtained by integrating spatial omics data with histology. **f**, Glomerulus region prediction result.

from expert annotations (**Figure 6b**). In contrast (**Figure 6b**), although we have made every effort to ensure fair comparisons with other baseline models and systems, baseline vision-language models were still unable to cope with the scale and complexity of spatial omics data, and consistently failed to complete the task, producing outputs that were either random or absent. Through TissueLab platform's visualization interface (**Figure 6c–f**), clinicians can directly inspect overlays of predictions with histological images, confirming that the identified glomerulus regions and counts are highly consistent with expert assessments.

These results demonstrate that TissueLab can still deliver expert-level performance in a setting where multi-modal data are available. By integrating molecular and morphological information, the system not only provides richer insights into kidney pathology but also establishes a scalable framework for multi-modal medical AI analysis.

3 Discussion

We bring the concept of "**laboratory intelligence**", devoting extensive engineering and design efforts in building TissueLab ecosystem and TLAgent and making them openly accessible to the community. TissueLab is fully open source, with installers for Windows, macOS, and Linux, and a freely accessible web portal (tissuelab.org). As one of the first agentic AI systems aimed at translating advances in medical imaging AI research into tools readily applicable to both day-to-day research and clinical practice, the TissueLab ecosystem and TLAgent offer a modular and extensible framework that enables clinicians, physician-scientists, lab technicians, and AI researchers to directly integrate state-of-the-art AI methods into their workflows. In doing so, TissueLab bridges the gap between algorithmic innovation and clinical/translational adoption, functioning as a programmable laboratory for medical imaging with modular, reconfigurable infrastructures that lower entry barriers and accelerate scientific discovery.

The strong performance of TissueLab is largely attributable to its unique ability to co-evolve with clinicians. The system allows user to insert a new or finetune an existing model, thereby rapidly adapt to new diseases with only minutes of annotation, an important feature that is critical for many research and clinical settings where AI can benefit largely on those difficult diesases that are rare to collect training data. Moreover, by establishing a shared ecosystem, TissueLab enables collaborative knowledge transfer across diverse tasks and base models. In this way, the system evolves from unseen cases to align with clinical demands, ensuring that each new interaction enhances its overall clinical intelligence.

Compared with prior foundation models and some other vision-language-based agentic AI systems, which often lack interpretability therefore prevents adoption in clinical setting, TissueLab is explicitly well-suited for translational applications. Most existing agentic AI systems operate as black boxes [1], offering limited insights into the biological mechanisms behind their predictions, which makes them difficult to trust and deploy in high-stakes medical contexts [43]. In contrast, TissueLab and TLAgent makes all intermediate data fully visualizable, allowing clinicians to inspect and interpret the reasoning process. Moreover, through its evolving memory layer, all intermediate outputs are not only transparent e but also editable, enabling users to directly trace, adjust, and fine-tune the reasoning process from iterative interaction (the "co-evolving" feature). In addition, TissueLab leverages the Model Context Protocol (MCP) to retrieve diagnosis in the latest and authoritative clinical guidelines, and because such guidelines evolve over time, this dynamic knowledge-retrieval ensures the system remains sustainable and aligned with the latest standards of care. This design ensures that the system remains transparent and continuously improves with minimal effort, while keeping clinicians in full control of the decision pathway. Furthermore, TissueLab supports the seamless integration of new models into its modular framework, thereby maintaining state-of-the-art performance while ensuring transparency and clinical alignment.

Through the online TissueLab ecosystem, we anticipate broader use of TLAgent, with user-provided knowledge, annotation, and model contributions continually enhancing its intelligence.

Despite these promising results, our study has several limitations. First, while there has been recent progress in self-evolving AI [44–47], TissueLab currently co-evolves with clinicians in the loop to ensure its human-aligned safety. This design minimizes risks but also increases the level of human supervision required. In the future, we hope to move toward a community-driven, self-evolving autonomous AI coscientist that can continuously improve under established safety constraints while reducing dependence on manual supervision. Second, although we have made every effort to ensure fair comparisons with other baseline models and systems, it is important to note that some baselines may not have access to the same high-resolution data (such as gigapixel pathology images) available to TLAgent. For instance, the pathology input to GPT-5-vision is restricted to the largest thumbnail the model can accept, whereas TLAgent operates directly on raw whole-slide images. We anticipate that future work could develop a GPT-5-vision-based pipeline for fairer comparison, though building such a system would itself require substantial effort. Lastly, integrating TissueLab into routine clinical workflows will require not only technical validation but also compliance with regulatory standards and interoperability with existing hospital systems. Addressing these limitations will be critical for translating TissueLab into a widely adopted and safe clinical intelligence system.

These limitations also motivate future work and underscore the potential of agentic AI systems to progress from "laboratory intelligence" towards "clinical intelligence". A natural next step is to further advance TissueLab into a community-driven self-evolving ecosystem [48, 49], where contributions from the broader TissueLab community collectively guide its evolution. By pooling diverse expertise and data within a controlled and transparent framework, TissueLab could continuously refine its capabilities by self-evolving, adapt more rapidly to emerging clinical challenges, and reduce reliance on manual oversight.

4 Methods

Adaptive and extensible agentic system for seamless model integration

The field of medical agentic AI is advancing rapidly, with successive generations of agents frequently exceeding the capabilities of their predecessors. Yet much of this apparent progress has come from continually replacing or updating the underlying tool models—a costly and ultimately unsustainable strategy. True evolution requires more than incremental tool updates; it demands the development of agentic AI systems that are inherently extensible and capable of self-adaptation.

To address these needs, TissueLab was designed as an adaptive and extensible framework. TissueLab agents can seamlessly integrate *any* model as it becomes available, without requiring modifications to the orchestration logic or reliance on a fixed toolbox. Mathematically, let $\mathcal{M}_t^{\text{any}}$ denote the set of all models integrated at time t. The effective capability of the system can be expressed as

$$\text{Capability}(t) \propto \sum_{m \in \mathcal{M}_t^{\text{any}}} \text{Perf}(m),$$

highlighting that system capacity expands monotonically with the integration of arbitrary models from the broader medical AI ecosystem.

Adaptivity in TissueLab is realized through a modular plugin architecture, inspired by design patterns in software engineering—particularly the *factory method*—that emphasize modularity and separation of concerns. Computational models are abstracted as task nodes within a directed acyclic workflow graph, making it straightforward to expand the system. Each node exposes a standardized interface with three essential functions: initialization to declare requirements and parameters, input handling to parse data into the expected format, and execution to perform model computation. The internal implementation of each model remains a black box to the agent, enabling uniform orchestration regardless of modality or complexity. This abstraction ensures that any state-of-the-art model can be mounted as a new node, allowing TissueLab to continuously incorporate advances in medical AI while preserving scalability and robustness.

Topological sorting enables parallel and distributed inference

TissueLab organizes computational models as modular task nodes within a directed acyclic workflow graph. By applying topological sorting [50], the system ensures that tasks with dependencies (e.g., segmentation before classification) are executed in sequence, while independent branches can be run in parallel.

This design improves scalability and reduces inference time, and also allows distributed execution across multiple GPU servers.

Importantly, dependencies are defined only in a pairwise manner (e.g., segmentation precedes classification), which avoids the complexity of reasoning over a full graph and reduces the likelihood of hallucinations by the LLM. This guarantees that clinically meaningful workflows remain both efficient and trustworthy, even in large-scale imaging studies. When executing planned workflows, the agent does not need to reason about the complete set of dependency relations among all models, which could otherwise increase complexity and risk hallucinations. Instead, only pairwise dependencies need to be specified; for example, segmentation should always precede classification. Based on this set of pairwise relations, the system performs a topological sorting to construct the workflow. The algorithm counts the in-degree of each task node, begins execution from nodes with zero in-degree, and after each node finishes, removes it from the graph while updating the in-degree of its successors. This procedure enables task nodes to be executed in parallel while preserving the logical order of dependencies. This design naturally accommodates both serial and parallel execution: tasks with strict dependencies (e.g., segmentation preceding classification) are executed sequentially, while independent branches can run concurrently to improve efficiency.

