Revolutionizing Medical Image Data Analysis: Uniting AI and Statistics for Breakthroughs and Challenges

University of North Carolina at Chapel Hill

Hongtu Zhu

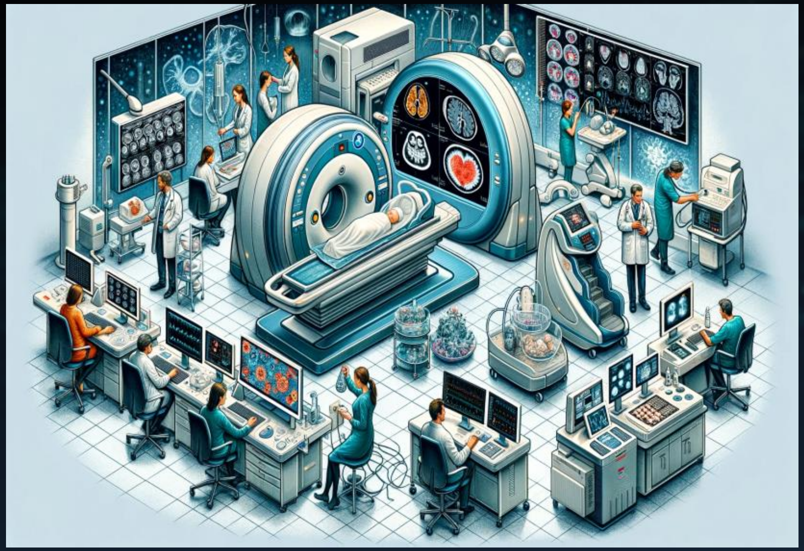
Thanks to Drs. Mingxia Liu, Xin Wang, Lijuan Liu, Gang Li, Hanchuan Peng, Wei Cheng, Marc Niethammer, Tengfei Li, and Bingxin Zhao for sharing their slides.

https://www.med.unc.edu/big-s2

CONTENTS

Part I

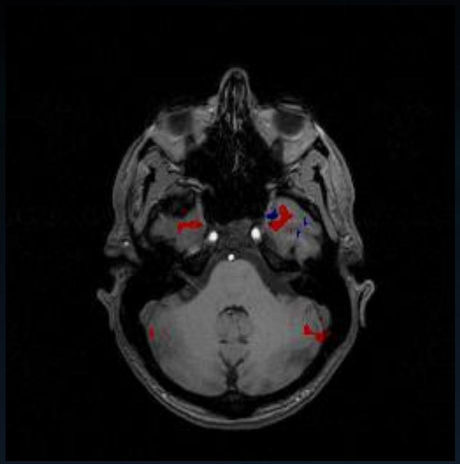
Introduction to Medical Image Data Analysis

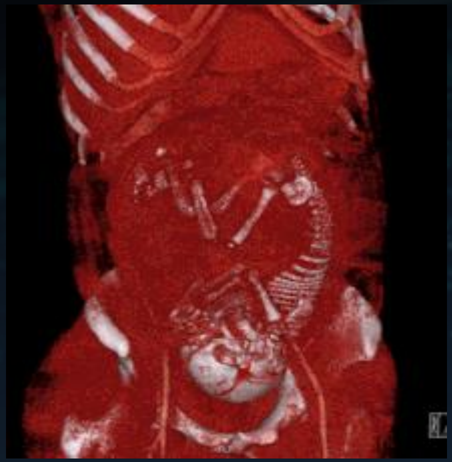


Medical Imaging

Medical imaging is the technique and process used to create images of the human body for clinical purposes or medical science. (https://en.wikipedia.org/)

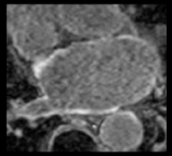
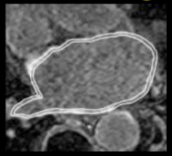
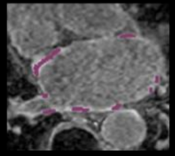
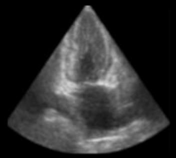
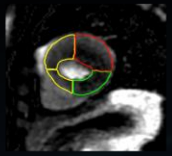
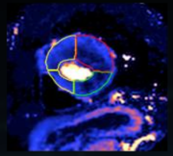
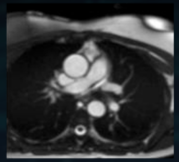
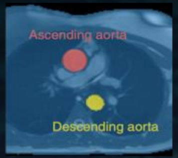
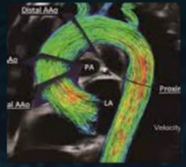
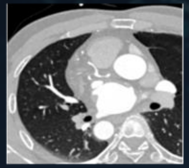
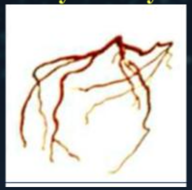
- ❑ These imaging methods are essential for delineating the **structure and functionality of organs and tissues**. Each modality employs a distinct targeting agent, generates data in varying dimensions, extracts unique features, and serves specific purposes within clinical and research contexts.

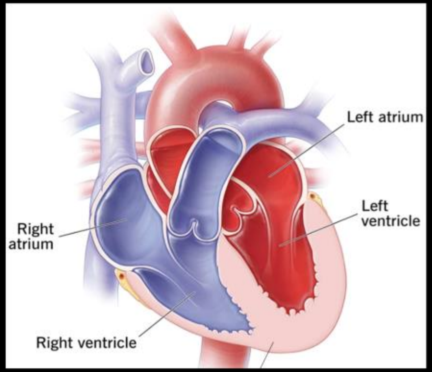


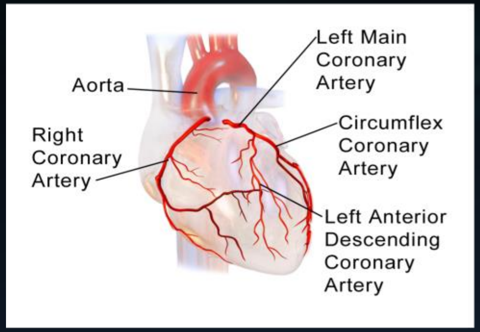


- X-ray radiography
- Computerized tomography (CT)
- Magnetic resonance imaging (MRI)
- Ultrasound
- Positron emission tomography (PET)
- ❖ Electroencephalography (EEG)
- ❖ Magnetoencephalography (MEG)
- ➢ Functional near-infrared spectroscopy (fNIRS)
- ➢ Mammography
- ➢ Light microscopy images
- ➢ Fluoroscopy
- ➢ Echocardiography

Cardiac Imaging

Heart	Target	Modality	Structure	Lesion/Function	Related Disease
Non-vessel	Atrium	**LEG MRI** 	LA Wall Seg 	LA fibrosis 	**Atrial Fibrillation**
	Ventricle	**Ultrasound** 	**Ventricle Seg** 	**Ventricle Function** 	**Ejection Fraction Estimation**
	Myocardium	**Myocardial Perfusion MRI** 	**Myocardium Seg** 	**Myocardium Function** 	**Ischemic Heart Disease**
Vessel	Aorta	**MRI** 	Aorta Seg 	Aorta Flow 	**Aorta Stenosis**
	Connery Arteries	**CTA** 	**Coronary Artery Seg** 	**Fractional Flow Reserve** 	**Coronary Artery Disease**





Wang, X. and Zhu, H (2024). Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook

Image Processing Analysis Methods

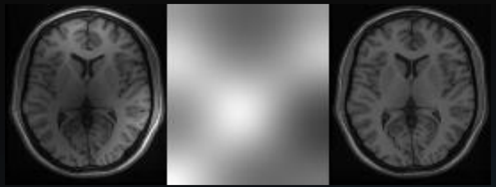
How to enhance and extract signals of interest in imaging data?

Deconvolution

$$\frac{S(\mathbf{q})}{S_0} = \int P(\mathbf{r}, \Delta) e^{i\mathbf{q}.\mathbf{r}} d\mathbf{r}; \quad \mathbf{q} = \gamma \delta g \mathbf{u}$$

➢ **Image Reconstruction**

➢ **Image Enhancement**



❖ **Image Segmentation**

❖ **Image Registration**

Structural Learning

Engineering

Statistics

Machine Learning

Mathematics

Structural Learning

Image Segmentation

- Organ parcellation
- Localization of pathology
- Surgical planning
- Image-guided interventions
- Computer-aided diagnosis
- Quantification of organ change

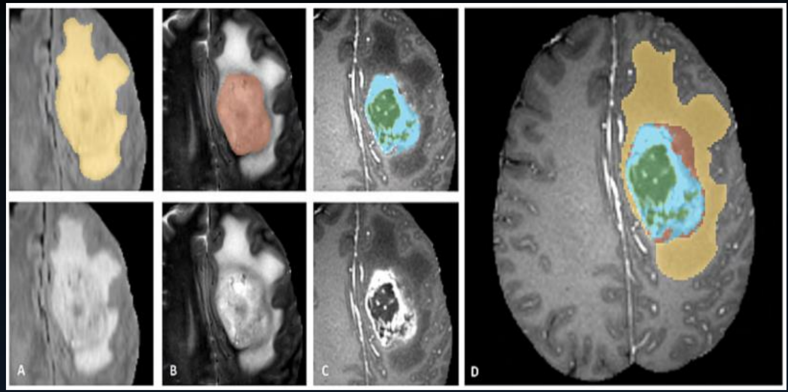
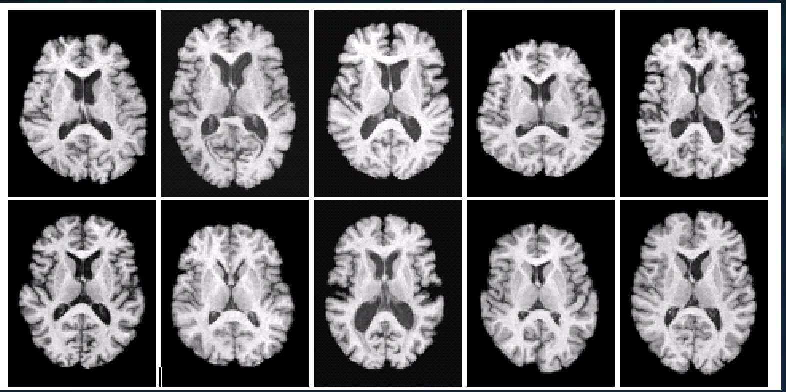
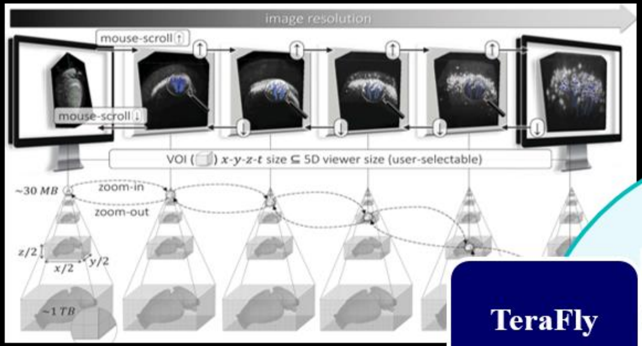


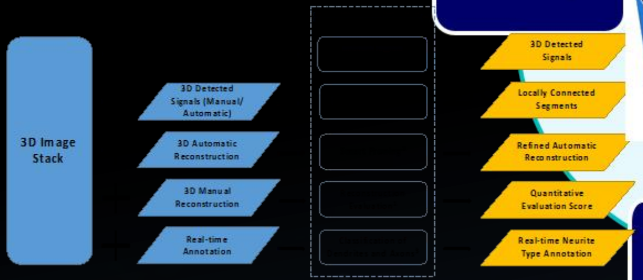
Image Registration

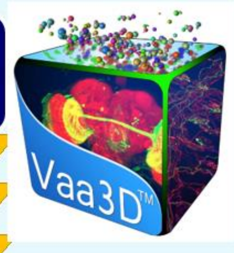
- Organ atlas
- Localization of pathology
- Automated image segmentation
- Multimodal fusion
- Population analysis
- Quantification of organ changes



