

# Programming for Biomedical Informatics

## Lecture 1 - Welcome & Getting Started

<https://github.com/biomedical-informatics/pbi>

Ian Simpson  
[ian.simpson@ed.ac.uk](mailto:ian.simpson@ed.ac.uk)

# Course Organisation

- **Lectures**

- Weeks 1-4, 6-11
- Tuesdays 09:10 - 11:00
- Thursdays 11:10 - 12:00
- Recorded and released same day
- All slides and videos will be put on GitHub

- **References & Reading**

- Relevant books, websites, papers, & any other sources will be in lecture slides and coding notebooks.
- These will be collated on dedicated pages on GitHub for easier access

- **Materials**

- Everything will be available via the course GitHub
- <https://github.com/biomedical-informatics/pbi>

- **Assessment**

- Formative (**not-assessed**) assignments on GitHub Classroom used to give you coding and resource practice to gain experience and to use in the coursework later in the semester
- These are predominantly coding problems/examples (and supposed to be fun !)
- Coursework (20% of overall course mark)
- Exam (80% of overall course mark)
- Will be introducing example exam questions and discussing model solutions through the course
- Exam prep session & Q&A in week 11

- **Communication**

- Piazza - all course discussion and questions (including private questions) please use it!

# Course Topics

<b>Week</b>	<b>Week Commencing</b>	<b>Weekly Topics (background, application)</b>
1	15th September	(L1) Course Introduction & Setup, (L2) Working with Notebooks & Git
2	22nd September	(L3) Introduction to the Biomedical Dataverse, (L4) Finding & Fetching Data
3	29th September	(L5) Mapping and Harmonisation, (L6) Data Integration & Summary Analysis
4	6th October	(L7) Biomedical Evidence, (L8) Mining & Analysing Biomedical Literature
5		BREAK
6	20th October	(L9) Measuring Gene Expression, (L10) Differential Gene Expression (GXD)
7	27th October	(L11) Biological Networks, (L12) Network Construction Techniques
8	3rd November	(L13) Essential Network Methods, (L14) Network Analysis in Practice
9	10th November	(L15) Structuring Biomedical Data with Ontologies (L16) Functional Analysis
10	17th November	(L17) Working with Multiple Data Modalities, (L18) Modelling at the Patient Level
11	24th November	Course Review, Exam Prep, and Q&A session

foundation

application

# Course Setup

## GitHub

biomedical-informatics / pbi

Type  to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

pbi Public

Edit Pins Watch 0 Fork 0 Star 0

initial-release 1 Branch 0 Tags Go to file Add file Code About

tisimpson updated week 0 readme 58151f5 · 2 weeks ago 287 Commits

python\_basics updated note last year

resources sorted banner last year

week0 updated week 0 readme 2 weeks ago

.gitignore updated last year

LICENSE updated last year

README.md update readme 2 weeks ago

README MIT license

**Programming for Biomedical Informatics (INFR11260)**

School of Informatics, The University of Edinburgh (2025)

Course Lecturer: Prof. Ian Simpson

Contact: [ian.simpson@ed.ac.uk](mailto:ian.simpson@ed.ac.uk)



- [Course GitHub Repository Cloning Link](#)
- [GitHub Classroom Assignments](#)
- [Course Discussion Boards](#) (you can login directly at Piazza or click through from LEARN)

About This is the website for the Programming for Biomedical Informatics Course (INFR11260) 2025

Readme MIT license Activity Custom properties 0 stars 0 watching 0 forks Report repository

Releases No releases published Create a new release

Packages No packages published Publish your first package

Languages Jupyter Notebook 100.0%

<https://github.com/biomedical-informatics/pbi>

# Course Setup

## GitHub Classroom

The screenshot shows the GitHub Classroom interface for the course "Programming for Biomedical Informatics 2025". The top navigation bar includes "Classrooms / Programming for Biomedical Informatics 2025" and the GitHub logo. Below the header, the course name "Programming for Biomedical Informatics 2025" and the repository name "biomedical-informatics" are displayed. The main navigation bar has tabs for "Assignments" (1), "Students" (33), "TAs and Admins" (1), and "Settings". The "Assignments" tab is active, showing a single assignment titled "Week 0 - Introduction to GitHub". The assignment status is "Active" and is categorized as "Individual assignment". A green button labeled "+ New assignment" is located in the top right corner of the assignments section. A blue notification bar at the bottom left encourages users to complete a survey.

<https://github.com/biomedical-informatics/pbi>

# Course Setup

## Noteable

Week 1

Week 2

Week 3

Week 4

Week 5

Week 6

Week 7

Week 8

Week 9

Week 10

Week 11

Week 12

Week 13

Week 14

Week 15

Week 16

Week 17

Week 18

Week 19

Week 20

Week 21

Week 22

Week 23

Week 24

Week 25

Week 26

Week 27

Week 28

Week 29

Week 30

Week 31

Week 32

Week 33

Week 34

Week 35

Week 36

Week 37

Week 38

Week 39

Week 40

Week 41

Week 42

Week 43

Week 44

Week 45

Week 46

Week 47

Week 48

Week 49

Week 50

Week 51

Week 52

Week 53

Week 54

Week 55

Week 56

Week 57

Week 58

Week 59

Week 60

Week 61

Week 62

Week 63

Week 64

Week 65

Week 66

Week 67

Week 68

Week 69

Week 70

Week 71

Week 72

Week 73

Week 74

Week 75

Week 76

Week 77

Week 78

Week 79

Week 80

Week 81

Week 82

Week 83

Week 84

Week 85

Week 86

Week 87

Week 88

Week 89

Week 90

Week 91

Week 92

Week 93

Week 94

[About Noteable](#) [Resources](#) [Pricing](#) [Status](#) [Contact Us ▾](#)

 THE UNIVERSITY OF EDINBURGH 

Create and share engaging coding lessons with Noteable, a cloud-based computational notebook service which works in your browser from any device.

Developed by [EDINA](#) at the University of Edinburgh, Noteable hosts your computational notebooks in one simple online hub and can integrate with your [VLE](#).

Streamline your preparation, teaching and marking and build a better learning experience for your students with Noteable.

[Get a Free Trial today](#)



**What our clients say**

Noteable supports teaching in a variety of subject areas and institutions.

“ For larger courses, or those containing students with less computing experience, this is undoubtedly a huge benefit of using the Noteable service.”

“ Interspersing live code blocks with narrative content makes for a wonderfully efficient and interactive classroom and online experience.”

“ Very easy to use and allowed me to deliver working examples live in lectures and for students to work with at home. No complicated set-up, works straight from a browser.”

<https://noteable.edina.ac.uk/login>

# Course Setup

## Visual Studio Code

The screenshot shows a Visual Studio Code interface with two main panes. The left pane displays a Jupyter Notebook cell containing Python code for comparing classification titles and printing proportions of rows where they differ. The right pane shows a Data Explorer view with a table of genomic and disease data.

```
#compare the contents of the column 'classification_title' and 'submitted_as_classification_name'
gencc_data['classification_title'].equals(gencc_data['submitted_as_classification_name'])

# print the proportion of rows where the two columns are not equal
print("Proportion of rows where the two columns are not equal: ", sum(gencc_data['classification_title'] != gencc_data['submitted_as_classification_name'])/len(gencc_data))

# print the rows where they are not equal
gencc_data[gencc_data['classification_title'] != gencc_data['submitted_as_classification_name']]
```