Formally, let the workflow be represented as a directed acyclic graph $G = (V, E)$ with in-degree function $d(v)$. At each iteration, all nodes $v \in V$ with $d(v) = 0$ are executed, then removed from G, and $d(u)$ is decremented for all $(v, u) \in E$:

$$\forall v \in V, \quad d(v) = 0 \;\; \Rightarrow \;\; \text{execute}(v),\; V \leftarrow V \setminus \{v\},\; d(u) \leftarrow d(u) - 1 \;\; \forall (v, u) \in E.$$

The procedure can equivalently be expressed as pseudocode:

Algorithm 3 Parallel Topological Sort

```
Input: Directed acyclic graph $G = (V, E)$
while $V$ is not empty do
    $S \leftarrow \{v \in V \mid indegree(v) = 0\}$
    execute all $v \in S$ in parallel
    for each $v \in S$ do
        remove $v$ from $V$
        for each $(v, u) \in E$ do
            $indegree(u) \leftarrow indegree(u) - 1$
        end for
    end for
end while
```

This procedure ensures that workflow execution remains deterministic and scalable: deterministic because the topological order is uniquely determined by the dependency graph, and scalable because independent branches of the workflow can be executed in parallel without violating dependencies. By reducing the need to reason over the entire dependency structure, the design also improves robustness and lowers the risk of hallucinations during orchestration.

Community-driven co-evolution

The co-evolving mechanism in TissueLab consists of two complementary components. First, human-in-the-loop feedback (e.g., clinician corrections on image patch-level classification) is captured as additional labeled data and incorporated into an active learning loop. These annotations are used to update lightweight modules (e.g., shallow classifiers such as XGBoost [51]), thereby improving reliability and clinical alignment. Because such modules can be retrained within minutes using a small number of user-provided annotations, the system can rapidly adapt to specific clinical scenarios without relying on large-scale pretraining or extensive datasets. Formally, if θ denotes the parameters of a lightweight model, then feedback data $\{(x_i, y_i)\}$ is used to solve

$$\theta^{\star} = \arg\min_{\theta} \sum_{i} \ell(y_i,\, f_\theta(x_i))\,,$$

where ℓ is the training loss and f_θ is the lightweight model. This update mechanism enables rapid personalization and clinician-aligned adaptation. Importantly, within the TissueLab ecosystem, these adapted modules can be shared across the community or further refined collaboratively, enabling collective improvement of classifiers and fostering continuous system-wide progress. In this way, contributions from individual users directly benefit the broader community.

Second, the agent maintains an extensible candidate pool of available models, which can incorporate any new models from the broader medical AI ecosystem on demand. Model selection or switching is driven by user evaluation of the outputs (e.g., whether results are deemed acceptable or inadequate), rather than performed autonomously. Formally, a large language model first retrieves a task-relevant candidate set $\mathcal{M}_t^{\text{cand}} \subseteq \mathcal{M}_t^{\text{any}}$. A feedback-guided policy π_ϕ, trained on accumulated feedback $\mathcal{F}_t$, then assigns rank scores to models in this candidate set:

$$s_m = \pi_\phi(m \mid \mathcal{M}_t^{\text{cand}}, \mathcal{F}_t), \qquad \mathcal{M}_t^{\star} = \text{Rank}\big(\mathcal{M}_t^{\text{cand}}, \{s_m\}\big).$$

This design ensures that model selection adapts dynamically with user feedback while leaving outputs unaltered. Furthermore, TissueLab provides a community-driven mechanism whereby AI researchers can seamlessly integrate their latest models into the agentic system, making cutting-edge research immediately available to clinicians in real workflows. This not only accelerates the translation of AI research into practice but also ensures that advances from the evolving AI ecosystem are leveraged safely, under explicit human oversight.

LLM orchestration

TissueLab is built upon a large language model (LLM) backbone. In our experiments, we employ chatgpt-4o-latest (OpenAI, 2025) as a representative implementation, though any future LLM can serve this role. The LLM is used only as an orchestration layer: it plans workflows, generates code, and invokes function calls, but it does not directly process raw medical images. All data analysis tasks (e.g., segmentation, classification, code analysis) are executed locally by domain-specific models and pipeline. This design avoids token overload and attention dilution when handling gigapixel whole-slide images or volumetric radiology data, while enabling the system to remain adaptive and extensible. Importantly, the modular design allows the base LLM to be replaced by future models without modifying the orchestration framework, making the system adaptive and future-proof.

Execution layer: interaction between agent and tool models

The execution layer consists of domain-specific models and pipelines, which are invoked as task nodes by the agent. all intermediate data states are persisted in a shared local storage (HDF5 database [52]), rather than being passed directly between tools. Each tool model is treated as a black box task node that exposes a standardized read() interface to retrieve the required inputs from the shared database, without needing to be aware of upstream or downstream dependencies. The orchestration layer (LLM) ensures that, at the time of execution, all necessary inputs for a given node have already been generated and stored. This design decouples tools from one another, simplifies integration of new models, and guarantees reproducibility by making all intermediate outputs explicitly available in the persistent storage.

Editable memory layer

To address the challenges of token overload and attention dilution [53] in long multi-step workflows, TissueLab implements an editable memory layer. Unlike conventional LLM-based systems that rely solely on limited token context, TissueLab employs persistent memory implemented through local structured storage

(HDF5 databases). All intermediate results, annotations, and execution logs are written into this memory layer, which can be accessed by downstream task nodes via standardized interfaces. Because the memory resides outside of the token context, its capacity is effectively limited only by hardware storage rather than by model constraints, enabling theoretically unbounded memory.

Importantly, the memory layer is fully visible to users. Clinicians or researchers can visualize the result of each step within the TissueLab platform for review, and they can refine performance by augmenting stored results and annotations. Such updates are directly incorporated into subsequent executions, thereby improving the agentâĂŹs performance over time. By storing explicit intermediate outputs that are both persistent and editable, the memory layer ensures that workflows remain transparent, auditable, and adaptable, with every decision traceable to concrete stored data rather than transient model states.

Semantic function-calling for data access

The Data Access Layer employs function calling to read the structure and content of locally stored HDF5 files. All datasets follow a semantic naming convention (e.g., `mask2D`, `volume3D`, `3Dmask_{timestamp}`, `4Dseries_{phase}`), and carry metadata such as `voxel_spacing`, `origin`, `orientation`, and `timestamp`. By inspecting both the dataset name and its structure, the LLM can infer the role of the data-for example, a dataset named `3Dmask` with three dimensions is recognized as a volumetric segmentation mask.

When a task requires specific inputs, the LLM issues function calls that query the HDF5 structure, retrieves the identified datasets together with their metadata, and generates local code to run the analysis pipeline. Computation is thus grounded in concrete structured data of arbitrary dimensionality, rather than within the token context of the model, avoiding attention dilution and token overload that commonly arise with gigapixel slides or volumetric scans.

Because access is explicit and aligned with the HDF5 schema, outputs are reproducible, auditable, and privacy-preserving, as sensitive medical data remain local. Extending the system to new modalities or dimensions only requires adding semantically named datasets, without modifying the orchestration logic, thereby further strengthening adaptivity. Together with the Editable Memory Layer, this design provides both persistent storage and semantically structured access, forming a reliable foundation for building co-evolving agentic AI workflows.

MCP for searching external criteria

In the medical domain, these challenges are further compounded by the necessity of clinical oversight: reliance on LLM-generated reasoning alone is insufficient, as clinical expertise remains indispensable for validating, correcting, and guiding AI outputs. Only through such integration can agentic systems mitigate hallucinations and ensure that their recommendations remain safe, reliable, and aligned with established standards of care. many tasks require reference to established guidelines or criteria, such as cancer staging systems, diagnostic thresholds, or treatment protocols. To ensure that agentic workflows are grounded in authoritative sources, TissueLab employs the Model Context Protocol (MCP) exclusively for external knowledge retrieval. MCP is not used for communication between local tool models, nor does the system rely on the LLM to guess criteria or on a fixed internal knowledge base. Instead, MCP treats the entire web as a knowledge repository, allowing the agent to continuously query up-to-date standards and guidelines as they evolve.

When the agent encounters a task that requires domain-specific thresholds or criteria (e.g., tumor staging according to AJCC, or risk stratification following TCGA-defined biomarkers), it issues an MCP query and incorporates the retrieved information into the workflow. This design guarantees that outputs are supported by traceable references, thereby reducing the risk of hallucination and enhancing clinical trust.

By restricting MCP to the retrieval of external evidence, TissueLab ensures that local computation remains efficient and privacy-preserving, while clinical outputs remain verifiable and aligned with evolving medical standards.

Benchmarking TissueLab agentic AI system on diverse clinical imaging cohorts

We evaluated TissueLab across diverse datasets spanning pathology, radiology, and spatial omics. Pathology tasks included whole-slide images of regional lymph node metastasis in colon adenocarcinoma (LNCO2 dataset, Region GÃd'vleborg Clinical Pathology and Cytology department) and prostate and colon cancer slides from VisiumHD. Radiology experiments were performed on multiple public cohorts, including fatty liver assessment (Kaggle: UNIFESP Chest CT Fatty Liver Competition), thoracic disease detection from chest X-rays (Kaggle: NIH Chest X-rays), cardiac hypertrophy analysis from the Sunnybrook Cardiac Data (SCD), and intracranial hemorrhage detection from the PhysioNet Computed Tomography Images for Intracranial Hemorrhage dataset. For spatial omics, we used VisiumHD kidney tissue samples, enabling cross-modal integration with corresponding H&E sections.

