

Enabling AI for Medicine with Open Source AI Models

INSERM

Kristin Lauter

Senior Director, FAIR

June 2025

Disclaimers:

1. I am not a physician and I am not giving any medical advice or recommending the use of any technology for medical purposes.
2. All information in this presentation is publicly available on websites or in the publications of the cited researchers or companies.

FAIR @Meta

Fundamental AI Research

2013 Founding of FAIR

FAIR Labs in Menlo Park, California, Seattle, NYC, Montreal, Pittsburgh, Paris, London...

2016 Introduction of Open Source Tools: PyTorch

FAIR introduced PyTorch, open source machine learning library, now widely used in the AI research community.

PyTorch Documentary, June 2024

Strong commitment to Open Source and history of releasing many AI Foundation models

Benefits of an open approach



Economic growth:

Broader accessibility
accelerates innovation



Quality & Safety:

Collaborative
approach to risk
mitigation



Advancing Science:

Build for the Ecosystem
adapt to specific needs.

2023

Foundation Models

Llama 1bn downloads

Toolformer

CM3Leon 230k downloads

I-JEPA, V-JEPA

Segment Anything 7k citations

DINOv2 2.4 M downloads, 4k citations

ImageBind

VoiceBox powered MovieGen

Massively Multilingual Speech

MusicGen

Seamless Time Magazine 2023

Code Llama 9m downloads



Llama

FEB 2023

For research use

7B/13B/33B/65B pretrained models

Fine-tuned models released by
researchers from Stanford and Berkeley



Llama 2

JULY 2023

For research and commercial use

7B/13B/70B family of models:

- Pretrained and fine-tuned
- CodeLlama, PurpleLlama

Meta Llama 3

April 2024

Open models available in 8B and 70B.

Pretrained, instruction fine-tuned versions
supporting a wide range of applications.

8B

70B

Trust and
safety

Llama 4

Multimodal Intelligence

April 2025

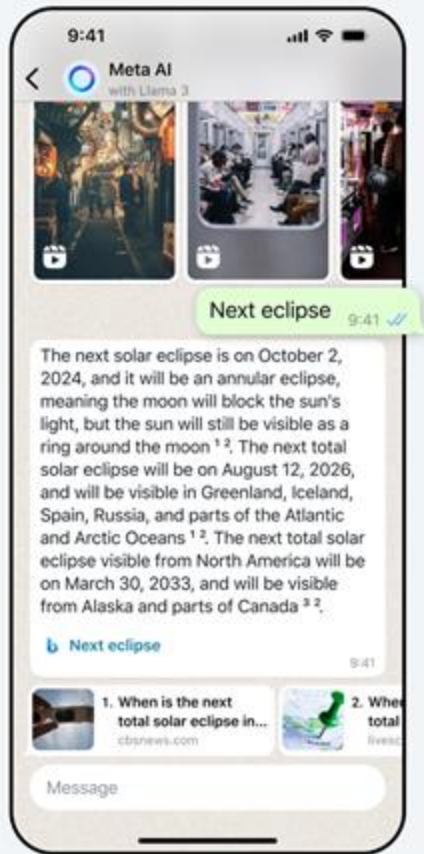
Three open models available in 2T, 400B,
109B parameters

Meta AI

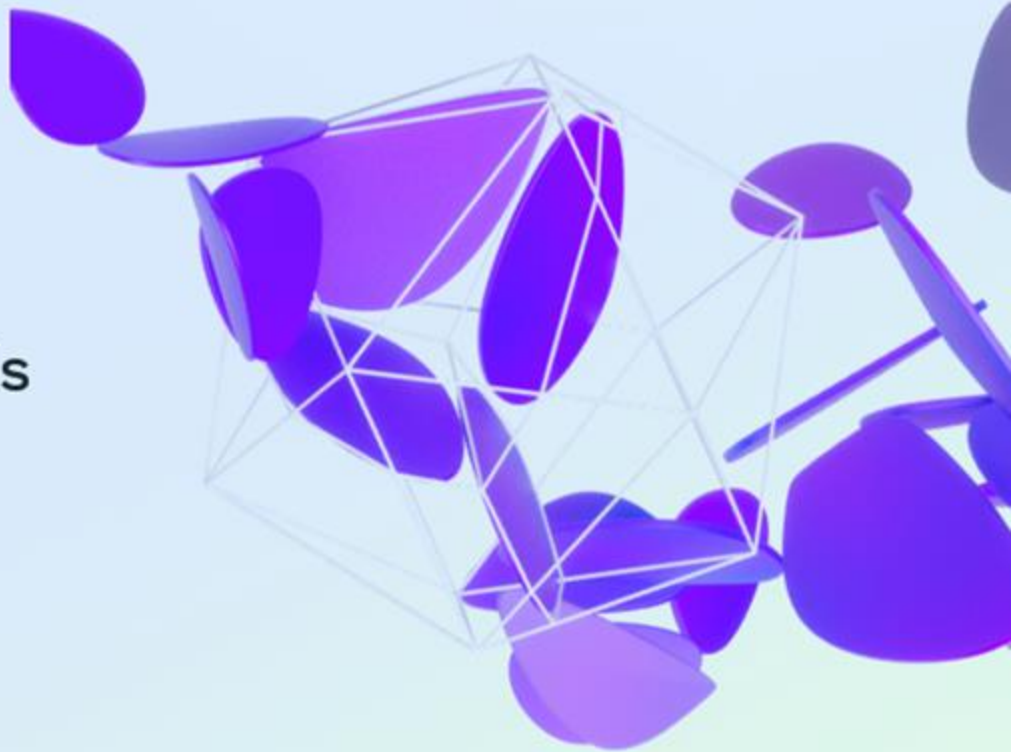
Built with Meta Llama 3, Meta AI is one of the world's leading AI assistants

Already on your phone, in your pocket for free. And it's starting to go global with more features.

You can use Meta AI on Facebook, Instagram, WhatsApp, Messenger, and on the web at meta.ai to get things done, learn, create and connect with the things that matter to you.



Medical Foundation Models built on Llama



Llama

Fine-tuned generative LLMs for medical advancement

MedAlpaca



LlaVa-Med



MediTron

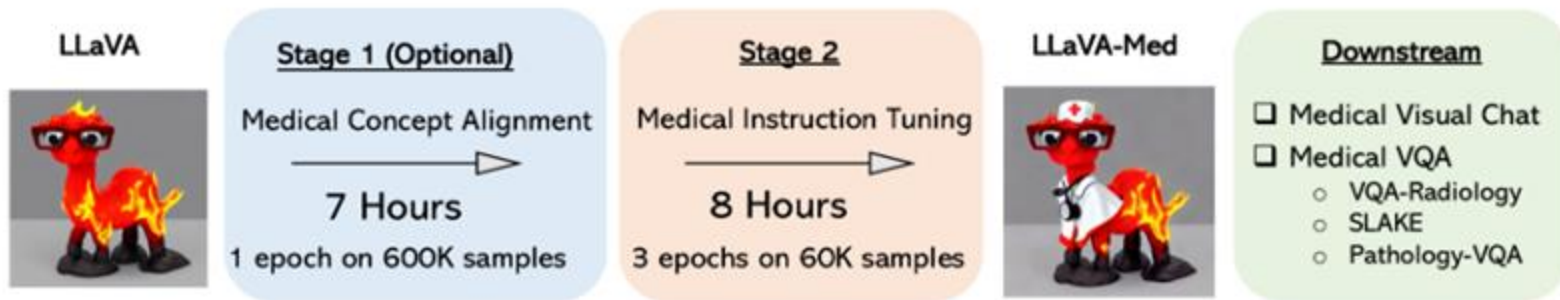


LLaVA-Med

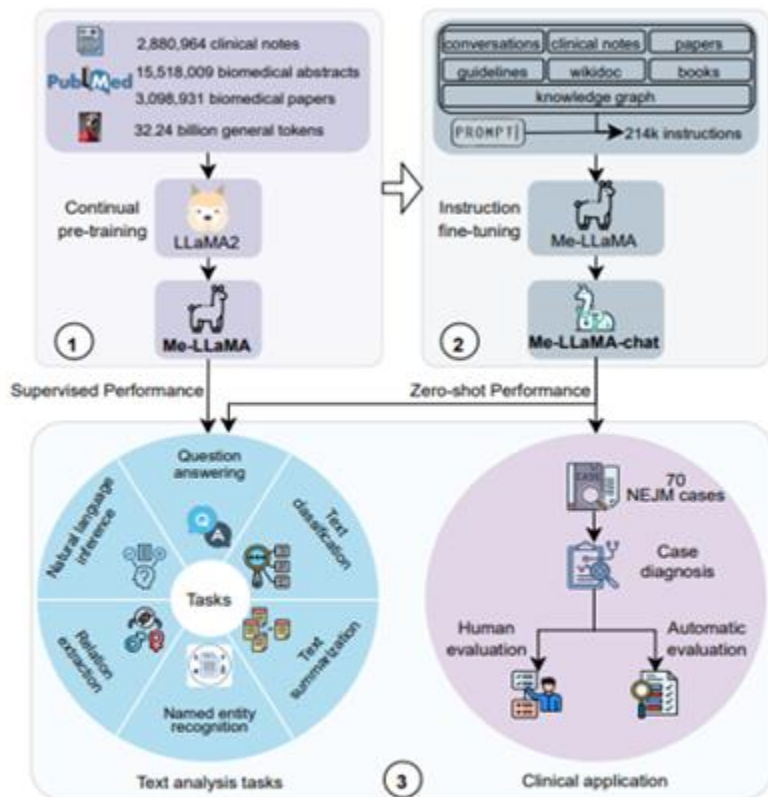
Visual question answering model

MD Anderson Cancer Center

- [June 1] 🔥 We released **LLaVA-Med: Large Language and Vision Assistant for Biomedicine**, a step towards building biomedical domain large language and vision models with GPT-4 level capabilities. Checkout the [paper](#)



Li, et al "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day" <https://arxiv.org/abs/2306.00890>



Me-LLaMA

Foundational medical LLM

Open source model

Open sourced datasets

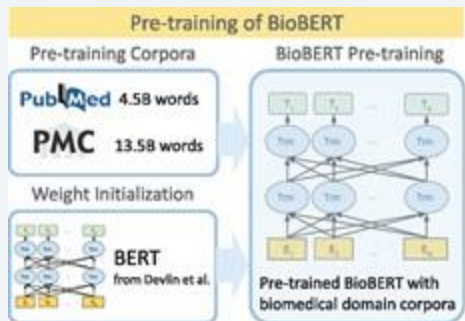
Xie, et al "Me-LLaMA: Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond"