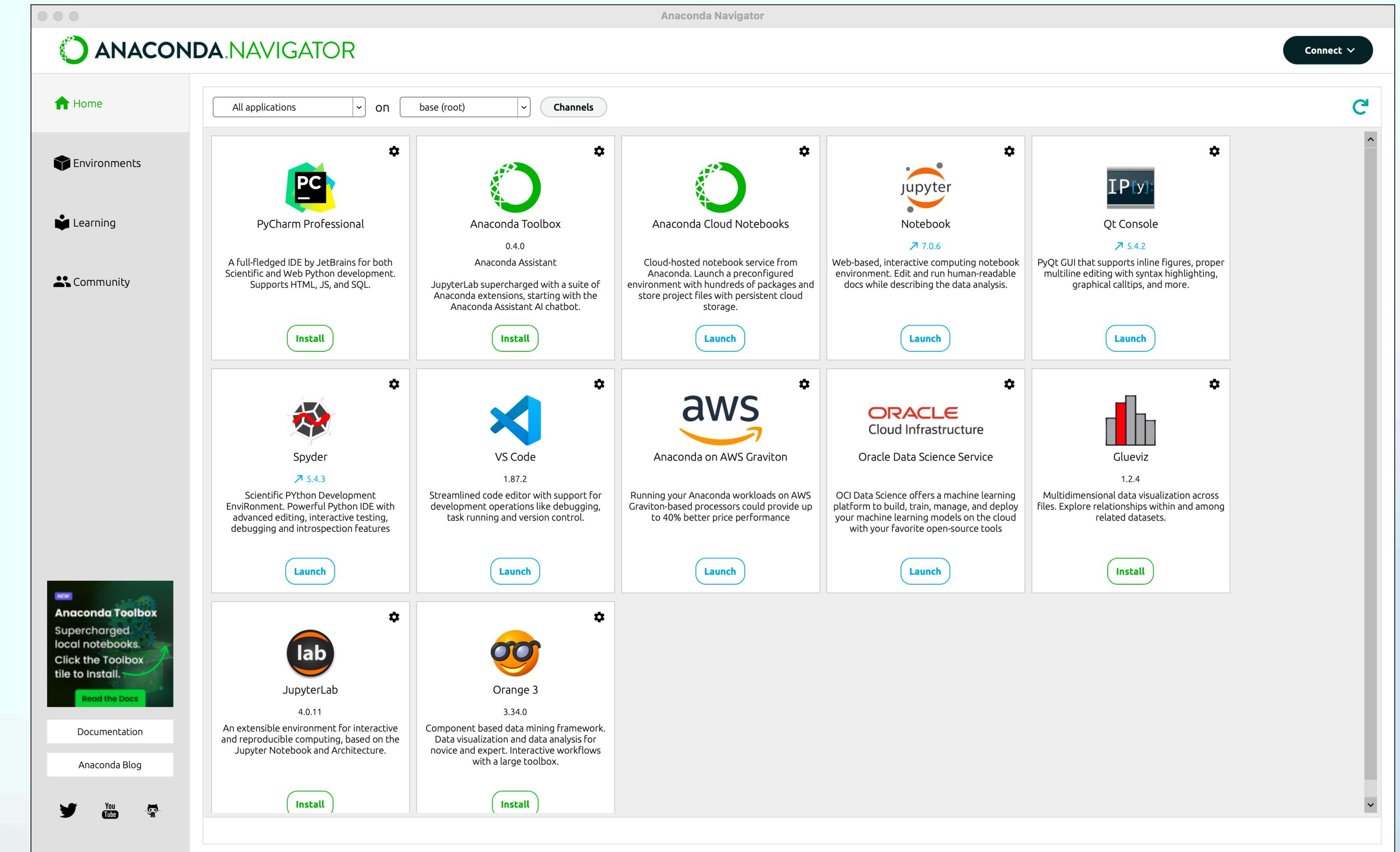
... Proportion of rows where the two columns are not equal: 0.18617596195417208

	uid	gene_curie	gene_symbol	disease_curie	disease_title	disease_original_curie	disease_original_title	classification_curie	classification_title	moi_curie
56	GENCC_000105-HGNC_9801-MONDO_0030913-HP_000000...	HGNC:9801	RAC1	MONDO:0030913	intellectual disability, autosomal dominant 48	MONDO:0030913	intellectual disability, autosomal dominant 48	GENCC:100002	Strong	HP:0000006
57	GENCC_000105-HGNC_7882-MONDO_0007318-HP_000000...	HGNC:7882	NOTCH2	MONDO:0007318	Alagille syndrome	MONDO:0007318	Alagille syndrome	GENCC:100003	Moderate	HP:0000006
87	GENCC_000105-HGNC_15766-MONDO_0014379-HP_000000...	HGNC:15766	ADNP	MONDO:0014379	ADNP-related multiple congenital anomalies - i...	MONDO:0014379	ADNP-related multiple congenital anomalies - i...	GENCC:100001	Definitive	HP:0000006
92	GENCC_000105-HGNC_1321-MONDO_0032634-HP_000000...	HGNC:1321	TIMMDC1	MONDO:0032634	mitochondrial complex 1 deficiency, nuclear ty...	MONDO:0032634	mitochondrial complex 1 deficiency, nuclear ty...	GENCC:100003	Moderate	HP:0000007
196	GENCC_000105-HGNC_2174-MONDO_0005258-HP_000000...	HGNC:2174	CNTN4	MONDO:0005258	autism spectrum disorder	MONDO:0005258	autism spectrum disorder	GENCC:100005	Disputed Evidence	HP:0000006
18484	GENCC_000112-HGNC_1750-OMIM_211380-HP_0000007...	HGNC:1750	CDH11	MONDO:0008885	Elsahy-Waters syndrome	OMIM:211380	Elsahy-Waters syndrome	GENCC:100001	Definitive	HP:0000007
18485	GENCC_000112-HGNC_9416-OMIM_619636-HP_0000007...	HGNC:9416	PRKG2	MONDO:0030553	acromesomelic dysplasia 4	OMIM:619636	Acromesomelic dysplasia 4	GENCC:100002	Strong	HP:0000007

<https://code.visualstudio.com/>

# Course Setup

## Anaconda/miniconda



<https://www.anaconda.com/>

<https://www.anaconda.com/docs/getting-started/miniconda/main>

# Applications in Biomedical & Health Informatics

## Opportunities

### Clinical & Health

- Administration Support
- Decision Support
- Patient Engagement
- Synthetic Data Generation
- Clinical Trial Design & Monitoring
- Population Level Modelling
- Professional Education

### Biomedical Science

- Drug Discovery and Design
- Protein Structure Prediction
- Biomedical Image Synthesis
- Patient Data Generation
- Drug Response Prediction
- Biological Sequence Generation
- Medical Text Generation
- Biomedical Signal Generation
- Disease Progression Modeling

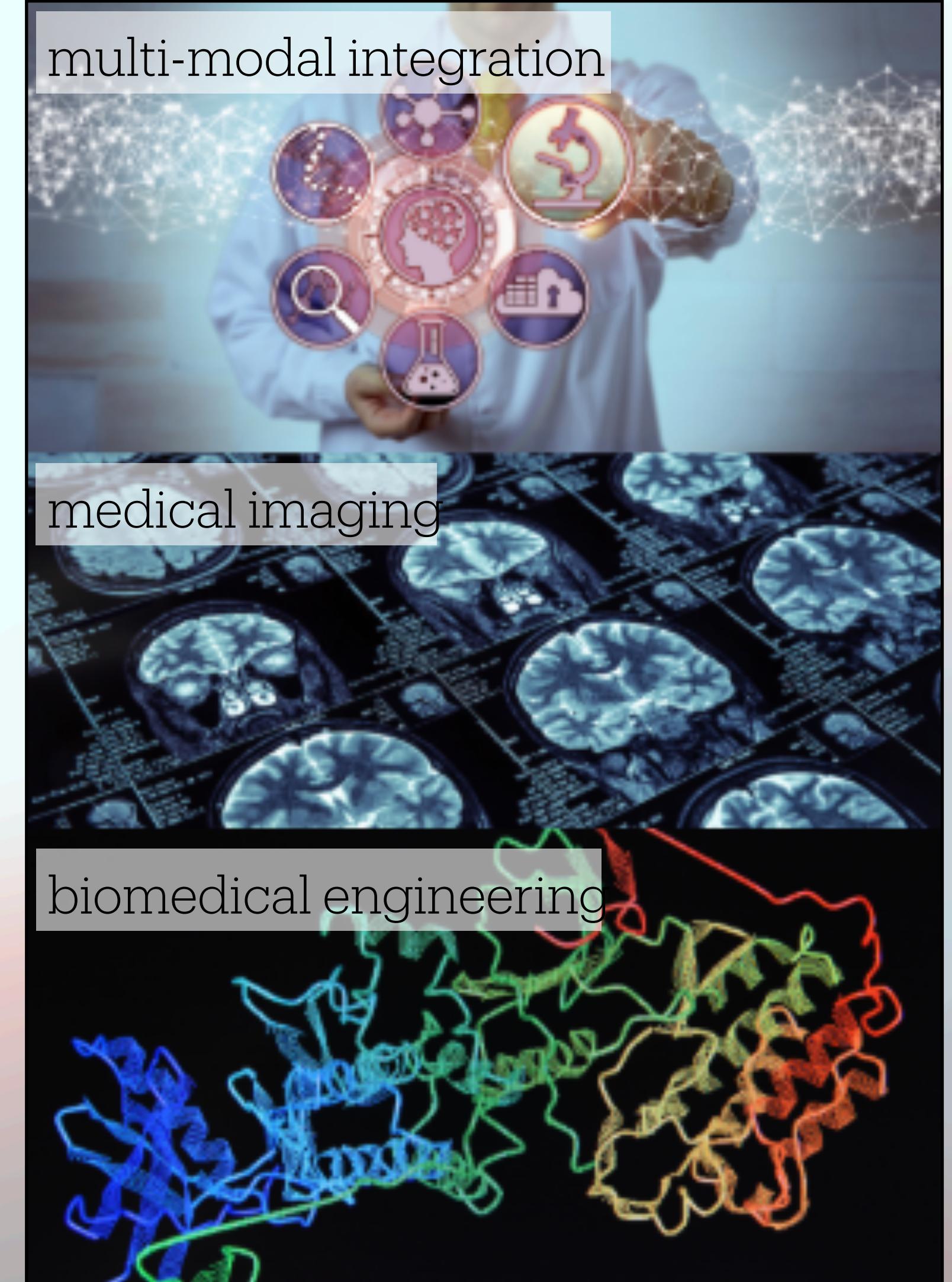
## Challenges

### Technical Challenges

- Unlabelled & Unstructured Data
- Missing Values
- Model & Data Bias
- Poor Longitudinal Coverage
- Scaling Problems
- Lack of Realistic Evaluation Benchmarks
- Explainability
- Data Availability & Inter-Operability