In the experiments related to LNCO2, we carefully constructed non-overlapping cohorts. Depth of invasion analysis was performed on 107 primary tumor site slides, while positive lymph node counting used 321 lymph node slides, with the two cohorts defined as mutually exclusive (428 slides excluded in total). For independent evaluation, we reserved a separate hold-out set of 50 slides from 6 patients, which were provided to pathologists for detailed annotation of tumor epithelium, stroma, and other compartments. To minimize potential confounding, these hold-out cases were chosen from samples not involved in any of the other experiments.

For fatty liver diagnosis, the dataset comprises 152 patients with chest CT scans (CS) and 112 patients with abdomen or combined abdomen-chest CT scans (AS), for a total of 264 patients. The data were split at the patient level into training and testing cohorts. The training set included 226 patients (114 CS and all 112 AS), while the testing set comprised the remaining 38 CS patients.

For VisiumHD, we directly applied TissueLabâĂŹs co-evolving framework, enabling clinicians to rapidly obtain new task-specific models from scratch through interactive feedback, without requiring large-scale pretraining.

To contextualize performance, we benchmarked TissueLab against state-of-the-art vision-language and agentic AI baselines. To ensure a fair comparison, all experiments were repeated five times, and mean values with standard deviations (mean ± s.d.) are reported; error bars in figures correspond to these standard deviations. For pathology, comparisons included LLaVA-Med [54], MedGemma [55], Quilt-LLaVA [56], GPT-4o-agent, GPT-4o-vision, and GPT-5-vision. For radiology, benchmarks included LLaVA-Med, MedGemma, M3D [57], GPT-4o-agent, GPT-4o-vision, and GPT-5-vision. In the case of spatial omics integration, existing VLM- and LLM-based agents lack the capacity to align gene expression data with H&E images and cannot handle high-dimensional expression matrices that exceed their token limits. As such, only comparisons with ground truth were possible in this domain.

In conclusion, these datasets cover a broad spectrum of tissue types, imaging modalities, and clinical tasks, providing a comprehensive benchmark for assessing the adaptability of our agentic AI system.

Tissue segmentation and classification

In pathology, there is currently no single universal model capable of performing accurate tissue segmentation and classification across all organ systems. For example, some models perform well for colorectal cancer slides but fail to segment and classify structures such as glomeruli in kidney biopsies. To address this limitation, TissueLab adopts a patch-level segmentation and classification strategy. Before patch-level analysis, TissueLab crop the Whole-slide images (WSIs) into multiple patches. Then perform background removal to ensure that only relevant tissue regions are retained. Specifically, low-resolution slide thumbnails are first processed with adaptive thresholding to distinguish tissue from background under varying staining and illumination conditions. The resulting masks are then refined using morphological closing and hole-filling operations to repair broken regions and fill internal gaps, followed by connected-component analysis to filter out small noisy fragments. Finally, boundary constraints are applied to remove common edge shadows and artifacts near the slide borders. This step-wise procedure progressively cleans artifacts, enhances mask continuity, and supports saving of intermediate results for debugging and visualization. Together, these operations yield accurate and stable tissue masks, ensuring that subsequent patch extraction and downstream analyses focus on true tissue regions rather than background noise. After preprocessing, TissueLab uses foundation models such as MUSK [58] or PLIP [59] to extract embeddings from tissue patches. Because these embeddings are trained with contrastive learning to capture transferable semantic representations, they can be applied in a zero-shot learning paradigm for initial segmentation and classification.

When the zero-shot predictions are insufficient, users can provide patch-level annotations that serve as feedback to the agentic system. Such annotations enable rapid retraining of lightweight models (e.g., linear heads) to specialize for the current task, often within minutes. Additionally, other tissue segmentation or tissue classification models developed by other researchers can be seamlessly integrated through the TissueLab ecosystem, further enhancing performance in domain-specific tasks. This multi-source and feedback-driven design allows the system to progressively improve agentic AI's accuracy in a co-evolving manner, while remaining adaptable to diverse tissue types and pathological contexts.

In contrast to pathology, the radiology domain already offers a broad range of high-performance pretrained models for segmentation and analysis. TissueLab leverages this ecosystem by integrating domain-specific models into its model factory as task nodes. Any pretrained radiology model can be registered in the factory. When a new query arrives, the LLM dynamically selects a context-relevant candidate pool for orchestration:

$$\mathcal{M}_t^{\text{cand}} = \text{LLM}\Big(\text{query}, \mathcal{M}_t^{\text{factory}}\Big) .$$

This design ensures that only appropriate domain-specific models are considered for execution, reducing unnecessary computation while maintaining flexibility.

As illustrative examples, TotalSegmentator [60] and BiomedParse [61] are incorporated as task nodes. TotalSegmentator provides fine-grained, high-accuracy segmentation of multiple anatomical structures across CT volumes, but is limited to a predefined set of organs and regions exposed through its interface. Meanwhile, BiomedParse enables query-driven segmentation of arbitrary user-specified structures, offering flexibility when the requested target is not part of the TotalSegmentator set.

This ensures that TissueLab leverages the strengths of the broader radiology AI ecosystem without duplicating effort, functioning as a flexible orchestration framework that can dynamically select domain-specific models to deliver more precise predictions in practical clinical workflows.

Cell segmentation and classification

To enable cell-level analysis, TissueLab integrates both external and in-house models within its agentic AI framework. For nucleus segmentation, TissueLab incorporates StarDist [62–64] as a tasknode, an external foundation model that provides polygonal instance segmentation with high accuracy and robustness across whole-slide images (WSIs). To ensure reliable detection of nuclei at patch boundaries, WSIs are processed using overlapping tiles. Overlapping regions are then deduplicated to prevent double counting, thereby maintaining accurate cell counts across the entire slide. This design ensures that segmentation remains scalable to gigapixel pathology images while preserving precision at the single-cell level.

Building on this segmentation backbone, TissueLab also integrates NuClass, an in-house vision–language foundation model designed for nuclei typing under a zero-shot paradigm. NuClass combines three key innovations: it leverages hierarchical biomedical ontologies and natural-language descriptions to enable agents to reason over semantic relationships between cell classes and adapt to unseen types; it adopts a nucleus-centered patch representation that naturally guides transformer attention to biologically meaningful regions, improving the precision of nucleus-level analysis; and it is pretrained on over 8.3 million nuclei patches from 11 diverse datasets, providing robust morphological priors across tissues that TissueLab agents can readily exploit in varied experimental settings.

Through this integration, TissueLab agents can perform scalable and ontology-aware cell classification, enabling downstream biomedical analyses such as cancer research, spatial gene expression studies, and discovery of disease-related cellular patterns.

Multi-modal analysis

TissueLab agentic AI system enables multi-modal analysis, where molecular signals from spatial omics can be interpreted together with histological features. By combining active learning outputs from cell-level classification with spatial gene expression data, the system provides richer and more comprehensive information for downstream analysis.

To support this workflow, TissueLab integrates CellCharter [65] for the identification, characterization, and comparison of spatial clusters from spatial omics data. While CellCharter focuses exclusively on spatial omics clustering, TissueLab links these clusters to cell-type information derived from NuClass-based classification or user-provided annotations collected through active learning. This correspondence allows users to interpret spatial transcriptomic clusters in terms of underlying histological cell types and biological functions, thereby advancing beyond modality-specific analyses.

In this way, TissueLab enables a multi-modal mapping between molecular clusters and cellular phenotypes, extending beyond single-modality processing to support multi-scale, multi-modal insights into tissue organization and disease mechanisms.

Evaluation metrics

We evaluate TissueLab using a broad spectrum of metrics to capture task completion, predictive accuracy, robustness, and clinical relevance.

Task-level metrics. We define model successfulness, or Task Completion Rate (TCR), as the proportion of tasks for which the agent produces a concrete and clinically relevant answer. Formally,

$$\mathrm{TCR} = \frac{N_{\text{success}}}{N_{\text{total}}},$$

where N_{success} denotes the number of queries for which the system output directly addresses the posed clinical question, and N_{total} is the total number of queries. A response is counted as successful only if it goes beyond generic or evasive statements and yields an actionable result that a clinician could use in practice. This metric is particularly critical for evaluating large language model (LLM)-based systems: unlike traditional accuracy measures, model successfulness explicitly penalizes hallucinated or superficial outputs, thereby ensuring that reported answers are both relevant and clinically interpretable. In this way, model successfulness captures the pragmatic utility of the agentic AI system in real-world workflows, rather than its performance on isolated benchmarks.