<https://arxiv.org/pdf/2402.12749>

BERT

understanding and analyzing medical text

BioBERT

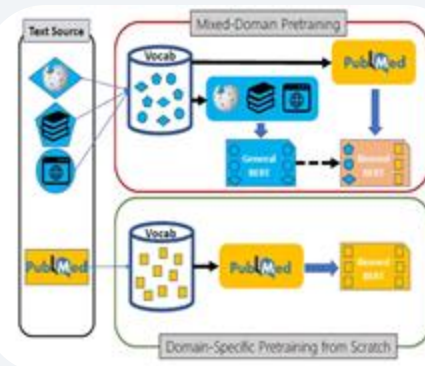


ClinicalBERT

Transformer
Encoders

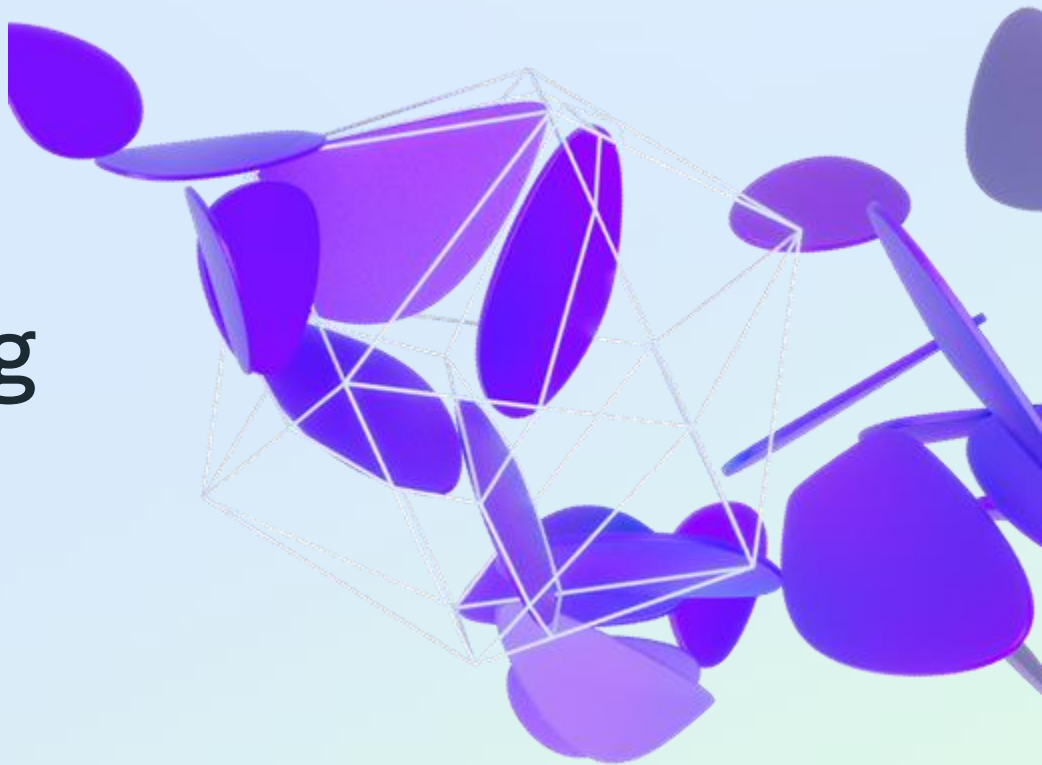


PubMedBERT



SAM: Segment Anything

Nikhila Ravi et al, Perception team, FAIR



Segmentation powers perception



VISUAL PERCEPTION



2012

ALEXNET
(U. Toronto)



2015

FASTER R-CNN
(Microsoft, Facebook)



2019

PANOTOPIC FPN
(Facebook)



2023

SEGMENT ANYTHING
(Meta)

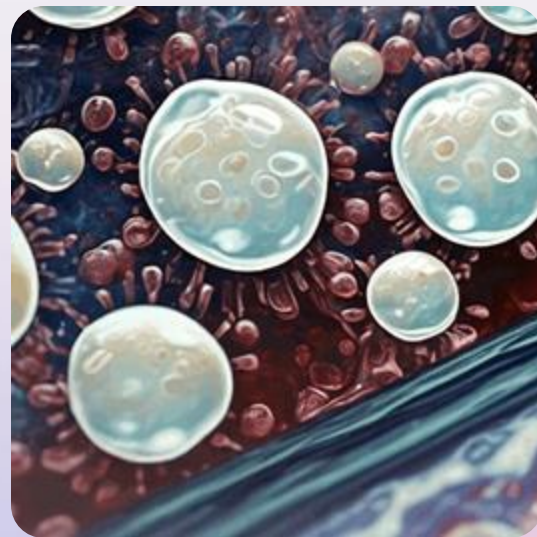
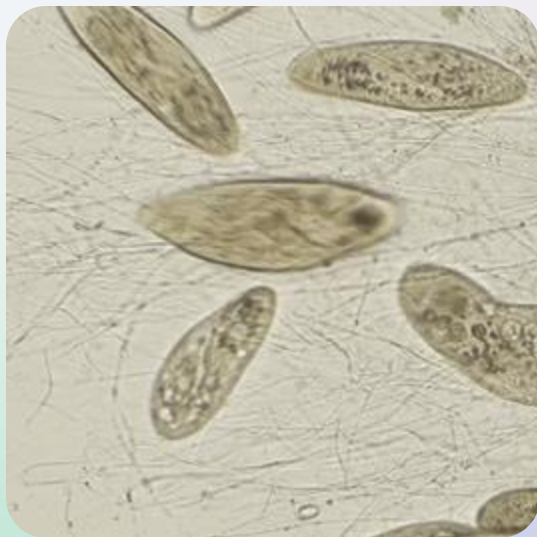
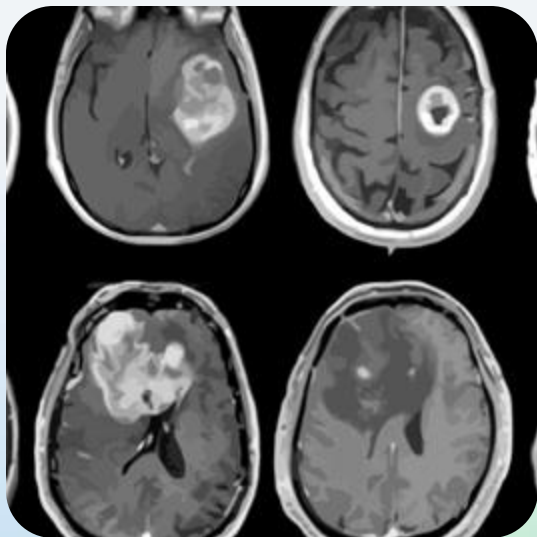


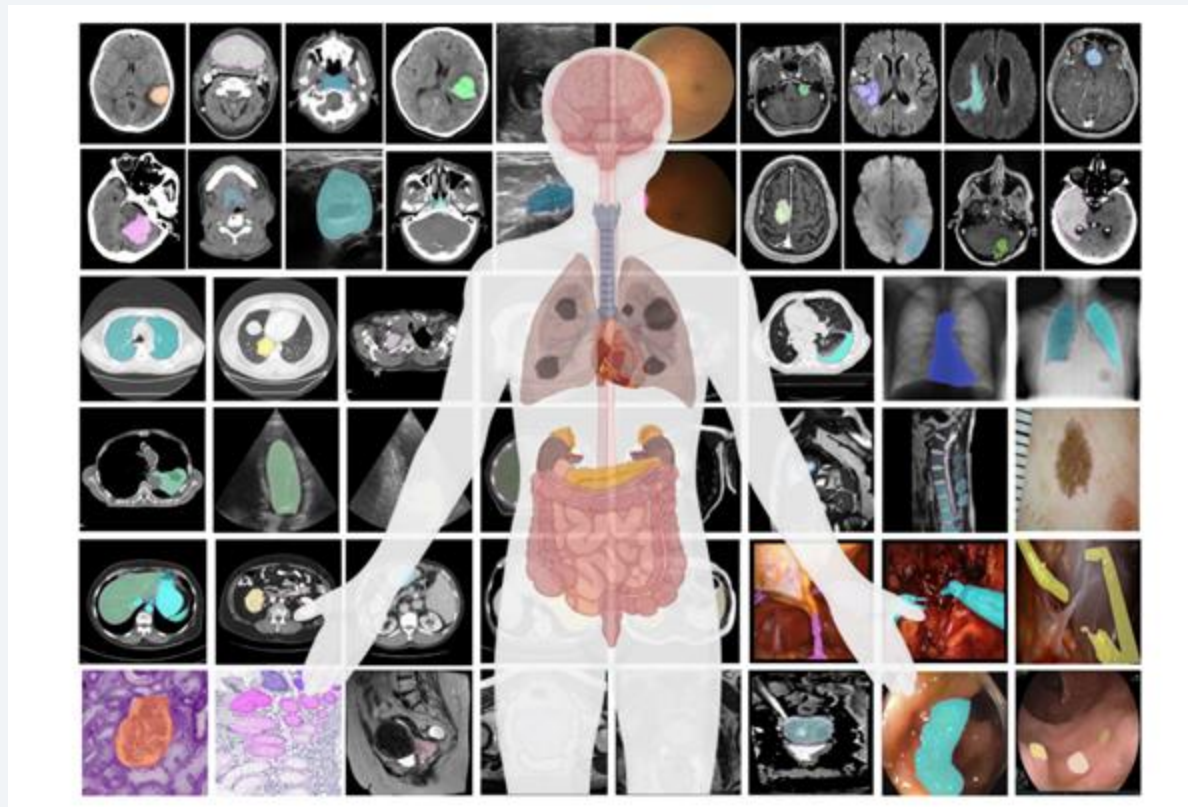
2024

SEGMENT ANYTHING 2
(Meta)

SAM 2

Real life impact in the fields of science and medicine



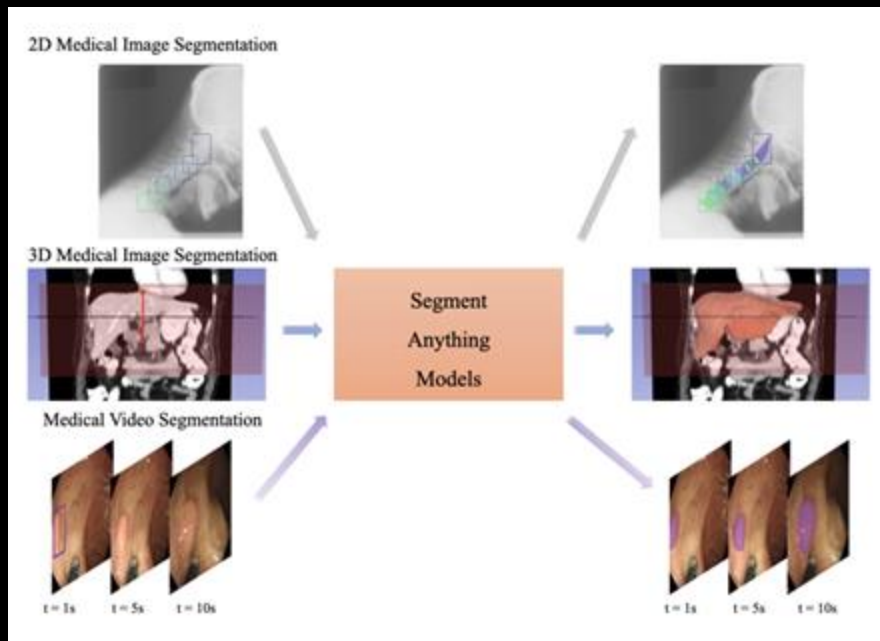


MedSAM

Foundation model for
medical image
segmentation and
analysis, covering 10
imaging modalities
over 30 cancer types

Ma, et al “Segment anything in medical images” <https://www.nature.com/articles/s41467-024-44824-z>

SAM MEDICAL APPLICATIONS: MEDSAM



Bo Wang

@BoWang87



🚀 The Segment Anything Model (SAM) has been upgraded to SAM2, featuring an efficient image encoder for segmenting images and videos. But does SAM2 outperform SAM1 in medical image and video segmentation?

We're thrilled to present our paper **"Segment Anything in Medical Images and Videos: Benchmark and Deployment"**! We comprehensively benchmark SAM2 across 11 medical image modalities and videos.

📄 Paper: arxiv.org/abs/2408.03322

📄 Code: [github.com/bowang-lab/Med...](https://github.com/bowang-lab/MedSAM)

Highlights:

1. SAM2 doesn't always outperform SAM1 in 2D medical images, but excels in video segmentation, making it more accurate and efficient for 3D images, such as CT and MR scans.
2. MedSAM still outperforms SAM2 on most 2D modalities, but SAM2 surpasses MedSAM for 3D image segmentation in a slice-by-slice approach.
3. Segmentation performance varies with model size; sometimes the smallest model outperforms larger ones.
4. Fine-tuning SAM2 significantly boosts its performance for medical image segmentation.