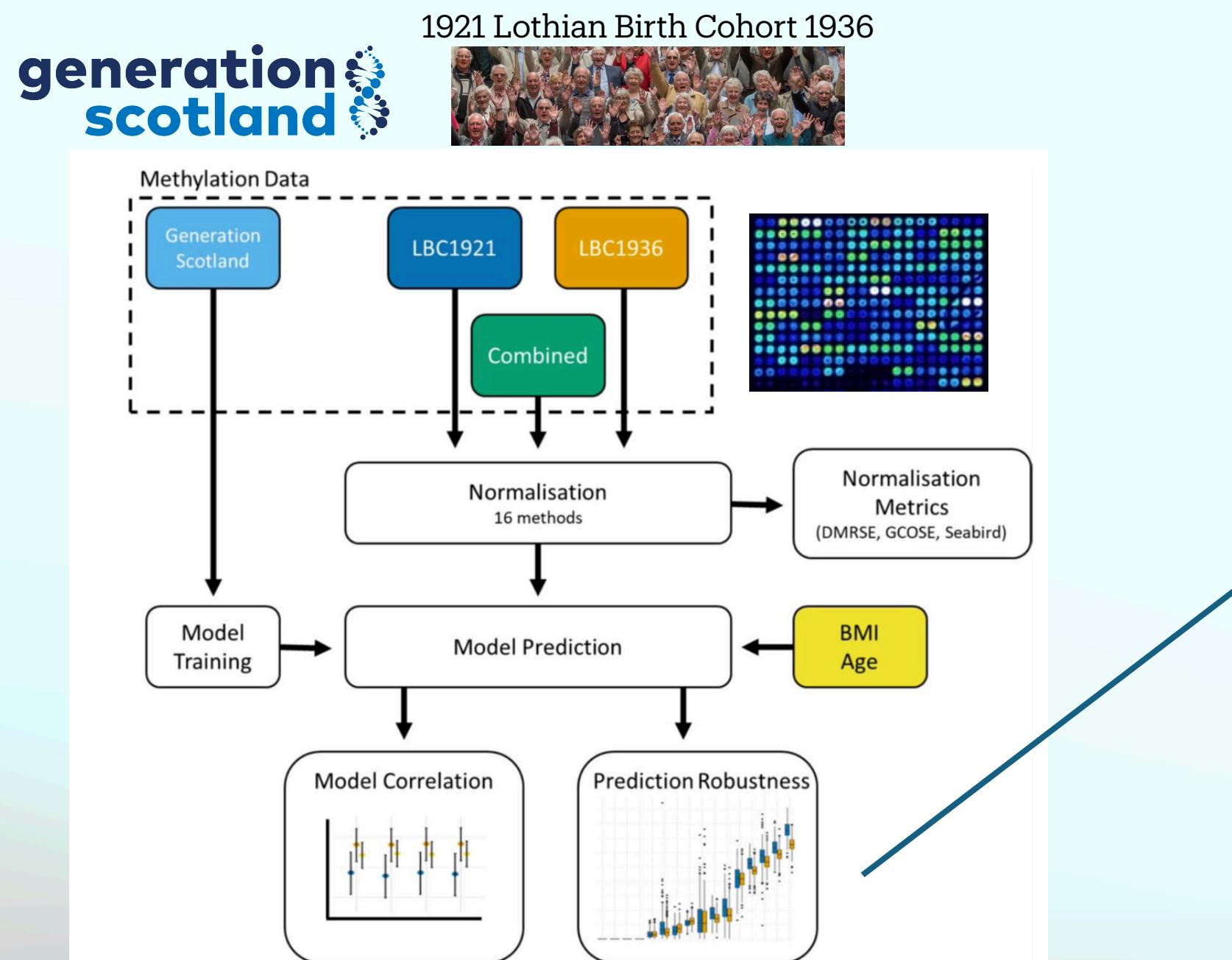
### Societal & Health Systems

- Clinical safety, Efficacy, & Reliability
- Evaluation, Regulation, & Certification
- Privacy
- Copyright & Ownership
- Implementation & Adoption

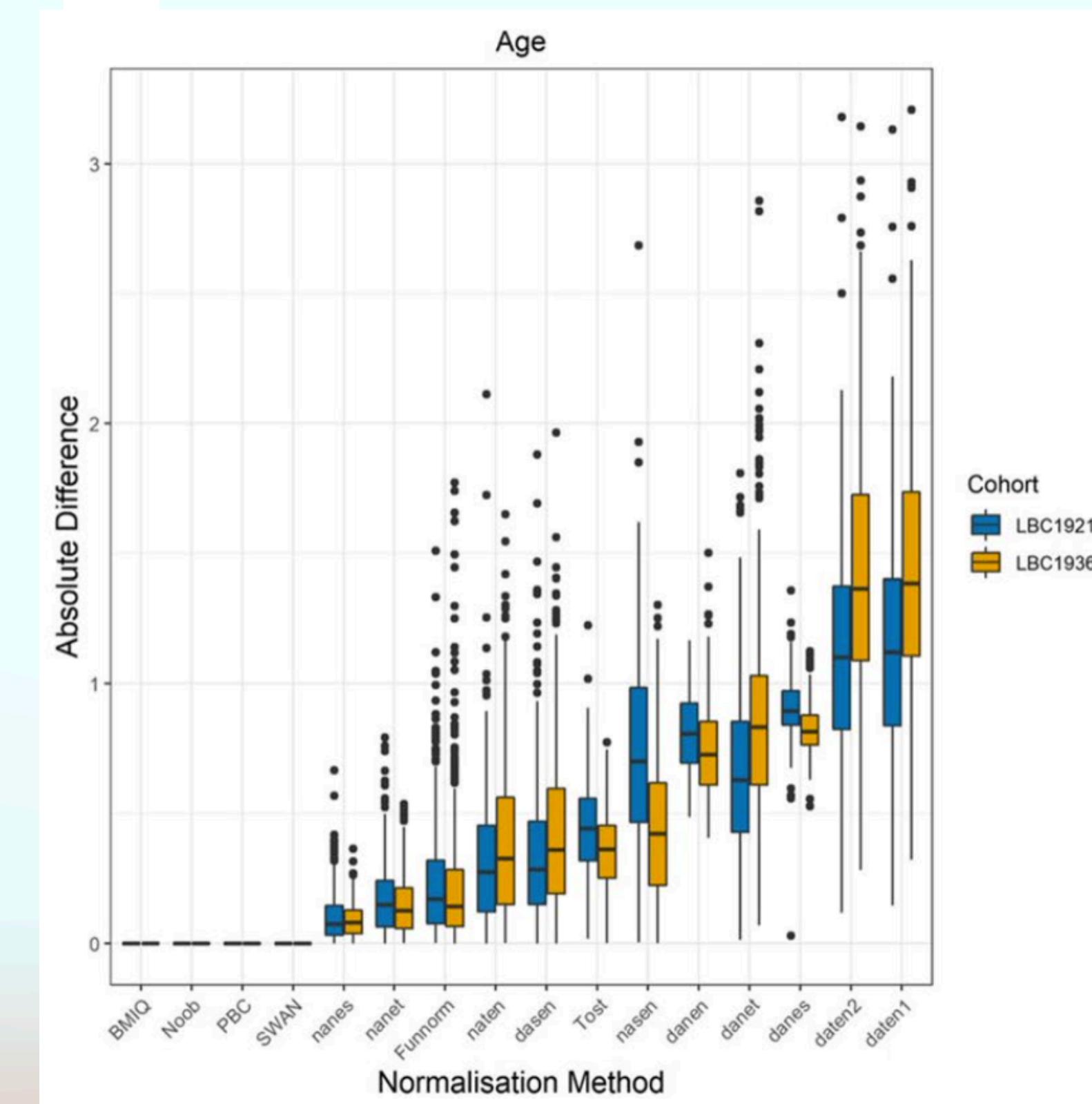


# Integration of Datasets for Individual Prediction of DNA Methylation-Based Biomarkers

## DNA Methylation Data



## Robustness of Predictions



- Projecting data from new datasets onto existing reference data is challenging
- Identification of technical vs biological variation
- Normalisation can help to resolve this but approach is critical.
- Epigenetic measures, such as EpiScore can be used to aid retention of biological variation during normalisation

Charlotte Merzbacher, Barry Ryan, Thibaut Goldsborough, Robert F Hillary, Archie Campbell, Lee Murphy, Andrew M McIntosh, David Liewald, Sarah E Harris, Allan F McRae, Simon R Cox, Timothy I Cannings, Catalina A Vallejos, Daniel L McCartney, Riccardo E Marioni (2023) Integration of DNA methylation datasets for individual prediction of DNA methylation-based biomarkers. *Genome Biology*. <https://doi.org/10.1186/s13059-023-03114-5>