Regression metrics. For continuous prediction tasks, we report the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The MAE,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{d}_i - d_i|,$$

quantifies the average absolute deviation between the predicted value $\hat{d}_i$ and the ground truth d_i across n samples. Because MAE is expressed in the same units as the target variable (e.g., millimeters for tumor invasion depth), it provides an intuitive measure of the typical error magnitude. The RMSE,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{d}_i - d_i)^2},$$

applies a quadratic penalty to deviations, making it more sensitive to large errors. This property highlights the occurrence of rare but clinically consequential misestimations, such as substantially underestimating tumor size or lesion volume. Together, MAE captures average-case reliability while RMSE emphasizes the risk of extreme errors, ensuring that both common and catastrophic failures are evaluated in clinical benchmarking.

Pearson correlation coefficient. The Pearson correlation coefficient (r) measures the strength of the linear association between predictions and ground truth values. It is defined as

$$r = \frac{\sum_{i=1}^{n}(d_i - \bar{d})(\hat{d}_i - \bar{\hat{d}})}{\sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2}\sqrt{\sum_{i=1}^{n}(\hat{d}_i - \bar{\hat{d}})^2}},$$

where d_i denotes the ground truth, $\hat{d}_i$ the prediction for sample i, and $\bar{d}$ and $\bar{\hat{d}}$ are their respective sample means. Values of r range from -1 (perfect inverse correlation) to $+1$ (perfect positive correlation), with 0 indicating no linear association. Unlike error-based metrics such as MAE or RMSE, r is invariant to scale and offset, and thus evaluates whether the relative ordering of samples is preserved. This property is especially important in clinical contexts such as risk stratification, where the correct ranking of patients (e.g., higher invasion depth implying higher risk) may be more consequential than the exact numerical estimate.

Classification metrics. For discrete prediction tasks, we report a suite of complementary classification metrics. Let TP denote true positives (correctly identified positive cases), TN true negatives (correctly identified negative cases), FP false positives (negative cases incorrectly predicted as positive), and FN false negatives (positive cases incorrectly predicted as negative).

Accuracy is the most widely used classification metric, defined as the proportion of correctly classified samples among all cases:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

It provides an intuitive single-number summary of overall correctness and is easy to interpret across different tasks. However, accuracy alone can be misleading in settings with severe class imbalance, as a model may achieve deceptively high scores by always predicting the majority class. For instance, in cancer screening with 95% healthy cases, a trivial model labeling all patients as healthy would still reach 95% accuracy while completely failing to detect diseased individuals. These limitations motivate the complementary use of precision, recall, and F1-score, which better capture diagnostic reliability under clinical imbalance.

To complement accuracy, we report precision, recall, and their harmonic mean (F1-score), which together provide a more nuanced assessment of diagnostic performance under class imbalance.

Precision quantifies the reliability of positive predictions, indicating the proportion of predicted positives that are truly positive.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

A high precision indicates that most predicted positives are true positives, thereby reducing the risk of over-diagnosis and unnecessary clinical interventions.

Recall (sensitivity) captures the modelâĂŹs ability to identify all true positive cases, thereby minimizing false negatives. This is crucial in clinical contexts such as cancer detection, where missed diagnoses can delay treatment.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

High recall is essential in clinical contexts such as early cancer detection, where missing positive cases (false negatives) could delay treatment and lead to adverse outcomes.

Because precision and recall often trade off against each other, we additionally report the F1-score, which summarizes their balance into a single value:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Clinically, this combined measure is valuable because high precision reduces the risk of over-diagnosis and unnecessary interventions, while high recall ensures that true disease cases are not missed. The F1-score therefore captures a clinically meaningful balance, rewarding models that can simultaneously avoid excessive false positives and minimize false negatives, which is particularly important in diagnostic workflows where both types of errors carry significant consequences.

For multi-class settings, we report the weighted F1-score:

$$\text{F1}_{\text{weighted}} = \frac{1}{N} \sum_{c=1}^{C} n_c \cdot \text{F1}_c,$$

where n_c is the number of samples in class c, F1_c is the F1-score for class c, and N is the total number of samples. This formulation accounts for class imbalance by ensuring that both majority and minority classes contribute proportionally to the overall evaluation.

For multi-class classification, we extend the F1-score by computing a weighted average across classes:

$$\text{F1}_{\text{weighted}} = \frac{1}{N} \sum_{c=1}^{C} n_c \cdot \text{F1}_c,$$

where n_c is the number of samples in class c, F1_c is the per-class F1-score, C is the total number of classes, and N is the overall sample size. This formulation accounts for class imbalance by assigning greater weight to more prevalent classes while still incorporating performance on minority classes. Clinically, the weighted F1-score ensures that rare but critical categories (e.g., high-grade tumors or rare genetic subtypes) are not overshadowed by abundant negative or benign cases, providing a more faithful reflection of real-world diagnostic performance across heterogeneous patient populations.

Advanced robustness metrics. To provide a more comprehensive view of classification performance, particularly under class imbalance and variable decision thresholds, we additionally report the Area Under the ROC Curve (AUC), the Matthews Correlation Coefficient (MCC), and CohenâĂŹs Kappa coefficient.

The AUC summarizes the Receiver Operating Characteristic curve, which plots the true positive rate (TPR $= TP/(TP + FN)$) against the false positive rate (FPR $= FP/(FP + TN)$) across all possible thresholds. AUC values range from 0.5 (no discriminative ability) to 1.0 (perfect separation). Unlike accuracy or F1, AUC captures global discriminative power independent of a single threshold, making it particularly informative for screening tasks where cut-offs may differ across clinical settings.

The MCC provides a single balanced statistic that incorporates all four outcomes (TP, TN, FP, FN):

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Ranging from -1 (inverse prediction) to $+1$ (perfect prediction), MCC is robust to class imbalance and is therefore well-suited for rare disease detection, where conventional accuracy may be inflated by negative cases.

Finally, CohenâĂŹs Kappa measures agreement between predicted and reference labels while correcting for chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement and p_e the expected agreement by chance. With values from -1 to $+1$, Kappa offers an interpretable analogue to inter-observer agreement among physicians, with $\kappa > 0.8$ typically regarded as strong concordance in clinical studies.

Summary. Together, these task-level, regression, and classification metrics provide a multifaceted evaluation of TissueLab. While accuracy and error-based measures reflect overall correctness and magnitude of deviation, precision-recall-F1 quantify diagnostic trade-offs, correlation captures consistency in relative ordering, and advanced metrics such as AUC, MCC, and Kappa ensure robustness under imbalance, threshold variability, and human-AI agreement. This spectrum of metrics thus ensures that model performance is not only statistically valid but also clinically meaningful.

Author Contributions

Conceptualization: Songhao Li, Zhi Huang. **Methodology:** Songhao Li, Jonathan Xu, Tiancheng Bao, Yuxuan Liu, Yuchen Liu, Yihang Liu, Lilin Wang, Wenhui Lei, Sheng Wang, Yinuo Xu, Yan Cui, Jialu Yao, Shunsuke Koga, Zhi Huang. **Experiments:** Songhao Li, Tiancheng Bao, Wenhui Lei, Yinuo Xu, Zhi Huang. **Data curation and acquisition:** Songhao Li, Jonathan Xu, Tiancheng Bao, Wenhui Lei, Shunsuke Koga, Zhi Huang. **Expert Pathology Annotation and Feedback:** Shunsuke Koga. **Manuscript writing:** Songhao Li, Zhi Huang. **Supervision:** Zhi Huang. All authors reviewed and approved the final manuscript.

Acknowledgements

We would like to thank other software engineers who have contributed to this software, including but not limited to, Steven Su, Qintian Huang.

Data Availability

Whole-slide images and expert annotations from the LNCO2 dataset are available via the AIDA Data Hub at https://datahub.aida.scilifelab.se/10.23698/aida/lnco2 upon application and approval. Public radiology datasets are accessible from their respective repositories: UNIFESP Chest CT Fatty Liver Competition at https://www.kaggle.com/competitions/unifesp-fatty-liver, NIH Chest X-rays at https://www.kaggle.com/datasets/nih-chest-xrays/data/data, Sunnybrook Cardiac Data (SCD) at https://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/, and PhysioNet Intracranial Hemorrhage CT dataset at https://physionet.org/content/ct-ich/1.3.1/. VisiumHD datasets of prostate, colon, and kidney tissues, including histology slides and matched spatial transcriptomics data, are publicly available from 10x Genomics at https://www.10xgenomics.com/spatial-transcriptomics.

