

OPEN PROBLEMS & FUTURE DIRECTIONS ON AI BIOMEDICINE APPLICATIONS

Lei Xing, PhD, DABR

Jacob Haimson & Sarah S. Donaldson Professor & Director of Medical Physics Division

Department of Radiation Oncology & Electrical Engineering (by courtesy)
Institute of Computational and Mathematical Engineering (iCME) & Molecular Imaging Program (MIPS)
Stanford University, Stanford, California

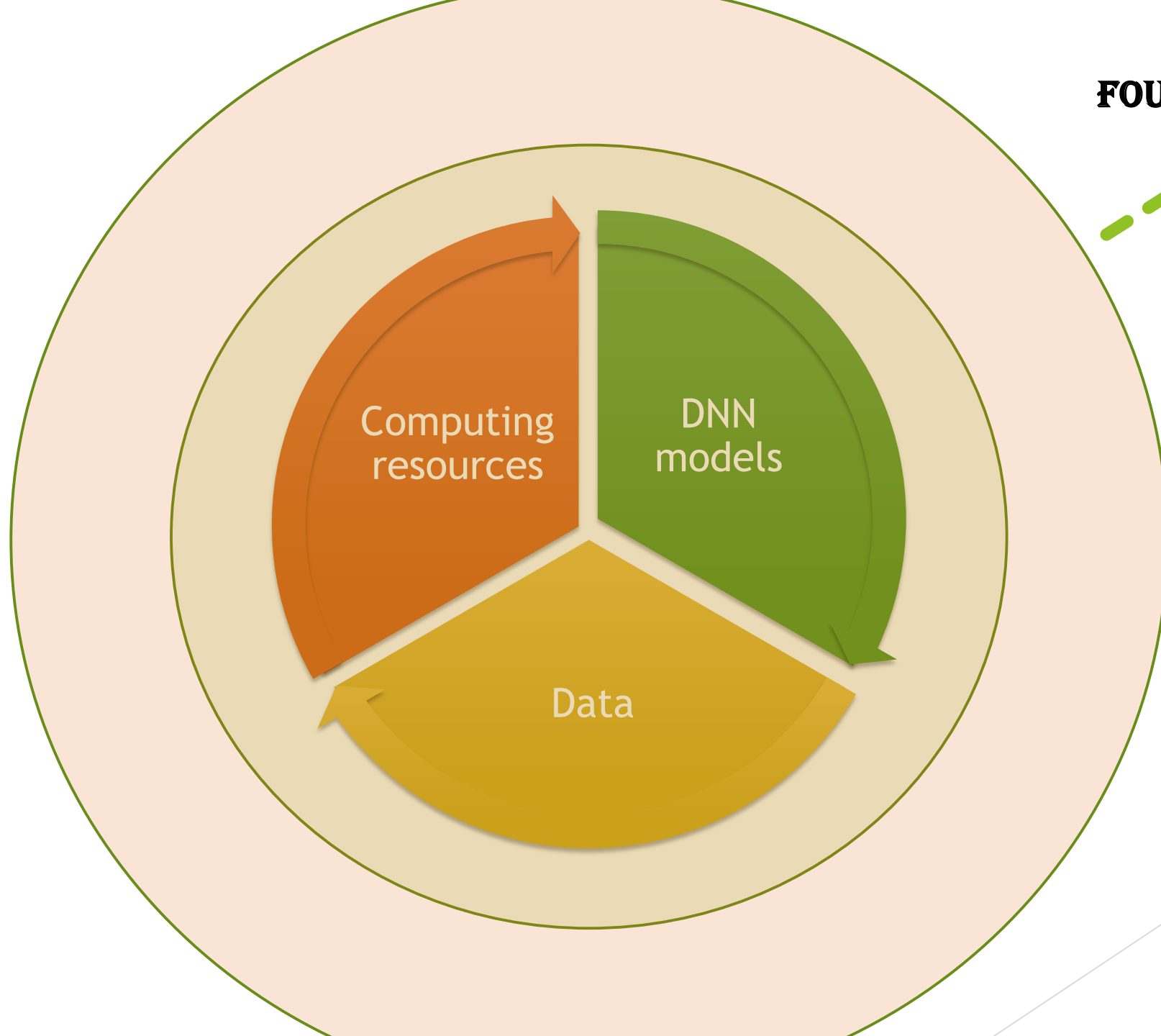
Stanford University

Department of Radiation Oncology
School of Medicine

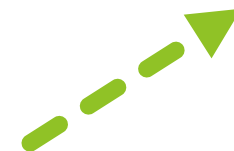


Problems & future research in AI

- ▶ Too brute-force
 - ▶ Too artificial
 - ▶ Deficient in reasoning & causal understanding
 - How to integrate symbolic reasoning with neural networks?
 - Can AI learn true cause-effect relationships from observational data?
 - How to enable counterfactual reasoning (“what if?” scenarios)?
- ▶ Generalization beyond i.i.d. data (learn from sparse, heterogenous data; generalize across tasks and domains without retraining)
 - ▶ Agentic AI, autonomous systems, multimodality, & world models
 - ▶ Gain true understanding instead of superficial pattern-matching
 - ▶ Robustness, ethics, & safety



FOUNDATION MODELS



Scaling Laws for Neural Language Models

Jared Kaplan *

Johns Hopkins University, OpenAI

jaredk@jhu.edu

Sam McCandlish*

OpenAI

sam@openai.com

Tom Henighan

OpenAI

henighan@openai.com

Tom B. Brown

OpenAI

tom@openai.com

Benjamin Chess

OpenAI

bchess@openai.com

Rewon Child

OpenAI

rewon@openai.com

Scott Gray

OpenAI

scott@openai.com

Alec Radford

OpenAI

alec@openai.com

Jeffrey Wu

OpenAI

jeffwu@openai.com

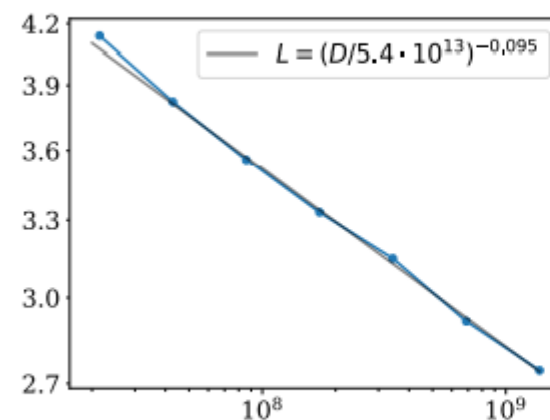
Dario Amodei

OpenAI

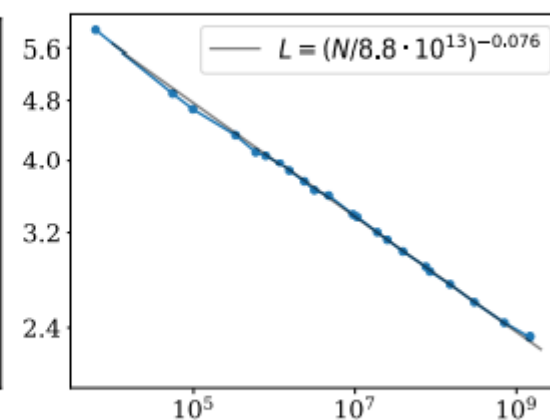
damodei@openai.com

Abstract

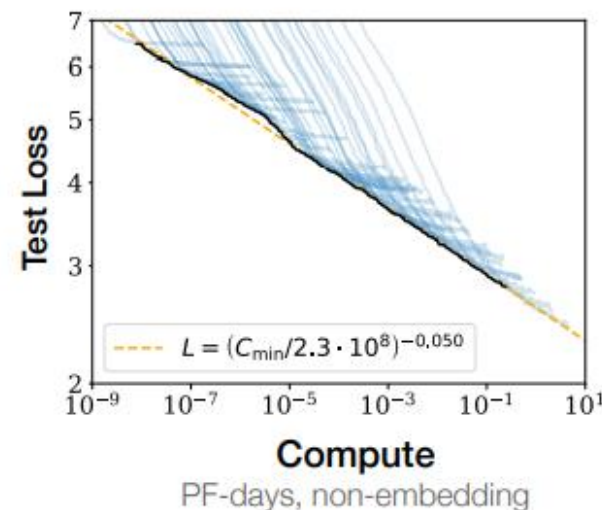
We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.



Dataset Size
tokens



Parameters
non-embedding

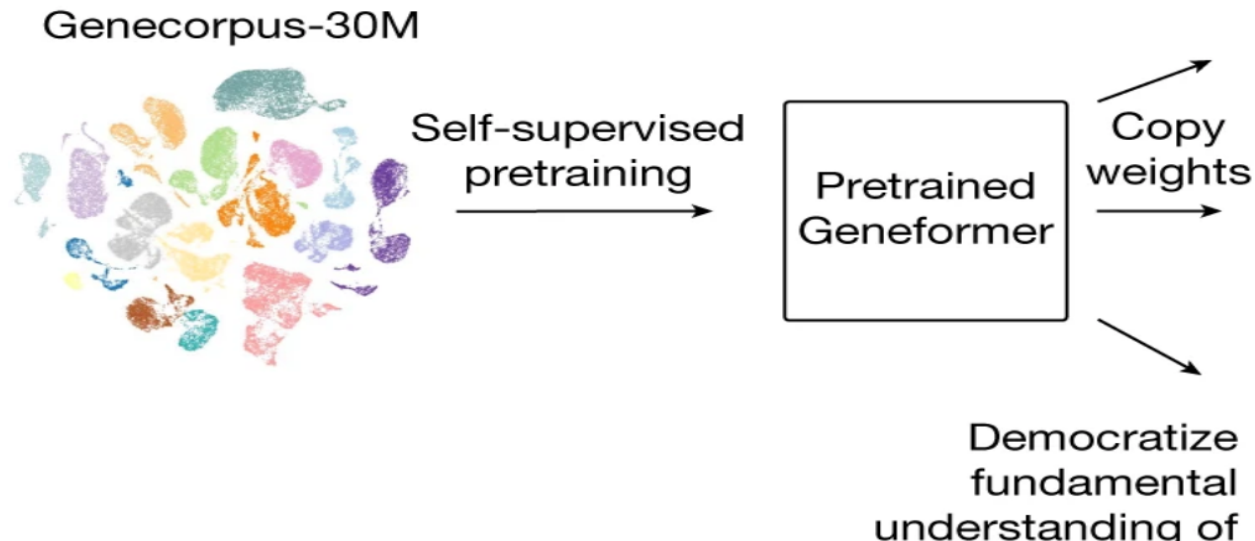


[nature](#) > [articles](#) > [article](#) > figure

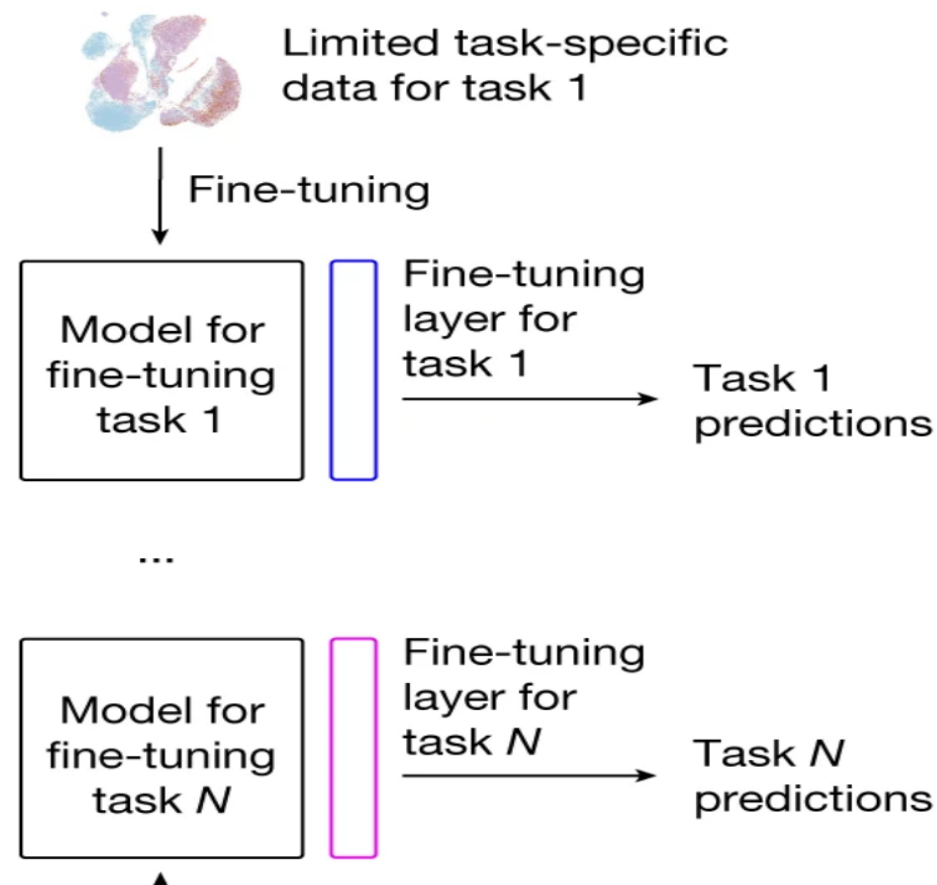
Fig. 1: Geneformer architecture and transfer learning strategy.