Ma, et al. "Segment Anything in Medical Images and Videos: Benchmark and Deployment." <https://arxiv.org/abs/2408.03322>

AI Co-pilots in surgery

Vision language models for real-time decision support in operating rooms

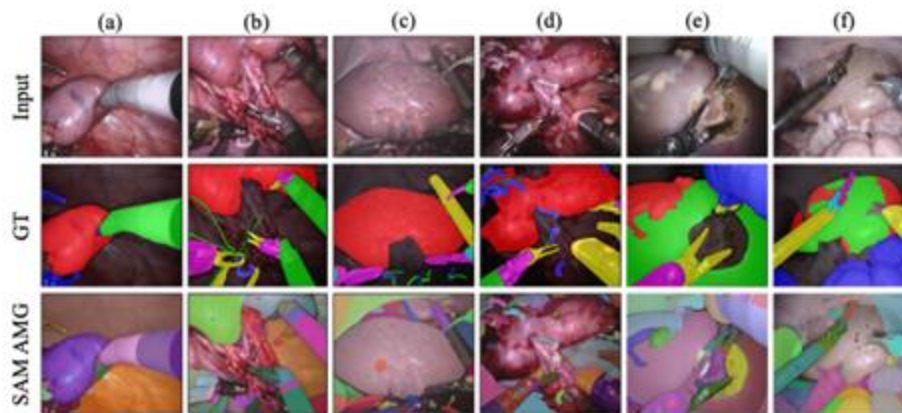
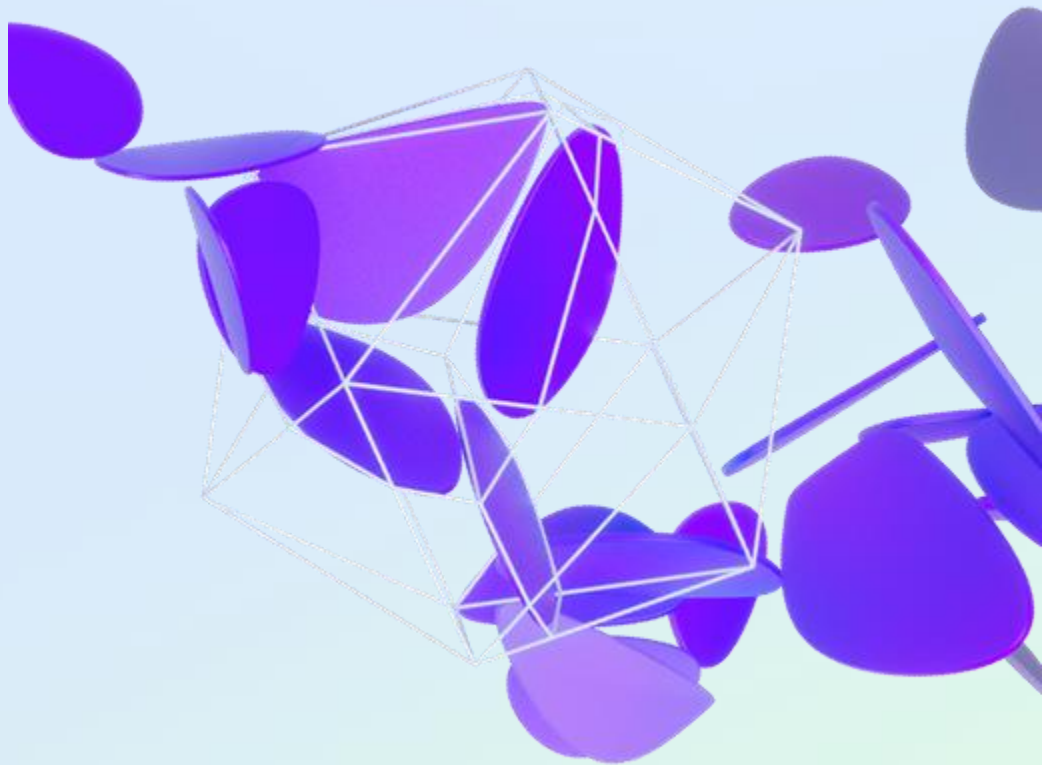


Fig. 3. Unprompted automatic mask generation for surgical scene segmentation.

Wang, et al “SAM Meets Robotic Surgery: An Empirical Study on Generalization, Robustness and Adaptation”

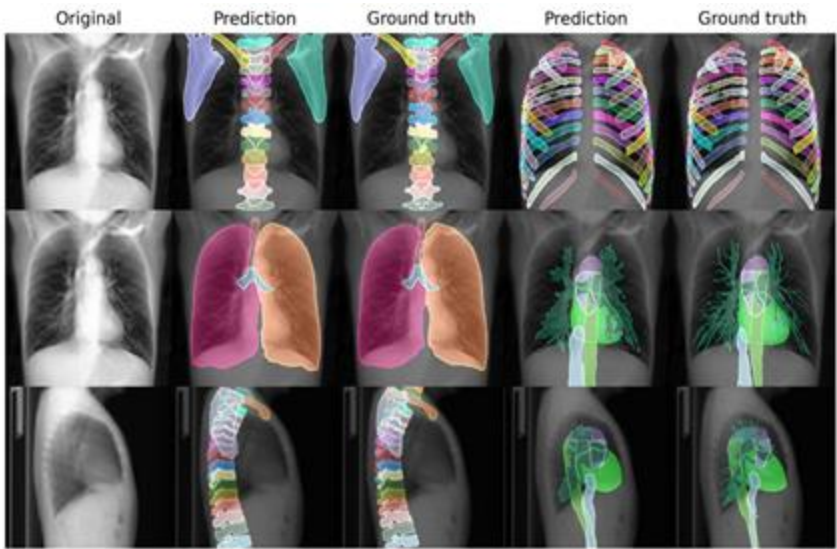
<https://arxiv.org/pdf/2308.07156>

DINO

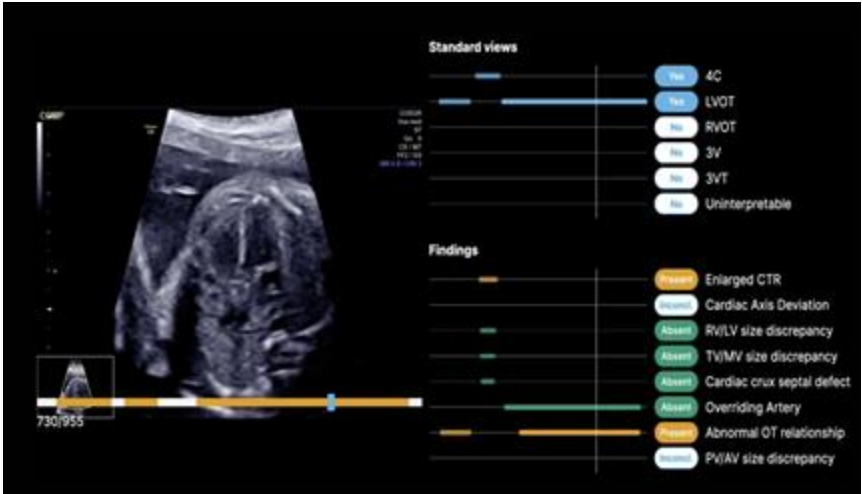


Patrick Labatut, Research Engineering Manager, DINO team, FAIR

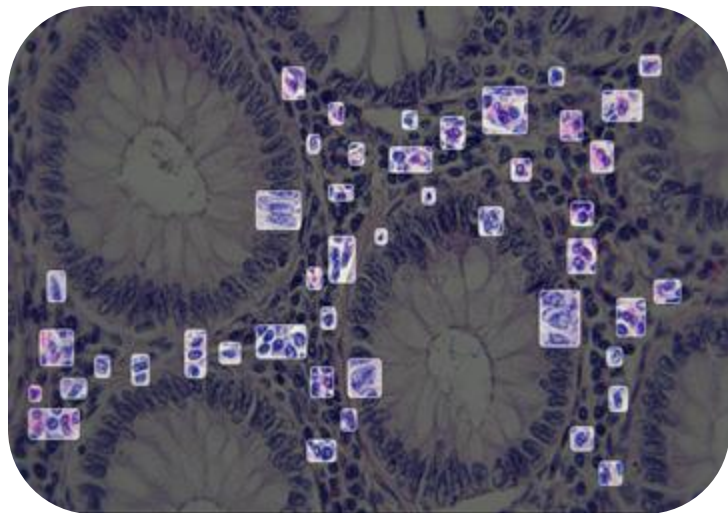
Medical Imaging:
X-rays analysis and disease detection
(in collaboration with MICS/AP-HP)



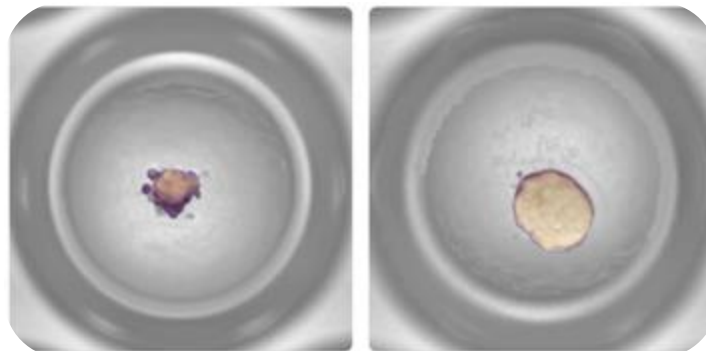
BrightHeart:
Fetal Heart Screenings



Mahmood Lab: Human Pathology



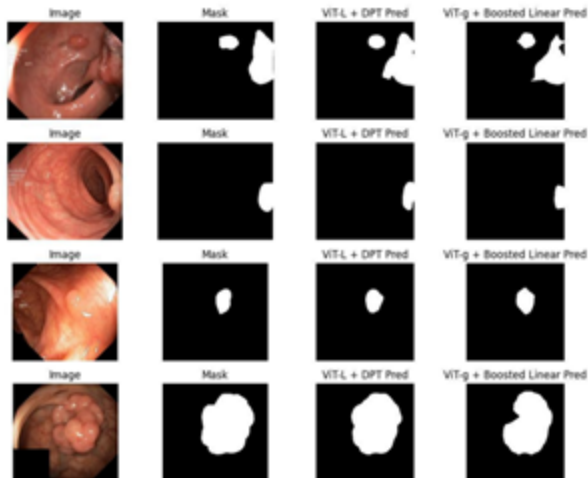
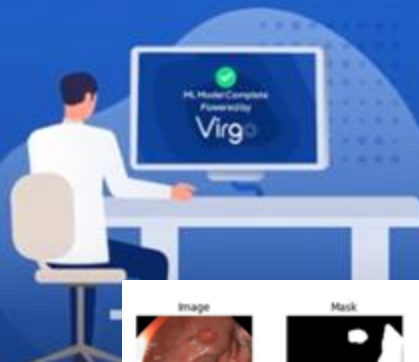
Orakl: Oncology



Virgo: Endoscopy



Changing how we
identify and enroll patients
for IBD clinical trials



Virgo Endoscopy

1.75 million procedure
videos to generate HIPAA-
compliant foundation
models for disease severity
scoring, identifying
patients to enroll in clinical
trials