Code Availability

TissueLab source code is publicly available at tissuelab.org, with installers for Windows, macOS, and Linux.

Conflict of Interests

None declared.

References

1. Ferber, D., El Nahhas, O. S. M., WÃűlflein, G., Wiest, I. C., Clusmann, J., LeÃ§mann, M.-E., Foersch, S., Lammert, J., Tschochohei, M., JÃd'ger, D., Salto-Tellez, M., Schultz, N., Truhn, D. & Kather, J. N. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer* **6.** Publisher: Nature Publishing Group, 1337–1349. ISSN: 2662-1347. https://www.nature.com/articles/s43018-025-00991-6 (2025) (Aug. 2025).
2. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine* **29.** Publisher: Nature Publishing Group, 2983–2984. ISSN: 1546-170X. https://www.nature.com/articles/s41591-023-02594-z (2025) (Dec. 2023).
3. Chen, C., Weishaupt, L. L., Williamson, D. F. K., Chen, R. J., Ding, T., Chen, B., Vaidya, A., Le, L. P., Jaume, G., Lu, M. Y. & Mahmood, F. *Evidence-based diagnostic reasoning with multi-agent copilot for human pathology* version: 1. June 26, 2025. arXiv: 2506.20964[cs]. http://arxiv.org/abs/2506.20964 (2025).
4. Anthropic, A. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Claude-3 Model Card* (2024).
5. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., *et al.* Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).
6. OpenAI *et al. gpt-oss-120b & gpt-oss-20b Model Card* Aug. 8, 2025. arXiv: 2508.10925[cs]. http://arxiv.org/abs/2508.10925 (2025).
7. *GPT-5 System Card* https://openai.com/index/gpt-5-system-card/ (2025).
8. Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K. & Press, O. *SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering* 2024.
9. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B. & Sánchez, C. I. A survey of deep learning in medical image analysis. *Medical Image Analysis* **42.** Survey / Review, 60–88 (2017).
10. Wang, H., He, Y., Coelho, P. P., Bucci, M., Nazir, A., Chen, B., Trinh, L., Zhang, S., Huang, K., Chandrasekar, V., Chung, D. C., Hao, M., Leote, A. C., Lee, Y., Li, B., Liu, T., Liu, J., Lopez, R., Lucas, T., Ma, M., Makarov, N., McGinnis, L., Peng, L., Ra, S., Scalia, G., Singh, A., Tao, L., Uehara, M., Wang, C., Wei, R., Copping, R., Rozenblatt-Rosen, O., Leskovec, J. & Regev, A. *SpatialAgent: An autonomous AI agent for spatial biology* Pages: 2025.04.03.646459 Section: New Results. Apr. 6, 2025. https://www.biorxiv.org/content/10.1101/2025.04.03.646459v1 (2025).
11. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature.* Publisher: Nature Publishing Group, 1–8. ISSN: 1476-4687. https://www.nature.com/articles/s41586-025-09442-9 (2025) (July 29, 2025).
12. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z. & Wang, Y. *MMedAgent: Learning to Use Medical Tools with Multi-modal Agent* Oct. 5, 2024. arXiv: 2407.02483[cs]. http://arxiv.org/abs/2407.02483 (2025).
13. Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. *Why Language Models Hallucinate* Sept. 4, 2025. arXiv: 2509.04664[cs]. http://arxiv.org/abs/2509.04664 (2025).
14. Das, A. B., Sakib, S. K. & Ahmed, S. *Trustworthy Medical Imaging with Large Language Models: A Study of Hallucinations Across Modalities* Aug. 9, 2025. arXiv: 2508.07031[eess]. http://arxiv.org/abs/2508.07031 (2025).
15. Wu, J., Wang, Y., Zhong, Z., Liao, W., Trayanova, N., Jiao, Z. & Bai, H. X. Vision-language foundation model for 3D medical imaging. *npj Artificial Intelligence* **1.** Publisher: Nature Publishing Group, 17. ISSN: 3005-1460. https://www.nature.com/articles/s44387-025-00015-9 (2025) (Aug. 6, 2025).
16. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err.
17. Gamma, E., Helm, R., Johnson, R. & Vlissides, J. *Design Patterns: Elements of Reusable Object-Oriented Software* ISBN: 0-201-63361-2 (Addison-Wesley, Reading, MA, 1994).

18. Huang, Z., Yang, E., Shen, J., Gratzinger, D., Eyerer, F., Liang, B., Nirschl, J., Bingham, D., Dussaq, A. M., Kunder, C., Rojansky, R., Gilbert, A., Chang-Graham, A. L., Howitt, B. E., Liu, Y., Ryan, E. E., Tenney, T. B., Zhang, X., Folkins, A., Fox, E. J., Montine, K. S., Montine, T. J. & Zou, J. A pathologistâĂŞAI collaboration framework for enhancing diagnostic accuracies and efficiencies. *Nature Biomedical Engineering* **9.** Publisher: Nature Publishing Group, 455–470. ISSN: 2157-846X. https://www.nature.com/articles/s41551-024-01223-5 (2025) (Apr. 2025).

19. Maras, G., Lindvall, M. & Lundstrom, C. *Regional lymph node metastasis in colon adenocarcinoma, second collection series (LNCO2)* 2020. https://datahub.aida.scilifelab.se/10.23698/aida/lnco2.

20. American Joint Committee on Cancer. *AJCC Cancer Staging System Version 9 Protocols* Version 9 official protocols. Released 2024-2025; replaces selected sites from the 8th Edition. 2024. https://www.facs.org/quality-programs/cancer-programs/american-joint-committee-on-cancer/version-9/.

21. National Comprehensive Cancer Network. *NCCN Clinical Practice Guidelines in Oncology: Colon Cancer. Version 3.2025* Accessed September 2025. https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf. 2025.

22. Gunderson, L. L., Jessup, J. M., Sargent, D. J., Greene, F. L. & Stewart, A. K. Revised TN categorization for colon cancer based on national survival outcomes data. *Journal of Clinical Oncology* **28,** 264–271 (2010).

23. Rosenberg, R., Friederichs, J., Schuster, T., Gertler, R., Maak, M., Becker, K., Greb, R. R., Ulm, K., HÃűfler, H., Nekarda, H. & Siewert, J. R. Prognosis of patients with colorectal cancer is associated with lymph node ratio: a single-center analysis of 3,026 patients. *Annals of Surgery* **248,** 968–978 (2008).

24. Anthropic. *Model Context Protocol (MCP) Specification* Open standard protocol specification. Released November 25, 2024; Accessed: 2025-09-12. 2024. https://modelcontextprotocol.io/specification/2025-03-26.

25. College of American Pathologists. *Cancer Protocol Templates* https://www.cap.org/protocols-and-guidelines. Accessed: 2025-09-12. 2023.

26. WHO Classification of Tumours Editorial Board. *WHO Classification of Tumours, 5th Edition, Volume 11: Haematolymphoid Tumours* ISBN: 978-92-832-4520-9 (International Agency for Research on Cancer (IARC), Lyon, France, 2024).

27. Chalasani, N., Younossi, Z., Lavine, J. E., Charlton, M., Cusi, K., Rinella, M., Harrison, S. A., Brunt, E. M. & Sanyal, A. J. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. *Hepatology* **67,** 328–357 (2018).

28. Lewis, P., Perez, E., Piktus, A., Karpukhin, V., *et al. Retrieval-augmented generation for knowledge-intensive NLP tasks* in *Advances in Neural Information Processing Systems (NeurIPS)* (2020). https://arxiv.org/abs/2005.11401.

29. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25.** Publisher: Nature Publishing Group. ISSN: 1546-170X. https://www.nature.com/articles/s41591-018-0300-7 (2025) (Jan. 2019).

30. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2097–2106 (2017).

31. Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y. & Rajpurkar, P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering* **6.** Publisher: Nature Publishing Group, 1399–1406. ISSN: 2157-846X. https://www.nature.com/articles/s41551-022-00936-9 (2025) (Dec. 2022).

32. Ma, D., Pang, J., Gotway, M. B. & Liang, J. A fully open AI foundation model applied to chest radiography. *Nature* **643.** Publisher: Nature Publishing Group, 488–498. ISSN: 1476-4687. https://www.nature.com/articles/s41586-025-09079-8 (2025) (July 2025).

33. Serpa, A. S. & FelipeKitamura MD, P. *UNIFESP Chest CT Fatty Liver Competition* https://kaggle.com/competitions/unifesp-fatty-liver. Kaggle. 2022.