From: [Transfer learning enables predictions in network biology](#)

a Self-supervised large-scale pretraining



Fine-tuning with limited task-specific data



AI Foundation Models

- **Versatility:** One model can be adapted for multiple tasks (e.g., language translation, summarization, question answering).
- **Efficiency:** Reduces the need for developing and training models from scratch for different tasks.

The deployment of FMs need to be cautious - more research is necessary to realize the potential of FMs.

Assessing the limits of zero-shot foundation models in single-cell biology

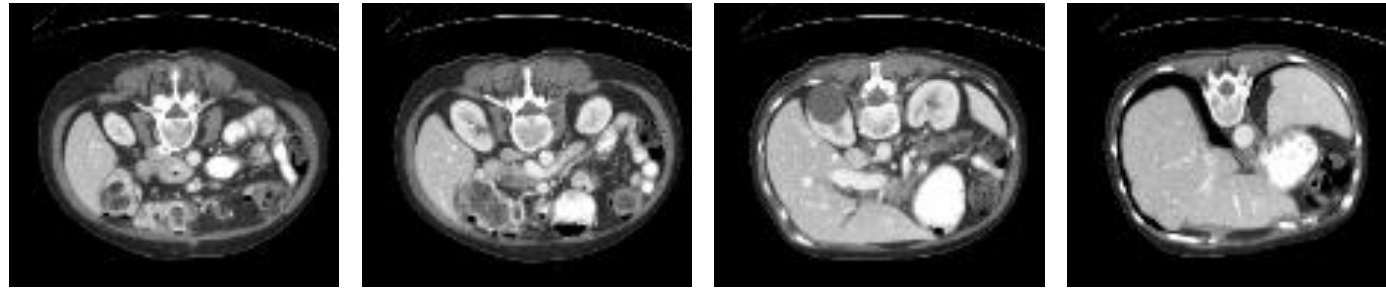
Kasia Z. Kedzierska^{1*}, Lorin Crawford², Ava P. Amini², Alex X. Lu²
¹University of Oxford, Oxford, UK; ²Microsoft Research, Cambridge, MA, USA
kasia@well.ox.ac.uk, {lcrawford, ava.amini, lualex}@microsoft.com

Abstract

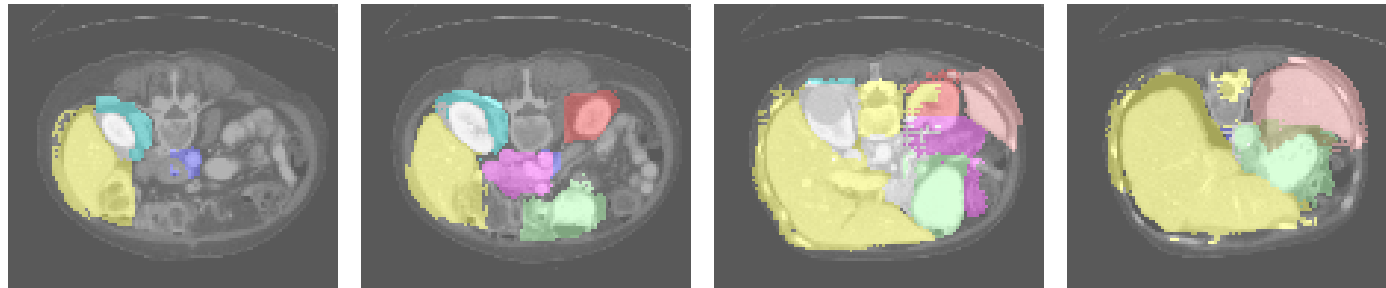
The advent and success of foundation models such as GPT has sparked growing interest in their application to single-cell biology. Models like Geneformer and scGPT have emerged with the promise of serving as versatile tools for this specialized field. However, the efficacy of these models, particularly in zero-shot settings where models are not fine-tuned but used without any further training, remains an open question, especially as practical constraints require useful models to function in settings that preclude fine-tuning (e.g., discovery settings where labels are not fully known). This paper presents a rigorous evaluation of the zero-shot performance of these proposed single-cell foundation models. We assess their utility in tasks such as cell type clustering and batch effect correction, and evaluate the generality of their pretraining objectives. Our results indicate that both Geneformer and scGPT exhibit limited reliability in zero-shot settings and often underperform compared to simpler methods. These findings serve as a cautionary note for the deployment of proposed single-cell foundation models and highlight the need for more focused research to realize their potential.²

Abdominal Organ Segmentation

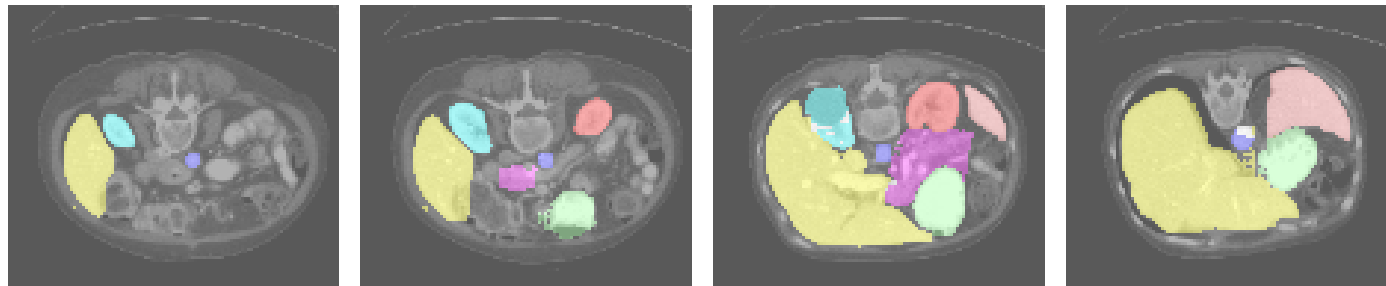
Image



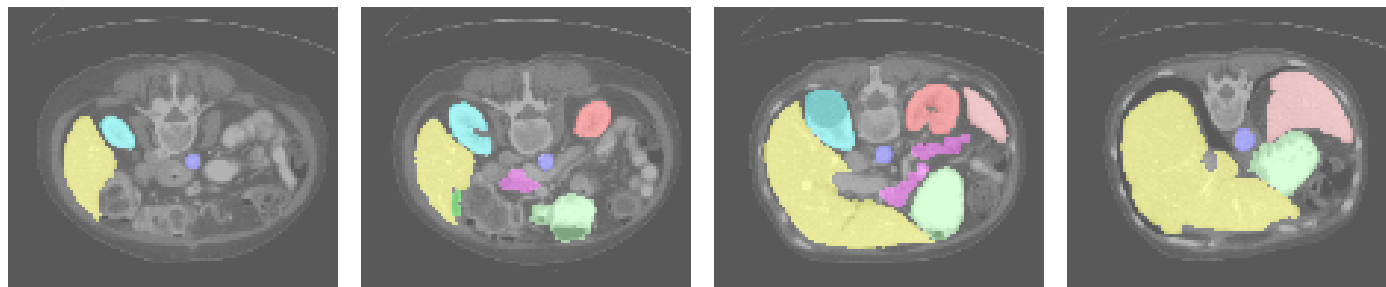
Med-SAM
Coarse box prompt



Med-SAM
Exact box prompt



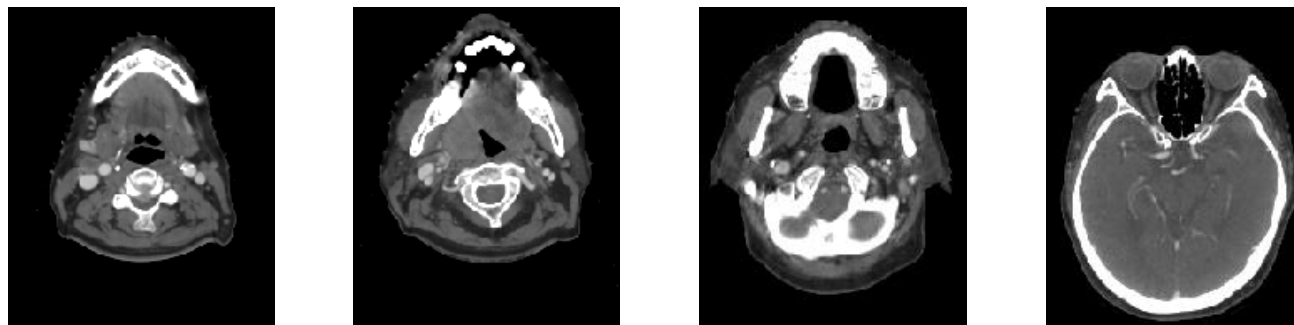
Label



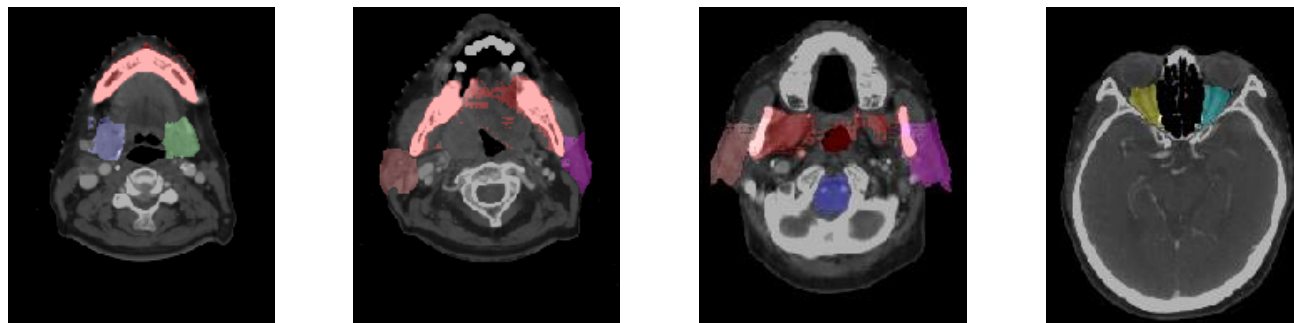
- Aorta
- Gallbladder
- Kidney_Left
- Kidney_Right
- Liver
- Pancreas
- Spleen
- Stomach

Head-and-neck Organ Segmentation

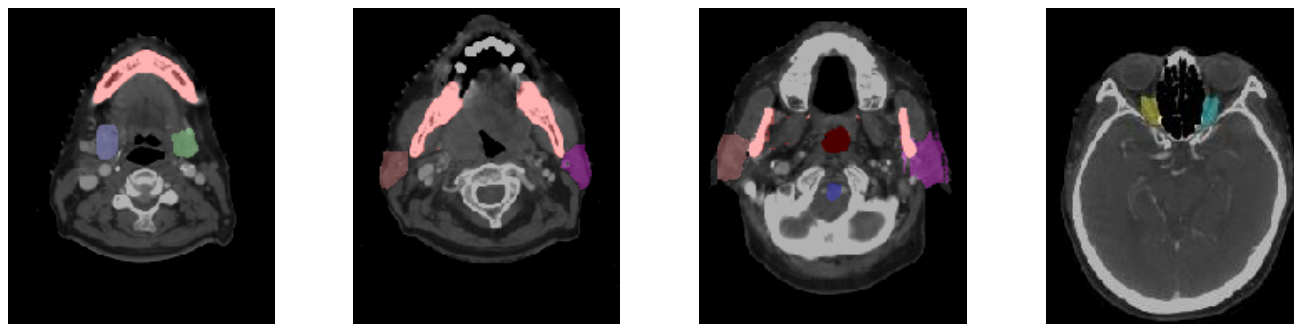
Image



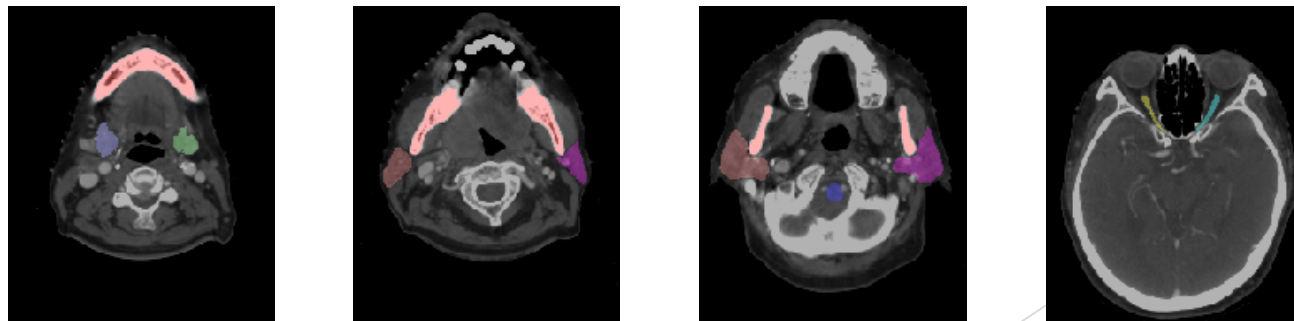
Med-SAM
Coarse box prompt



Med-SAM
Exact box prompt



Label



- Mandible
- Chiasm
- Brainstem
- Optic Nerve_Left
- Optic Nerve_Right
- Parotid_Left
- Parotid_Right
- Submandibular_Left
- Submandibular_Right