UNI

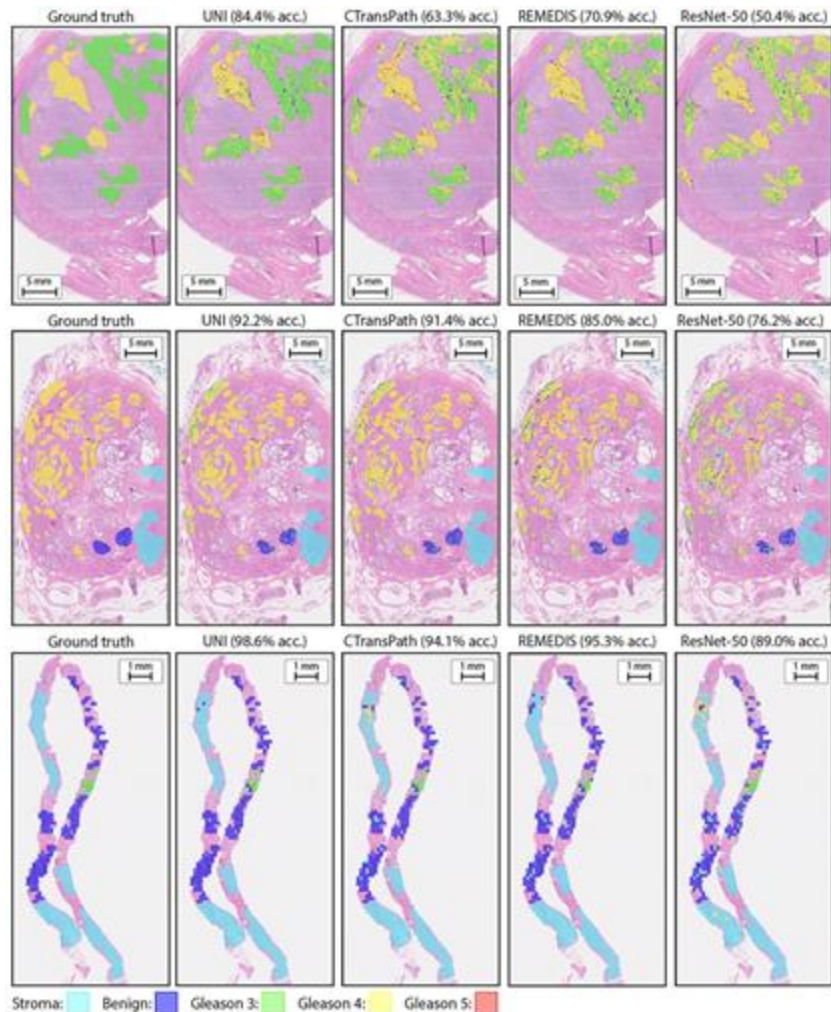
Universal model for histopathology

Based on pretrained DINOv2

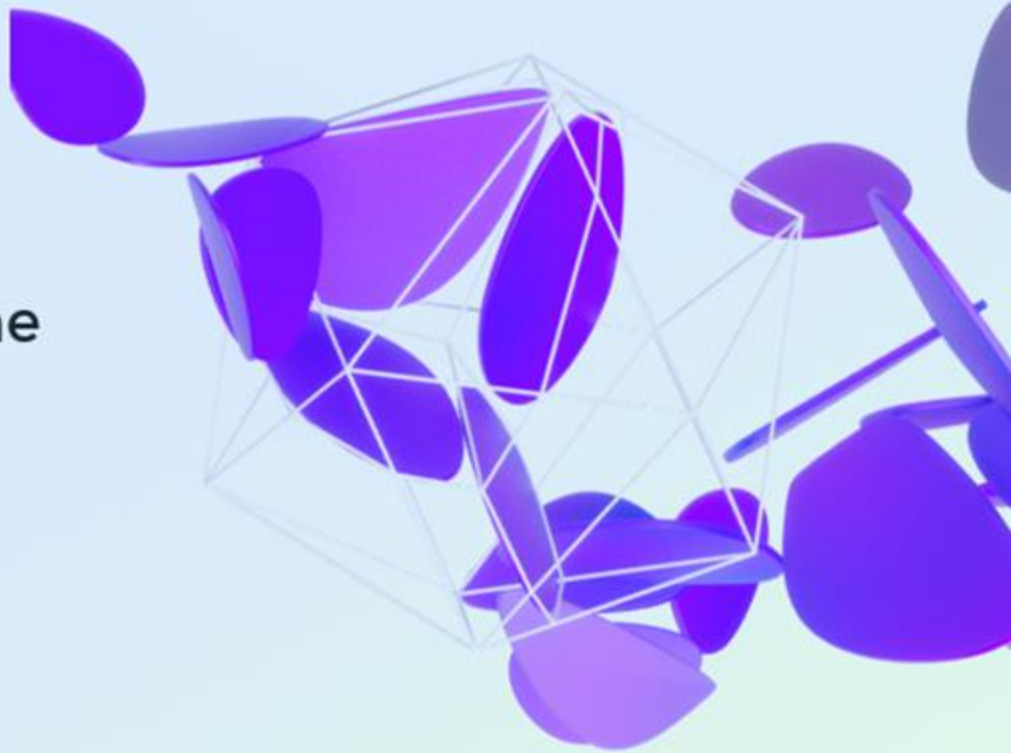
Mahmood AI Lab, Harvard University

Chen et al. “Towards a general-purpose foundation model for computational pathology”

<https://pubmed.ncbi.nlm.nih.gov/38504018/>

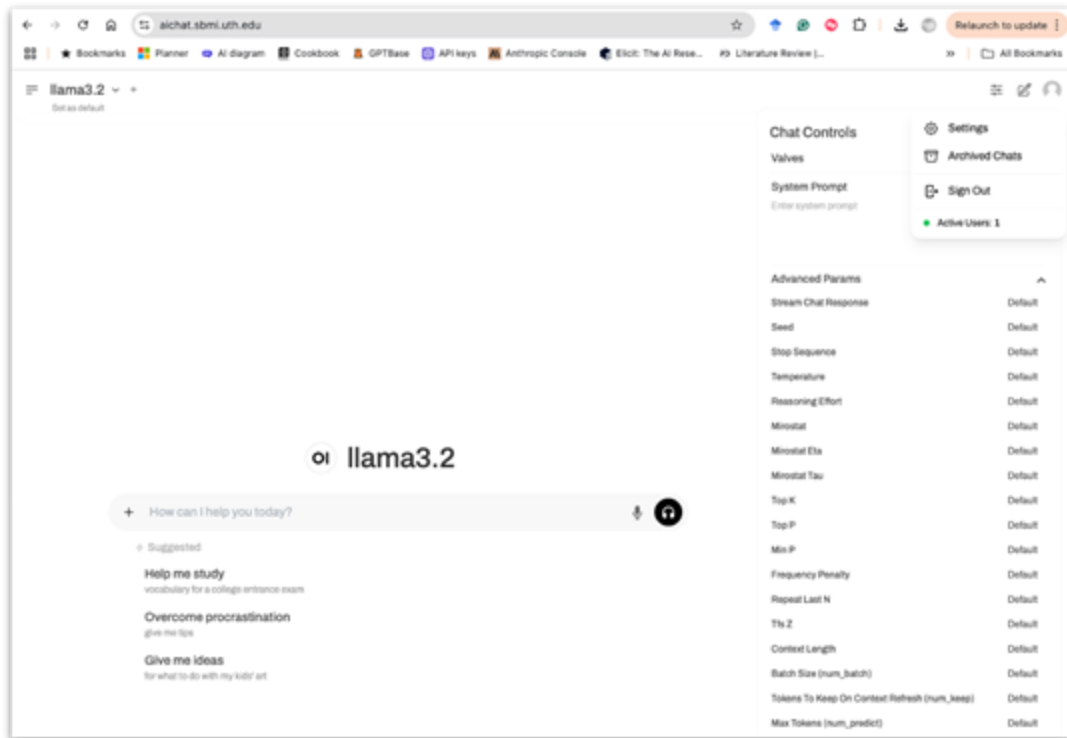


UT Houston Applications of AI in medicine



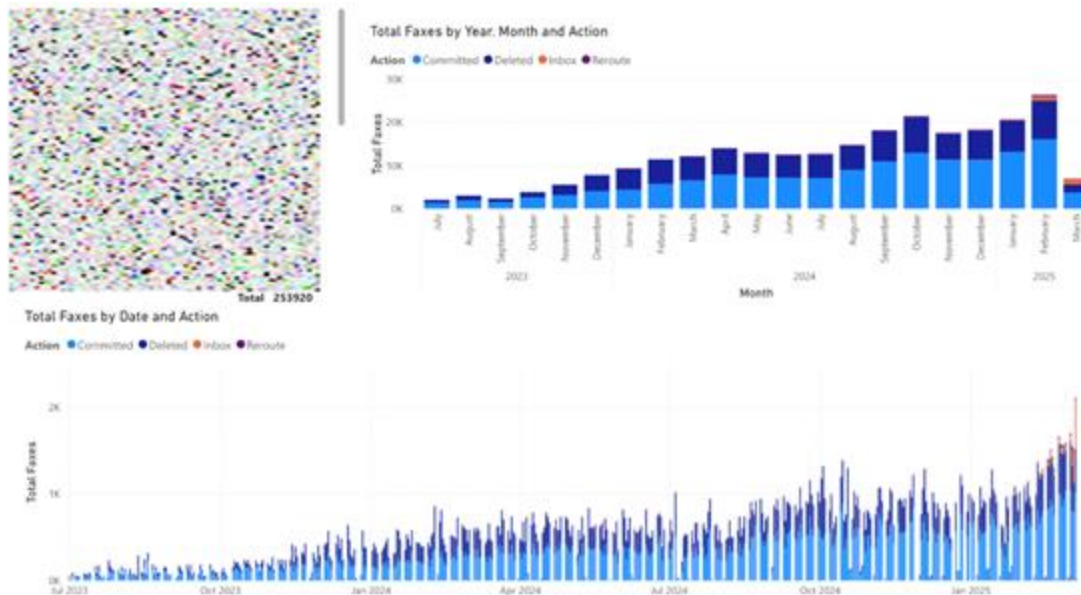
AIChat: UT Houston LLM Platform

Secure, campus-wide platform
deploying Llama to power student
and faculty research, rapid
prototyping, and accelerating AI-
driven innovation



IDFax

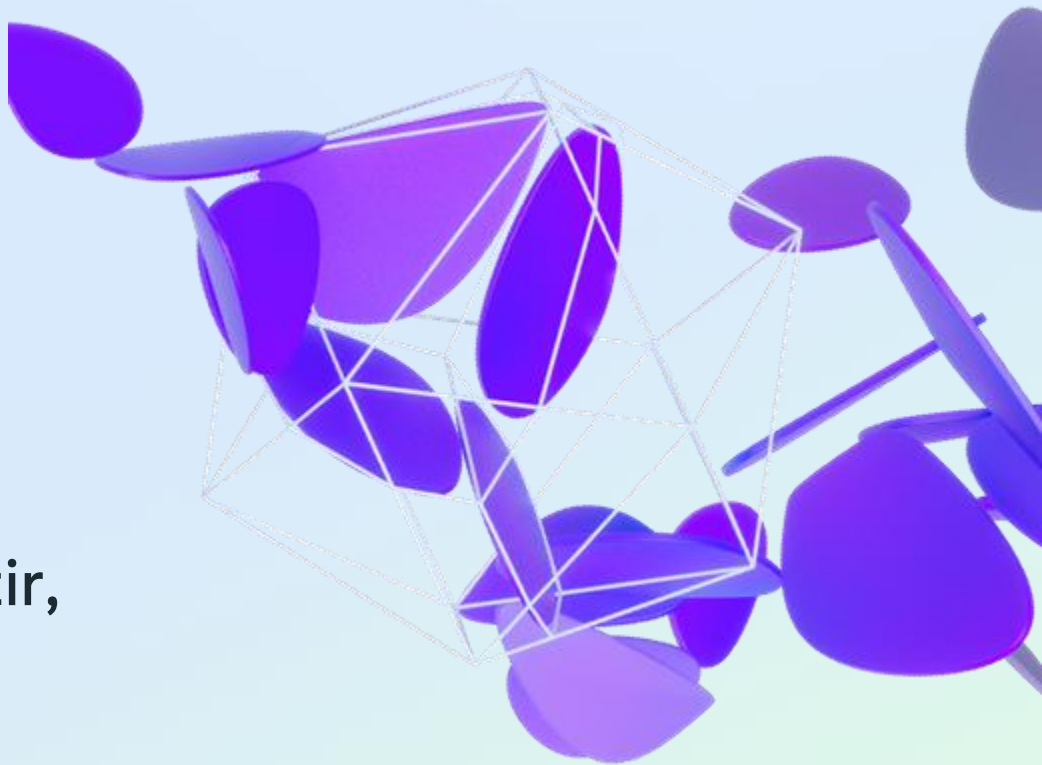
Automate and streamline fax processing in healthcare processing 15,000 faxes monthly, reducing processing times by 50%, and saving \$2M USD annually



UTHealth
Houston

Commercial Developments

Epic, HippocraticAI, Palantir,
Zauron Labs, ...



Hippocratic AI

Hippocratic AI has developed a safety-focused Large Language Models for the healthcare industry, a constellation of 22 models.

Key features and capabilities of Hippocratic AI's LLM:

Patient-facing interactions

Empathy and communication

Non-diagnostic tasks: explaining complex diagnoses, scheduling follow-ups, ...

Safety and security

Enhanced health monitoring: remind patients to take medication, prompt early interventions

Provider efficiency: automates routine tasks

Collaboration with clinicians: Hippocratic AI allows licensed clinicians to co-develop AI agents through an "AI Agent App Store"

Zaaron Labs: Guardian

Locally deployed, fine-tuned Llama to identify high-probability radiology exams that would benefit from second review

Peer Review - 5 cases to review.



To Clark, Kal L



Mon 4:01 PM

Hello, Kal

For this week, we have identified 5 cases assigned to you for review.

[Click Here](#) to review your cases.

[Click Here](#) to manually review a case.