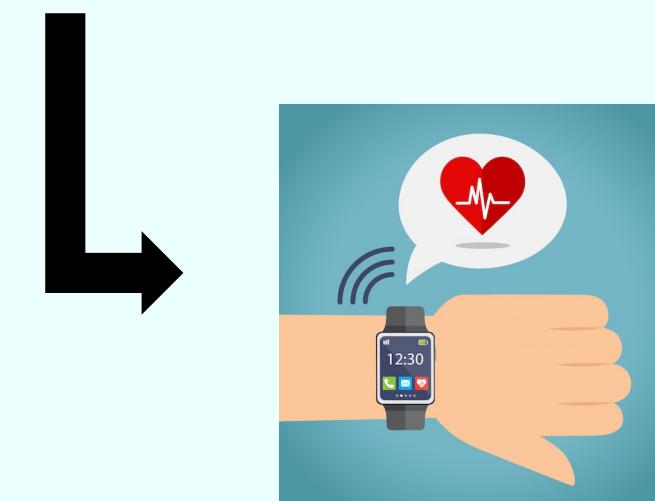
Charlotte  
Merzbacher

Barry  
Ryan

Thibaut  
Goldsborough

# Automated Mood Disorder Symptoms Monitoring From Multivariate Time-Series Sensory Data: Getting the Full Picture Beyond a Single Number

Psychiatric assessments are scheduled infrequently and rely on self-reported experiences

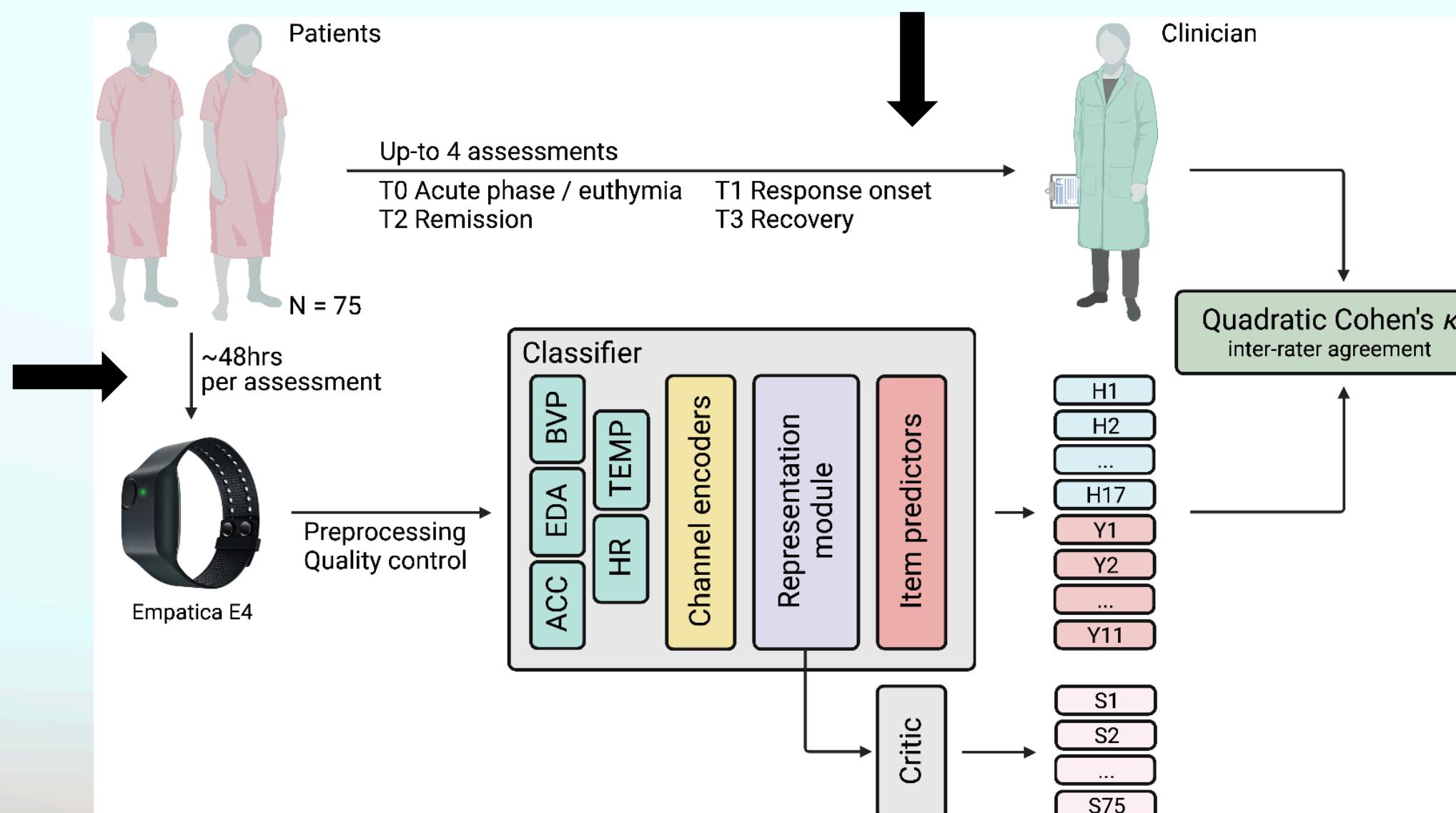
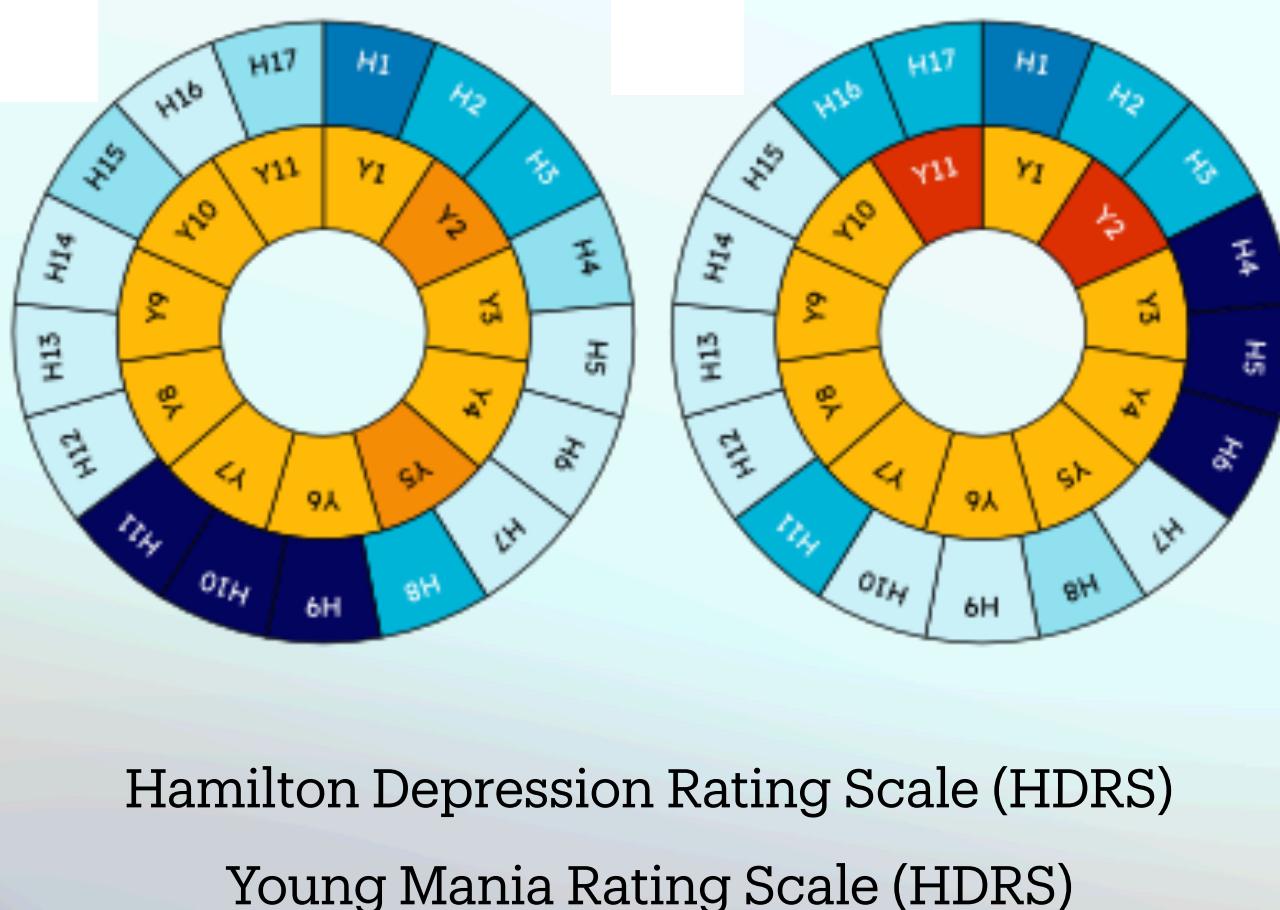