AI Foundation Models & Reinforcement learning from Human Feedback

- Fine-tuning
- prompting
- In-context learning
- Chain-of-Thought (CoT)

:2311.06668v2 [cs.LG] 16 Nov 2023

In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering

Sheng Liu[†], Lei Xing[†], James Zou[†]

[†] Stanford University

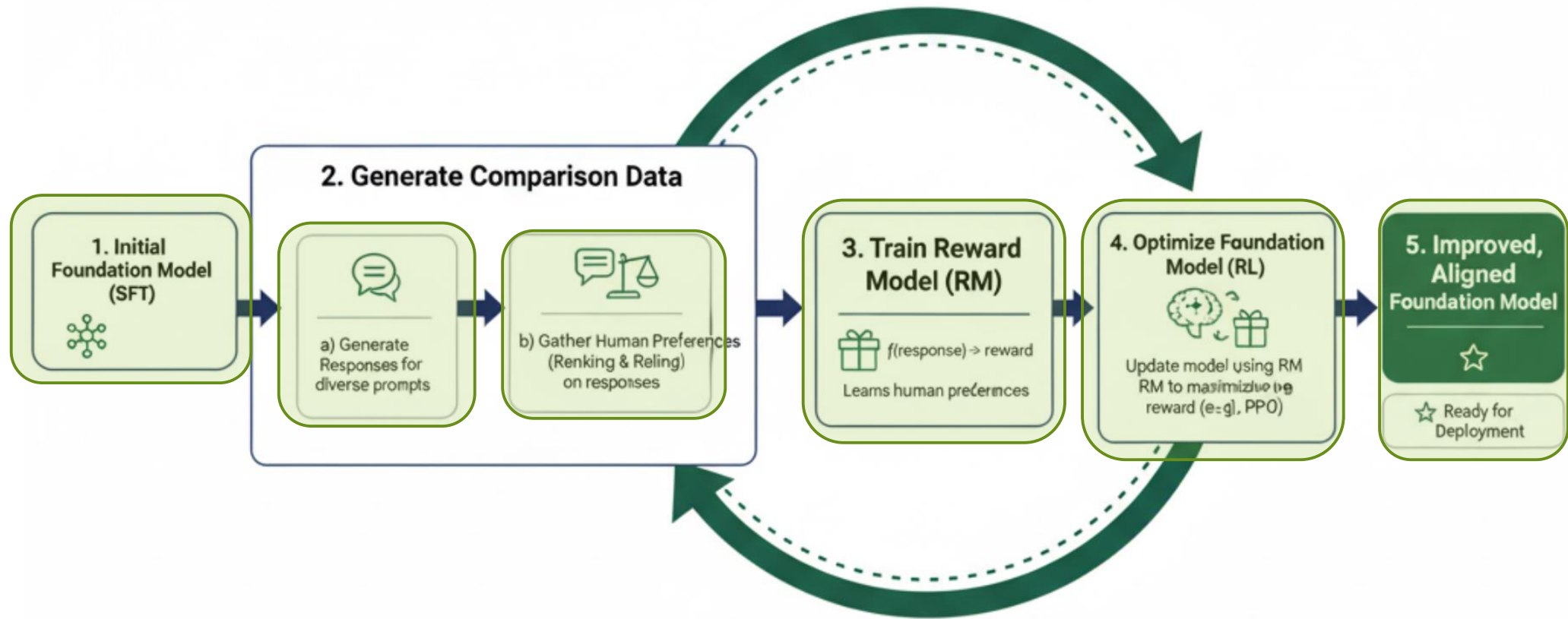
{shengl, lei, jamesz}@stanford.edu

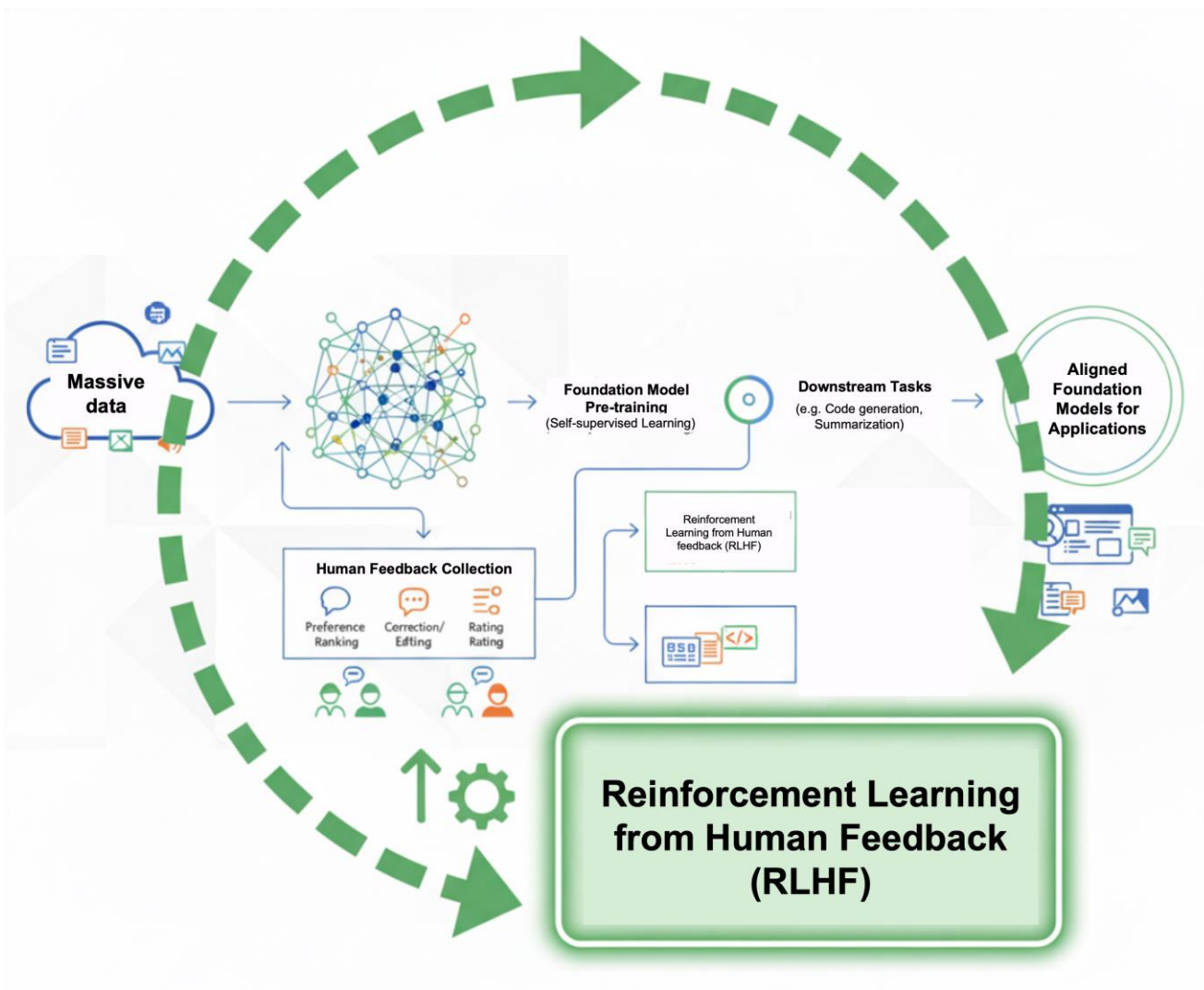
Warning: This paper includes examples and model-generated content that may be deemed offensive.

Abstract

Large language models (LLMs) demonstrate emergent in-context learning capabilities, where they adapt to new tasks based on example demonstrations. However, in-context learning has seen limited effectiveness in many settings, is difficult to quantitatively control and takes up context window space. To overcome these limitations, we propose an alternative approach that recasts in-context learning as *in-context vectors* (ICV). Using ICV has two steps. We first use a forward pass on demonstration examples to create the in-context vector from the latent embedding of the LLM. This vector captures essential information about the intended task. On a new query, instead of adding demonstrations to the prompt, we shift the latent states of the LLM using the ICV. The ICV approach has several benefits: 1) it enables the LLM to more effectively follow the demonstration examples; 2) it's easy to control by adjusting the magnitude of the ICV; 3) it reduces the length of the prompt by removing the in-context demonstrations; 4) ICV is computationally much more efficient than fine-tuning. We demonstrate that ICV achieves better performance compared to standard in-context learning and fine-tuning on diverse tasks including safety, style transfer, role-playing and formatting. Moreover, we show that we can flexibly teach LLM to simultaneously follow different types of instructions by simple vector arithmetics on the corresponding ICVs.

AI Foundation Models & Reinforcement learning from Human Feedback





nature

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 17 September 2025

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

[Daya Guo](#), [Dejian Yang](#), [Haowei Zhang](#), [Junxiao Song](#), [Peiyi Wang](#), [Qihao Zhu](#), [Runxin Xu](#), [Ruoyu Zhang](#), [Shirong Ma](#), [Xiao Bi](#), [Xiaokang Zhang](#), [Xingkai Yu](#), [Yu Wu](#), [Z. F. Wu](#), [Zhibin Gou](#), [Zhihong Shao](#), [Zhuoshu Li](#), [Ziyi Gao](#), [Aixin Liu](#), [Bing Xue](#), [Bingxuan Wang](#), [Bochao Wu](#), [Bei Feng](#), [Chengda Lu](#), ... [Zhen Zhang](#) [+ Show authors](#)

[Nature](#) **645**, 633–638 (2025) | [Cite this article](#)

199k Accesses | **4** Citations | **767** Altmetric | [Metrics](#)

Abstract

General reasoning represents a long-standing and formidable challenge in artificial intelligence (AI). Recent breakthroughs, exemplified by large language models (LLMs)^{1,2} and chain-of-thought (CoT) prompting³, have achieved considerable success on foundational reasoning tasks. However, this success is heavily contingent on extensive human-annotated demonstrations and the capabilities of models are still insufficient for more complex problems. Here we show that the reasoning abilities of LLMs can be incentivized through pure reinforcement learning (RL), obviating the need for human-labelled reasoning trajectories. The proposed RL framework facilitates the emergent development of advanced reasoning patterns, such as self-reflection, verification and dynamic strategy adaptation. Consequently, the trained model achieves superior performance on verifiable tasks such as mathematics, coding competitions and STEM fields, surpassing its counterparts trained through

Dealing with increased complexity and multimodality data integration (data are expensive, but paired data are even more so). com

- ▶ Vision-language models
- ▶ Multi-omics integration
- ▶ Multimodality imaging
- ▶ ...Multimodal