Interesting Cases From this Week

1. [Interesting Case](#): Interval increase in size of a centrally necrotic soft tissue nodules with internal enhancing septations at the left posterior lateral abdominal wall measuring 3.4 x 2 x 3.9 cm and the left lateral wall measuring 3.7 x 2.1 x 3.3 cm.
2. [Interesting Case](#): Increased size of a large hypermetabolic right hilar mass, now measuring up to 10 cm, with resultant obstruction of the right upper lobe bronchus and encasement of the bronchus intermedius with associated volume loss is consistent with progression of primary malignancy.
3. [Interesting Case](#): Dotatate avid left pleural thickening and bilateral airspace

Data Security & Privacy?



Security and Privacy notions

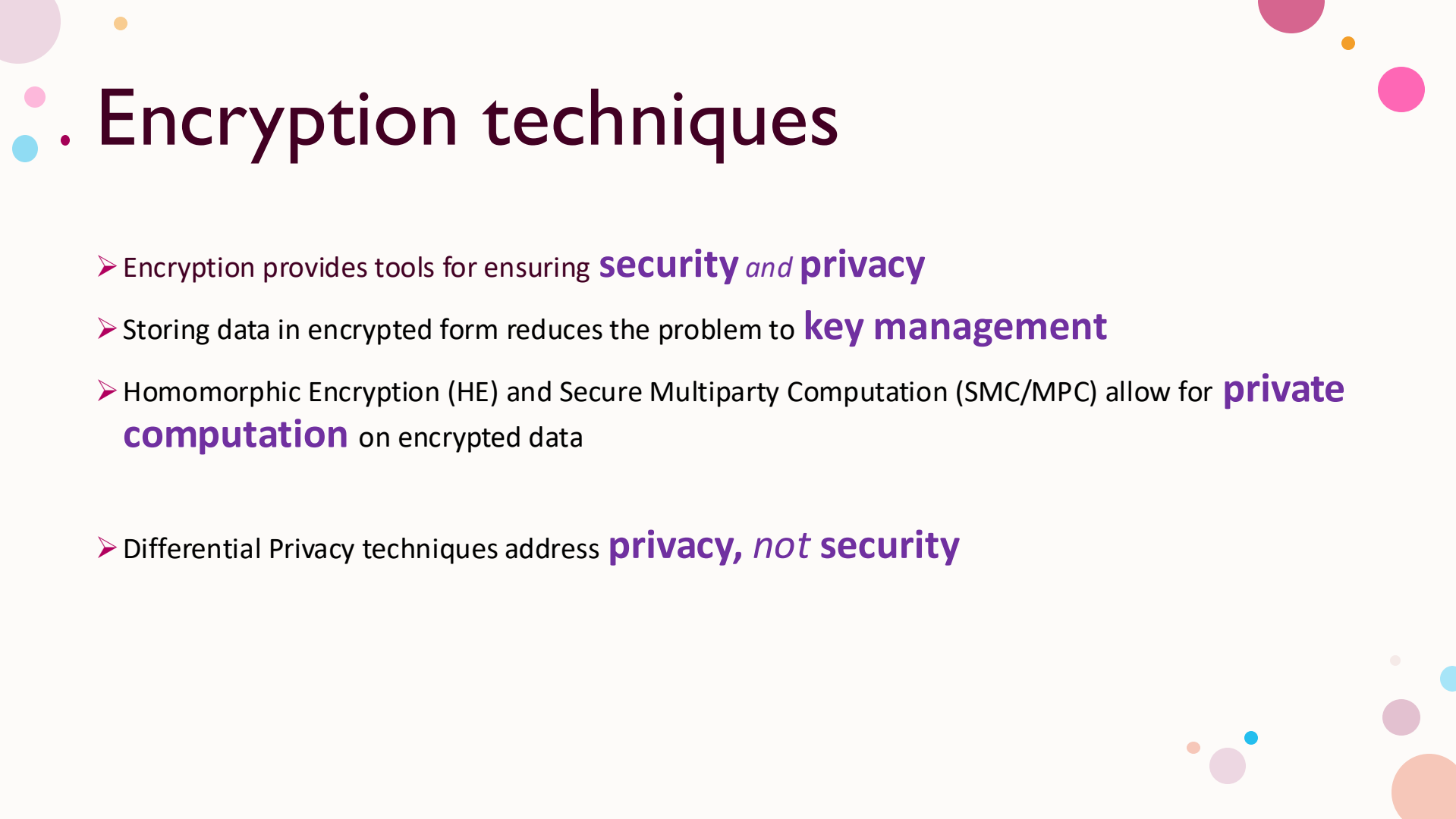
Security – to secure access to the data. No unauthorized access means confidentiality of the data restricted to the “authorized” data curators.

Privacy – limit the use of data to the expressed wishes of the data owner.



Overview of Privacy Enhancing Technologies (PETs)

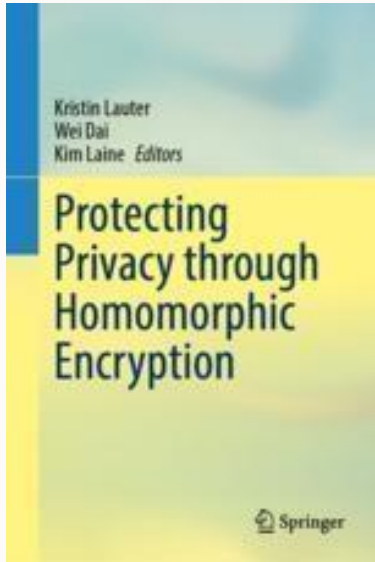
- Homomorphic Encryption (HE)
- Secure Outsourcing: Storage & Computation
- Private compute
- Secure Multiparty Computation (MPC)
- Collaborative Learning, Private Set Intersection, Secure Data Exchange
- Differential Privacy (DP)
- Query answering to protect from reidentification
- Secure Hardware (TEE/SGX)
- Outsourcing secure computation



Encryption techniques

- Encryption provides tools for ensuring **security** *and* **privacy**
- Storing data in encrypted form reduces the problem to **key management**
- Homomorphic Encryption (HE) and Secure Multiparty Computation (SMC/MPC) allow for **private computation** on encrypted data
- Differential Privacy techniques address **privacy, not security**

. Homomorphic encryption volume



Contains:

1. Homomorphic Encryption Standard (HES)
2. Overview of Schemes
3. Applications developed at events

10 years of iDASH benchmarks

Secure Genome Analysis competitions

funded by NIH


Dr. Xiaoqian Jiang (UTHealth), Arif Harmanci (UTHealth), Haixu Tang (IUB), XiaoFeng Wang (IUB), Lucila Ohno-Machado (UCSD), T-T Kuo (UCSD), Miran Kim (UNIST)

- Homomorphic Encryption (HE)
- Secure Multi-Party Computation (MPC/SMP)
- Differential Privacy (DP)
- Hardware solution (SGX)

Protecting genomic data analytics in the cloud: state of the art and opportunities

Haixu Tang Xiaoqian Jiang, Xiaofeng Wang, Shuang Wang, Heidi Sofia, Dov Fox, Kristin Lauter, Bradley Malin, Amalio Telenti, Li Xiong and Lucila Ohno-Machado. BMC Medical Genomics BMC series 2016 9:63 <https://doi.org/10.1186/s12920-016-0224-3>

Summary of Challenges and Tasks



2015	12 Teams	<ul style="list-style-type: none">• HE-based genome analysis (DNA sequence comparison)• MPC-based genome analysis
2016	50 Teams	<ul style="list-style-type: none">• Testing for genetic disease on homomorphically encrypted genomes• MPC-based privacy-preserving search of similar cancer patients across organizations• Protecting queries in Beacon service
2017	65 Teams	<ul style="list-style-type: none">• Homomorphic Logistic Regression Training• Secure record de-duplication• Secure GWAS using SGX
2018	64 Teams	<ul style="list-style-type: none">• Secure Parallel Genome Wide Association Studies using Homomorphic Encryption• Blockchain-based immutable logging and querying for cross-site genomic dataset access audit trail• MPC-based secure search of DNA segments in large genome databases
2019	105 Teams	<ul style="list-style-type: none">• Secure Genotype Imputation using Homomorphic Encryption (54 teams)• Distributed Gene-Drug Interaction Data Sharing based on Blockchain and Smart Contracts• Privacy-preserving Machine Learning as a Service on SGX• MPC-based Secure Collaborative Training of Machine Learning Model

Secure Genotype Imputation

Cell Systems





Available online 30 August 2021

In Press, Corrected Proof [?](#)



Methods

Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation

Miran Kim ^{1, 16}, Arif Ozgun Harmanci ^{2, 16, 17}  , Jean-Philippe Bossuat ³, Sergiu Carpov ^{4, 5}, Jung Hee Cheon ^{6, 7}, Ilaria Chillotti ⁸, Wonhee Cho ⁶, David Froelicher ³, Nicolas Gama ⁴, Mariya Georgieva ⁴, Seungwan Hong ⁶, Jean-Pierre Hubaux ³, Duhyeong Kim ⁶, Kristin Lauter ⁹, Yiping Ma ¹⁰, Lucila Ohno-Machado ¹¹, Heidi Sofia ¹², Yongha Son ¹³ ... Xiaoqian Jiang ¹⁵  

Private AI Demos History

- 2014: Heart attack risk, personal health data ~ 1 second
- 2015: CryptoNets demo showing neural net prediction: MNIST data set ~80 seconds
- 2016: Genomics predicting flowering time from 200K SNPs ~ 1 second
- 2016: Pneumonia mortality risk: intelligible models ~ 8 seconds for 4,000 predictions
- 2017: Personalized medicine: predicting drug response ~ 1 second
- 2018: Twitter sentiment analysis (150K text features) ~ less than a second
- 2018: cat/dog image classification ~ less than a second
- 2019: Asure Run (Private Fitness App)
- 2019: Chest Xray diagnostics
- 2019: Secure Weather prediction

Trusted monitoring service



***HEAR: Human Action Recognition
via Neural Networks on
Homomorphically Encrypted Data***



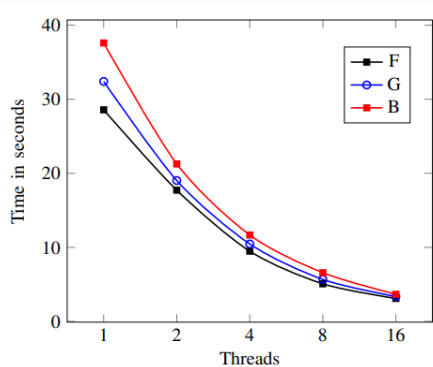
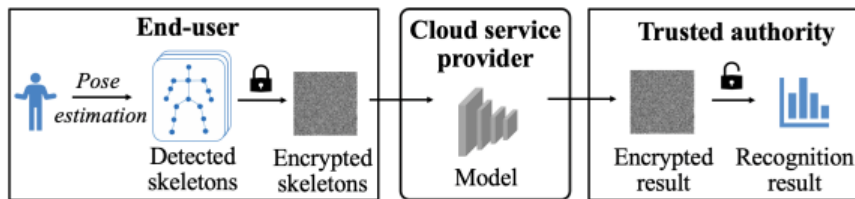
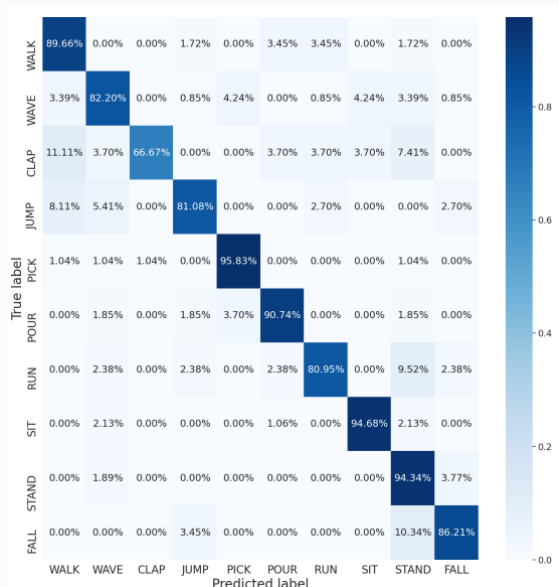
Miran Kim, Xiaoqian Jiang, Kristin
Lauter, Elkhan Ismayilzada, Shayan
Shams



Can monitor encrypted video
footage for falls in near real-time.