Light Microscopy Imaging at Single Cell







UltraTracer

Virtual Reality

TeraFly

Data Protocols

Artificial Intelligence

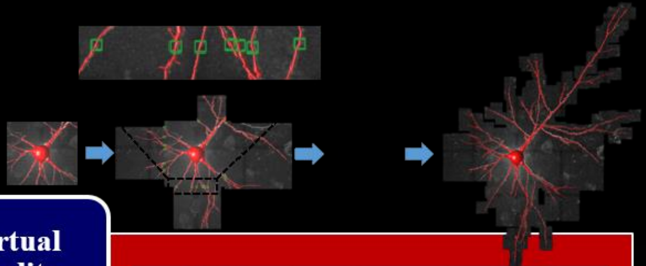
•••

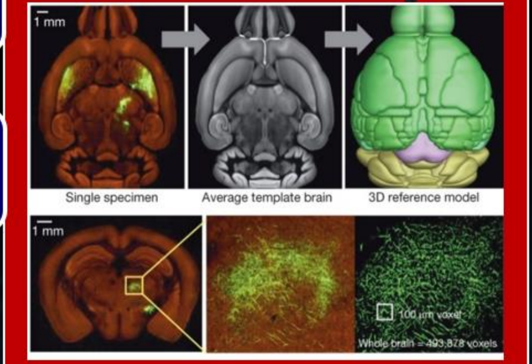
UltraTracer: Nature Methods 2017
TeraFly: Nature Methods, 2016
DeepNeuron: Brain Informatics 2018

Wang, et al. *Nature Commu* (2019)

Qu, et al. *Nature Methods*, (2022)

Han, X., et al. *Sci Adv.*, (2023)

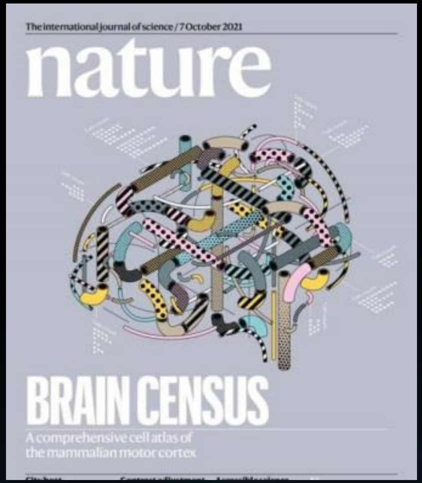


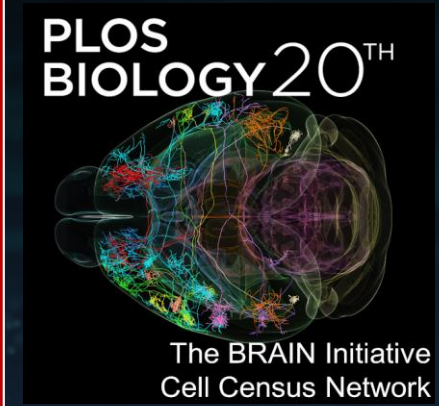


"Top 10 life scientific advances of 2021" China

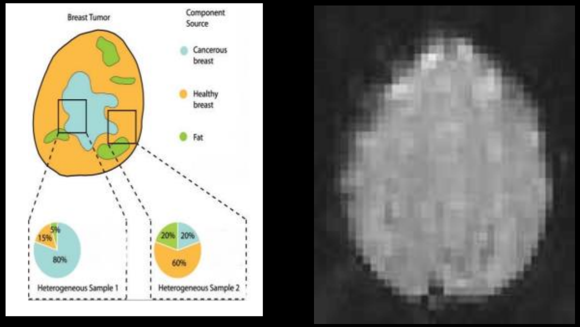
Morphological diversity of single neurons in molecularly defined cell types

Peng, H., et al. *Nature*, (2021)

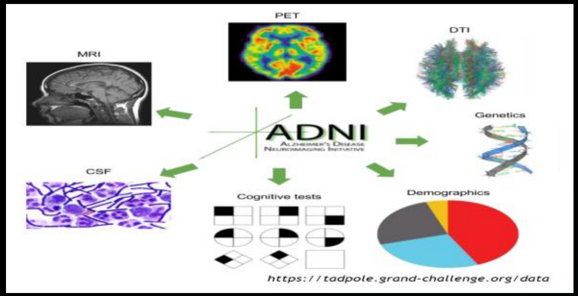




Ecological Layout for Imaging-based Analysis

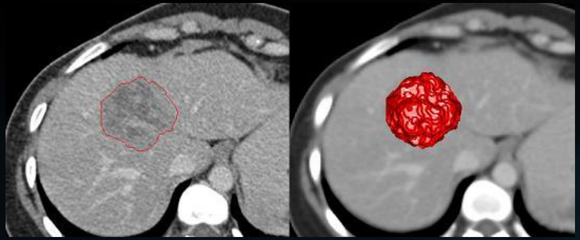


Deconvolution

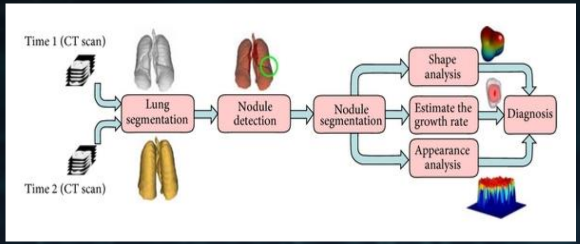


Integration

Imaging-related Database



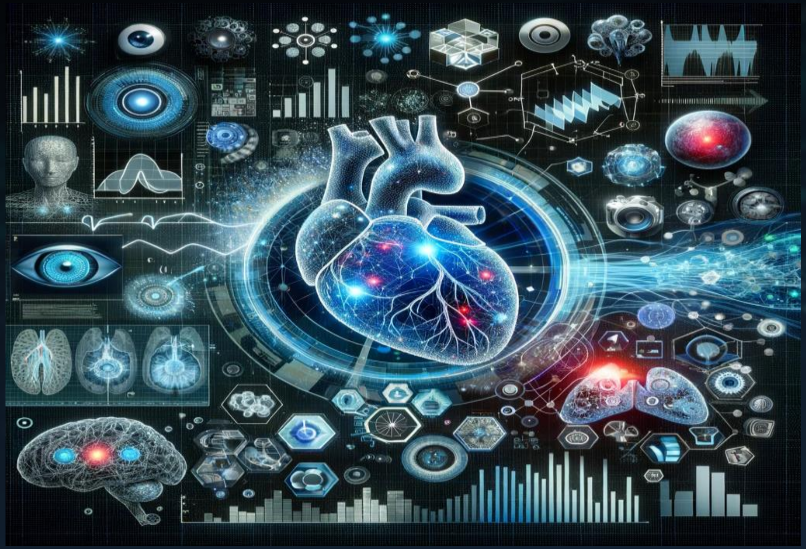
Structural Learning



Prediction

Part II

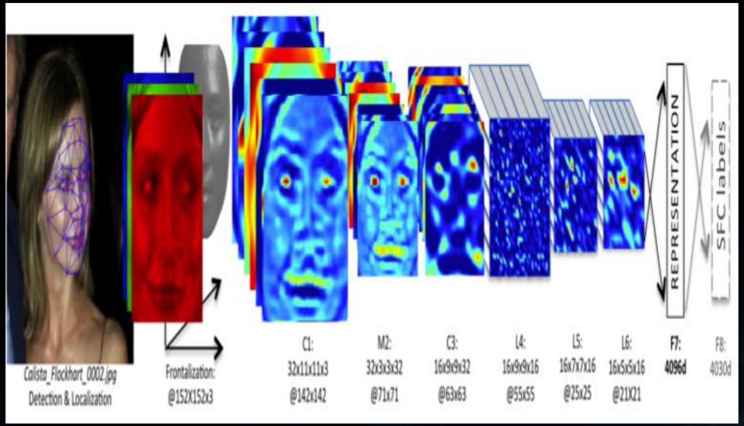
State-of-the-Art AI Applications in Medical Imaging and Statistical Challenges