Spatiotemporal signal analysis

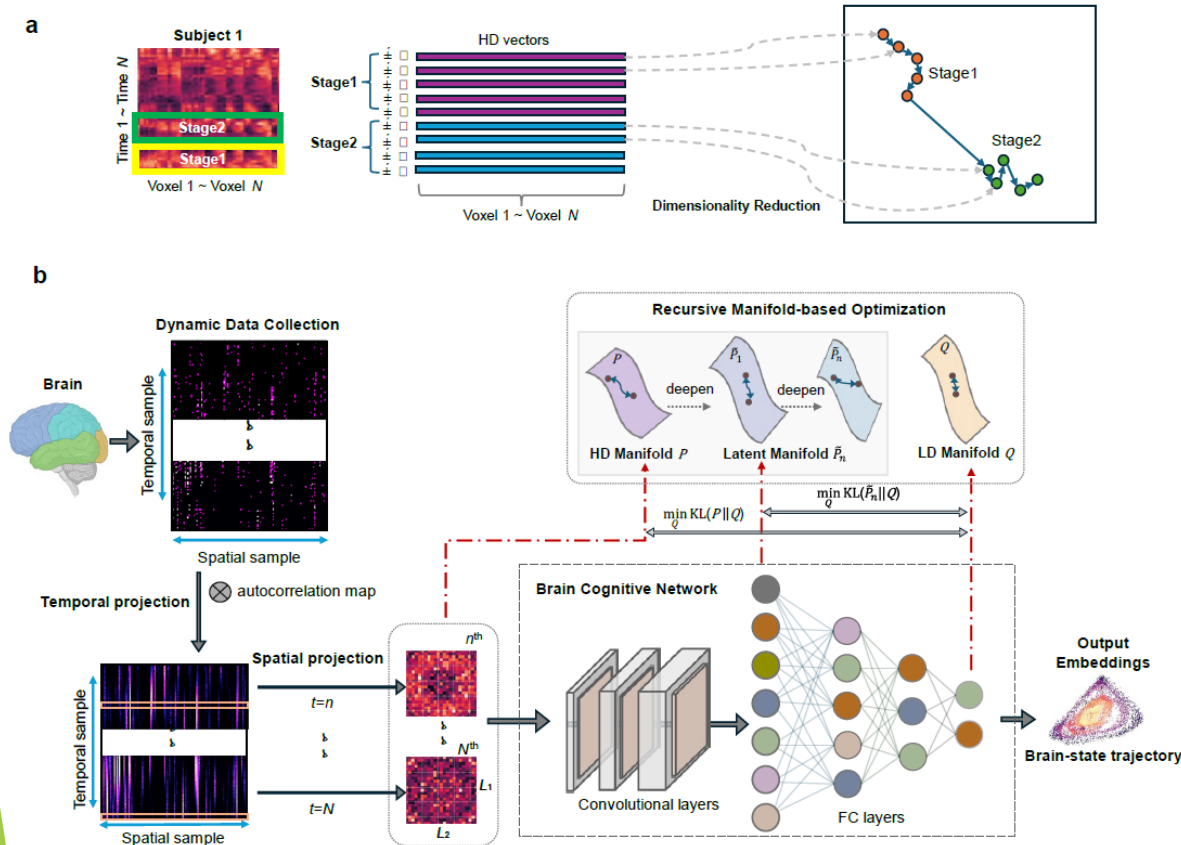


Figure 1. A schematic overview of the BCNE approach. **a.** A brief overview on how dimensionality reduction generates time trajectories, revealing individual brain responses during similar stage transitions. **b.** BCNE calculation process. Initially, the dynamic brain data are input into the temporal and spatial projection modules to generate reconfigured correlative data. Subsequently, BCN is applied to the preprocessed data, learning brain cognitive mapping by minimizing the KL divergence loss between the pairwise distance matrices built on the high-dimensional (HD) and low-dimensional manifolds. Finally, a recursive optimization strategy is employed to gain dynamic insights. This strategy involves recursively replacing the pairwise distance matrix P built on the HD manifold with pairwise distance matrices $\bar{P}$ based on latent manifolds of increasing depth.

Nature Computational Science, in press, 2025

Revealing Neurocognitive Patterns by Self-Supervised Learning of Manifold Embedding from Dynamic Brain Data

Zixia Zhou¹, Junyan Liu¹, Wei Emma Wu¹, Sheng Liu¹, Qingyue Wei¹, Rui Yan¹, Yi Guo², Qian Tao³, Yuanyuan Wang², Md Tauhidul Islam^{1*}, and Lei Xing^{1*}

¹ Department of Radiation Oncology, Stanford University School of Medicine, Stanford, California, 94305, USA

² Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China

³ Department of Imaging Physics, Delft University of Technology, Lo-rentzweg 1, 2628 CJ Delft, the Netherlands

Abstract

Dynamic brain data, teeming with biological and functional insights, are becoming increasingly accessible through advanced measurements, providing a gateway to understanding the inner workings of the brain in living subjects. However, the vast size and intricate complexity of the data also pose a daunting challenge in reliably extracting meaningful information across various data sources. This paper introduces a generalizable self-supervised deep manifold learning for exploration of dynamic patterns. Unlike existing methods that extracting patterns directly from the input data as in the existing methods, the proposed brain cognitive network embedding (BCNE) seeks to capture the brain-

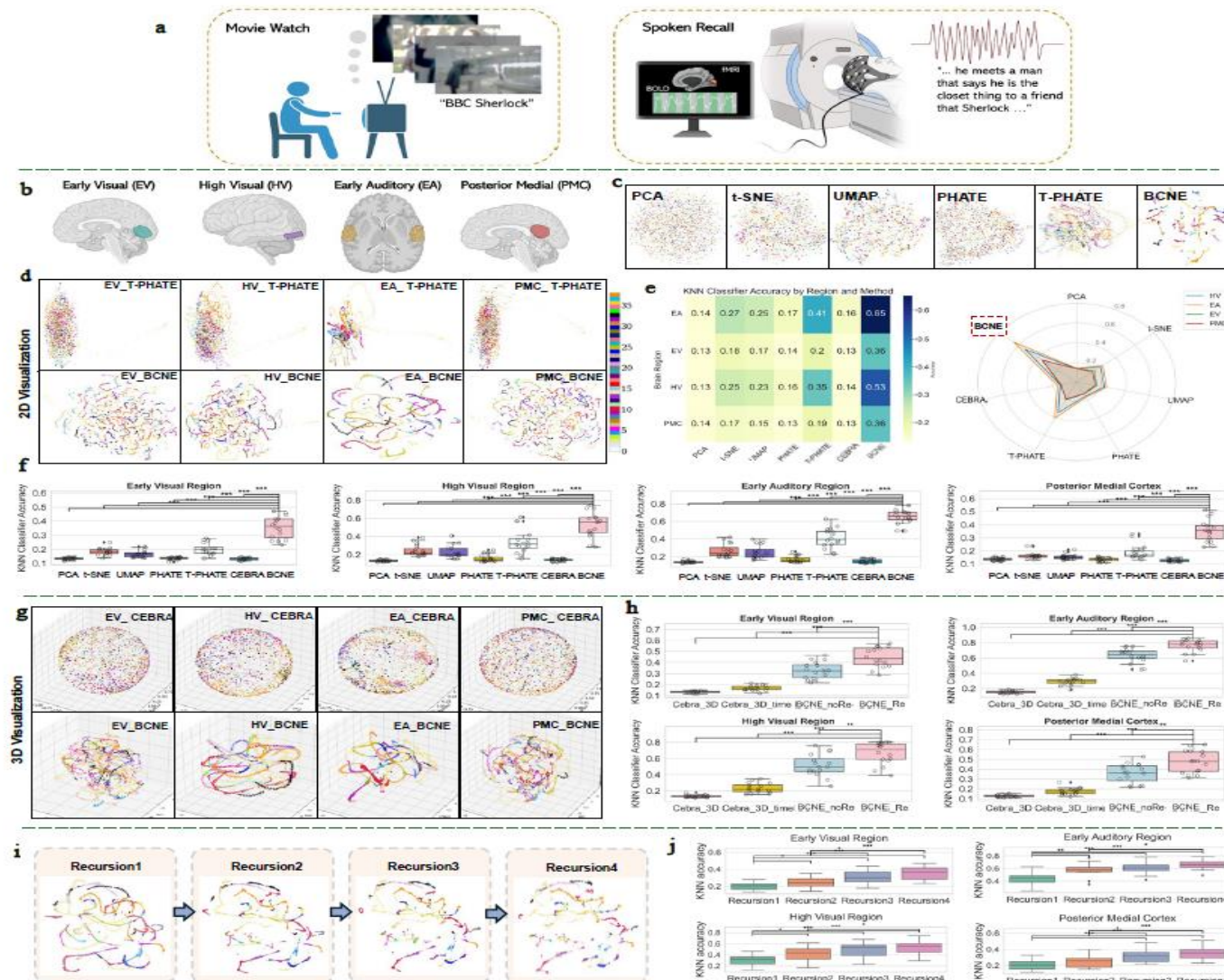
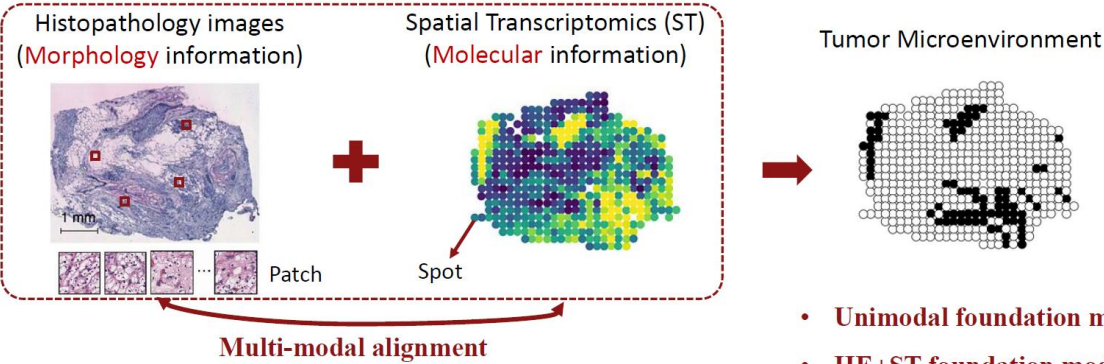


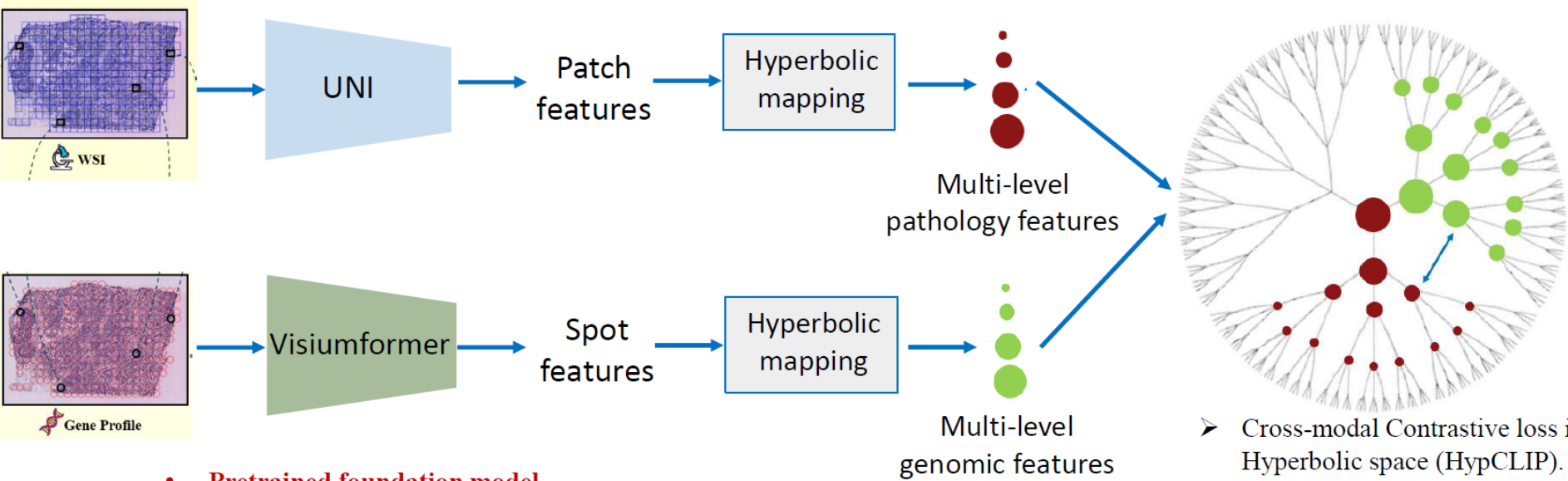
Figure 2. Results of the Sherlock fMRI dataset. (a) Illustration of the Sherlock dataset acquisition process. Sixteen subjects were tested, who watched the BBC Sherlock movie, then recalled the movie and verbalized their memories. (b) The selected ROIs for separate embedding and visualization. (c) Visualization comparison of hippocampal ventral (HV) region data generated by PCA, t-SNE, UMAP, PHATE, T-PHATE, and BCNE. (d) 2D visualization comparison between T-PHATE and BCNE for entorhinal ventral (EV), hippocampal ventral (HV), entorhinal anterior (EA), and posterior medial ventral (PMV) regions. (e, f) Heatmaps, radar maps, and boxplots for KNN classifier accuracy for the 39-event classification task calculated by embeddings generated by PCA, t-SNE, UMAP, PHATE, T-PHATE, CEBRA (2D), and BCNE. (g) 3D visualization comparison between CEBRA and BCNE for EV, HV, EA, PMV regions. (h) Boxplots for KNN classifier accuracy calculated by embeddings generated by CEBRA (3D), CEBRA (3D with temporal correlation projection), BCNE (no recursion), and BCNE. (i) 2D visualizations generated by BCNE with varying recursion depths. (j) Boxplots for KNN classifier accuracy calculated by embeddings generated by BCNE with different recursion depths. Statistical significance was indicated by asterisks: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



Histopathology + Spatial Transcriptions (ST)



- Unimodal foundation model ✓
- HE+ST foundation model ?