HEAR: Human Action Recognition via Neural Networks on Homomorphically Encrypted Data

- The sensitivity and specificity of our model for fall detection are 86% and 98.3%, respectively.
- Near real-time inference (3.1 seconds per 32 frames) on encrypted data

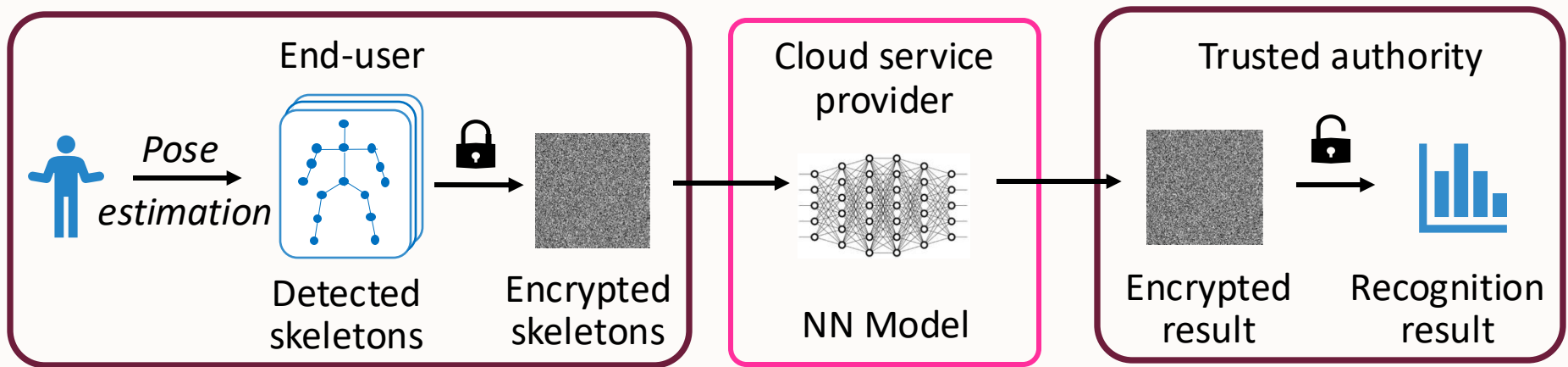


Method	Hoisting	Lazy rescaling	Latency	Speedup
F	✓		9.748 sec	1.02×
	✓	✓	3.567 sec	2.78×
	✓	✓	3.091 sec	3.21×
G	✓		9.606 sec	1.01×
	✓	✓	3.375 sec	2.87×
	✓	✓	3.357 sec	2.89×
B	✓		10.030 sec	0.99×
	✓	✓	3.670 sec	2.73×
	✓	✓	3.661 sec	2.74×

2D CNN based HEAR action recognition

Fast HEAR model time efficiency with different implementation methods

HE-based Private Action Recognition: Scenario



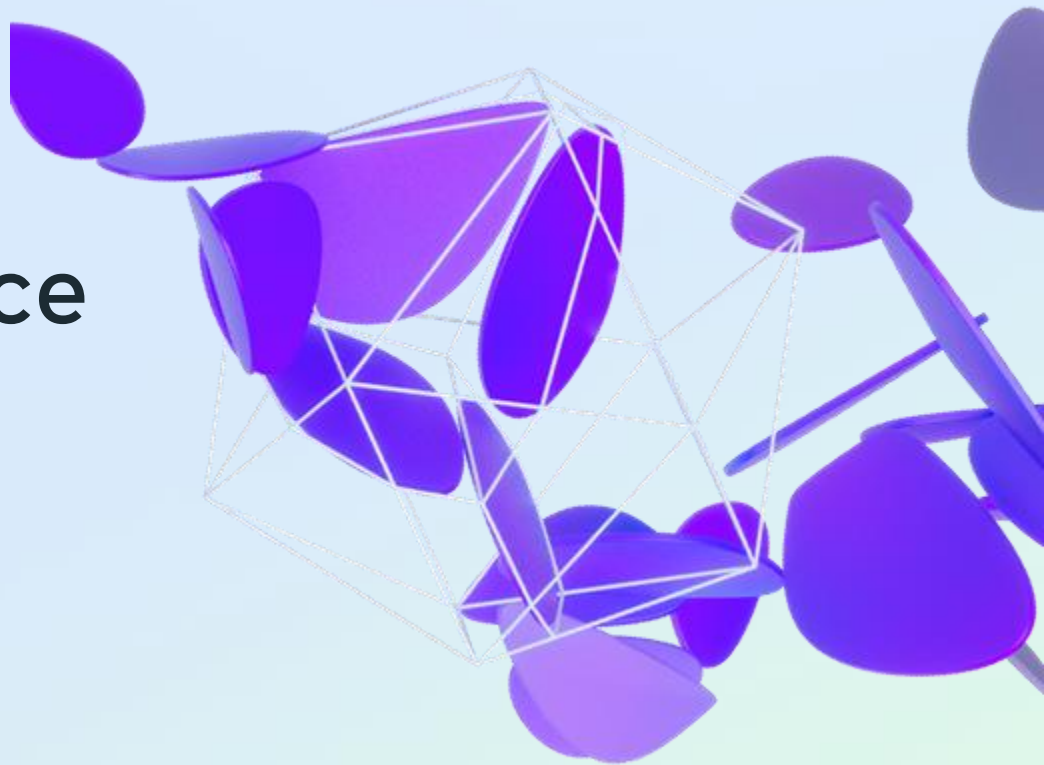
- 1) Encrypt the detected skeletons.
- 2) Send the encrypted tensor to the cloud.

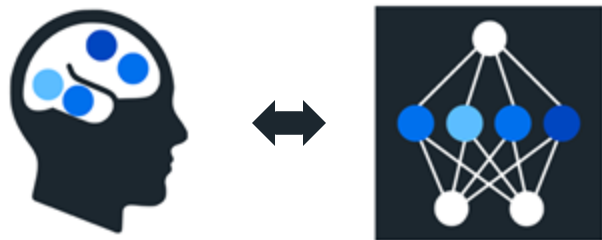
Predict the motion on the encrypted tensor

- 1) Decrypt the result.
- 2) If the result needs immediate intervention, alert 911 and hospital.
- 3) Else, save the action and time stamp for feature report.

AI for Neuroscience

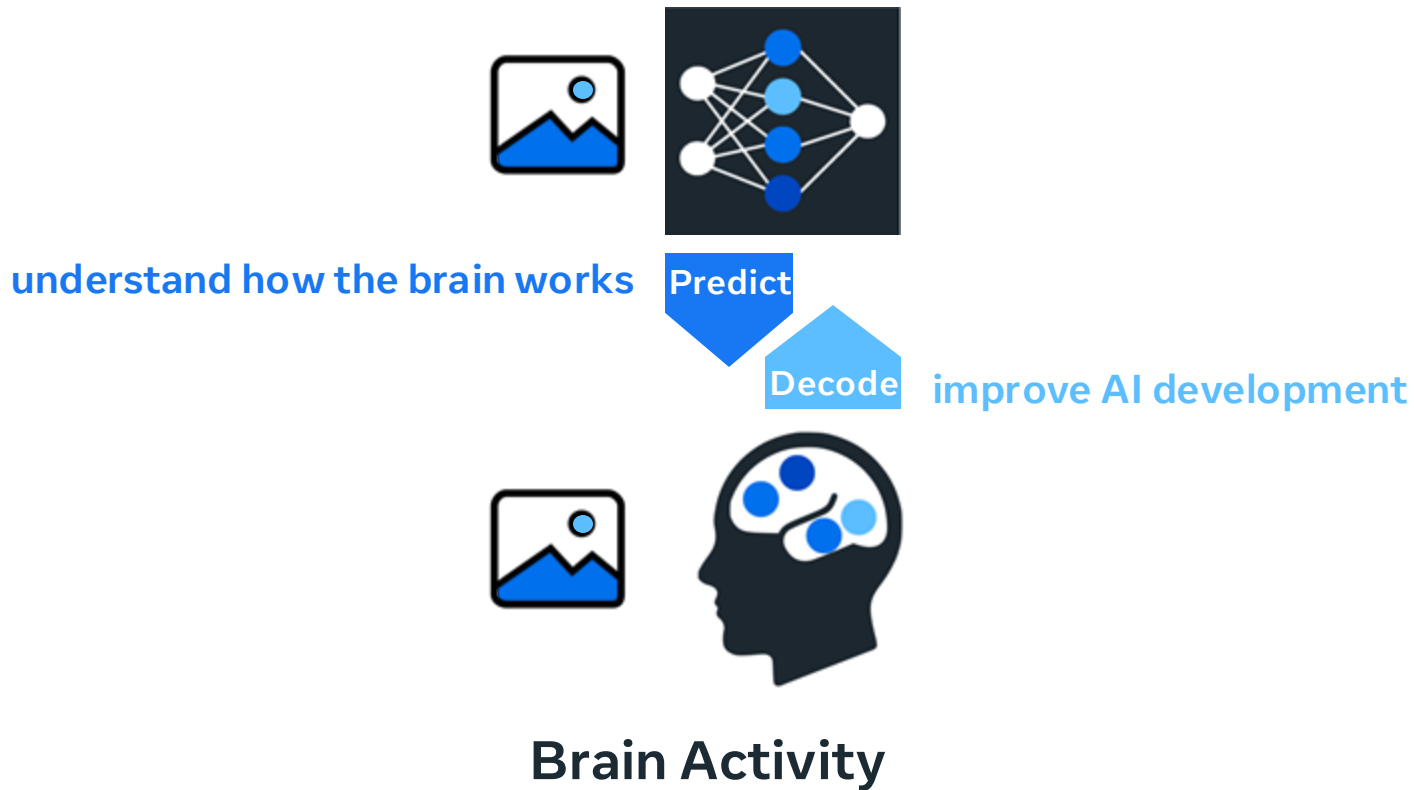
Jean Remi King, Research Scientist
BRAIN & AI team, FAIR





 **Brain**
Model the brain with ai.

Artificial Intelligence



2022



wav2vec 2.0



2023



DINO V2



2024



Llama 2



2025



Brain2Qwerty



February
2025 release



2022



wav2vec 2.0



2023



DINO V2



2024



Llama 2



2025



Brain2Qwerty



February
release



2022



wav2vec 2.0



2023



DINO V2



2024



Llama 2



2025



Brain2Qwerty



Today



Dynadiff



May release: Dynadiff

A single stage model to decode images from
real-time fMRI

<https://github.com/facebookresearch/dynadiff>

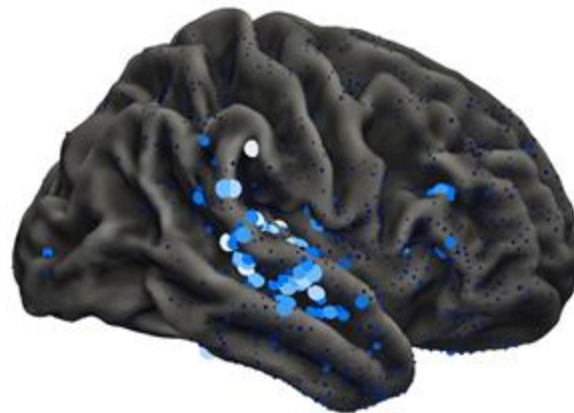


viewed image



brain decoding

The Emergence of Language in the Developing Brain.