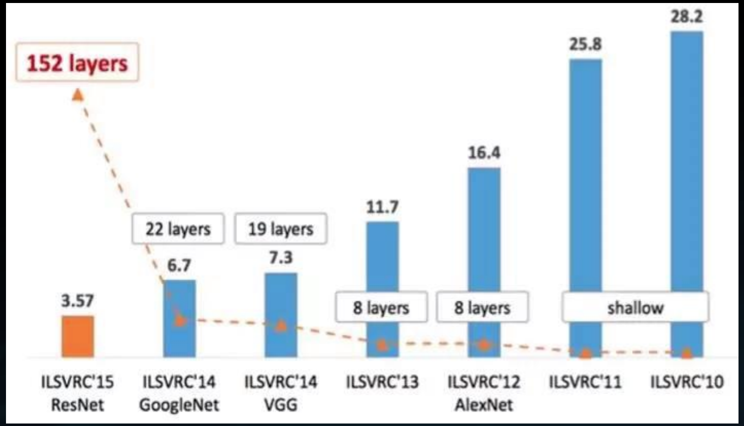


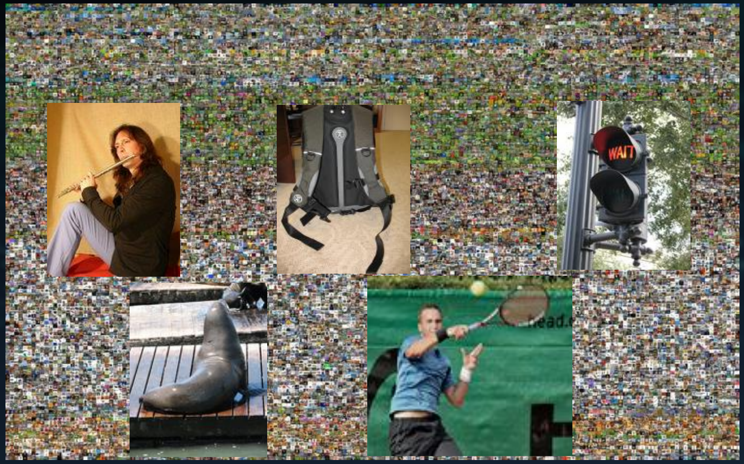
AI Milestones

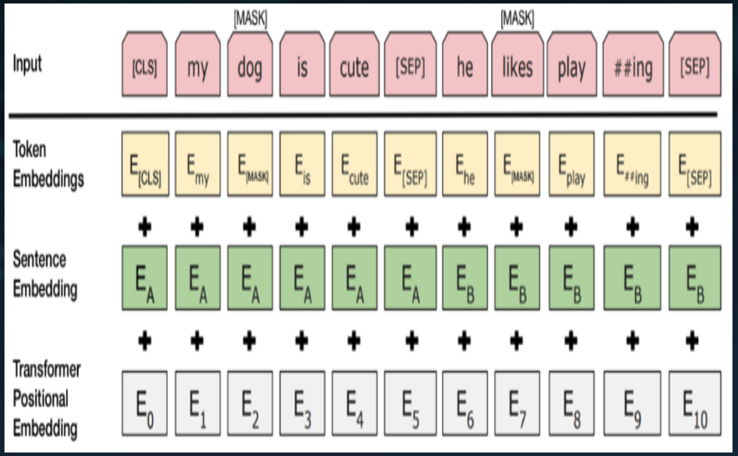
Annotated Datasets

Deep Learning



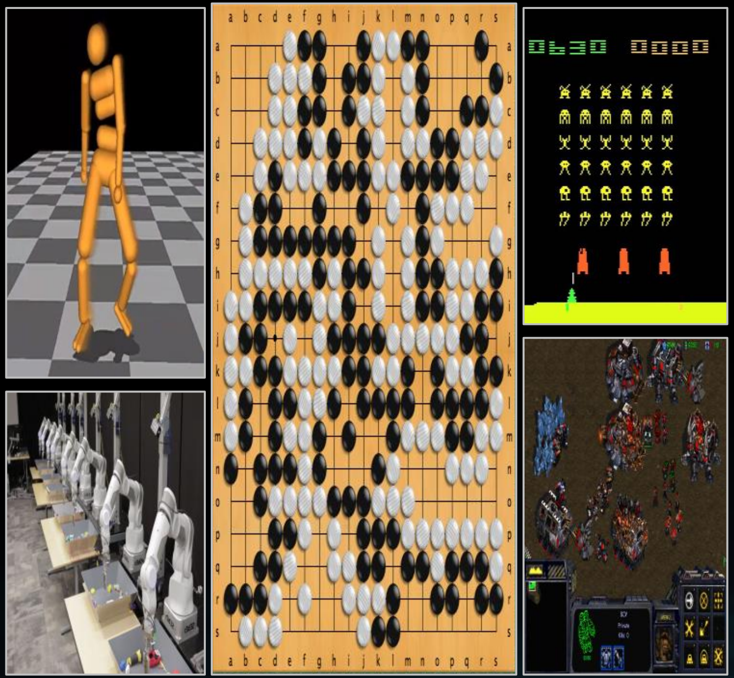




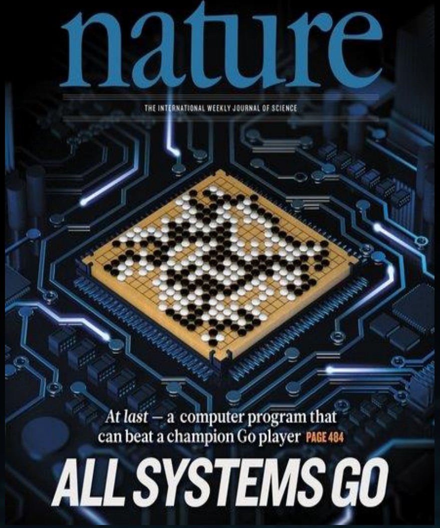


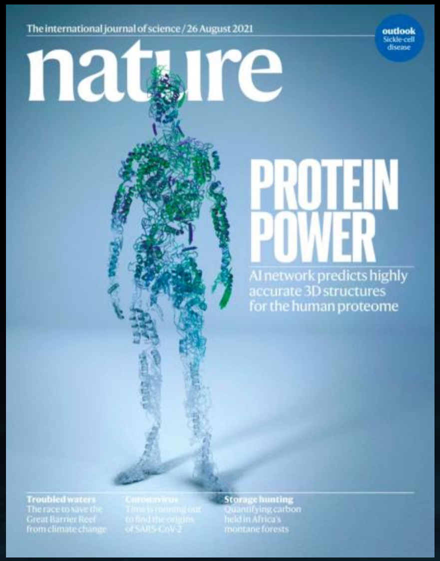
AI Milestones

Reinforcement Learning



AI Products

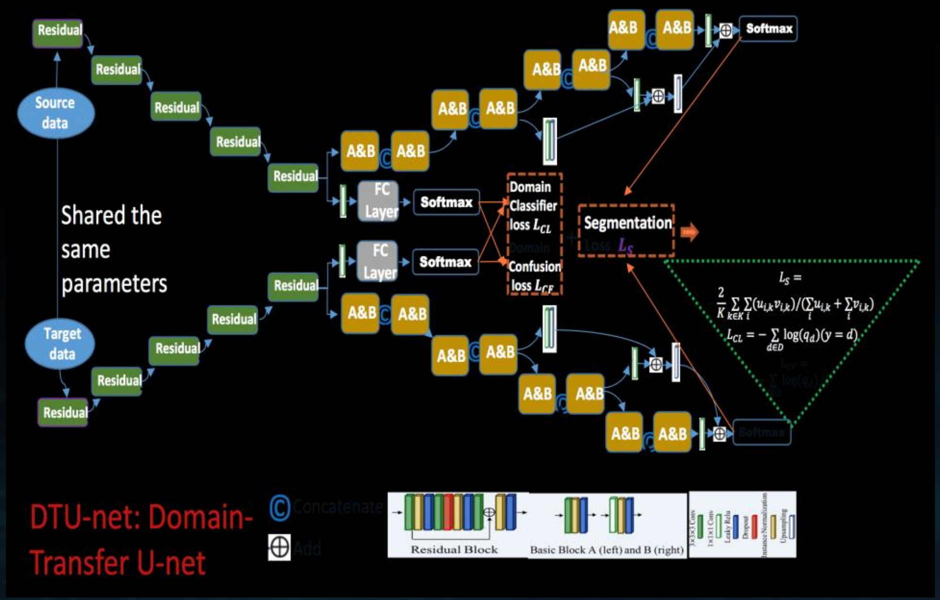




AI for Image Segmentation

Segmentation Annotation

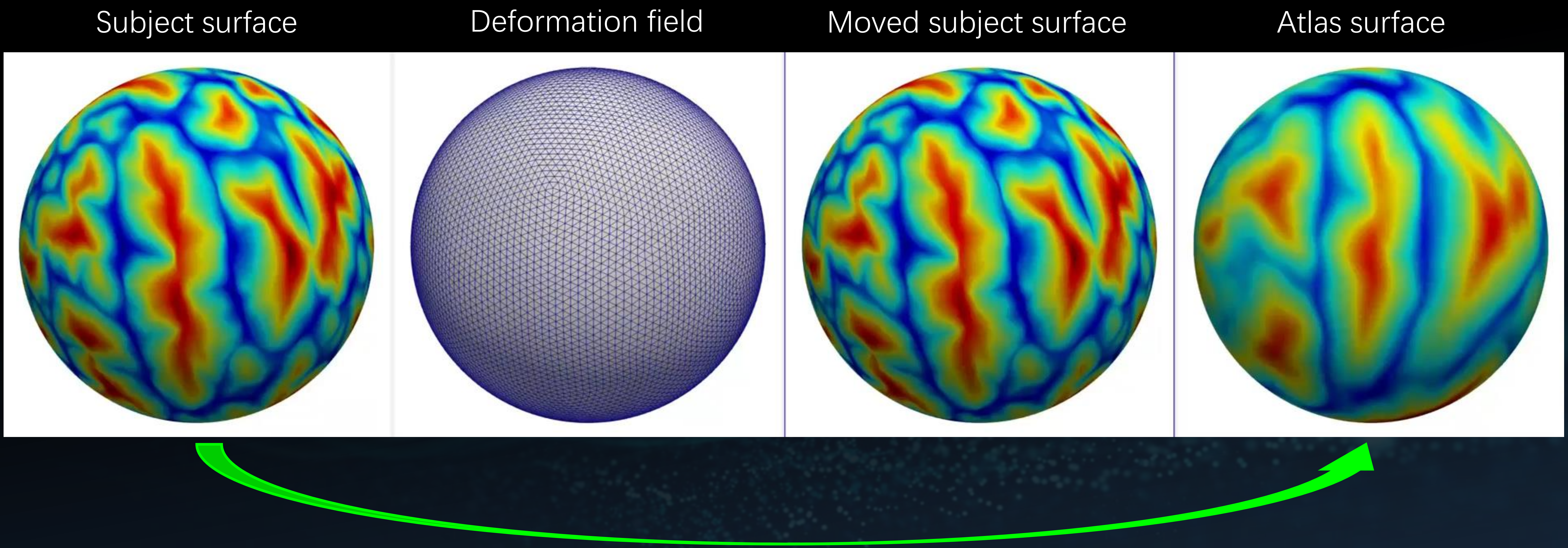
U-Nets



Liu, Q., Xu, Z., Bertasius, G., & Niethammer, M. (2023). SimpleClick: Interactive Image Segmentation with Simple Vision Transformers. ICCV., 22290-22300. 2023.

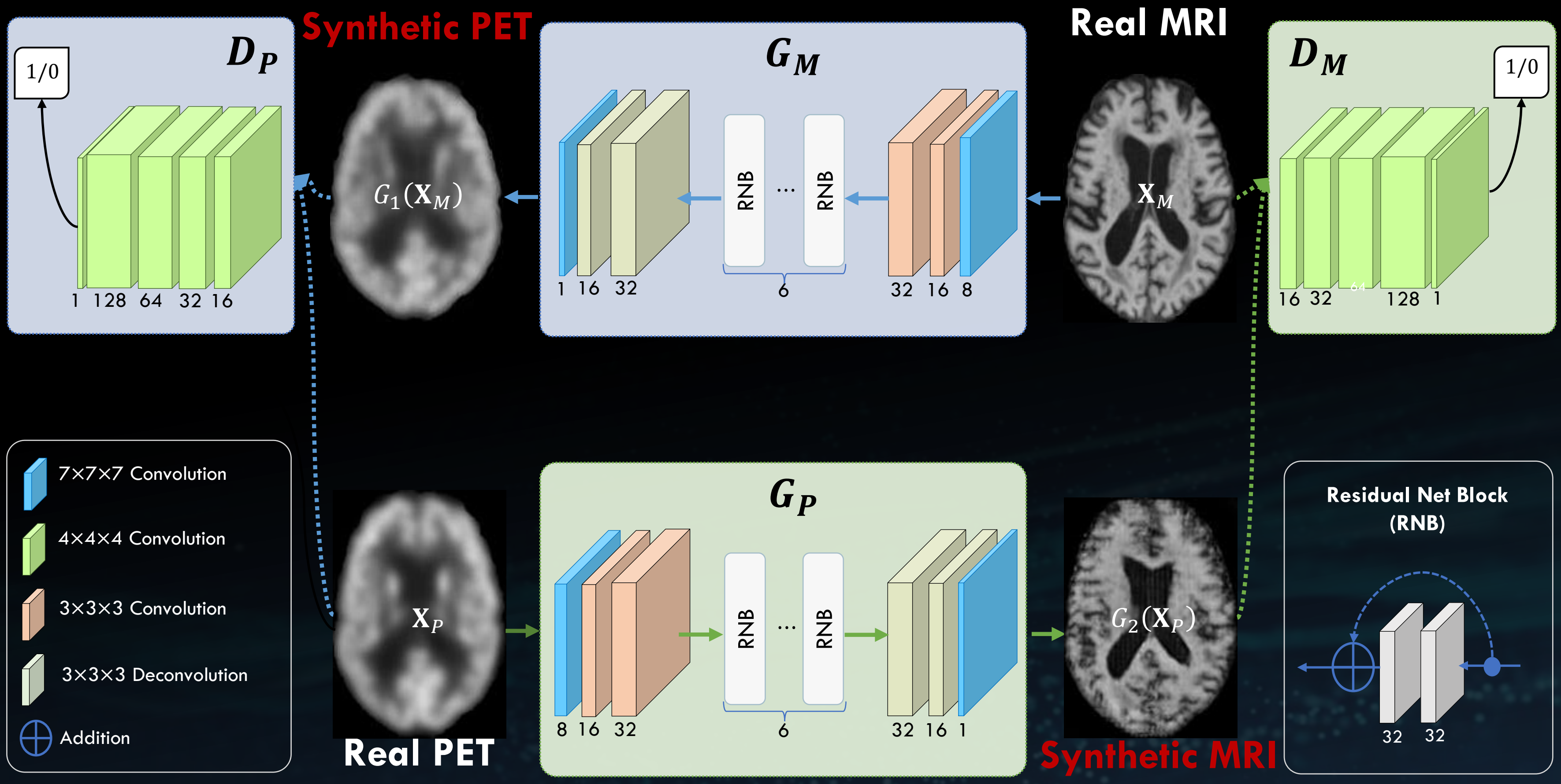
R. Azad *et al.*, “Medical Image Segmentation Review: The success of U-Net.” arXiv, Nov. 27, 2022.
Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE PAMI* 44.7 (2021): 3523-3542.

Superfast Spherical Surface Registration

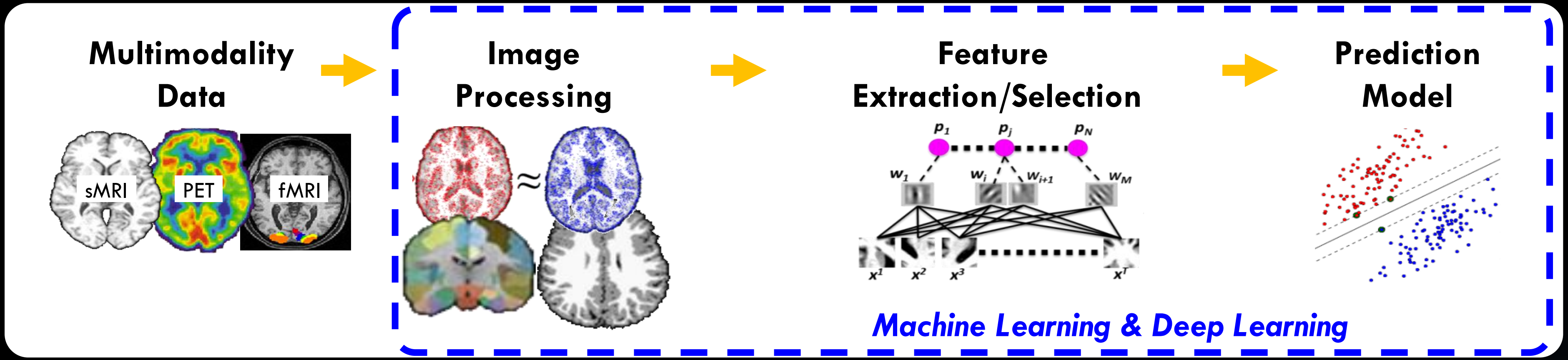


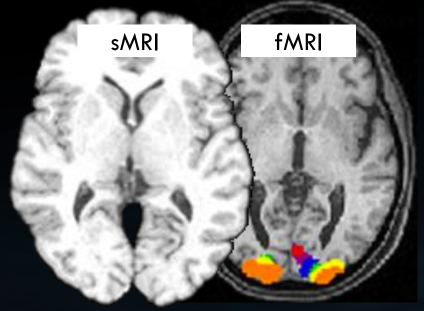
Zhao F, Wu Z, Wang F, Lin W, Xia S, Shen D, Wang L, Li G. S3Reg: Superfast Spherical Surface Registration Based on Deep Learning. IEEE Trans Med Imaging 2021; 40(8): 1964-1976.