34. Hssayeni, M. D., Croock, M. S., Salman, A. D., Al-khafaji, H. F., Yahya, Z. A. & Ghoraani, B. Intracranial Hemorrhage Segmentation Using A Deep Convolutional Model. *Data* **5,** 14 (2020).

35. Hssayeni, M. *Computed Tomography Images for Intracranial Hemorrhage Detection and Segmentation (version 1.3.1)* RRID:SCR_007345. 2020. https://doi.org/10.13026/4nae-zg36.

36. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R. & Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101.** RRID:SCR_007345, e215–e220 (2000).

37. Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A. J. & Wright, G. A. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal–Cardiac MR Left Ventricle Segmentation Challenge* **49,** 1–8. https://hdl.handle.net/10380/3070 (2009).

38. Ommen, S. R., Mital, S., Burke, M. A., Day, S. M., Deswal, A., Elliott, P., Evanovich, L. L., Hung, J., Joglar, J. A., Kantor, P., Kimmelstiel, C. & et al. 2020 AHA/ACC Guideline for the Diagnosis and Treatment of Patients With Hypertrophic Cardiomyopathy. *Journal of the American College of Cardiology* **76,** e159–e240 (2020).

39. Lang, R. M., Badano, L. P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F. A., Foster, E., Goldstein, S. A., Kuznetsova, T. & et al. Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Journal of the American Society of Echocardiography* **28,** 1–39 (2015).

40. 10x Genomics. *Visium HD Spatial Gene Expression* https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression. Accessed: 2025-09-13. 2023.

41. Oliveira, M. F. d., Romero, J. P., Chung, M., Williams, S. R., Gottscho, A. D., Gupta, A., Pilipauskas, S. E., Mohabbat, S., Raman, N., Sukovich, D. J., Patterson, D. M. & Taylor, S. E. B. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics* **57.** Publisher: Nature Publishing Group, 1512–1523. ISSN: 1546-1718. https://www.nature.com/articles/s41588-025-02193-3 (2025) (June 2025).

42. Rao, V. M., Hla, M., Moor, M., Adithan, S., Kwak, S., Topol, E. J. & Rajpurkar, P. Multimodal generative AI for medical image interpretation. *Nature* **639.** Publisher: Nature Publishing Group, 888–896. ISSN: 1476-4687. https://www.nature.com/articles/s41586-025-08675-y (2025) (Mar. 2025).

43. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. & consortium, P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* **20,** 310. https://doi.org/10.1186/s12911-020-01332-6 (Nov. 2020).

44. Kuba, J. G., Gu, M., Ma, Q., Tian, Y. & Mohan, V. *Language Self-Play For Data-Free Training* Sept. 9, 2025. arXiv: 2509.07414[cs]. http://arxiv.org/abs/2509.07414 (2025).

45. Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J. & Weston, J. *Self-Rewarding Language Models* Mar. 28, 2025. arXiv: 2401.10020[cs]. http://arxiv.org/abs/2401.10020 (2025).

46. Li, Z., Yu, W., Huang, C., Liu, R., Liang, Z., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H. & Yu, D. *Self-Rewarding Vision-Language Model via Reasoning Decomposition* Aug. 27, 2025. arXiv: 2508.19652[cs]. http://arxiv.org/abs/2508.19652 (2025).

47. Sun, Z., Liu, Z., Zang, Y., Cao, Y., Dong, X., Wu, T., Lin, D. & Wang, J. *SEAgent: Self-Evolving Computer Use Agent with Autonomous Learning from Experience* Aug. 12, 2025. arXiv: 2508.04700[cs]. http://arxiv.org/abs/2508.04700 (2025).

48. RitorÃĺ, Ã., JimÃĺnez, C. M., GonzÃąlez, J. L., RejÃşn-Parrilla, J. C., HervÃąs, P., Toro, E., Parra-CalderÃşn, C. L., Celi, L. A., TÃžnez, I. & Armengol de la Hoz, M. Ã. The role of Open Access Data in democratizing healthcare AI: A pathway to research enhancement, patient well-being and treatment equity in Andalusia, Spain. *PLOS Digital Health* **3,** e0000599 (Sept. 2024).

49. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. v., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. M. *HuggingFace's Transformers: State-of-the-art Natural Language Processing* July 14, 2020. arXiv: 1910.03771[cs]. http://arxiv.org/abs/1910.03771 (2025).

50. Kahn, A. B. Topological Sorting of Large Networks. *Communications of the ACM* **5,** 558–562 (1962).

51. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 785–794.

52. Folk, M., Heber, G., Koziol, Q., Pourmal, E. & Robinson, D. An overview of the HDF5 technology suite and its applications. *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases,* 36–47 (2011).

53. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. & Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* **12.** Place: Cambridge, MA Publisher: MIT Press, 157–173. https://aclanthology.org/2024.tacl-1.9/ (2025) (2024).

54. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H. & Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day.

55. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., RamÃľ, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D. F., Kirmizibayrak, C., Pilgrim, R., Golden, D. & Yang, L. *MedGemma Technical Report* July 12, 2025. arXiv: 2507.05201[cs]. http://arxiv.org/abs/2507.05201 (2025).

56. Seyfioglu, M. S., Ikezogwo, W. O., Ghezloo, F., Krishna, R. & Shapiro, L. *Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos* Jan. 13, 2025. arXiv: 2312.04746[cs]. http://arxiv.org/abs/2312.04746 (2025).

57. Bai, F., Du, Y., Huang, T., Meng, M. Q.-H. & Zhao, B. *M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models* Mar. 31, 2024. arXiv: 2404.00578[cs]. http://arxiv.org/abs/2404.00578 (2025).

58. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.-H., Willens, S., Olguin, F. M., Nirschl, J. J., Neal, J., Diehn, M., Yang, S. & Li, R. A visionâĂŞlanguage foundation model for precision oncology. *Nature* **638.** Publisher: Nature Publishing Group, 769–778. ISSN: 1476-4687. https://www.nature.com/articles/s41586-024-08378-w (2025) (Feb. 2025).

59. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visualâĂŞlanguage foundation model for pathology image analysis using medical Twitter. *Nature Medicine* **29.** Publisher: Nature Publishing Group, 2307–2316. ISSN: 1546-170X. https://www.nature.com/articles/s41591-023-02504-3 (2025) (Sept. 2023).

60. Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M. & Segeroth, M. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. *Radiology: Artificial Intelligence* **5,** e230024. ISSN: 2638-6100. arXiv: 2208.05868[eess]. http://arxiv.org/abs/2208.05868 (2025) (Sept. 1, 2023).