Wearables can enable near-continuous remote monitoring leveraging passively collected physiological data

Reducing monitoring to acute episode detection misses on actionable clinical information

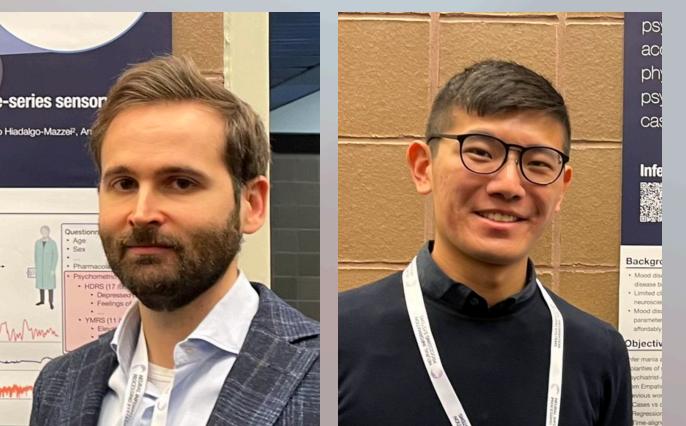
Timely Interventions  
Better Outcomes

Novel Task: Inferring all items from YMRS and HDRS (two popular standardized psychometric scales)



Model	ACC		F <sub>1</sub> score		
	segment	subject	segment	subject	
SL	ENET	66.38	71.88	66.54	70
	KNN	70.37	82.81	71.34	80.6
	SVM	71.25	81.25	71.63	78.81
	XGBoost	72.02	82.81	71.72	82.03
	E4mer	75.35	81.25	74.39	81.33
SSL	MP (LR)	77.53	87.5	77.87	88.3
	MP (FT)	<b>81.23</b>	<b>90.63</b>	<b>81.45</b>	<b>91.47</b>
	TP (LR)	71.16	81.25	72.06	82.37
	TP (FT)	75.69	84.38	75.1	83

Filippo Corponi, Bryan M. Li, Gerard Anmella, Ariadna Mas, Isabella Pacchiarotti, Marc Valentí, Iria Grande, Antoni Benabarre, Marina Garriga, Eduard Vieta, Stephen M. Lawrie, Heather C. Whalley, Diego Hidalgo-Mazzei, Antonio Vergari (2024) Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. Nature: Translational Psychiatry. <https://doi.org/10.1038/s41398-024-02876-1>