Wav2Vec

2018

Self-supervised
learning from raw audio

Contrastive learning
and masked prediction

HuBERT

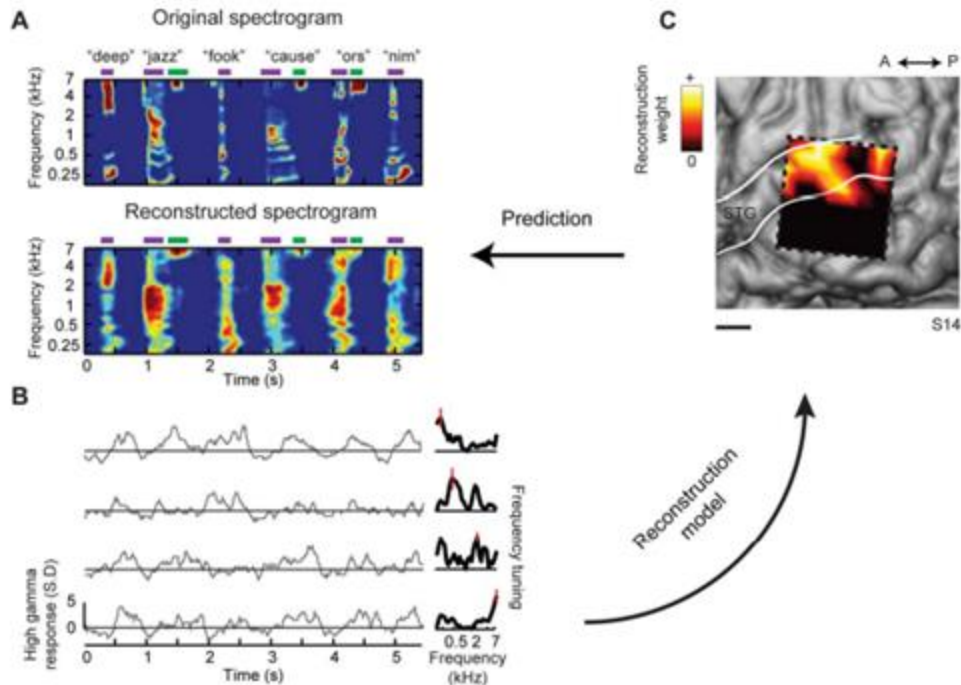
2021

Self-supervised speech
representation learning

clustering and masked
prediction

Reconstructing Speech using HuBERT

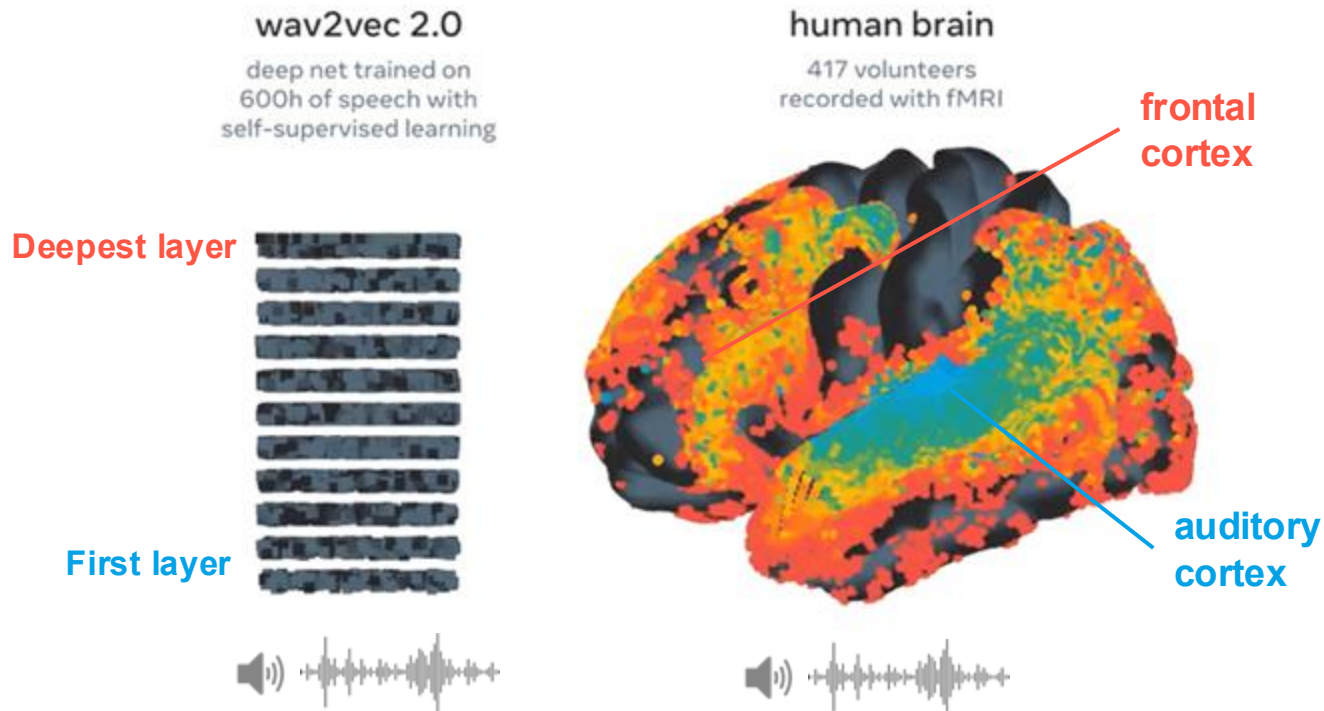
Decoding neural activity in real time to help epileptic patients speak at UCSF



Pasley et al (2012) "Reconstructing Speech from Human Auditory Cortex" <https://pubmed.ncbi.nlm.nih.gov/22303281/>

Speech Processing

Foundation models
spontaneously build
brain-like representations



Millet et al (2022) "Toward a realistic model of speech processing in the brain with self-supervised learning"