Cross-Modality Image Synthesis

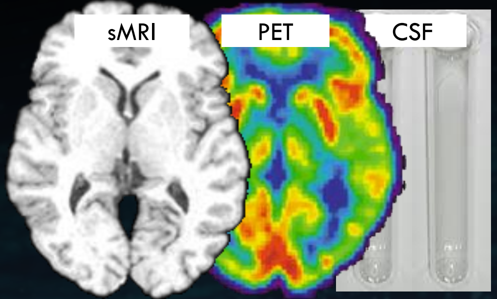


Computer-Aided Medical Data Analysis

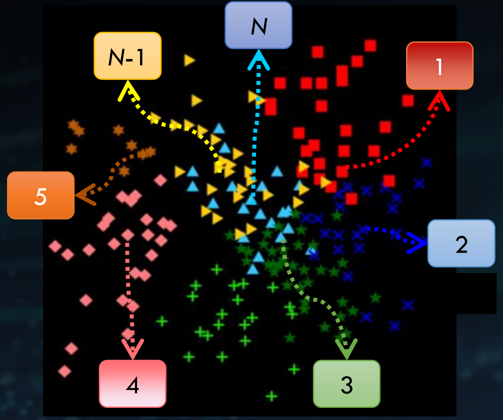




Neuroimage Representation Learning



Multimodality Data Fusion



Multi-Site Data Adaptation

Major Challenges

Complex Organs and Tissues

Heterogeneity within Individual Subjects and across Centers/Studies

Image=

f*(*age, gene, race, disease, others,

device, acquisition, noises)

- **There is no publicly available, high-quality imaging datasets with detailed annotation information that cover a large spectrum of segmentation tasks in healthcare.**
- **How to quantify the uncertainty and generalizarability of atlases as well as deconvolution and structural learning methods and results?**
- **How to develop RL method for various segmentation and registration tasks?**

Part III

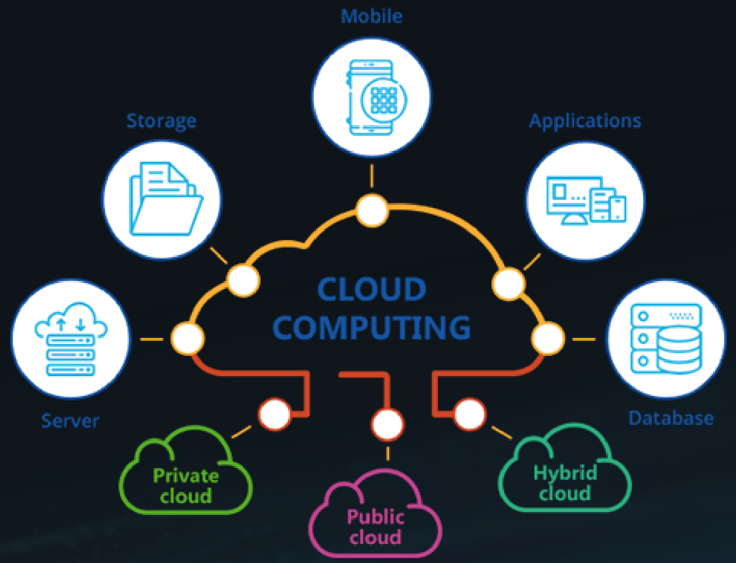
Opportunities for Statisticians in Advancing Medical Imaging Data Analysis

Application to ABC

Big Data

http://medium.com

Application



Computing

Analytical Tools

Applied Mathematics

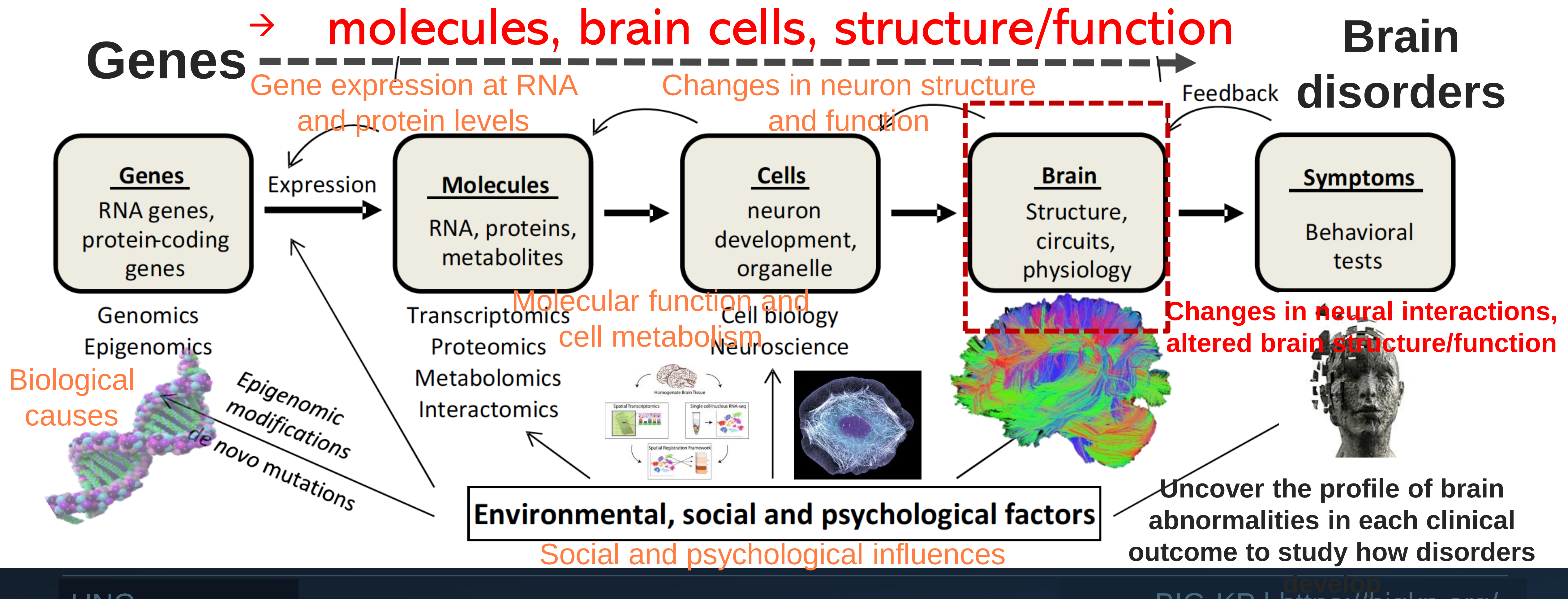
Statistics

Machine Learning

Engineering

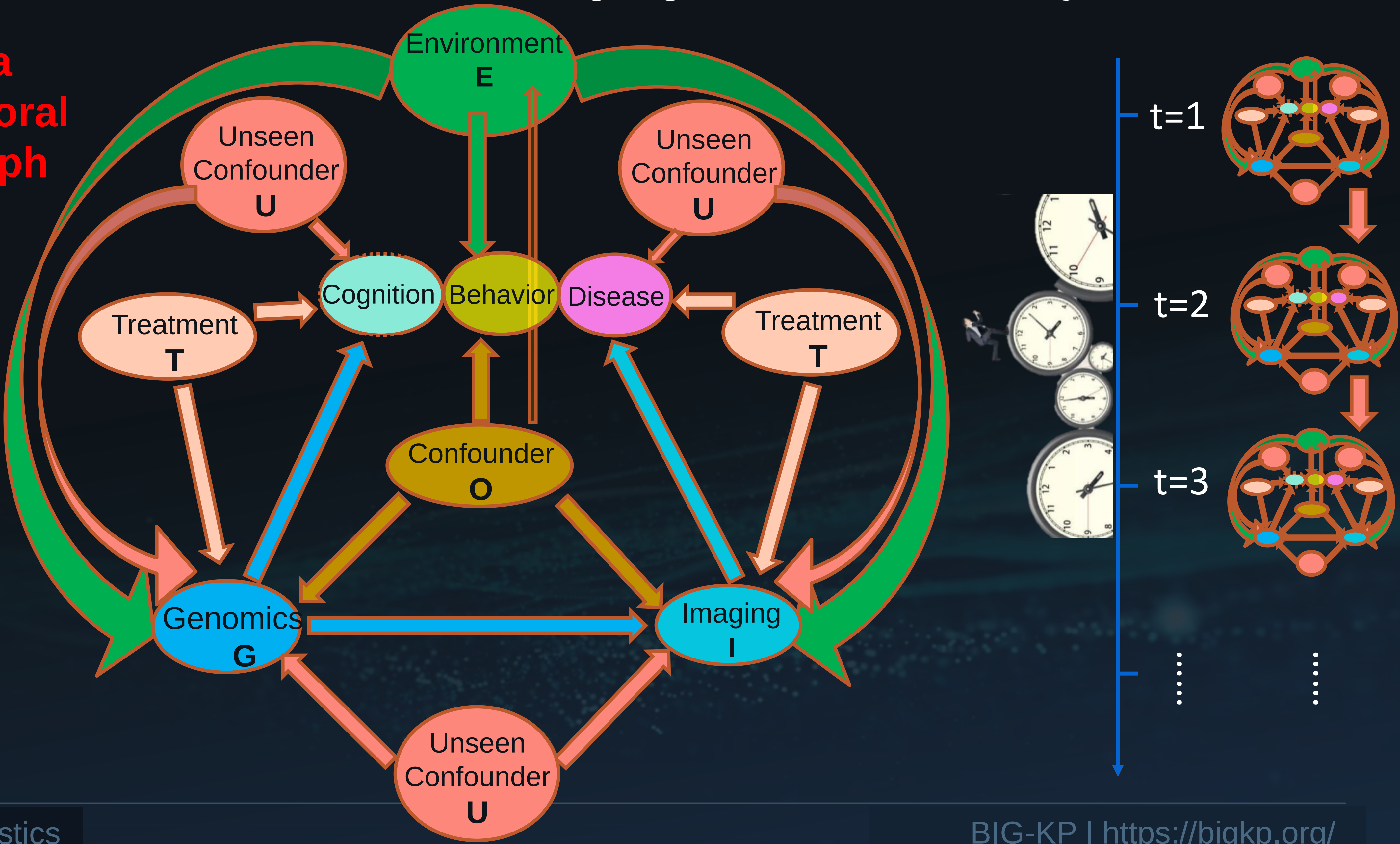
Brain Imaging Genetics Paradigm

Neuroimaging: an important component to help understand the complex biological pathways of brain disorders



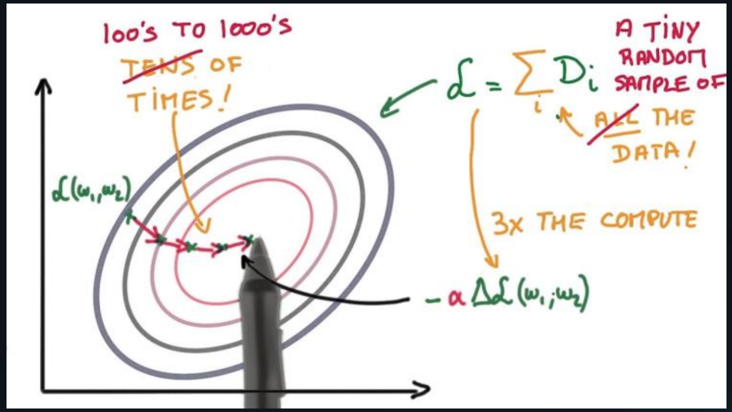
Causal Genetics Imaging Clinical Pathway

CGIC is a spatiotemporal causal graph



Methodological Challenges

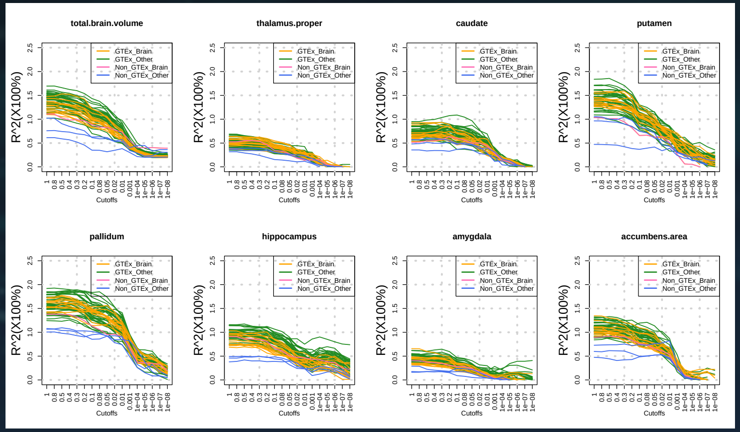
Multiple Biobanks/Trials Integration
(e.g., Heterogeneity in global populations)