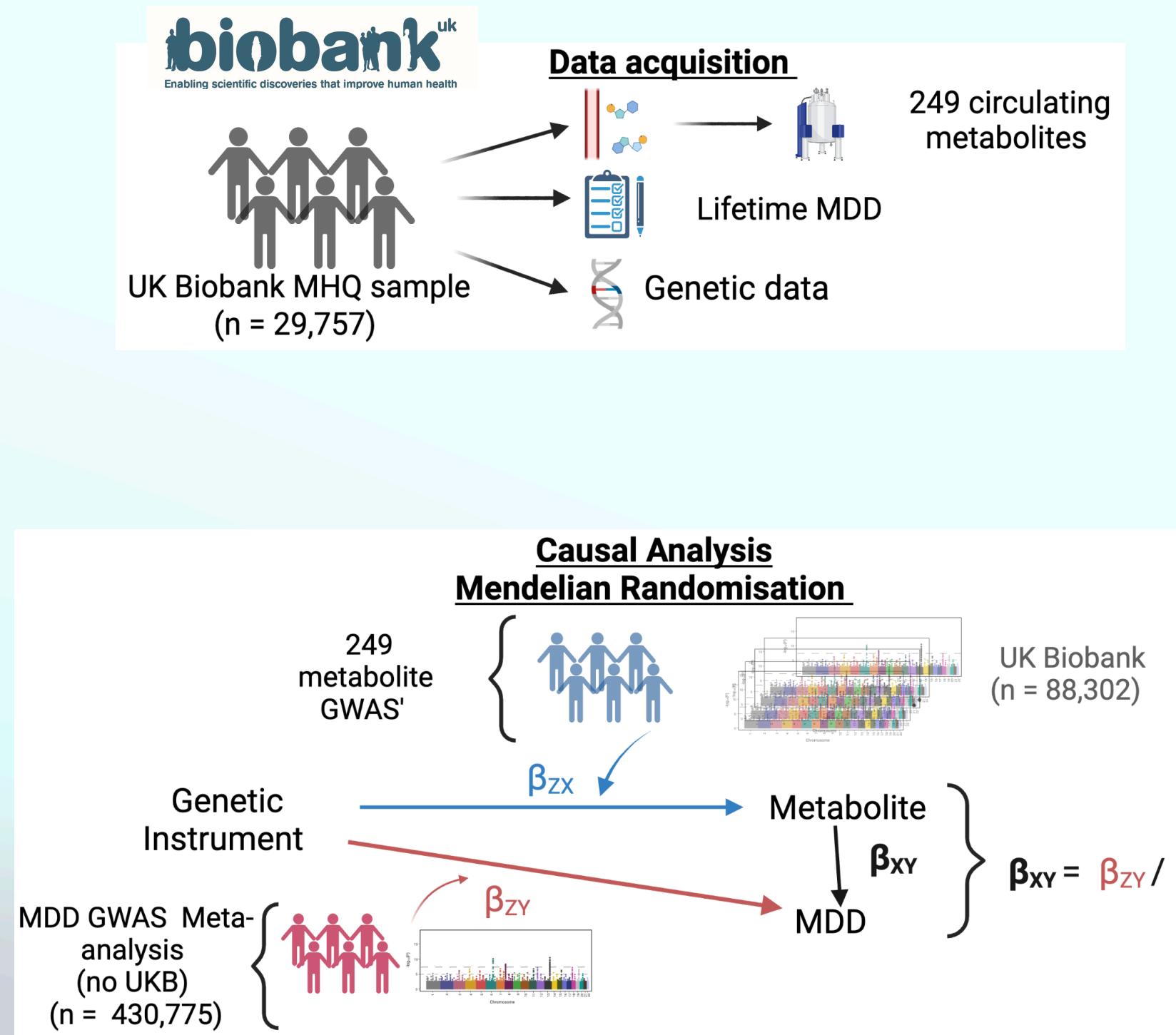


Filippo  
Corponi

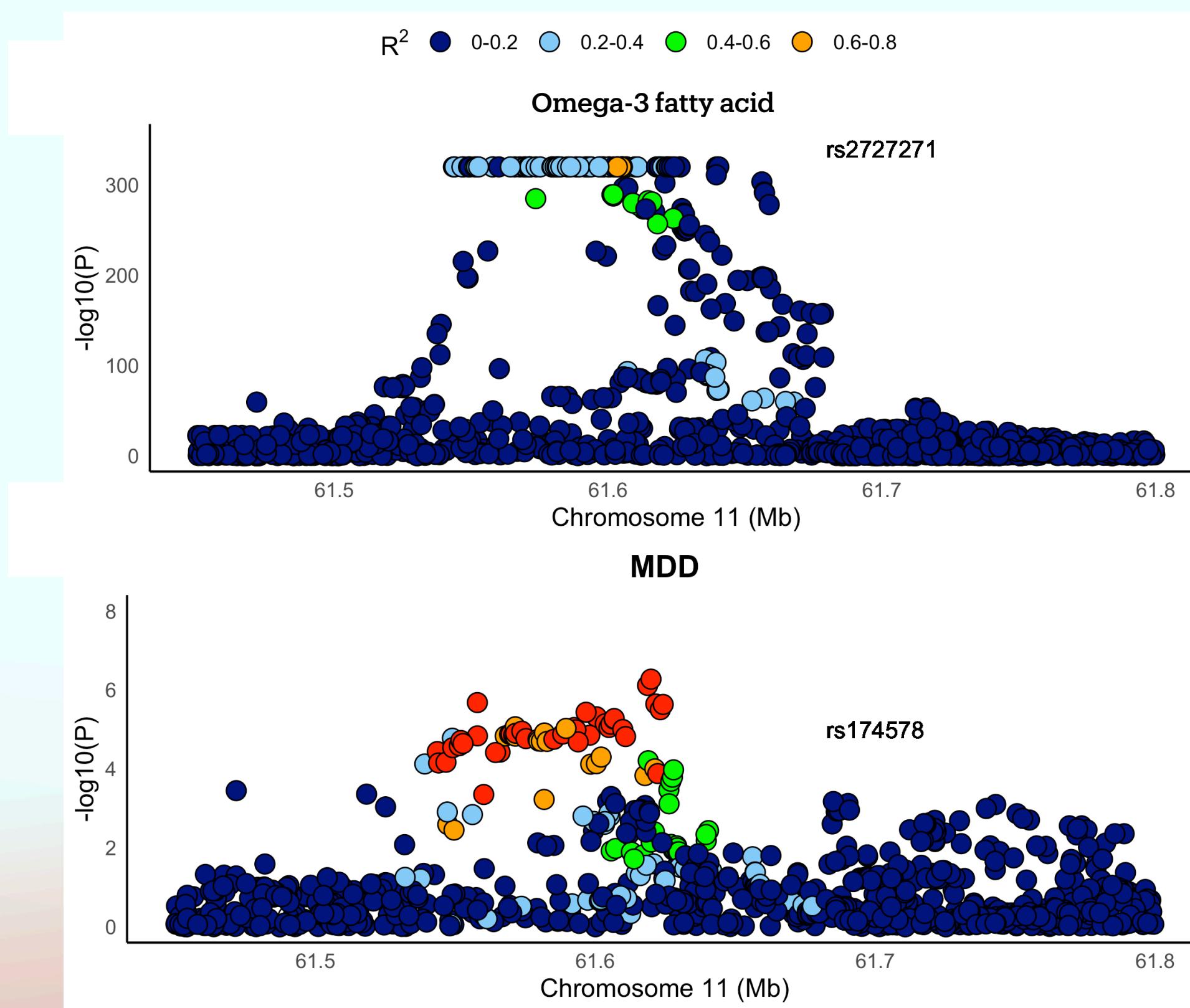
Bryan  
Li

# Metabolites and Major Depressive Disorder in the UK Biobank

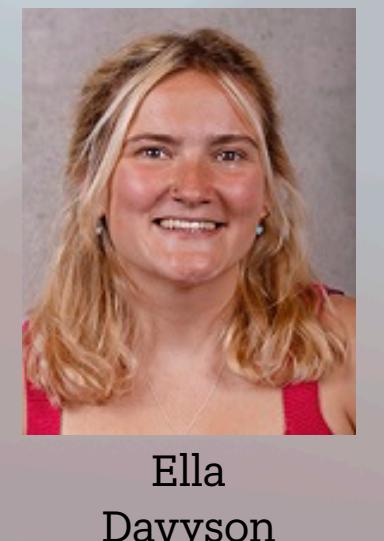
Most metabolites were significantly associated with depression.



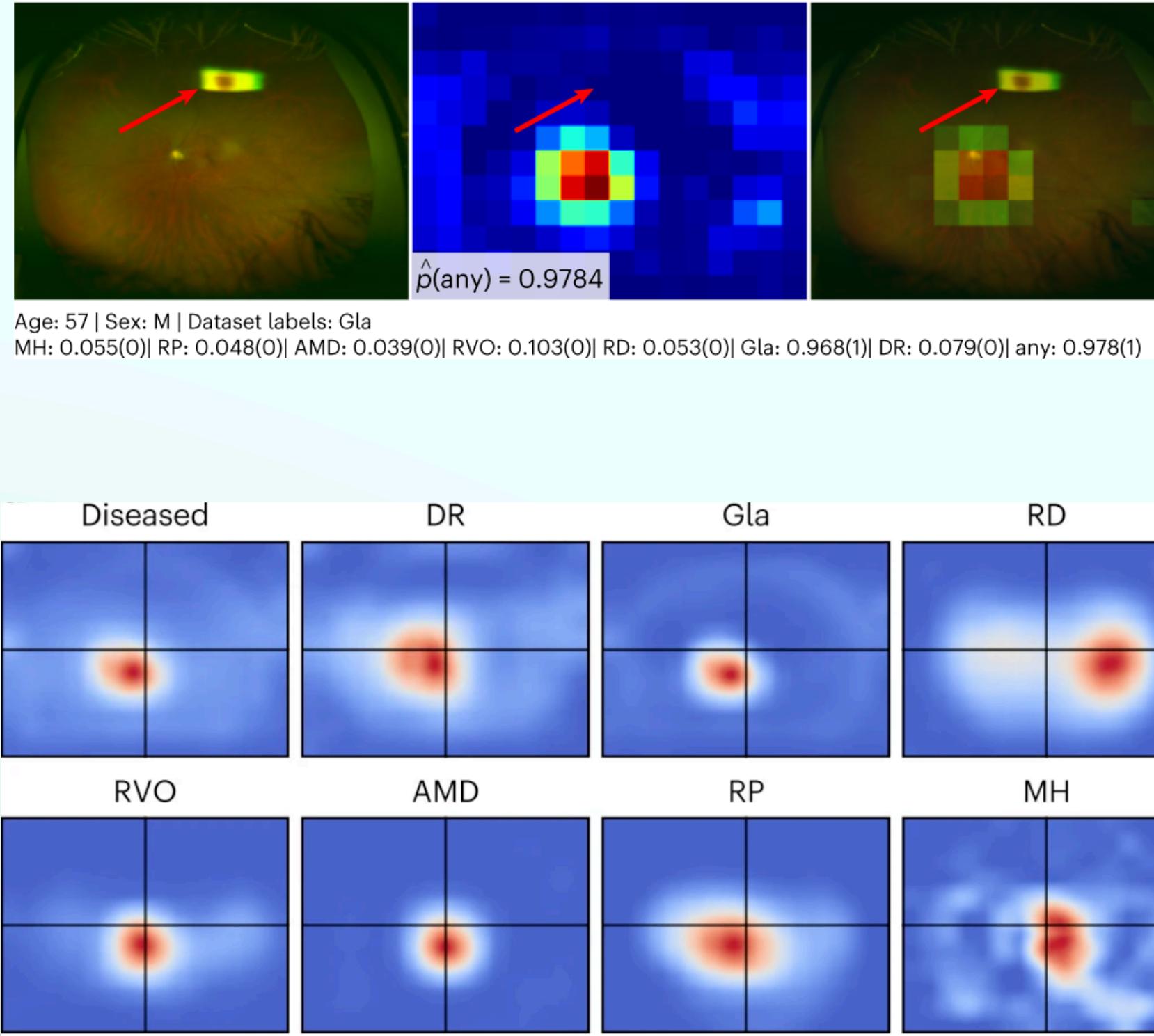
Evidence of causality between lowered omega-3 and higher omega-6:omega-3 fatty acid ratio with depression.



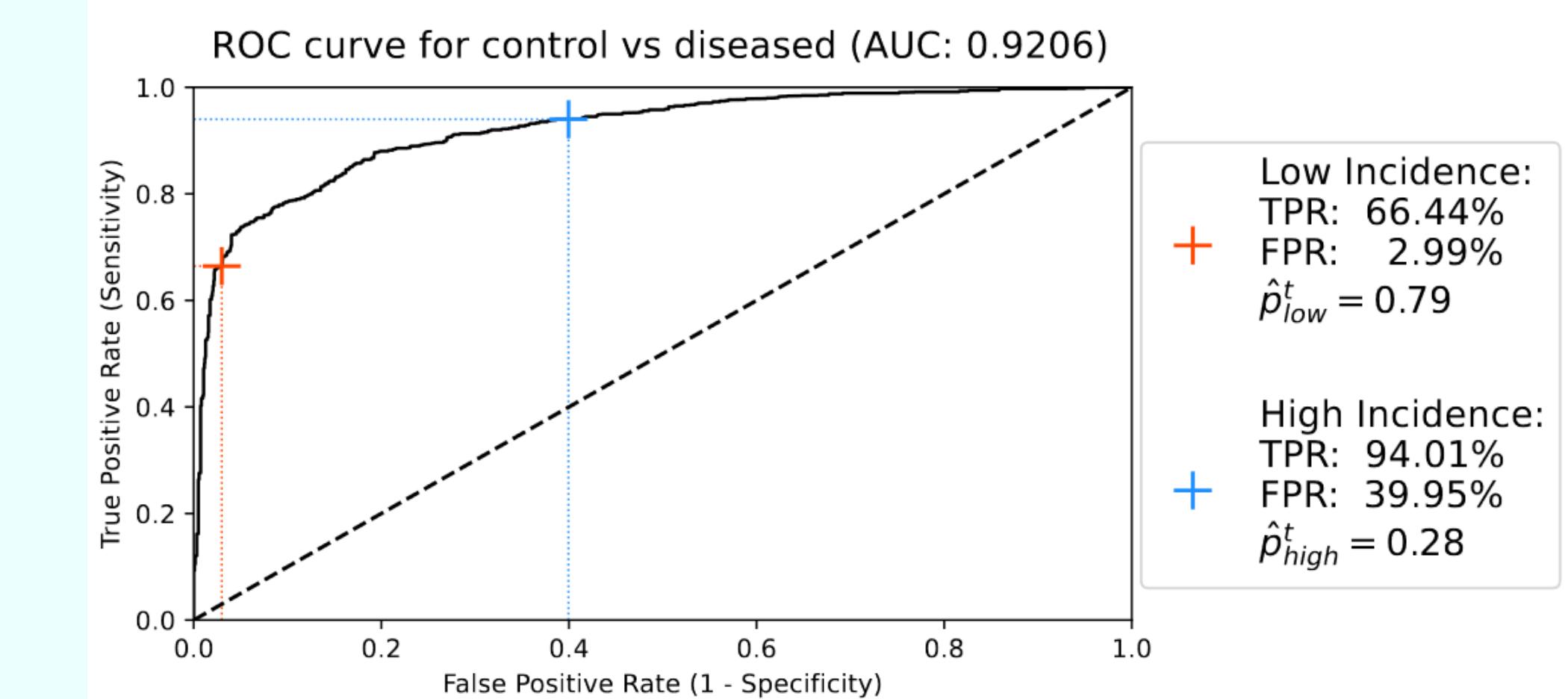
**Eleanor Davyson, Xueyi Shen, Danni A. Gadd, Elena Bernabeu, Robert F. Hillary, Daniel L. McCartney, Mark Adams, Riccardo Marioni, and Andrew M. McIntosh**  
(2023) Metabolomic Investigation of Major Depressive Disorder Identifies a Potentially Causal Association With Polyunsaturated Fatty Acids.  
Society of Biological Psychiatry. <https://doi.org/10.1016/j.biopsych.2023.01.027>