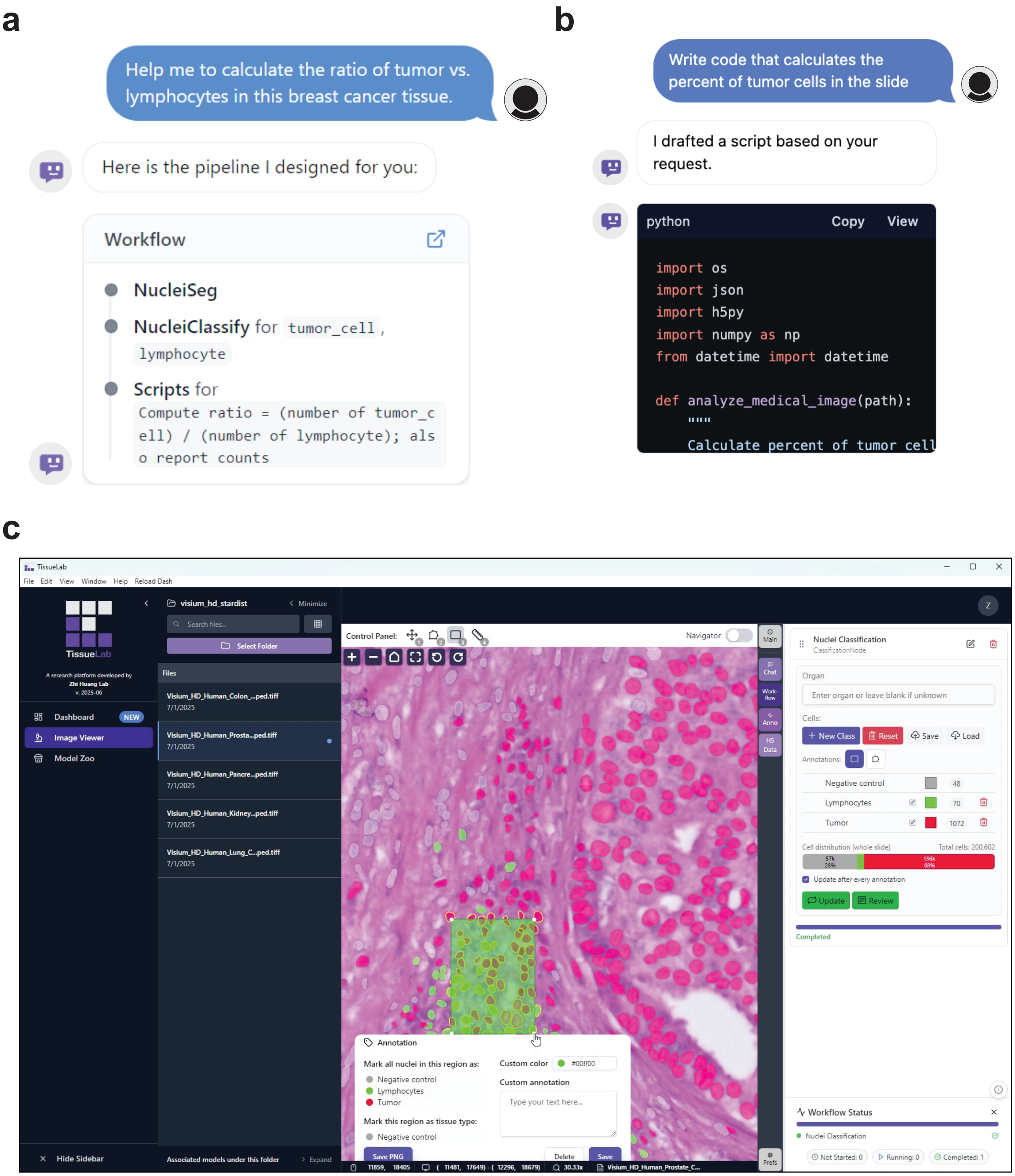
61. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H. & Wang, S. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods* **22.** Publisher: Nature Publishing Group, 166–176. ISSN: 1548-7105. https://www.nature.com/articles/s41592-024-02499-w (2025) (Jan. 2025).

62. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. *Cell Detection with Star-Convex Polygons* in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II* (2018), 265–273.

63. Weigert, M., Schmidt, U., Haase, R., Sugawara, K. & Myers, G. *Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy* in *The IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020).

64. Weigert, M. & Schmidt, U. *Nuclei Instance Segmentation and Classification in Histopathology Images with Stardist* in *The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)* (2022).

65. Varrone, M., Tavernari, D., Santamaria-Martinez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics* **56.** Publisher: Nature Publishing Group, 74–84. ISSN: 1546-1718. https://www.nature.com/articles/s41588-023-01588-4 (2025) (Jan. 2024).


a
b
c
Workflow
python
Copy
View
TissueLab
File Edit View Window Help Reload Dash
TissueLab
A research platform developed by
Zhi Huang Lab
v. 2025-06
Dashboard NEW
Image Viewer
Model Zoo
Hide Sidebar
visium_hd_stardist
Minimize
Search files...
Select Folder
Files
Visium_HD_Human_Colon_...ped.tiff 7/1/2025
Visium_HD_Human_Prosta...ped.tiff 7/1/2025
Visium_HD_Human_Pancre...ped.tiff 7/1/2025
Visium_HD_Human_Kidney...ped.tiff 7/1/2025
Visium_HD_Human_Lung_C...ped.tiff 7/1/2025
Associated models under this folder
Expand
Control Panel:
Navigator
Main
Chat
Work-flow
Anno
H5 Data
Prefs
Z
Save PNG
Delete
Save
11859, 18405
(11481, 17649) - (12296, 18679)
30.33x
Visium_HD_Human_Prostate_C...

Help me to calculate the ratio of tumor vs. lymphocytes in this breast cancer tissue.

Here is the pipeline I designed for you:

- NucleiSeg
- NucleiClassify for `tumor_cell`, `lymphocyte`
- Scripts for `Compute ratio = (number of tumor_cell) / (number of lymphocyte); also report counts`

Write code that calculates the percent of tumor cells in the slide

I drafted a script based on your request.

```python
import os
import json
import h5py
import numpy as np
from datetime import datetime

def analyze_medical_image(path):
    """
    Calculate percent of tumor cell
```

Annotation

Mark all nuclei in this region as:
- ● Negative control
- ● Lymphocytes
- ● Tumor

Mark this region as tissue type:
- ● Negative control

Custom color: #00ff00

Custom annotation: Type your text here...

Nuclei Classification
ClassificationNode

Organ: Enter organ or leave blank if unknown

Cells: + New Class | Reset | Save | Load

Annotations:

Negative control	▨	48
Lymphocytes	🟩	70
Tumor	🟥	1072

Cell distribution (whole slide) Total cells: 200,602

57k 28% | 136k 68%

☑ Update after every annotation

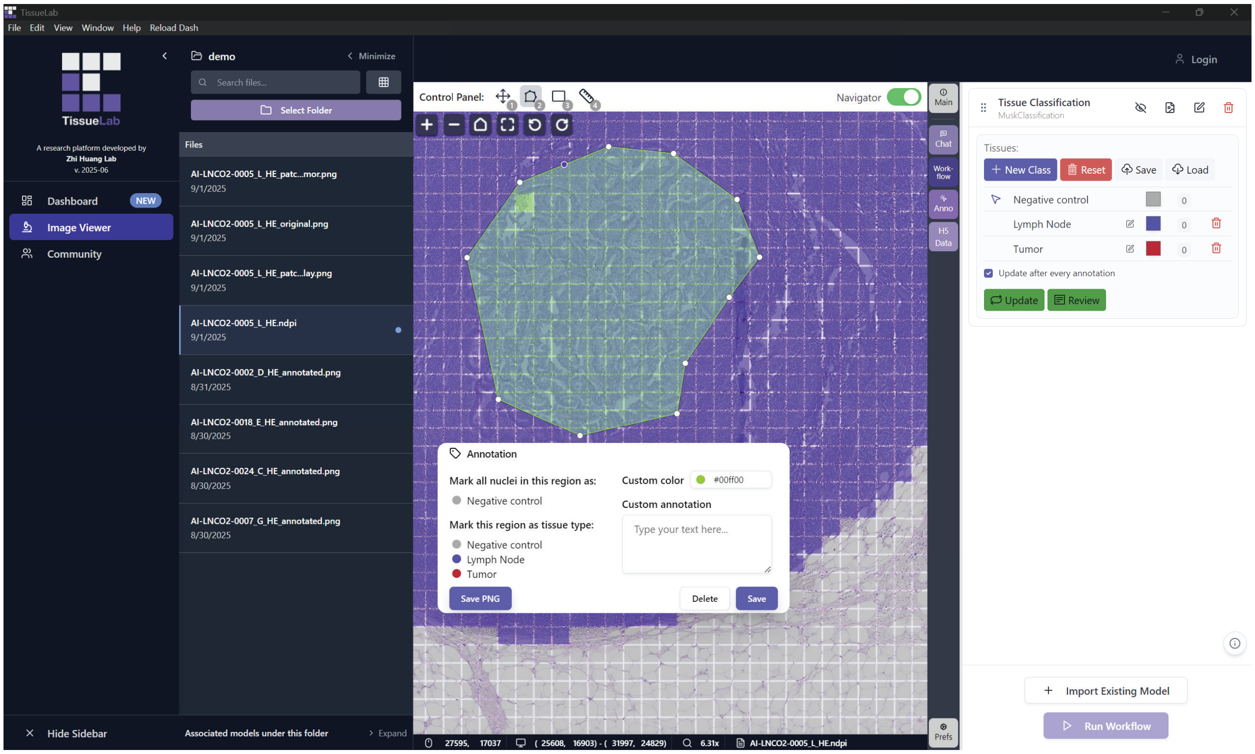
Update | Review

Completed

Workflow Status
- Nuclei Classification ✓

Not Started: 0 | Running: 0 | Completed: 1

Supplementary Figure 1: TissueLab agentic AI system in use. **a**, Conversational interface of the agentic AI system, in which the agent produces a corresponding workflow that can be executed to obtain the cell proportion. **b**, Code-inspection interface for reviewing and editing generated code. **c**, User interface showing a whole-slide image (WSI) with cell segmentation results, where the user performs active learning-based classification to refine model predictions.



Annotation

Mark all nuclei in this region as:

- ● Negative control

Mark this region as tissue type:

- ● Negative control
- ● Lymph Node
- ● Tumor

Custom color ● #00ff00

Custom annotation

Type your text here...