New Computational Tools
(e.g., challenge of dense signal in biobank-scale database)

Database/
Tool/Theory for
Brain Imaging
Genetics

Omics Data Integration
(e.g., new tech, biological pathway)



Advanced Methods for Dense Signals
(e.g., deep learning)

Important Statistical Topics

- ❖ Experimental Design
- ❖ Statistical Parametric Mapping
- ❖ Object Oriented Data (OOD) Analysis
- ❖ Imputation Methods
- ❖ Data Integration Methods

- ➢ Dimension Reduction Methods
- ➢ Image Genetics
- ➢ Causality Research
- ➢ Predictive Analysis
- ➢ Knowledge-based Methods
- ➢ Reinforcement Learning

Zhu, H., Li, T., & Zhao, B. Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science, Volume 6, Issue 1, 2023.*

Other Important Topics

AD and ADRD Related Datasets

	ADSP	ARIC	ADNI	ADGC	UKB	CHS	FHS	HRS	GR@ACE	NACC-UDS	ROSMAP	MSBB	A4	WRAP	OASIS3
Samples (k)	22.8	15.7	2.2	22.6	500	5.8	10.4	15.7	7.4	43.9	3.6	0.37	6.9	1.7	1.09
AD Cases (k)	11.0	3.2	0.4	11.8	3.1	0.4	0.2	1.1	4.1	17.9	0.8	0.28	-	0.02	0.47
Other ADRD (k)	-	-	-	-	3.8	-	-	-	-	0.9	-	-	-	-	-
AD Candidate (k)	-	-	-	-	62.1[1]	-	-	-	-	-	-	-	1.1[2]	-	-
MCI Cases (k)	-	1.3	1.0	-	-	0.5	0.1	-	1.5	7.7	0.4	0.05	-	0.12	-
Longitudinal	N	Y	Y	N	Y	Y	Y	Y	N	Y	Y	N	N	Y	Y
Female (%)	61.1	55.0	47.0	59.6	51.6	62.0	52.2	53.6	60.4	57.2	72.7	64.2	57.7	70.2	55.6
Race (%)															
White	72.7	73.0	91.3	92.2	94.4	88.3	100	83.7	100	79.2	92.9	82.0	88.9	83.7	80.9
Black	13.4	27.0	5.0	-	1.5	11.7	-	16.3	-	12.7	5.9	10.3	4.9	11.7	14.9
Other	13.9	-	3.7	7.8	4.1	-	-	-	-	8.1	1.2	7.7	6.2	4.6	4.2
Ethnicity (%)															
Non-Hispanic	83.1	100	95.6	92.2	100	100	100	90.7	100	91.5	94.9	92.8	95.0	97.8	95.8
Age range	32-89	45-84	54-91	60+	44-82	65+	30-62	51-61	65+	36+	54+	61+	65-85	43-90	42-96
Genetic/Omics data	WGS, WES	WGS, Protein and RNA	GWAS	GWAS	GWAS, WES, WGS	GWAS	GWAS	GWAS, WES	GWAS	GWAS	Protein and RNA, WGS	WES, Protein and RNA, WGS	GWAS	WGS	N
Imaging data (# of subjects)	N	Brain MRI (2k), PET scan (0.3k)	Brain MRI (1.7k), PET scan (1.3k)	N	Brain MRI (45k)	Brain MRI (3.6k)	Brain MRI (5.2k)	N	N	Brain MRI (6k)	N	N	Brain MRI (1.8k), PET scan (4.4k)	Brain MRI (k), PET scan (0.45k)	Brain MRI (0.91k), PET scan (0.78k)
Other variables/ phenotypes	N	Cardiovascular risk factors/biomarkers, Neurologic function, Echo/Electro-cardiography, Functional status, β-amyloid, Total/phosphorylated tau	Cardiovascular risk factors/biomarkers, Neurologic function	Braak staging	Cardiovascular risk factors/biomarkers, Neurologic function, Functional status, Echo/Electro-cardiography	Cardiovascular risk factors/biomarkers, Neurologic function, Functional status, Echo/Electro-cardiography, β-amyloid	Cardiovascular risk factors/biomarkers, Echo/Electro-cardiography, Neurologic function, β-amyloid, Total/phosphorylated tau	Cardiovascular risk factors/biomarkers, Functional status, Neurologic function	N	β-amyloid, Total/phosphorylated tau	Braak staging, Neurologic function	Braak staging, Cardiovascular risk factors/biomarkers	Functional status, Neurologic function, β-amyloid, Total/phosphorylated tau	Neurologic function, Functional status	Functional status, Neurologic function

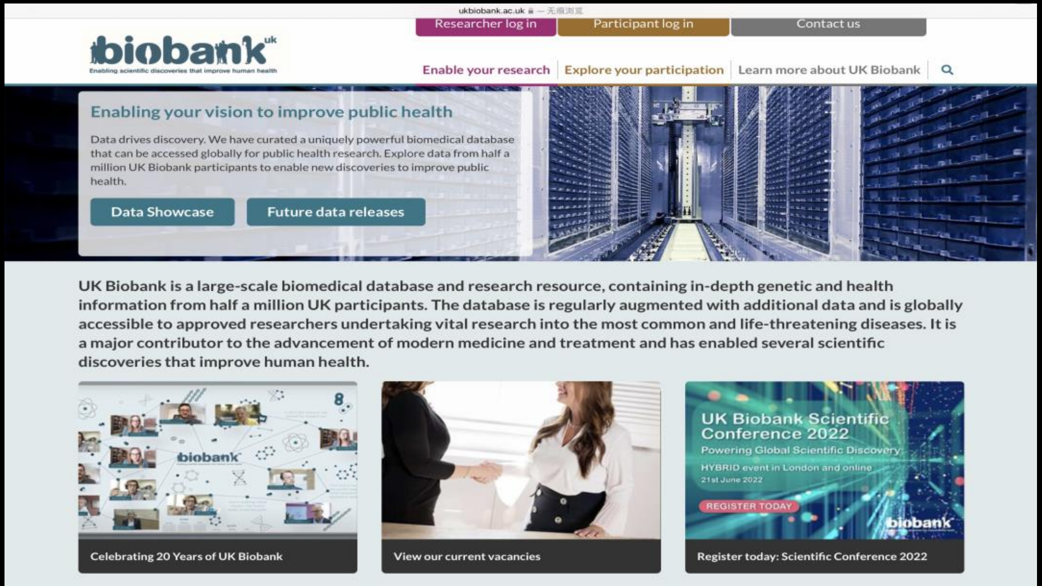
[1]: Subjects with AD-proxy for UKB (if either parent has AD) [2]: AD candidates (for A4) with an "elevated" level of amyloid plaque detected from the PET scan

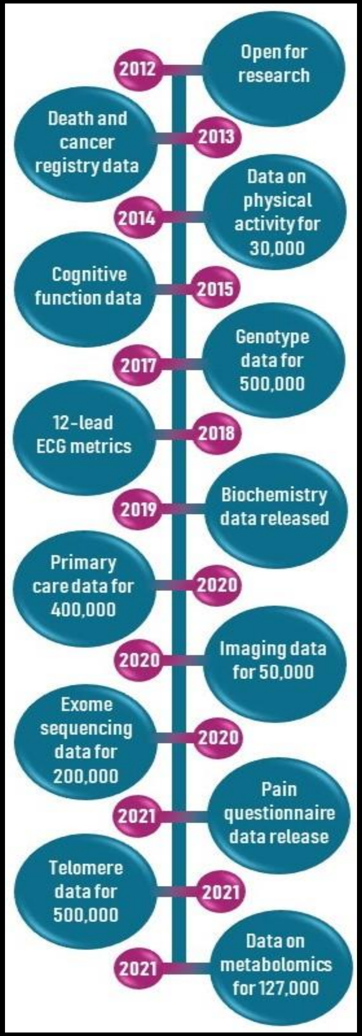
GWAS; Protein and RNA; Brain MRI; Cardiovascular risk factors/biomarkers; Neurologic function; β-amyloid; Total/phosphorylated tau
WES; WGS; PET scan; Braak staging; Functional status; Echo/Electro-cardiography

Figure 2. Summary information on 15 studies of the AD-related database (sample size, age, sex, race, data type, etc.)

The UK Biobank Study

UK Biobank has collected and continues to collect extensive environmental, lifestyle, and genetic data on half a million participants.





2006-now

- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Genotyping, whole exome sequencing & whole genome sequencing for all participants.
- **Health linkages:** Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.
- **Biomarkers:** Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.
- **Activity monitor:** Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.
- **Online questionnaires:** Data on a range of exposures and health outcomes that are difficult to assess via routine health records, including diet, food preferences, work history, pain, cognitive function, digestive health and mental health.
- **Repeat baseline assessments:** A full baseline assessment is undertaken during the imaging assessment of 100,000 participants.
- **Samples:** Blood & urine was collected from all participants, and saliva for 100,000.