# Disease Detection in Ultra-Wide-Field Retinal Images



AMD: Age-related Macular Degeneration, RVO: Retinal Vein Occlusion, Gla: Glaucoma, MH : Macular Hole, DR : Diabetic Retinopathy, RD : Retinal Detachment, RP : Retinitis Pigmentosa



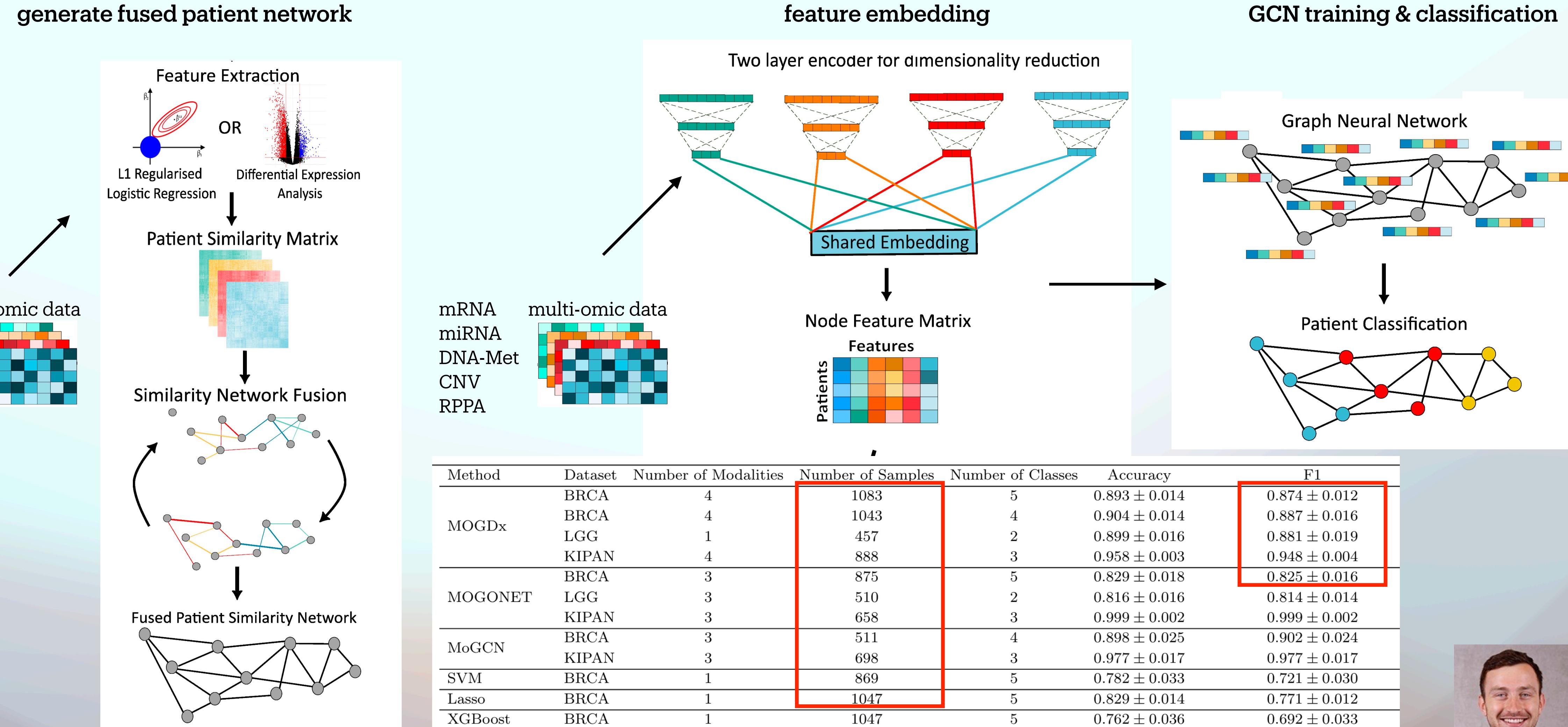
	Diseased	DR	Gla	RD	RVO	AMD	RP	MII
Logistic Regression with Age + Sex	0.5964	0.5988	0.5155	0.7676	0.4892	0.8021	0.6776	0.5625
Ensemble of Experts (binary DL models + balanced data)	0.8318 *	0.8432	0.9141	0.9217	0.8996	0.7113	<b>0.9490</b>	0.6454
Ours (Single multi-label DL model + realistic data)	<b>0.9206</b>	<b>0.9125</b>	<b>0.9422</b>	<b>0.9753</b>	<b>0.9468</b>	<b>0.9510</b>	0.9438	<b>0.7987</b>

Justin Engelmann, Alice D. McTrusty, Ian J. C. McCormick, Emma Pead, Amos Storkey & Miguel O. Bernabeu (2022) Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. Nature machine intelligence. <https://doi.org/10.1038/s42256-022-00566-5>



Justin  
Engelmann

# Cancer Classification From Multi-Modal Fused Patient Networks



Ryan B, Marioni RE, Simpson TI. Multi-Omic Graph Diagnosis (MOGDx) : A data integration tool to perform classification tasks for heterogeneous diseases. Bioinformatics 2024 <https://doi.org/10.1093/bioinformatics/btae523>



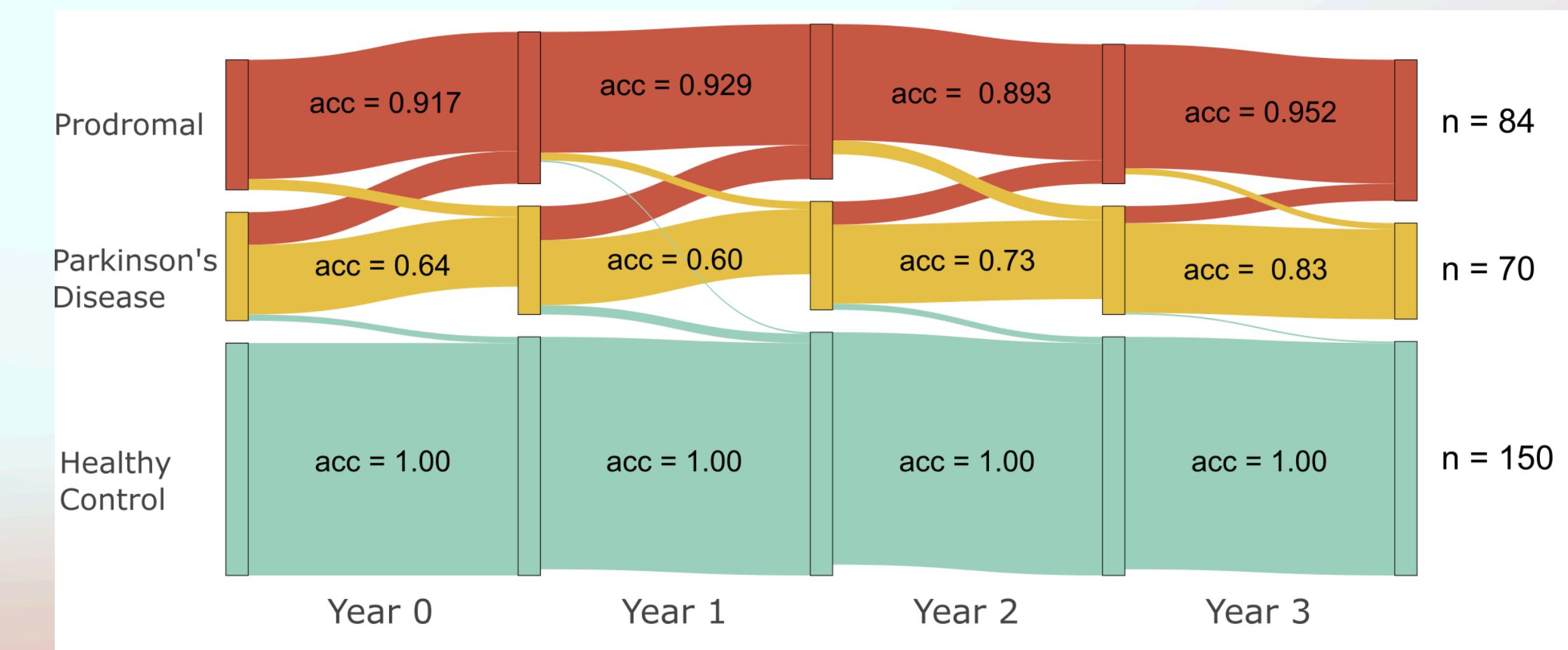
Barry  
Ryan

# Informative Modalities Vary During Parkinson's Disease Progression

- Cross-Sectional performance of MOGDx when stratifying participants into
  - Parkinson's Disease
  - Prodromal (early indicators of disease but no clinical diagnosis)
  - Healthy Control