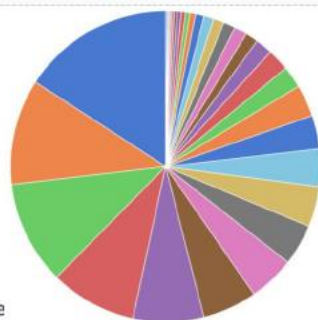
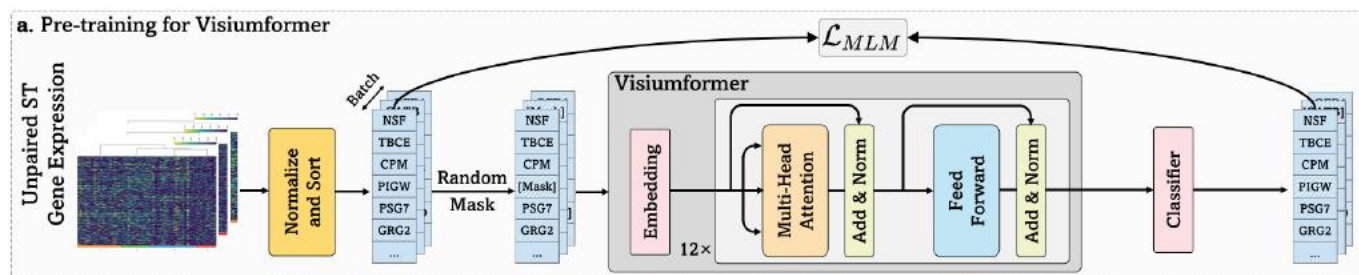
- Pretrained foundation model
- A unified tokenization method

- Cross-modal Contrastive loss in the Hyperbolic space (HypCLIP).
- Intra-modal hierarchical loss.

Visiumformer

- **Architecture:** Transformer backbone with 12 blocks.
- **Training data:** 3.94 million ST spots from 1,363 slices, span across 30 tissue/organ types.
- **Tokenization:** Sort the gene expressions and look for the gene index from a vocabulary including 20,310 genes.
- **Pre-training:** Masked training strategy.

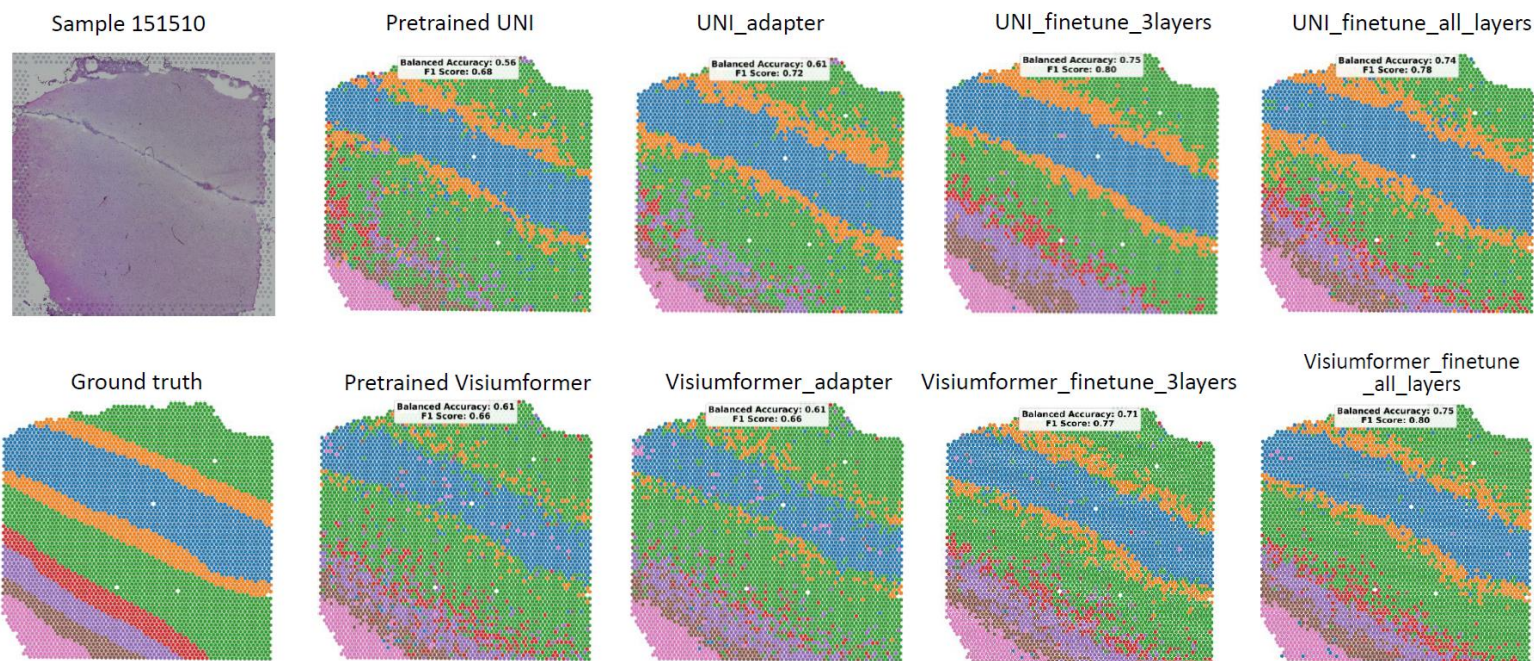
$$T_i = \{id(ep_i^0), id(ep_i^1), \dots, id(ep_i^n) : ep_i^k \geq ep_i^{k+1}\}$$



Tissue/Organ	Count	Percentage
Brain	215	(15.9%)
Lung	149	(11.0%)
Skin	146	(10.8%)
Breast	122	(9.0%)
Kidney	100	(7.4%)
Liver	77	(5.7%)
Pancreas	65	(4.8%)
Prostate	58	(4.3%)
Ovary	57	(4.2%)
Spinal cord	53	(3.9%)
Bowel	47	(3.5%)

Tissue/Organ

Experimental results – Linear Probing Visualization



Multi-Sensor Learning Enables Information Transfer Across Different Sensory Data and Augments Multi-Modality Imaging

L. Zhu, Y. Chen, L. Liu, L. Xing*, L. Yu*, IEEE T-PAMI, 2024.

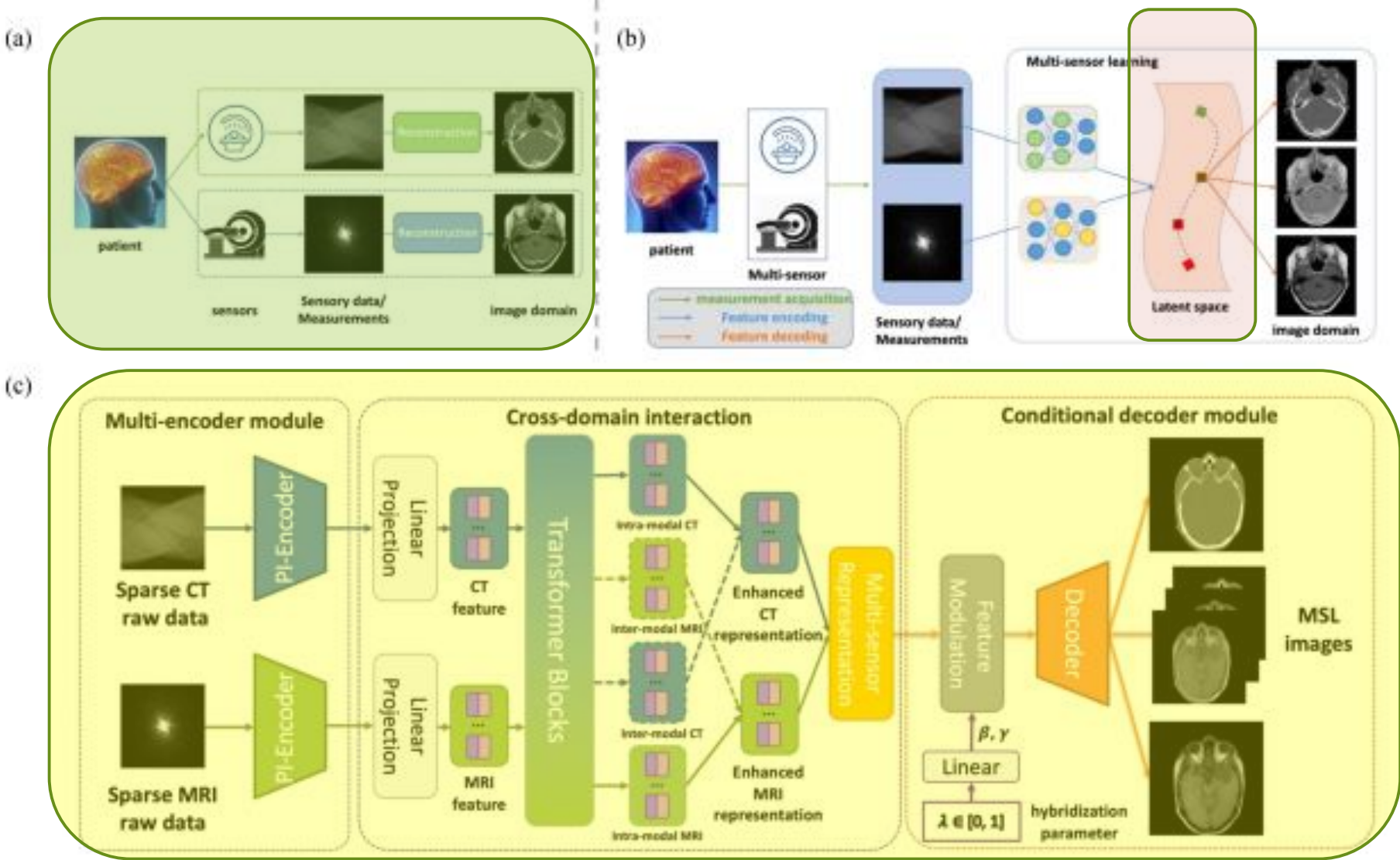


Fig. 1. Data-driven multi-modality imaging with multi-sensor learning. (a) Conventional approach reconstructs independent images of different modalities and fuses them post hoc without any information transfer across different modalities. (b) The proposed DMI strategy enables synergetic imaging by sharing inter-modality features in a latent representation space. (c) Schematic diagram of our MSL framework, featuring a multi-domain encoder-decoder design and a

Multi-sensor Learning Enables Information Transfer across Different Sensory Data and Augments Multi-modality Imaging