Data Preprocessing and Data Modeling

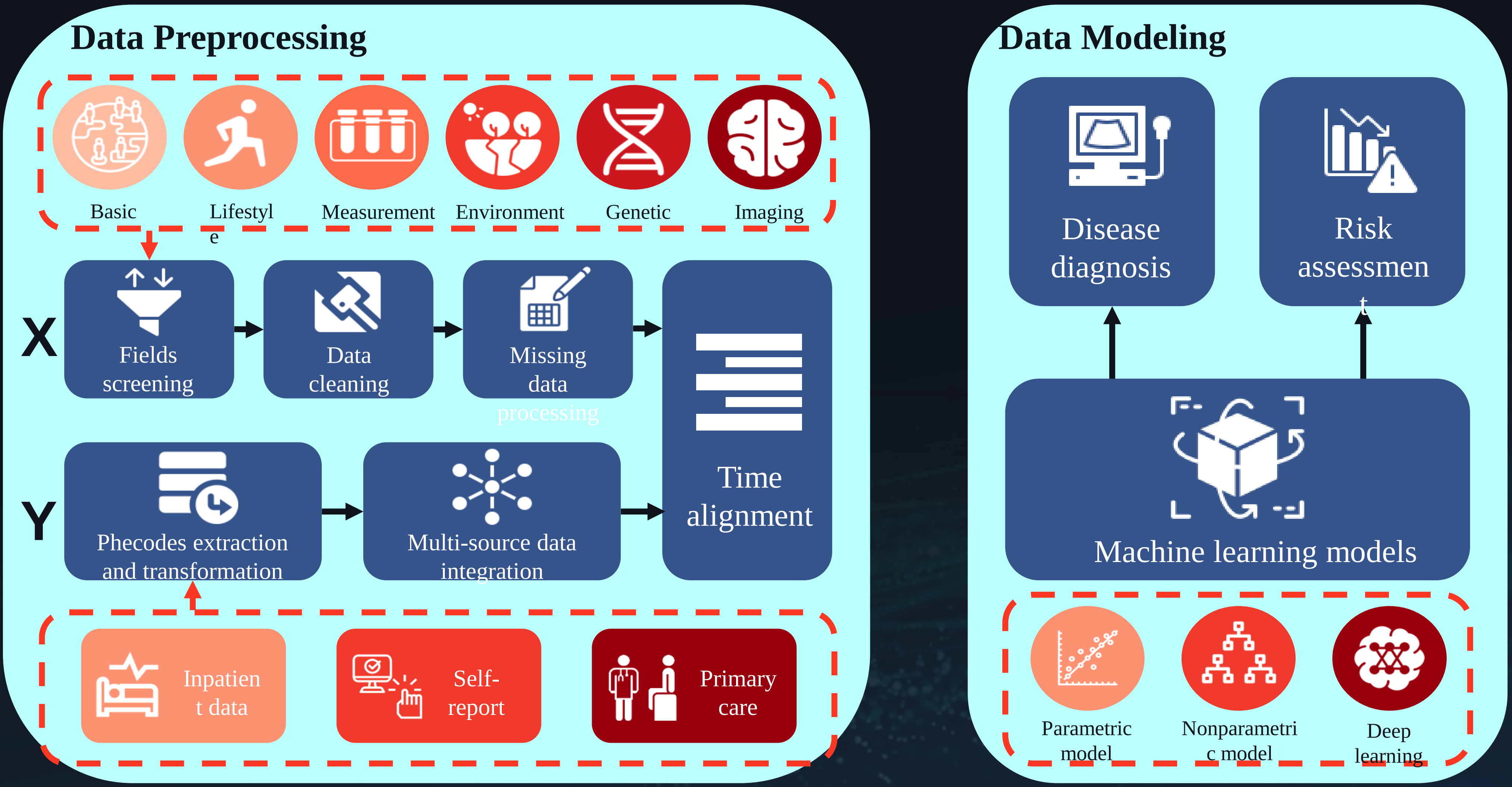
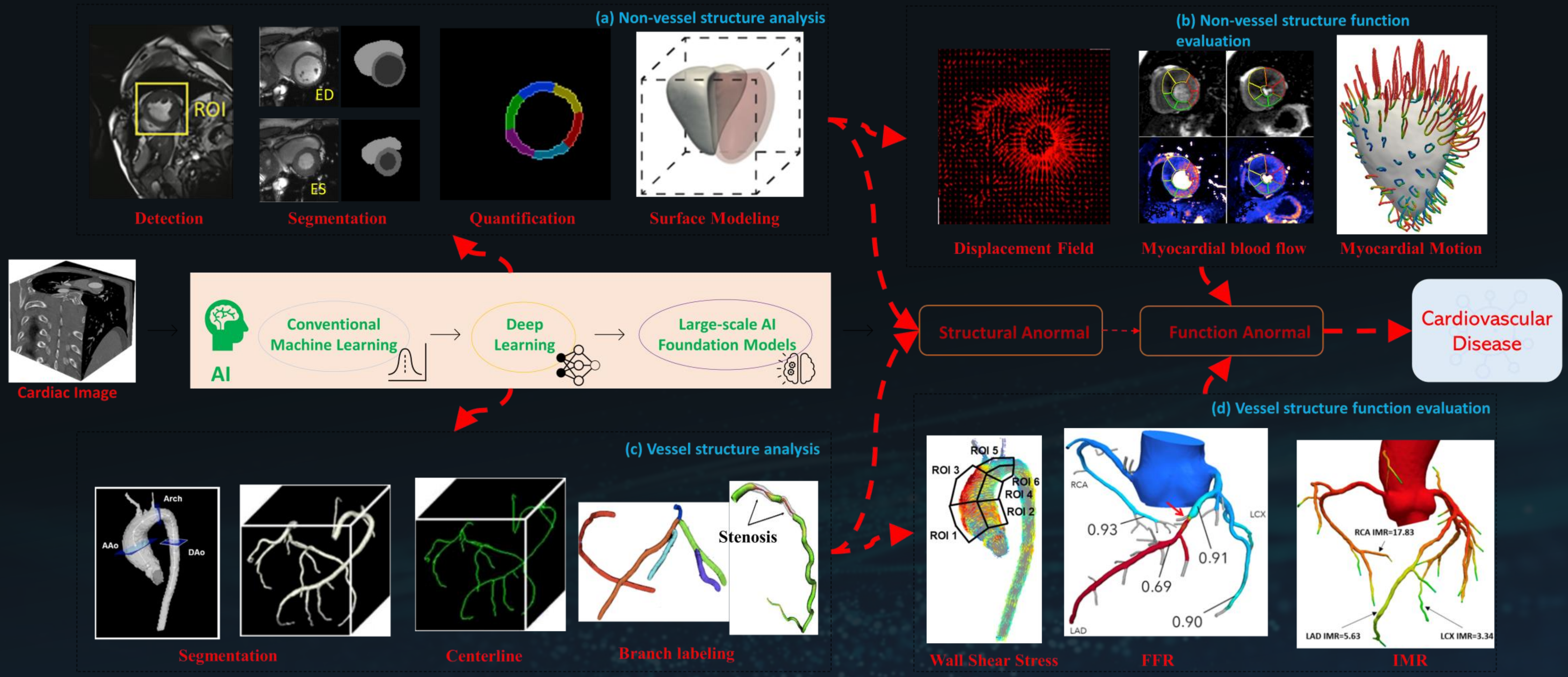


Image Analysis Pipeline



Prediction Models

Preprocess

Baseline	Baseline	Baseline
Lifestyle	Lifestyle	Lifestyle
Measurement	Measurement	Measurement

Purpose determined data input

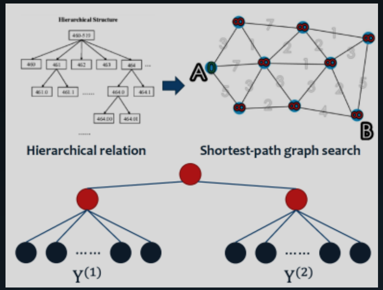
Baseline Imaging Environment
Genetic Lifestyle Disease Diagnosed Measurement

Baseline Imaging Environment
Genetic Lifestyle Disease Diagnosed Measurement

Model

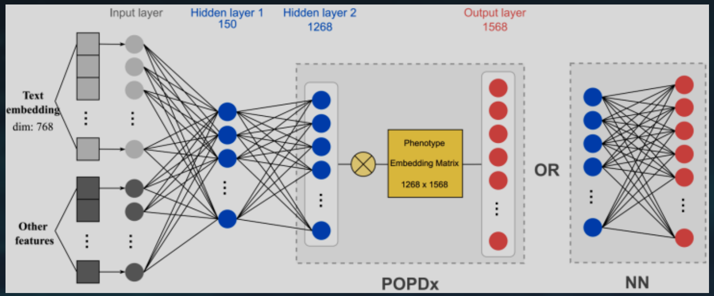
Boosting | POPDX | Neural network | Logistic regression | ...

CoxPH | Deep Surv | Deep Hit | POPDx Surv | ...



Disease diagnosis & Risk assessment

- Individual disease
- Categorized disease
- Joint diagnosis and evaluation

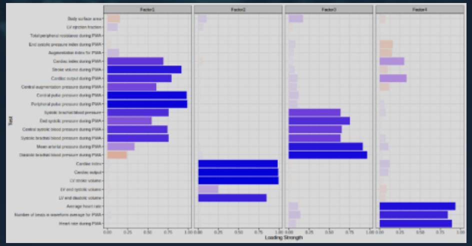


Post analysis

Feature importance extraction

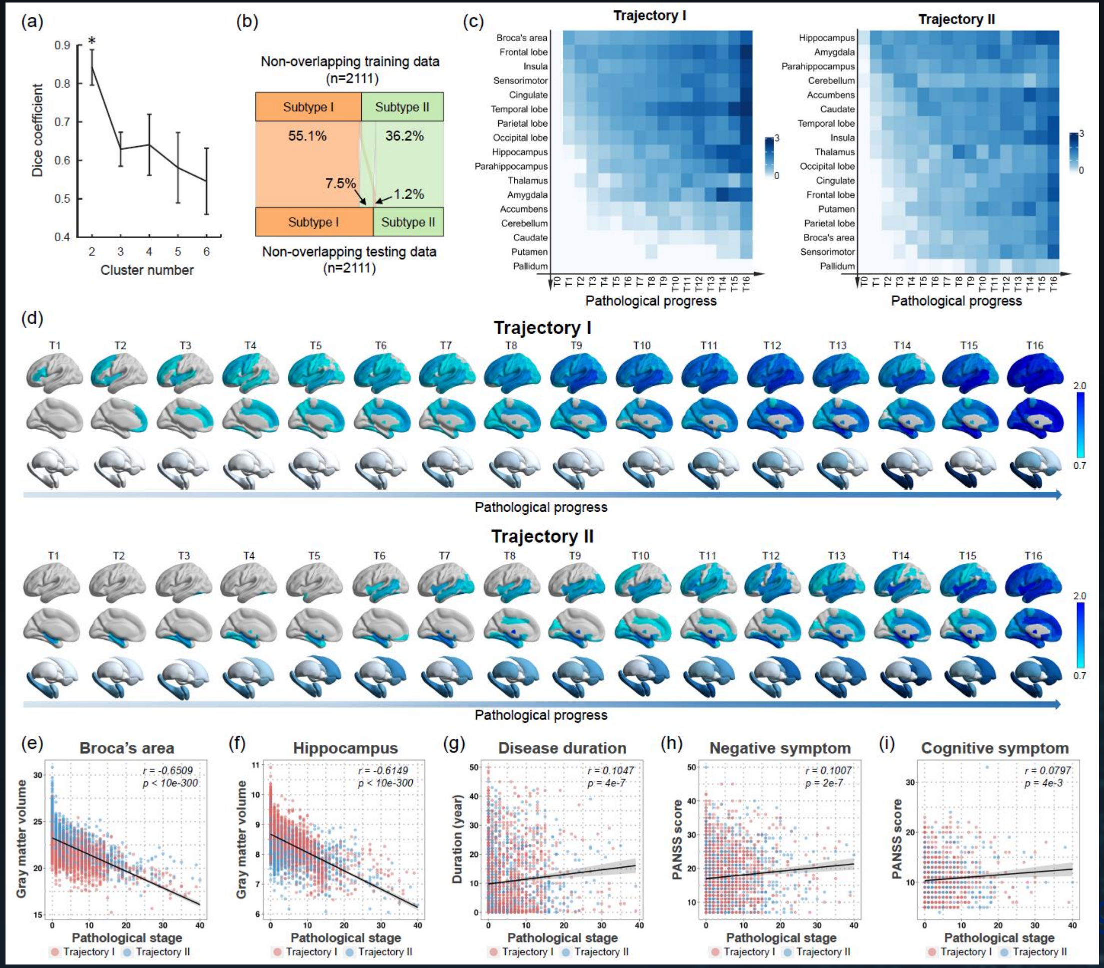
- Deep Lift
- SHAP
- Integrated Gradients

Data-driven dimension reduction

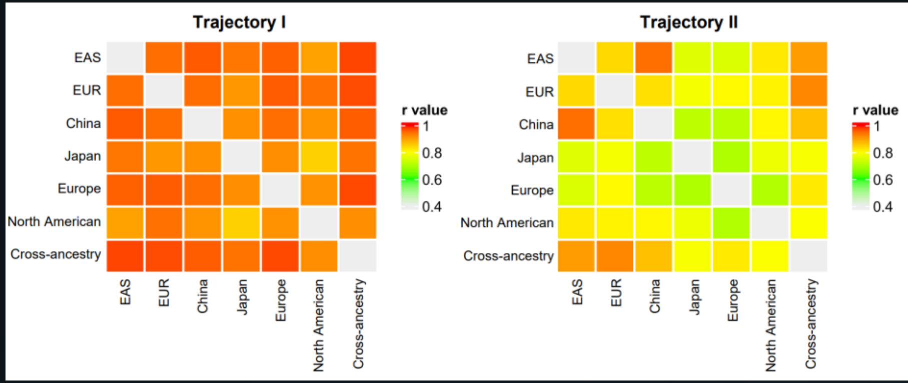


Neuroimaging Biomarkers for Subtypes of Schizophrenia

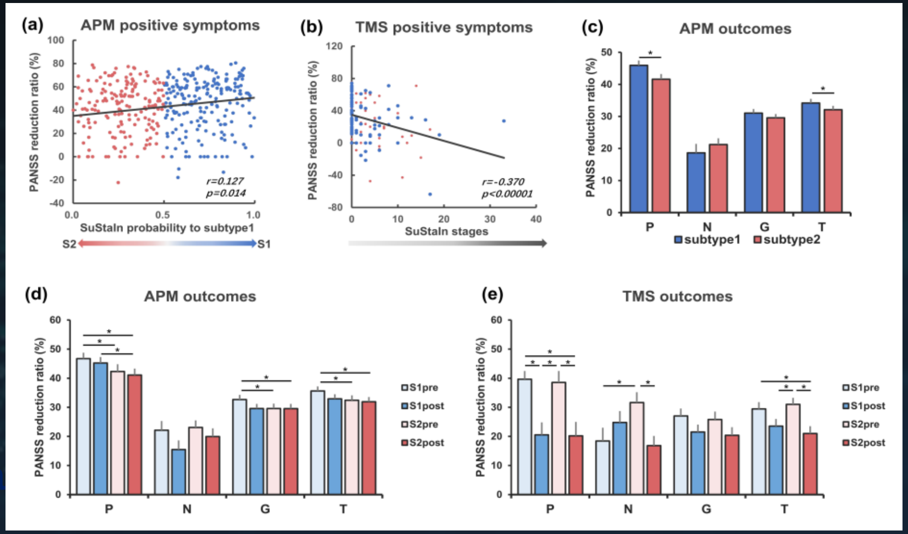
Two pathophysiological progression trajectories in schizophrenia