<https://arxiv.org/pdf/2206.01685>

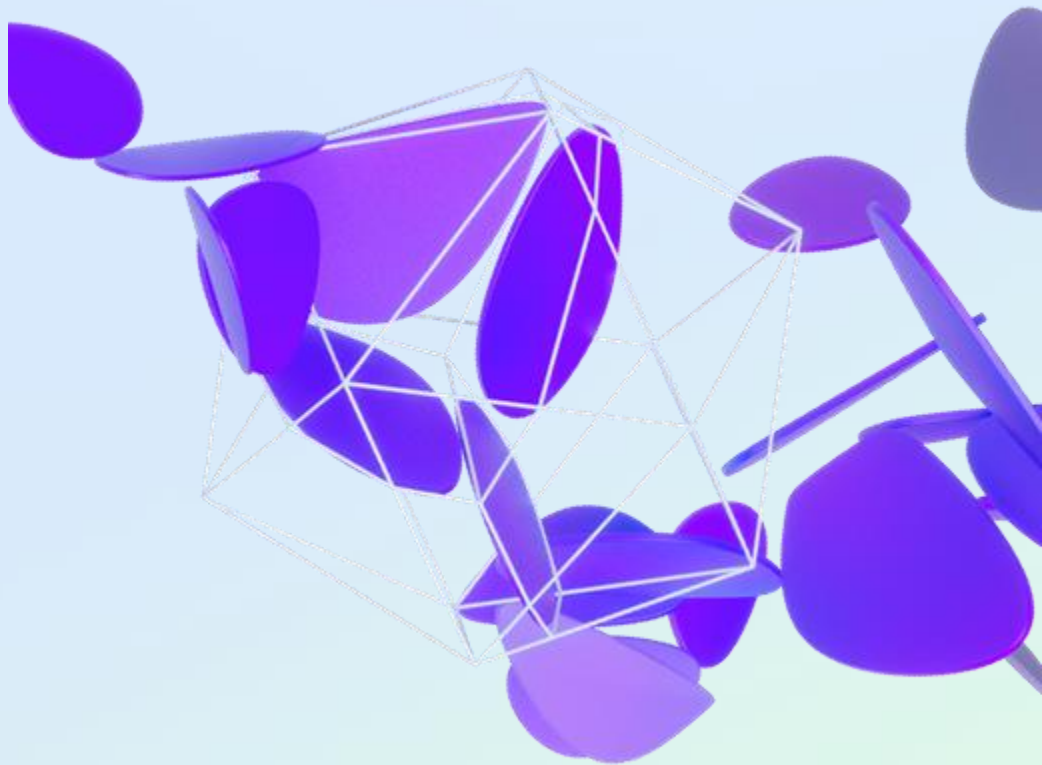
FAIR Chemistry

OMol Dataset

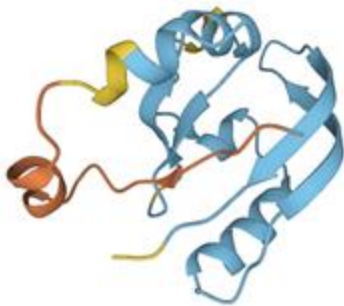
UMA Model Releases

Larry Zitnick, Director, Research Scientist

Zach Ulissi, Research Manager



Solving many real-world problems requires working at the atomic level



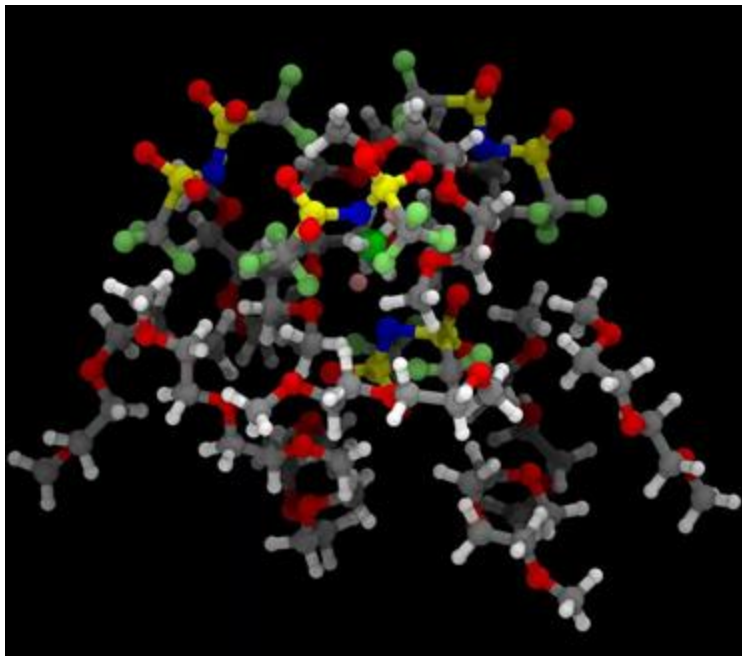
Drug Discovery



Climate



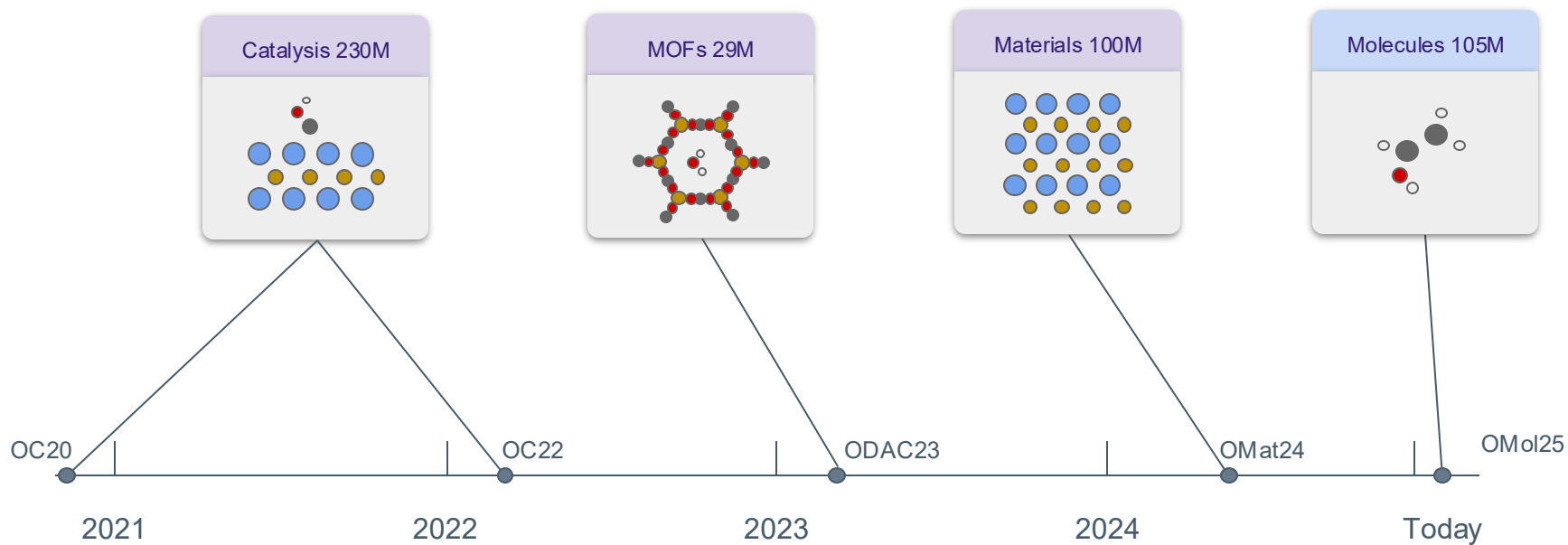
Augmented Reality



Density Functional
Theory (DFT) is slow

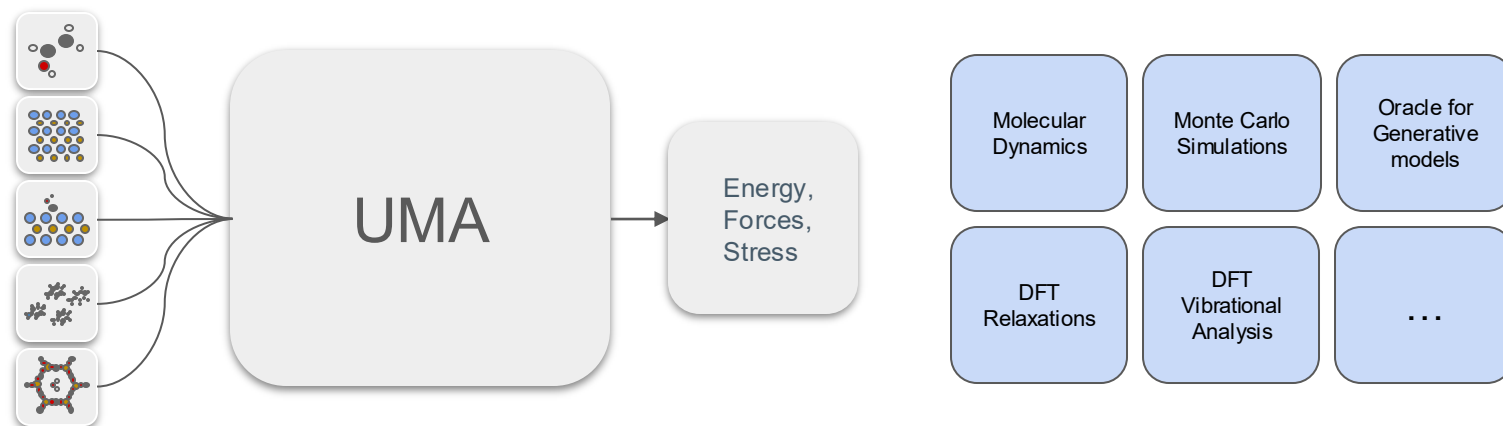
From days to seconds with AI...

Data



Universal Model for Atoms (UMA)

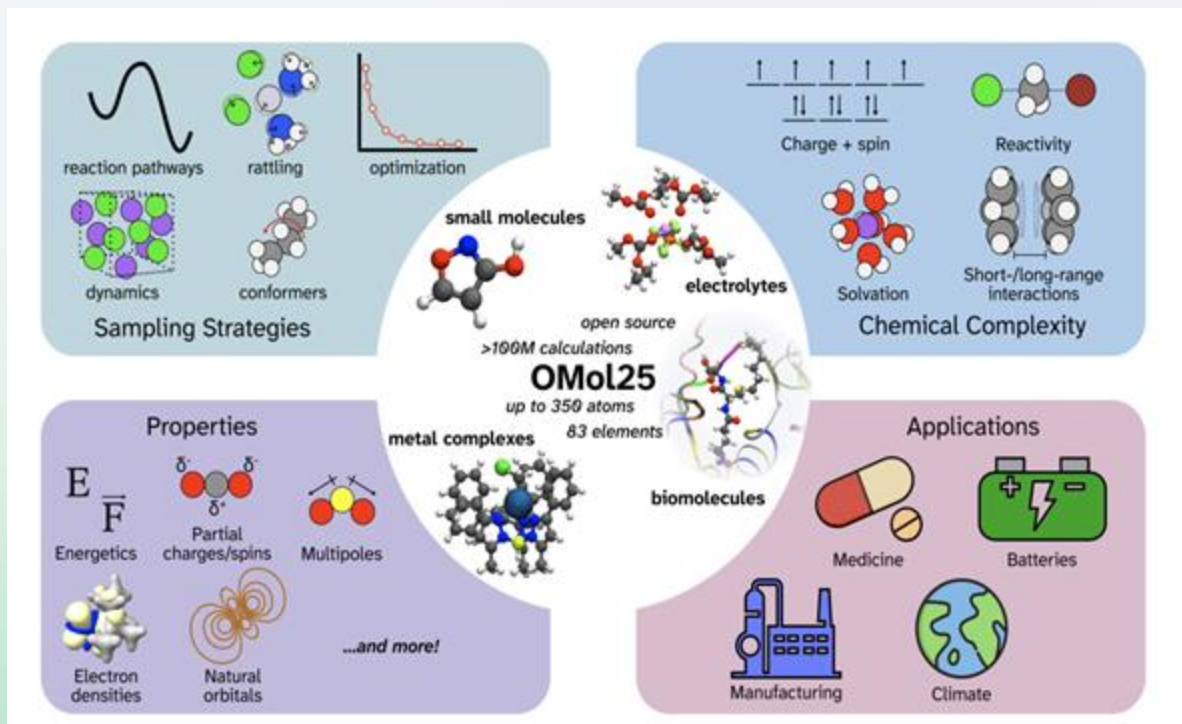
10,000x faster than DFT



Open Molecules (OMol) is comprised of 83M unique molecular systems, now the most chemically diverse and accurate molecular dataset

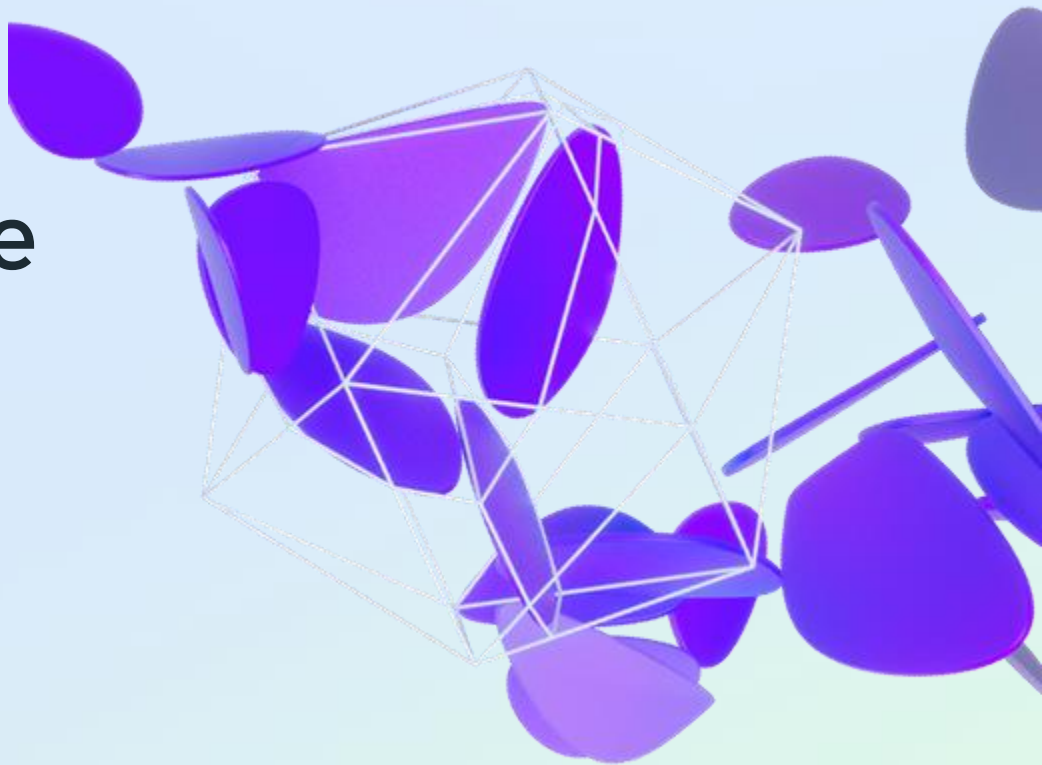
Universal Model for Atoms (UMA) enables researchers to compute properties from atomic simulations

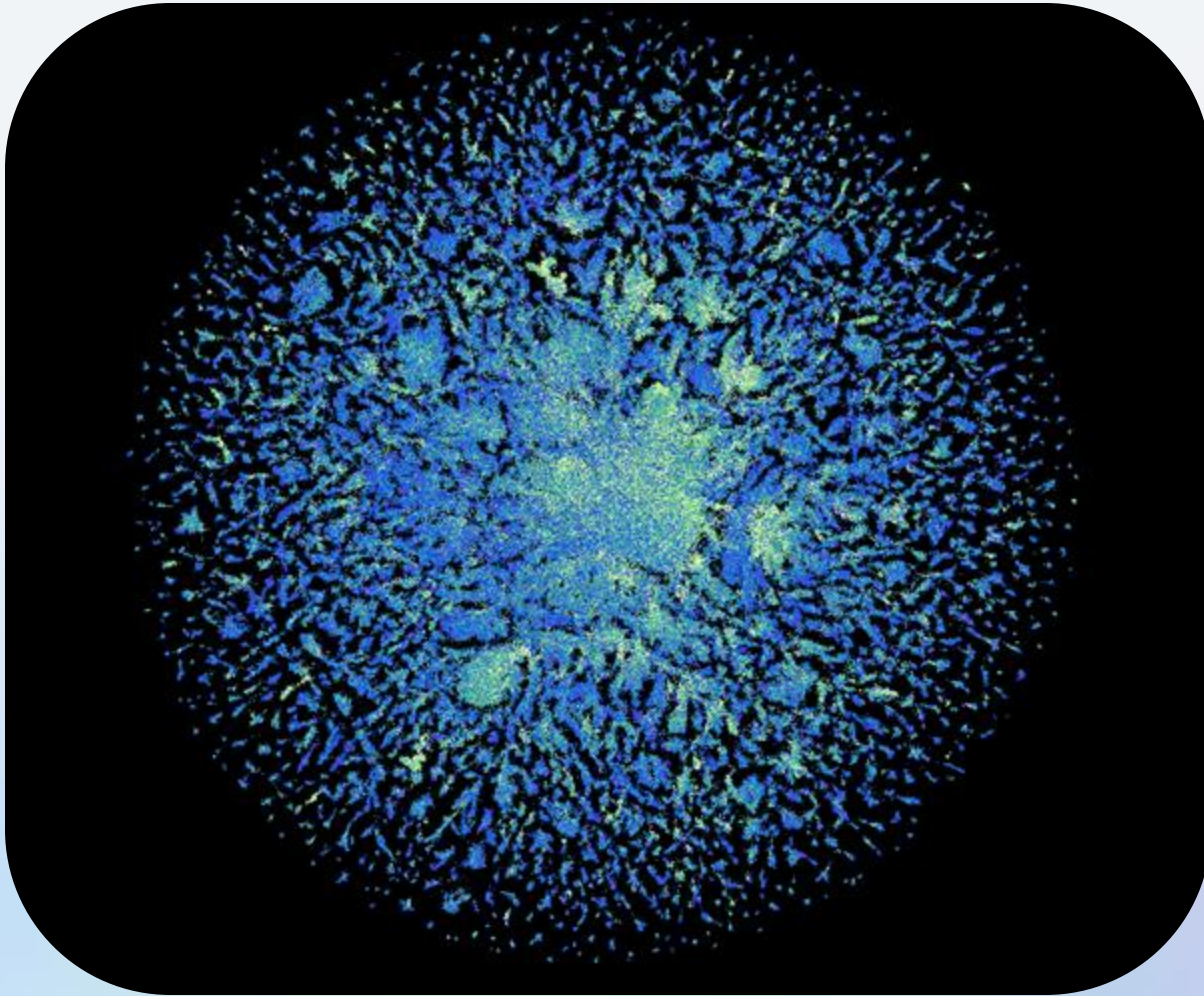
Enables AI-driven exploration of novel biomolecules



Levine et al, The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models" <https://arxiv.org/abs/2505.08762>

Evolutionary Scale Modeling (ESM)





Dataset of 600 million metagenomic structures, and **open source model** for protein structure prediction

Enabling antibody engineering and drug discovery

<https://ai.meta.com/blog/protein-folding-esmfold-metagenomics/>

Questions?

THANK YOU