L. Zhu, Y. Chen, L. Liu, L. Xing*, L. Yu*, IEEE T-PAMI, 2024.

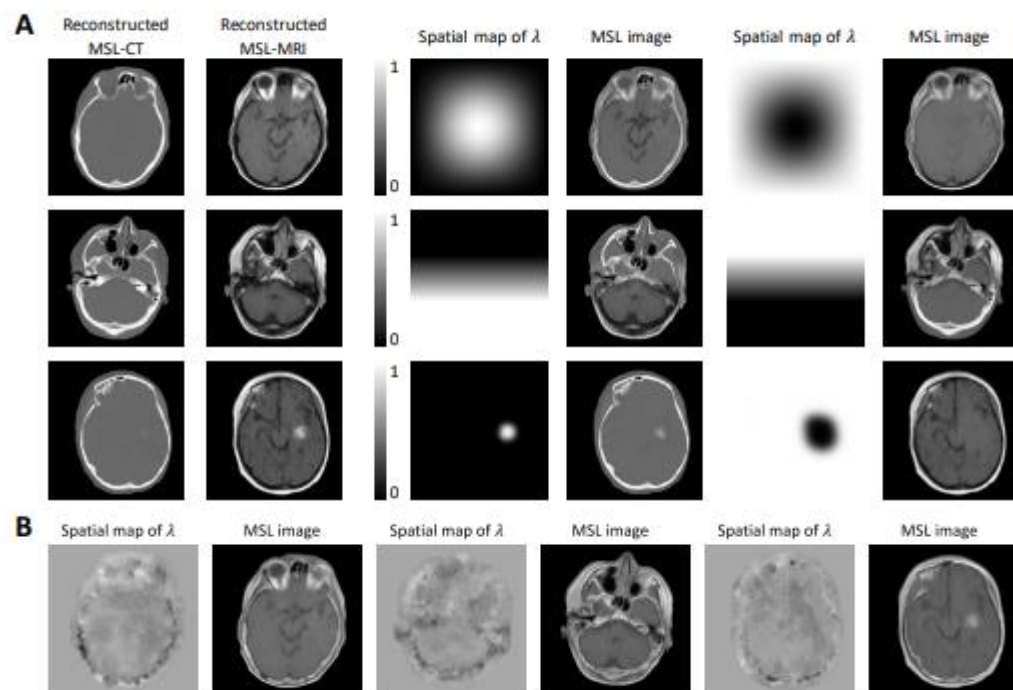
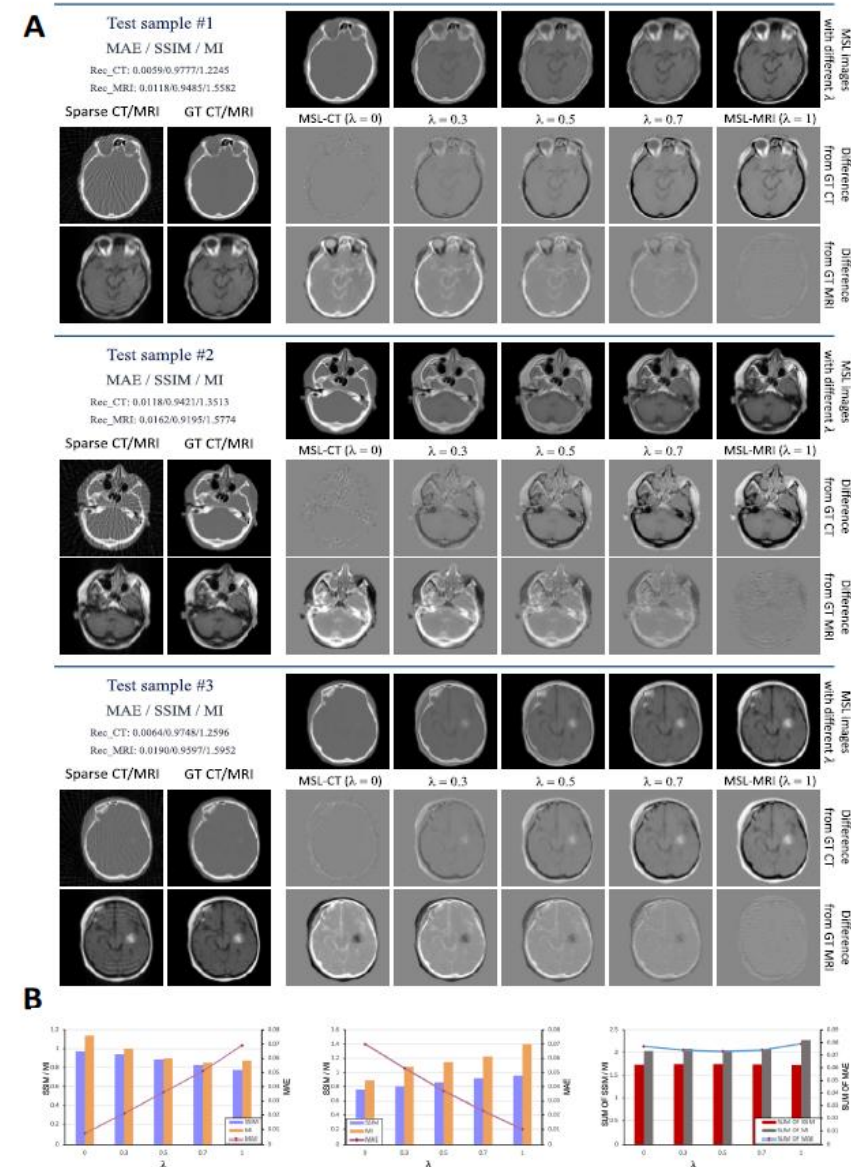
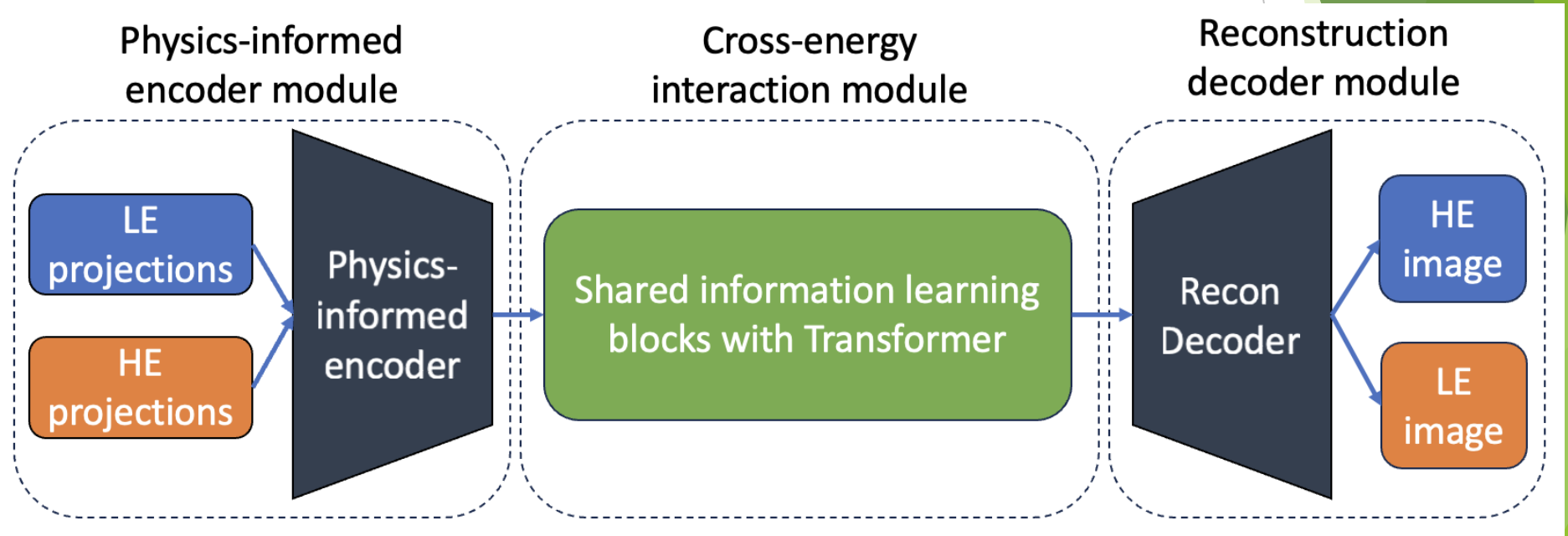


Fig. 3. **MSL imaging with local contrast tuning.** (A) The left panel displays the MSL-CT ($\lambda = 0$) and MSL-MRI ($\lambda = 1.0$) images for three different samples. The corresponding MSL images generated with six different spatial λ distributions are shown in the middle and left panels (where the white and black denote $\lambda=1$ and 0, respectively). (B) The optimized spatial maps of λ and the corresponding MSL images for the three samples shown in A are presented.



Qualitative and quantitative analyses of the MSL images. (A) MSL imaging results of three different samples. Sparse images from conventional reconstruction algorithms, ground truth (GT) images, MSL images, and difference images between the MSL image and GT CT and MRI images are shown. (B) Quantitative evaluation of MSL imaging. The MAE, SSIM, and MI metrics of the MSL image with respect to the GT CT and MRI images at different λ are shown in the left and middle panels, respectively. The right panel shows the summation of the paired metrics, which remains almost unchanged as λ changes.

Multisensor learning (MSL) for dual energy CT

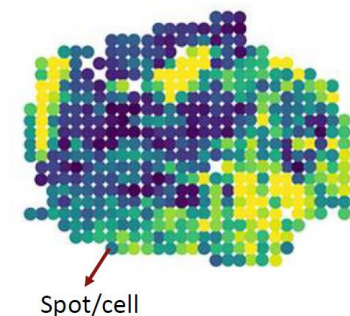
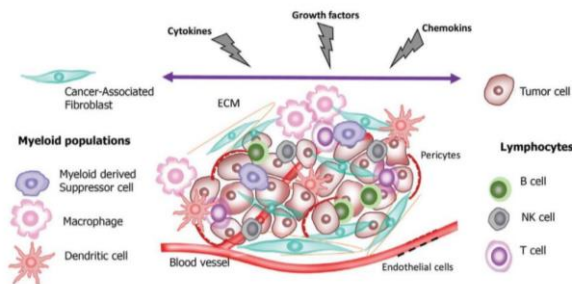
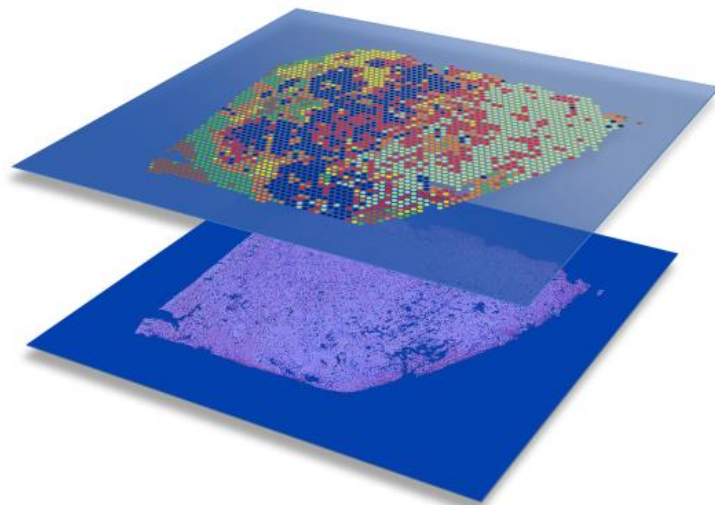


- Dual energy CT data for training
- Single energy CT data for testing

L. Zhu, Y. Chen, L. Liu, L. Xing*, L. Yu*, IEEE T-PAMI, 2024.

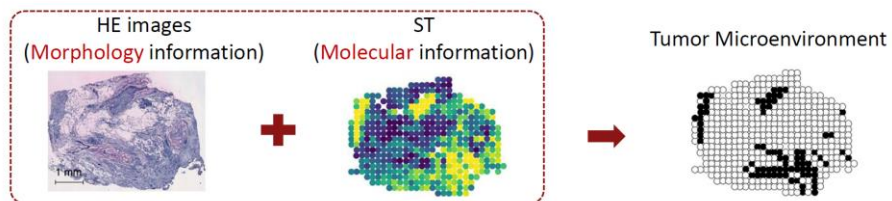
A. Li, L. Zhu, Y. Chen, L. Liu, L. Xing*, L. Yu*, AAPM 2025.

Spatial transcriptomics



<https://www.scdiscoveries.com/blog/knowledge/spatial-transcriptomics-5-key-applications/>

HE + ST integration

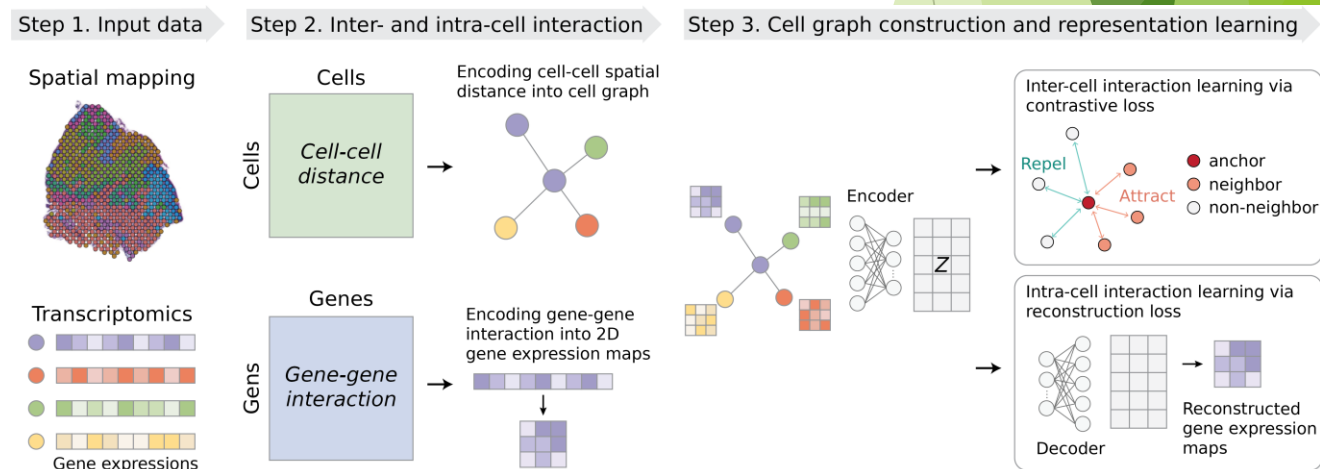


Spot-level Tasks (Biology):