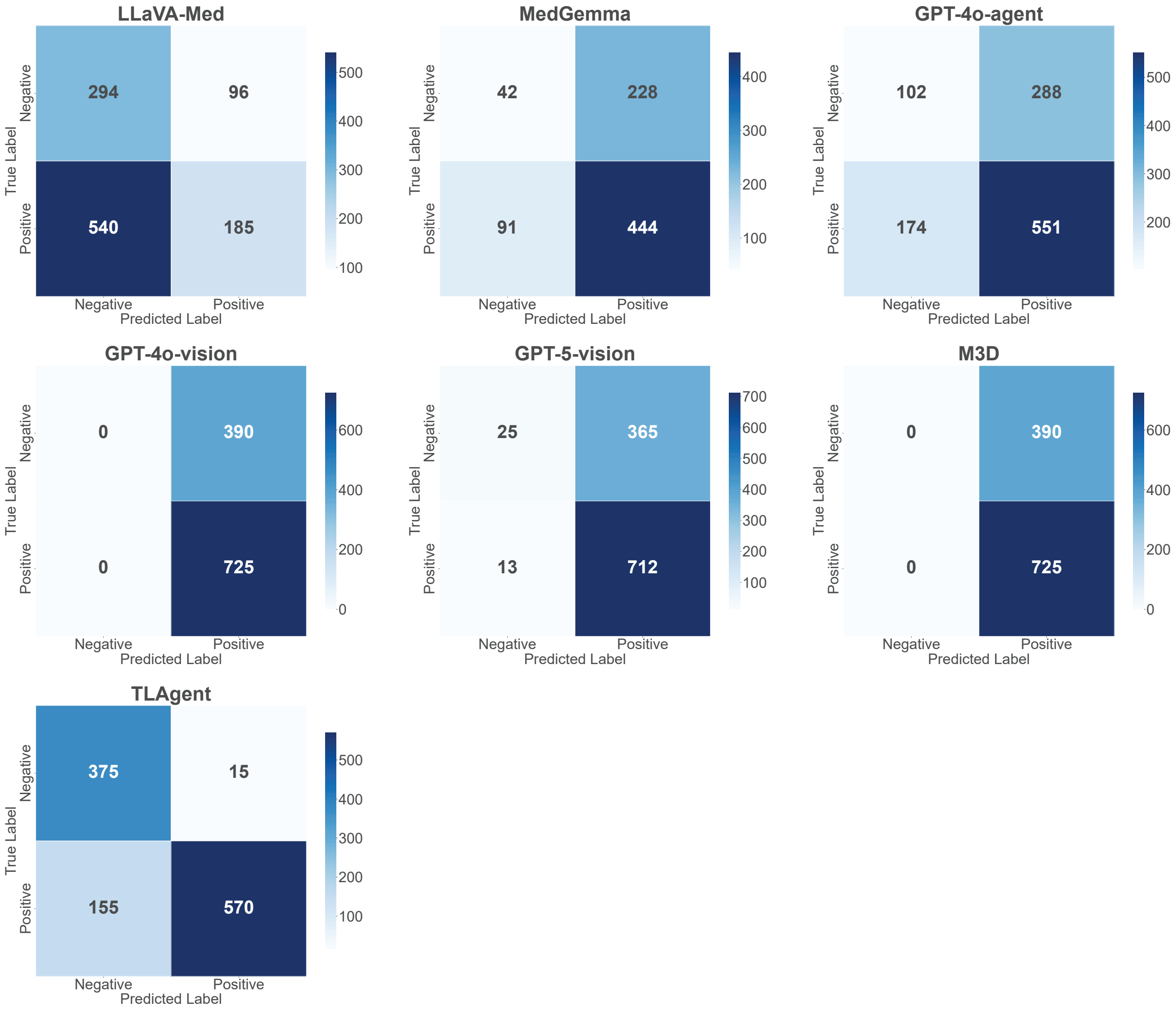
Tissues:

Negative control	▨	0	
Lymph Node	🟦	0	
Tumor	🟥	0	

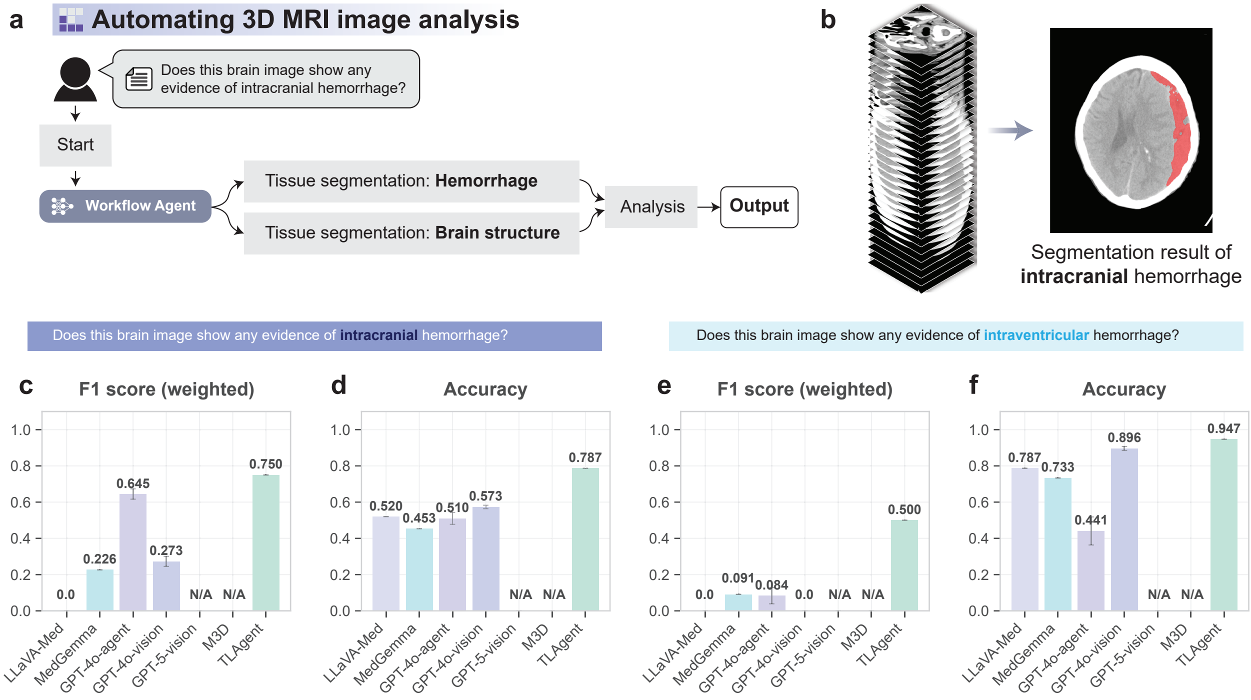
☑ Update after every annotation

27595, 17037 (25608, 16903) - (31997, 24829) 6.31x AI-LNCO2-0005_L_HE.ndpi

Supplementary Figure 2: Tissue segmentation interface. User interface showing a whole-slide image (WSI) with tissue segmentation results, where the user performs active learning-based classification to refine model predictions.



Supplementary Figure 3: Confusion matrices for fatty liver diagnosis across models. Shown are confusion matrices for LLaVA-Med, MedGemma, GPT-4o agent, GPT-4o vision, GPT-5 vision, M3D, and the TissueLab agentic AI system (TLAgent). While some baseline vision-language models appear to achieve reasonable overall metrics, their confusion matrices reveal systematic misclassifications and failure to identify the correct categories. In contrast, TLAgent demonstrates consistent and clinically reliable predictions across classes, underscoring the importance of evaluating performance not only by aggregate scores but also by class-wise diagnostic fidelity.



Supplementary Figure 4: Automating 3D MRI image analysis for intracranial hemorrhage. **a**, Workflow generated by the agent using segmentation of hemorrhage and brain structures as the basis for diagnosis, followed by code-based quantitative analysis. **b**, Visualization of segmentation results: the left panel shows the full stack of slices, while the right panel displays a representative slice with hemorrhage regions highlighted in red. **c**, F1-score comparison across models for the primary diagnostic question: *"Does this brain image show any evidence of intracranial hemorrhage?"* **d**, Accuracy comparison for the same primary diagnostic task. **e**, F1-score comparison for the fine-grained sub-question: *"Does this brain image show any evidence of intraventricular hemorrhage?"* **f**, Accuracy comparison for the sub-question task. Collectively, these panels show how the TissueLab agentic AI system automates end-to-end 3D MRI analysis-from workflow construction and visualization to quantitative benchmarking-while delivering more consistent and clinically reliable hemorrhage detection than baseline models.