Trajectories are reproducibility for samples from different locations of the world



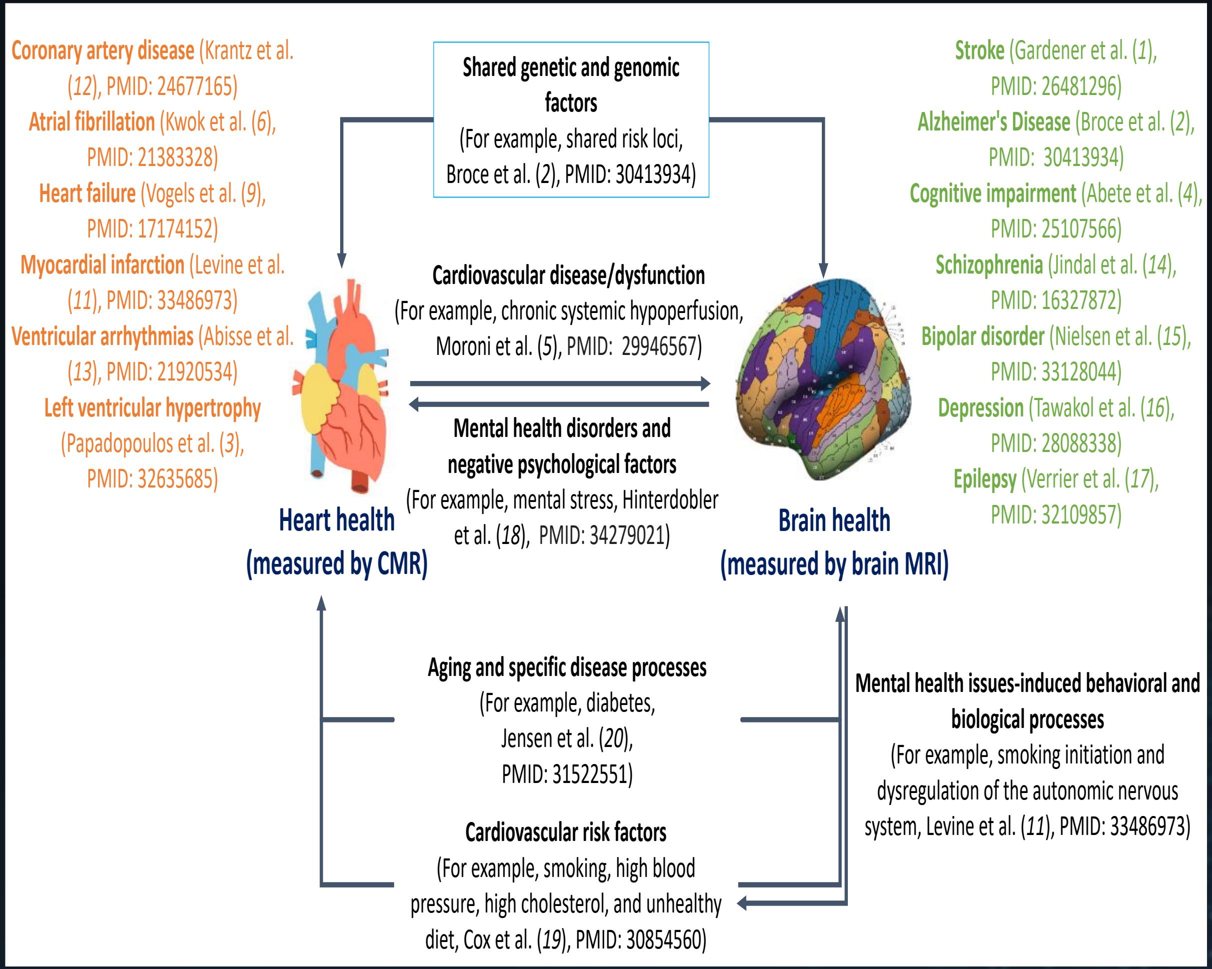
Treatment Outcomes in Subtypes of Schizophrenia

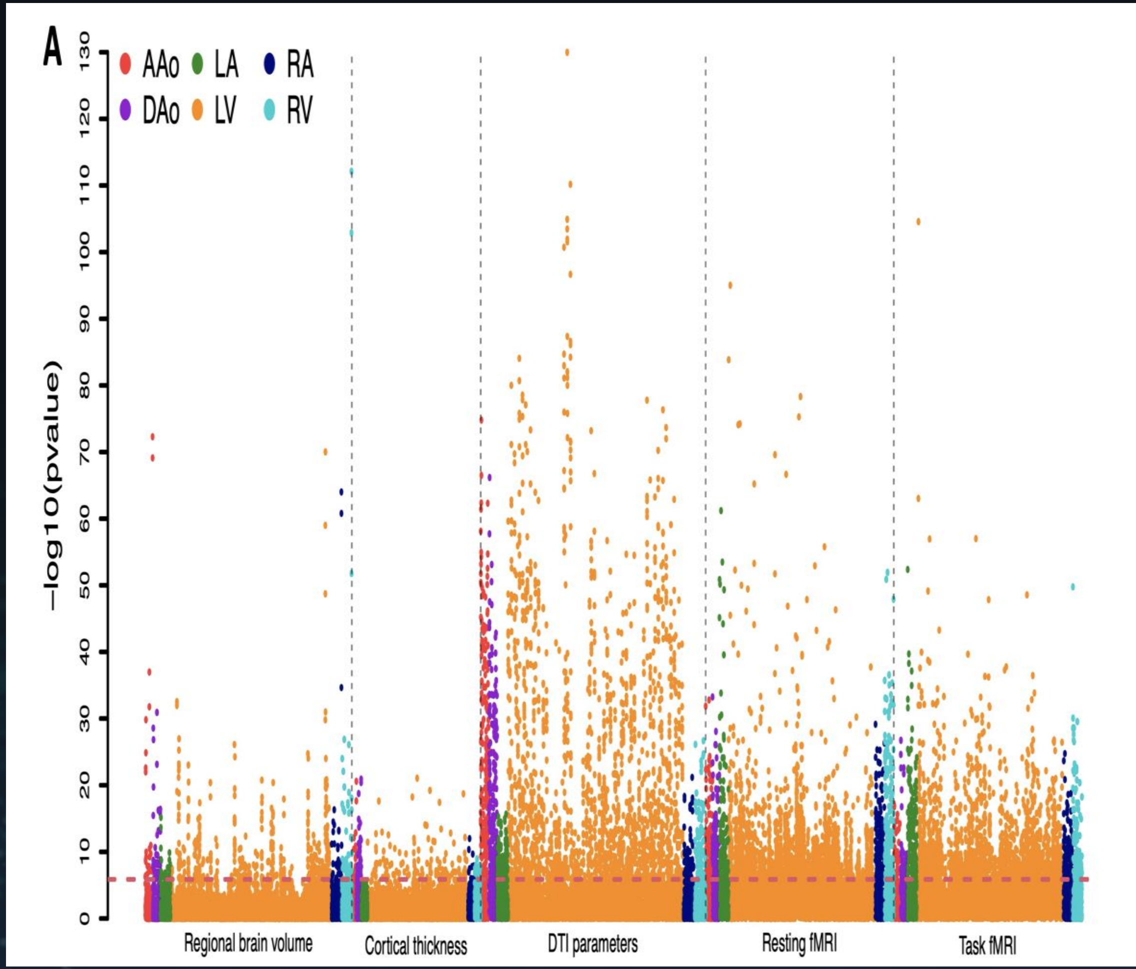


YC Jiang, et al, 2023, *Nature Mental Health*

YC Jiang, et al, *Nature Communications*, Under revision

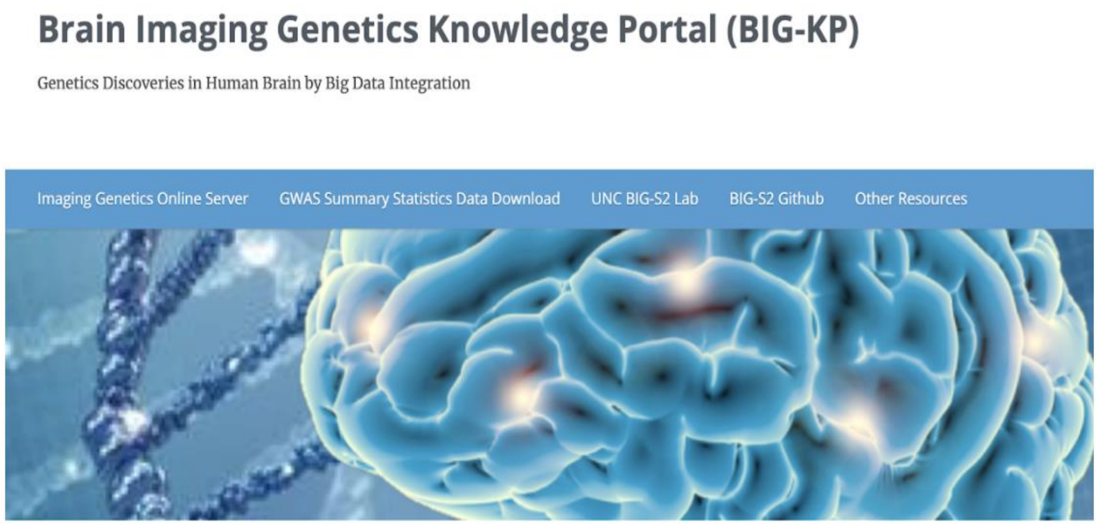
Heart-Brain Connections

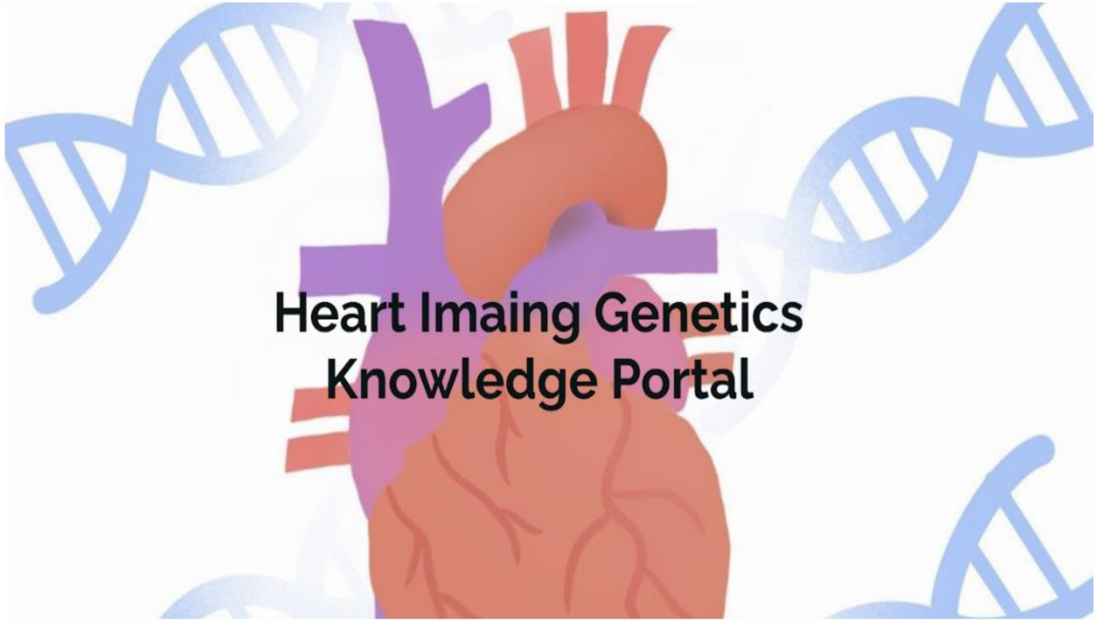




Zhao, B., Li, T., …., Stein, J. L., & Zhu, H. Heart-brain connections: Phenotypic and genetic insights from magnetic resonance images. *Science*, 380(6648), abn6598, 2023.