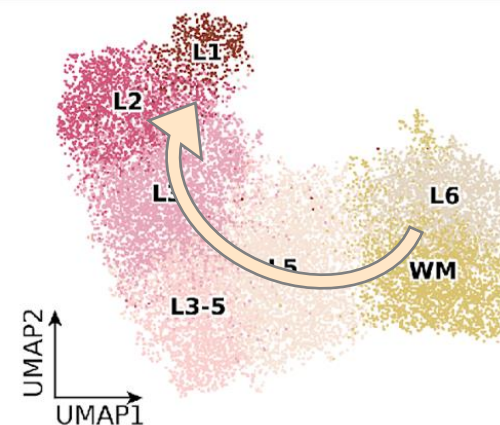
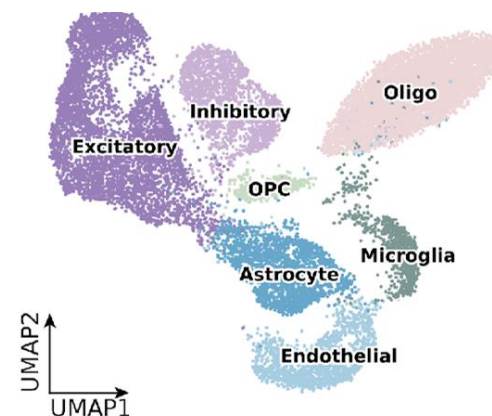
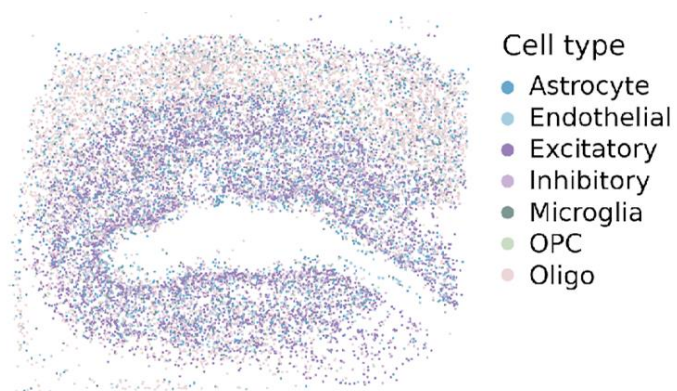
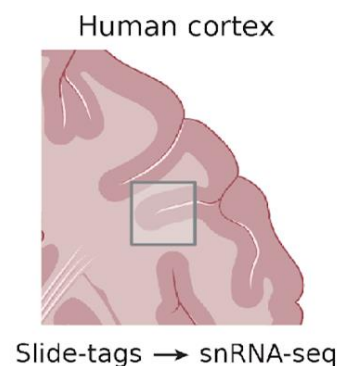
- Spatial domain detection.
- ST super-resolution.
- Cell type decomposition/deconvolution.
-

Patient-level Tasks (Clinical):

- Cancer subtyping.
- Gene mutation prediction.
- Response to immunotherapy.
-



Cell type and spatial domain identification



THE KEY ETHICAL CHALLENGES

- AI is not just a technological tool; it's a social and economic force.
- Ethical issues arise from the entire lifecycle of AI: from framework design, data curation, model training, to deployment and impact.
- **Key areas of concern are:**
 1. **Bias & Fairness:** The risk of perpetuating or amplifying societal biases.
 2. **Privacy & Data Security:** The use of vast amounts of data and the potential for misuse.
 3. **Accountability & Transparency:** The "black box" problem of understanding and assigning responsibility for AI decisions.
 4. **Socioeconomic Impact:** The effects on employment, inequality, and human agency.

ENHANCING ACCOUNTABILITY & TRANSPARENCY OF AI MODELS

- **Explainable AI (XAI) Research:**

- **Post-hoc Explanations:** Develop methods to explain the decisions of complex "black box" models *after* they've made a prediction (e.g., SHAP, LIME).
- **Interpretable Models:** Design inherently interpretable AI models (e.g., decision trees, linear models) for high-stakes applications where transparency is paramount.
- **Visualizations & User Interfaces:** Create intuitive interfaces that help users understand AI reasoning, even for complex models.

RELATIONSHIP

PERTURBATIONS

FEATURE VISUALIZATION

INCLUSION OF PRIORS

FDA'S AI SAFETY INITIATIVES

The U.S. **Food and Drug Administration (FDA)** announced its extended public list of **AI-enabled medical devices**, increasing transparency and awareness.

To increase openness and understanding regarding regulated artificial intelligence (AI) technology in healthcare, the U.S. Food and Drug Administration (FDA) published an enlarged public list of AI-enabled medical devices on **July 10, 2025**.

The FDA defines **AI-enabled devices** as those leveraging artificial intelligence or machine learning (AI/ML) for one or more functions integral to clinical care. These devices are either classified as software intended to operate as a medical device (Software as a Medical Device, or SaMD) or are embedded within broader systems.

The AI-Enabled Medical Device List is not merely a catalogue—it is a strategic tool developed to serve multiple stakeholders:

- **For developers and digital health innovators:** The list offers visibility into how the FDA is evaluating AI technologies, helping shape future innovations and regulatory submissions.
- **For hospitals and healthcare providers:** It enhances transparency in medical AI, enabling better-informed decisions when adopting new technologies.
- **For patients and the public:** It clearly signals when medical devices incorporate AI, fostering trust in safety, accuracy, and oversight.

Emphasis on Transparency, Lifecycle, and Foundation Models



Article

<https://doi.org/10.1038/s41467-023-43958-w>

Revealing hidden patterns in deep neural network feature space continuum via manifold learning

Received: 17 May 2023

Accepted: 24 November 2023

Published online: 21 December 2023

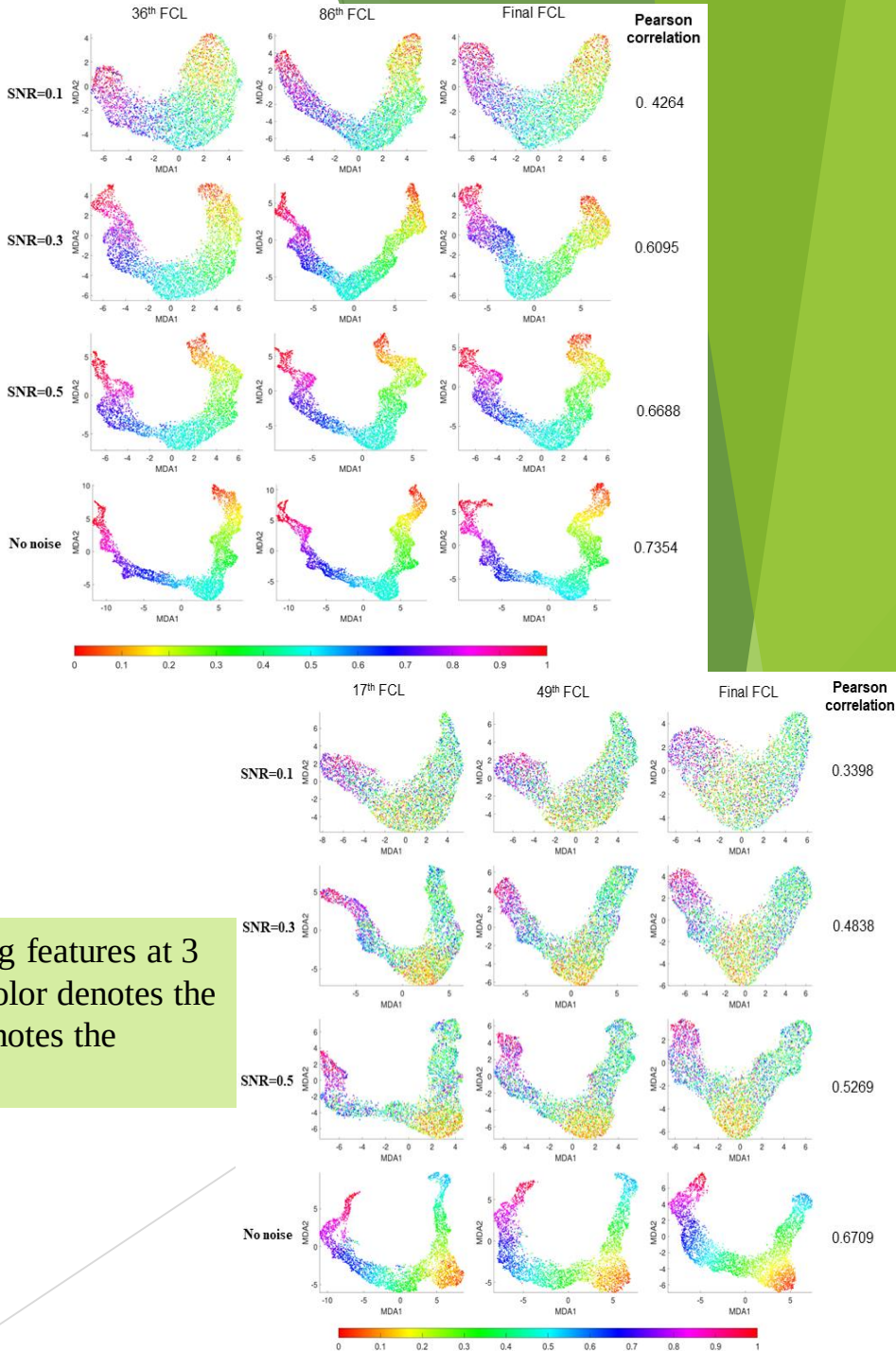
Check for updates

Md Tauhidul Islam^{1,4}, Zixia Zhou^{1,4}, Hongyi Ren¹, Masoud Badiei Khuzani¹, Daniel Kapp¹, James Zou², Lu Tian², Joseph C. Liao³✉ & Lei Xing¹✉

Deep neural networks (DNNs) extract thousands to millions of task-specific features during model training for inference and decision-making. While visualizing these features is critical for comprehending the learning process and improving the performance of the DNNs, existing visualization techniques work only for classification tasks. For regressions, the feature points lie on a high dimensional continuum having an inherently complex shape, making a meaningful visualization of the features intractable. Given that the majority of deep learning applications are regression-oriented, developing a conceptual



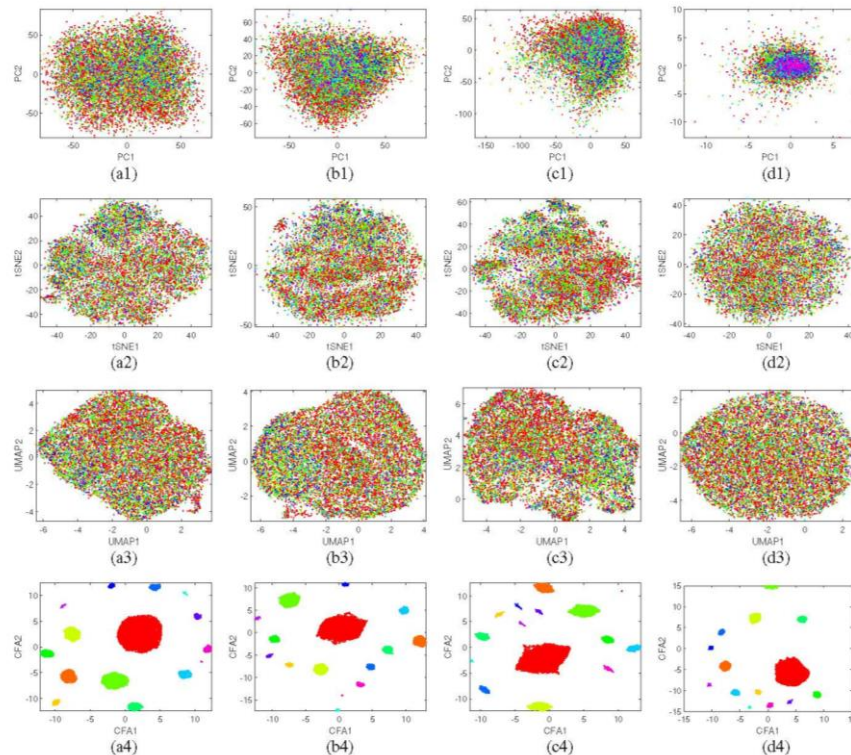
MDA visualization of GPT-2 and BERT finetuning features at 3 different layers at different SNRs. Red to violet color denotes the survival days from 0 to 10,000 days. Colorbar denotes the normalized survival days.



Nature Communications 14, 8506, 2023

Islam M.T., Xing L, "Deciphering the feature representation of deep neural networks for high performance biomedical AI", TPAMI, 2024.