	Modalities	Number of Participants	Accuracy	F1 score	Improvement in Accuracy
Genetic + Idiopathic (All)	Year 0 DNAm + SNP + mRNA + miRNA	1515	0.630 ± 0.019	0.665 ± 0.017	0.110 ± 0.018
	Year 1 DNAm	548	0.624 ± 0.020	0.667 ± 0.032	0.111 ± 0.02
	Year 2 Clinical + DNAm	542	0.694 ± 0.037	0.717 ± 0.034	0.166 ± 0.037
	Year 3 DNAm	493	0.712 ± 0.018	0.699 ± 0.048	0.146 ± 0.018
Genetic	Year 0 DNAm + SNP	489	0.789 ± 0.036	0.753 ± 0.04	0.419 ± 0.036
	Year 1 DNAm + SNP	443	0.867 ± 0.018	0.835 ± 0.02	0.472 ± 0.018
	Year 2 DNAm + SNP	432	0.866 ± 0.031	0.837 ± 0.032	0.477 ± 0.031
	Year 3 DNAm + SNP	365	0.841 ± 0.034	0.811 ± 0.038	0.403 ± 0.034
Idiopathic	Year 0 SNP + miRNA	667	0.681 ± 0.031	0.752 ± 0.008	0.069 ± 0.031
	Year 1 CSF + DNAm + SNP	582	0.720 ± 0.039	0.776 ± 0.035	0.122 ± 0.039
	Year 2 CSF + Clinical + DNAm	399	0.805 ± 0.022	0.770 ± 0.022	0.246 ± 0.022
	Year 3 CSF + DNAm	360	0.764 ± 0.022	0.721 ± 0.021	0.183 ± 0.022

- Strong disease signature found in integration of SNP and DNAm modalities for individuals with a genetic association
- Models trained later in the disease course are more accurate
- Epigenetic modifications are informative throughout the disease course



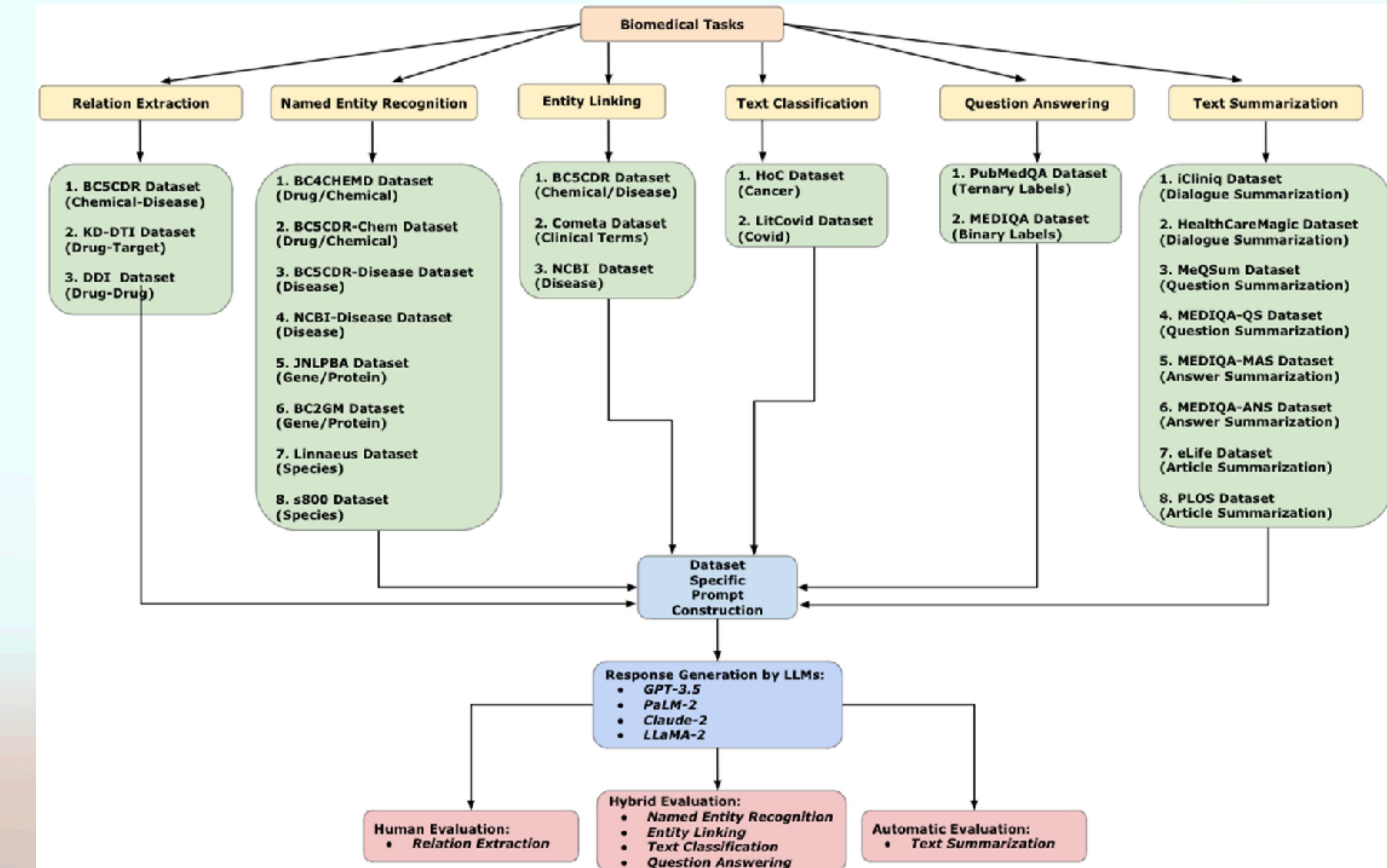
Ryan B, Marioni R, Simpson TI. An integrative network approach for longitudinal stratification in Parkinson's disease.  
PLOS Computational Biology 2025 21(3): e1012857. <https://doi.org/10.1371/journal.pcbi.1012857>

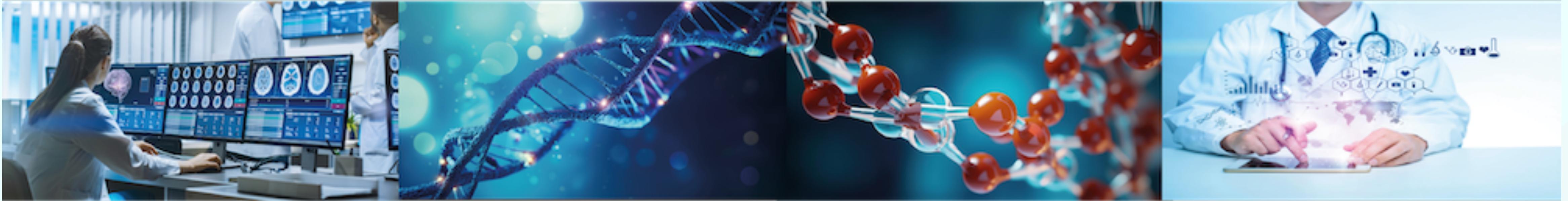


# Large Language Models for Biomedical Text

Large Language Models (LLMs), such as OpenAI's GPT (Generative Pre-trained Transformer) and its variants, are increasingly being applied in biomedical text processing due to their ability to understand and generate human-like text.

- **Summarisation** - LLMs can summarise biomedical texts and medical records and generate synthetic biomedical text for data augmentation.
- **Literature Mining** - identify trends, patterns, and associations between biomedical concepts. This includes identifying novel drug candidates, predicting disease risk factors, and discovering potential therapeutic targets.
- **Clinical Decision Support** - question answering, reference recommendations, possible treatment options, prediction of patient outcomes, automated report generation.





# Programming for Biomedical Informatics

Next Lecture this Thursday - “Working with Notebooks & Git”

**Please Bring your Laptop!**

**Ask Questions on the Piazza Discussion Board**

<https://github.com/biomedical-informatics/pbi>