Brain- Heart Imaging Genetics Knowledge Portal



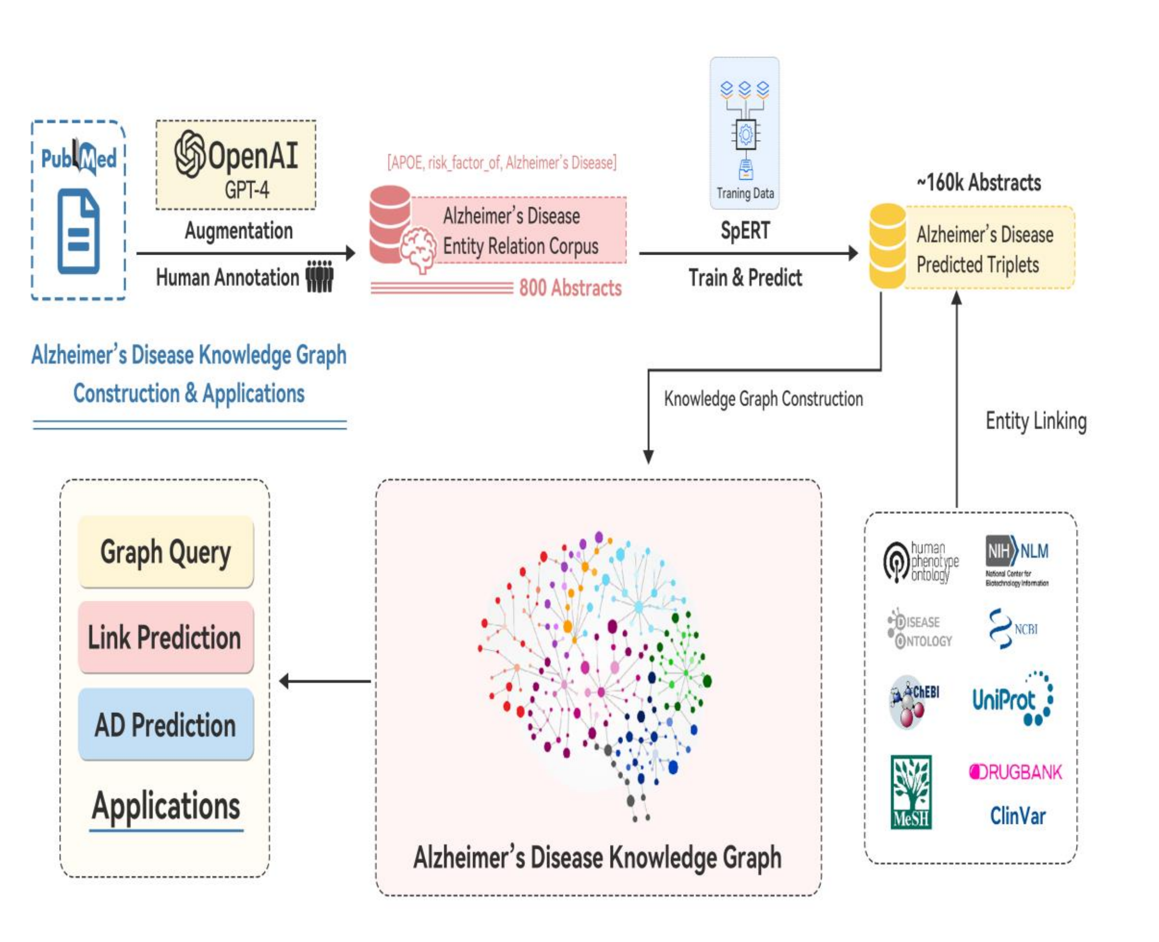


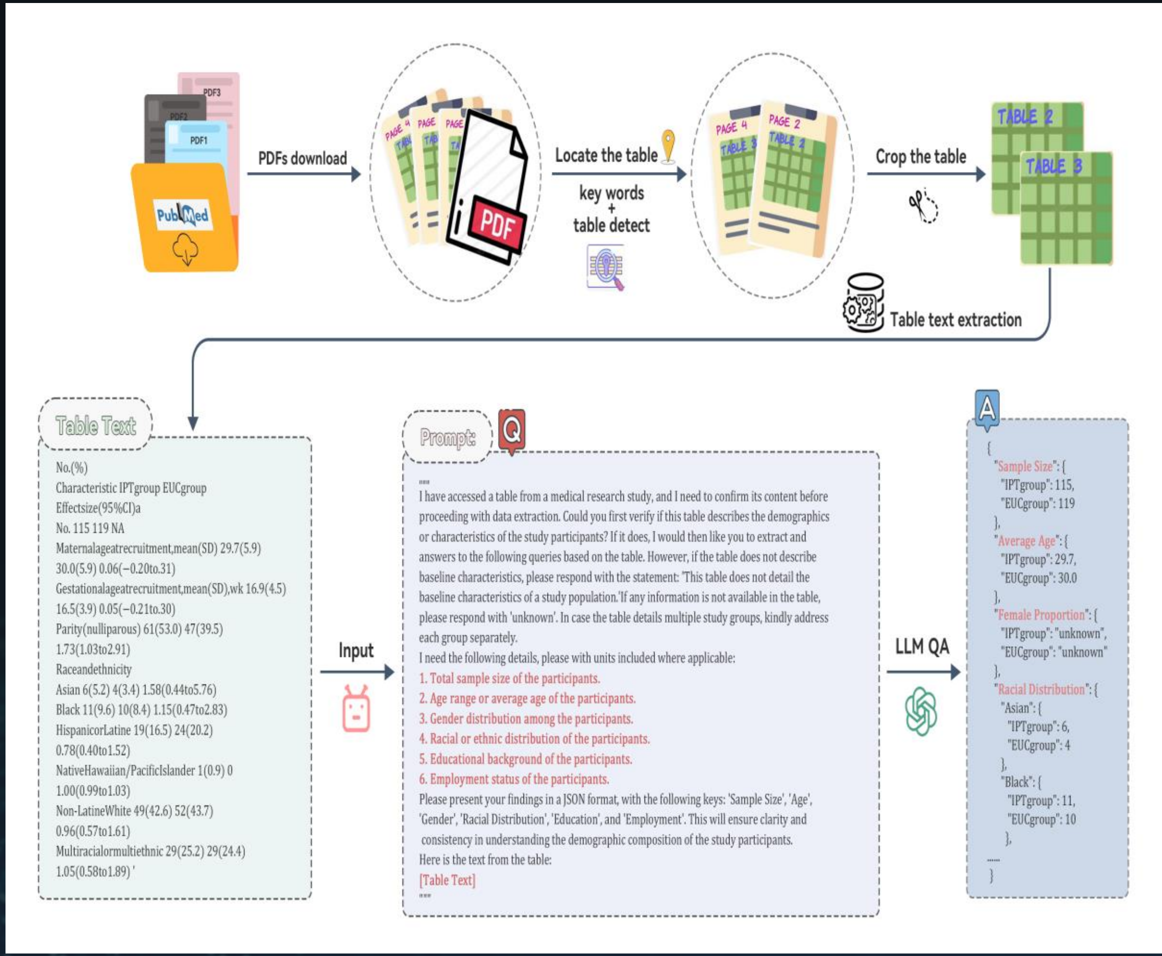
<u>**Brain Imaging Genetics Knowledge Portal (BIG-KP)**</u>

<u>**Heart Imaging Genetics Knowledge Portal (Heart-KP)**</u>

Aim to build the best knowledge database of neuroimaging genetics

Knowledge Graph Construction





Yang et al., Alzheimer's Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction.

Foundation Models for GMAI

Perspective

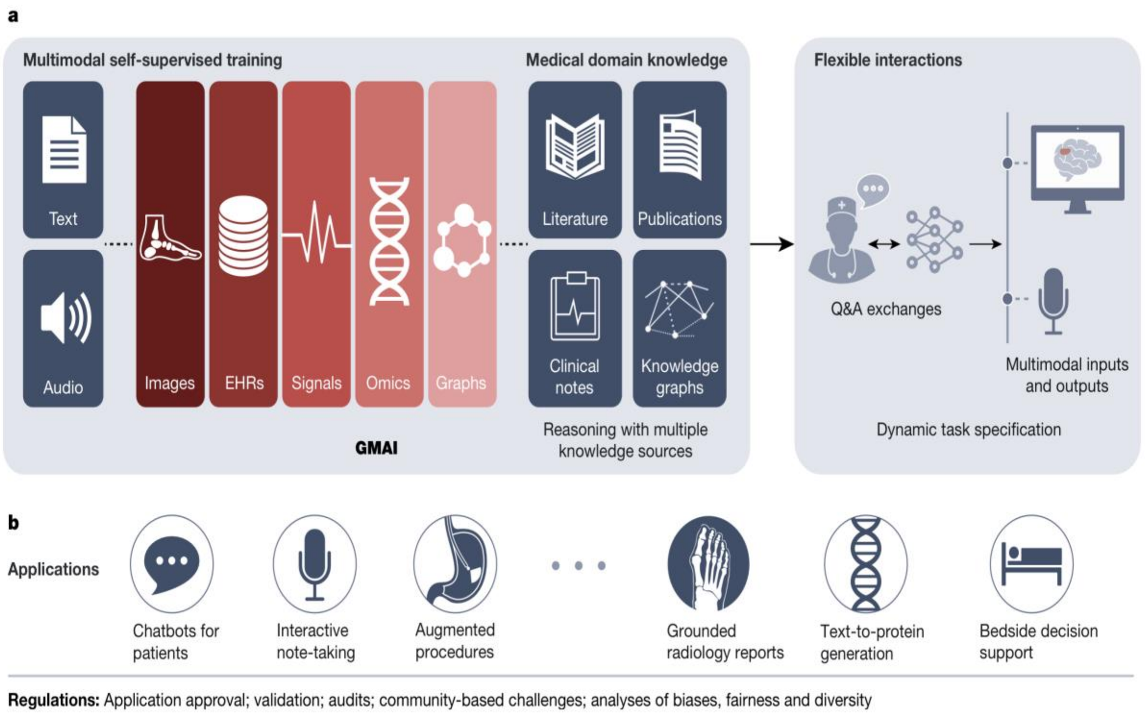


Fig. 1 | Overview of a GMAI model pipeline. a, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or speech data. Next, the GMAI model needs to access various sources of medical knowledge to carry out medical reasoning tasks, unlocking a wealth of capabilities that can be used in downstream applications. The resulting GMAI model then carries out tasks that the user can specify in real time. For this, the GMAI model can retrieve contextual information from sources such as knowledge graphs or databases, leveraging formal medical knowledge to reason about previously unseen tasks. **b,** The GMAI model builds the foundation for numerous applications across clinical disciplines, each requiring careful validation and regulatory assessment.

SS
Estimate the risk (in percentages) of developing a cardiovascular disease within 10 years fo the person below.
57 year old female, without diabetes, without hypertension, non smoker, total cholesterol 194.6 mg/dL, HDL 58.6 mg/dL, LDL 119.0 mg/dL, triglyceride 63.3 mg/dL, systolic blood pressure 137 mmHg, diastolic blood pressure 86 mmHg, BMI 20.72
Please answer exactly in the format below, without blank lines, and no further information or answer is required.
Risk percentage=(in percentages, round to one decimal place)

Risk percentage=8.2%

Fig. 2 | Example of a ChatGPT prompt and response for risk stratification. Tabular data extracted from the UK biobank and KoGES were organized and queried into a sentence format like the example above. The 10-year CVD risk percentage was extracted using regular expressions from the corresponding answers.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.22.23289842; this version posted May 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 2 | Performance comparison of Framingham, Bard, and ChatGPT Risk Score

	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
UK biobank						
GPT-4	**0·834**	0·393	**0·849**	**0·084**	0·975	**0·138**
GPT-3·5	0·674	**0·598**	0·677	0·061	**0·980**	0·111
Bard	0·702	0·447	0·711	0·052	0·973	0·093
Framingham	0·773	0·508	0·782	0·076	0·978	0·132
KoGES						
GPT-4	**0·902**	0·153	**0·926**	0·062	0·972	0·088
GPT-3·5	0·836	0·273	0·854	0·056	0·974	0·093
Bard	0·779	0·307	0·794	0·045	0·973	0·079
Framingham	**0·874**	**0·278**	0·893	**0·077**	**0·975**	**0·120**

PPV: positive predictive value, NPV: negative predictive value. Bold font indicates the highest value of the corresponding metric.

Moor, M., … ., Rajpurkar, P. (2023) Foundation models for generalist medical artificial intelligence. *Nature*.

Han,C.., … ., Yoon, D. (2023) **Large-language-model-based 10-year risk prediction of cardiovascular disease: insight from the UK biobank data.** *medRxiv*

Acknowledgement

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Brain Imaging Genetics Knowledge Portal (BIG-KP)

Genetics Discoveries in Human Brain by Big Data Integration

bigkp.org

Funding: U.S. NIH Grants MH086633, MH116527, and NIA-AG082938-01

Pictures: Copyrights belong to their own authors and/or holders.

Data: We thank Bingxin Zhao, Tengfei Li and other members of the UNC BIG-S2 lab (https://med.unc.edu/bigs2/) for processing the neuroimaging data.

UK Biobank resource application number: 22783.