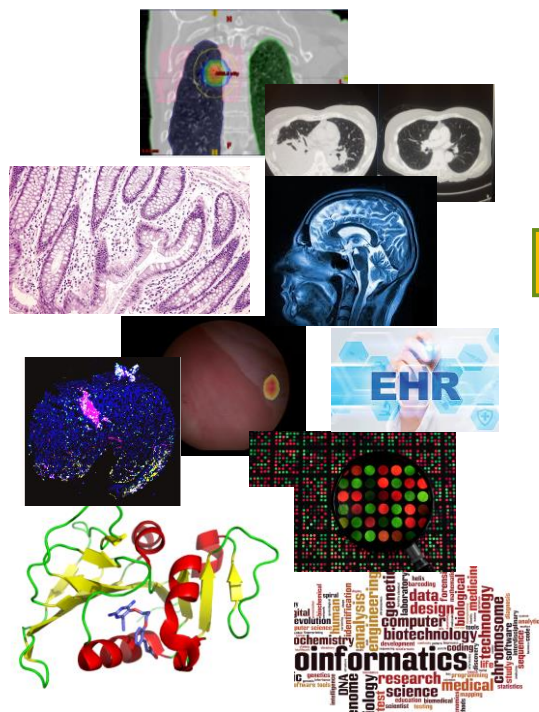
Visualization & analysis of deep learning features



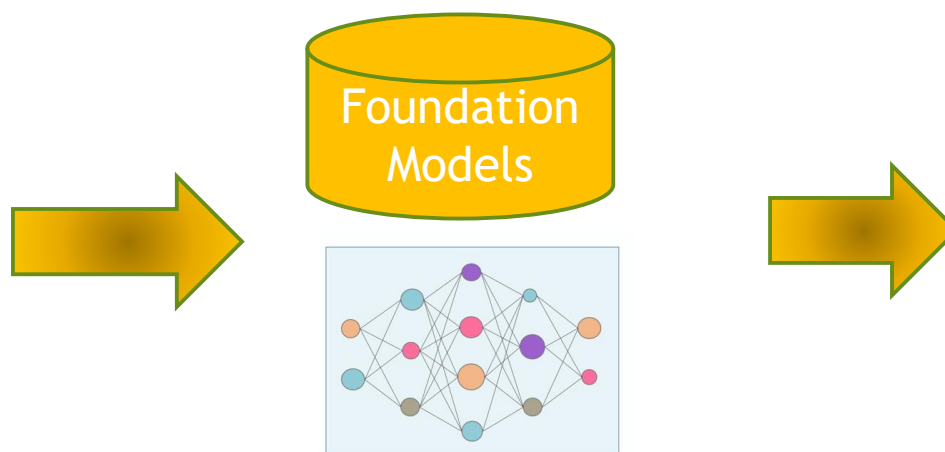
PCA (1), t-SNE (2), UMAP (3) and CFA (4) visualization of features of chest x-ray training data at layer 32 (a), 63 (b), 116 (c) and 142 (d) of GoogLeNet. The color codes are: 1) red-normal, 2) yellowish red-hernia, 3) yellow-P.T, 4) greenish yellow-Fibrosis, 5) light green-Emphysema, 6) deep green-Edema, 7) bright green-Consolidation, 8) cyanish green-Pneumothorax, 9) cyan-Pneumonia, 10) light blue-Nodule, 11) deep blue-Mass 12) blueish magenta-Infiltration, 13) light magenta-Effusion, 14) deep magenta-Cardiomegaly, 15) redish magenta Atelectasis.

AI FOUNDATION MODELS FOR PRECISION MEDICINE

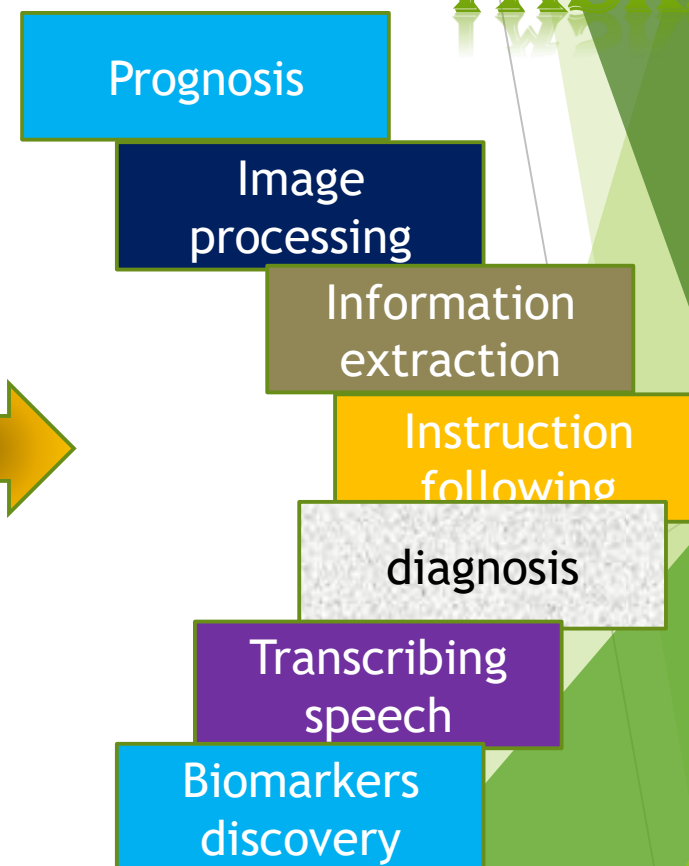
DATA



MULTIMODALITY & MULTISCALE FOUNDATION MODEL

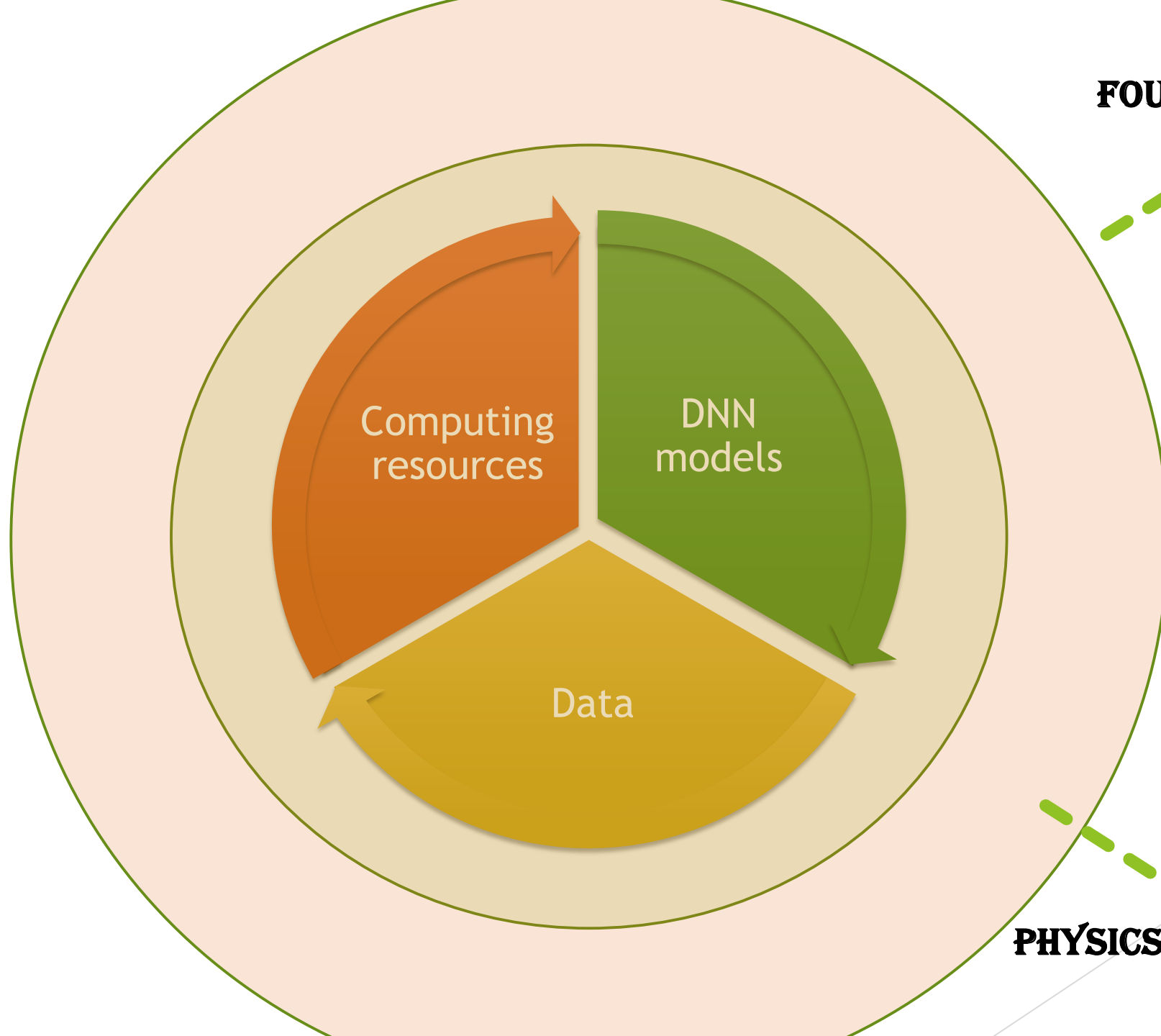


TASKS

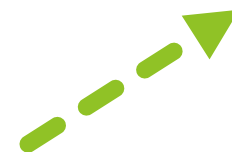


Agentic AI





FOUNDATION MODELS



DATE-EFFICIENT MODELS

PHYSICS/BIOLOGY-INFORMED MODELS

Give me sufficient **CLEAN** data, I can solve all inference problems



- Archimedes



Give me a firm place to stand and a lever and I can move the Earth

- Archimedes



SUMMARY

Biomedical AI Foundation Models

1



AI is changing the landscape of biomedicine

2



Prior physics/biology knowledge

3



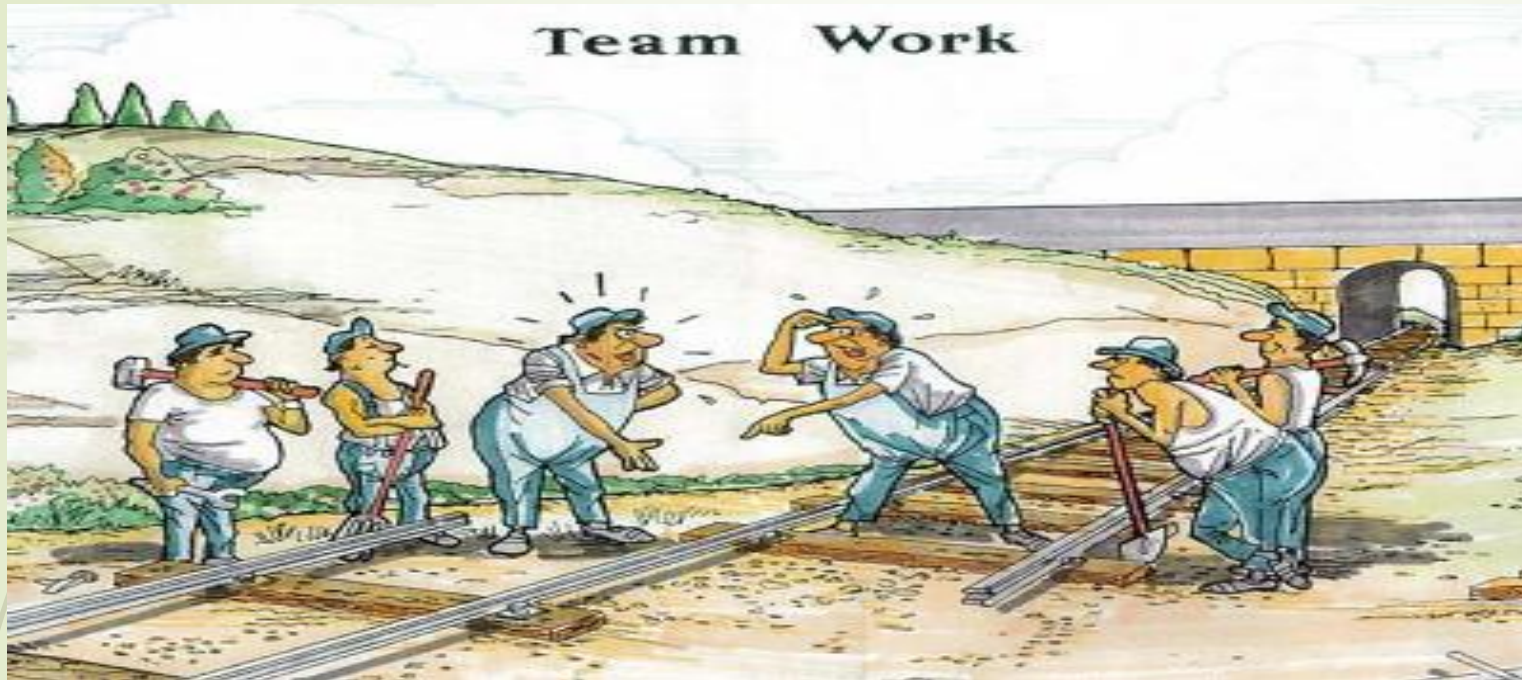
foundation models for various biomedical applications
–need biology/physics knowledge integration

4



Multi-modality & Multi-scale FM

Acknowledgements



- ❖ NIH: 1R01CA223667, 1R01CA227713, and 1R01CA256890
- ❖ Google Inc.
- ❖ Mevion Medical Systems

