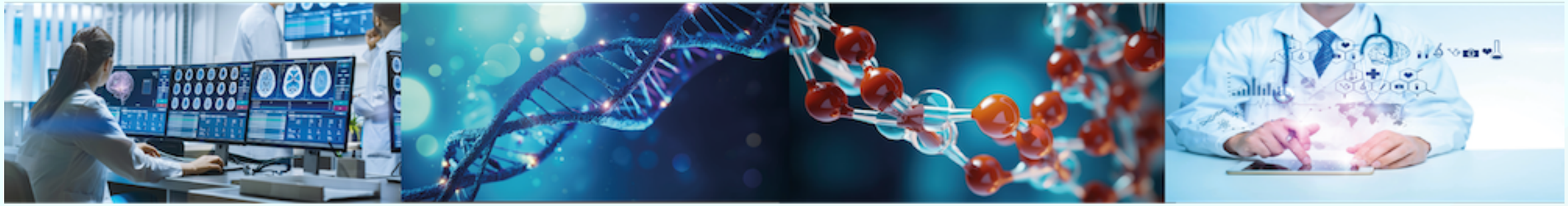


Programming for Biomedical Informatics

Lecture 1 - Welcome & Getting Started

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

Ian Simpson
ian.simpson@ed.ac.uk

THE UNIVERSITY of EDINBURGH
informatics

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

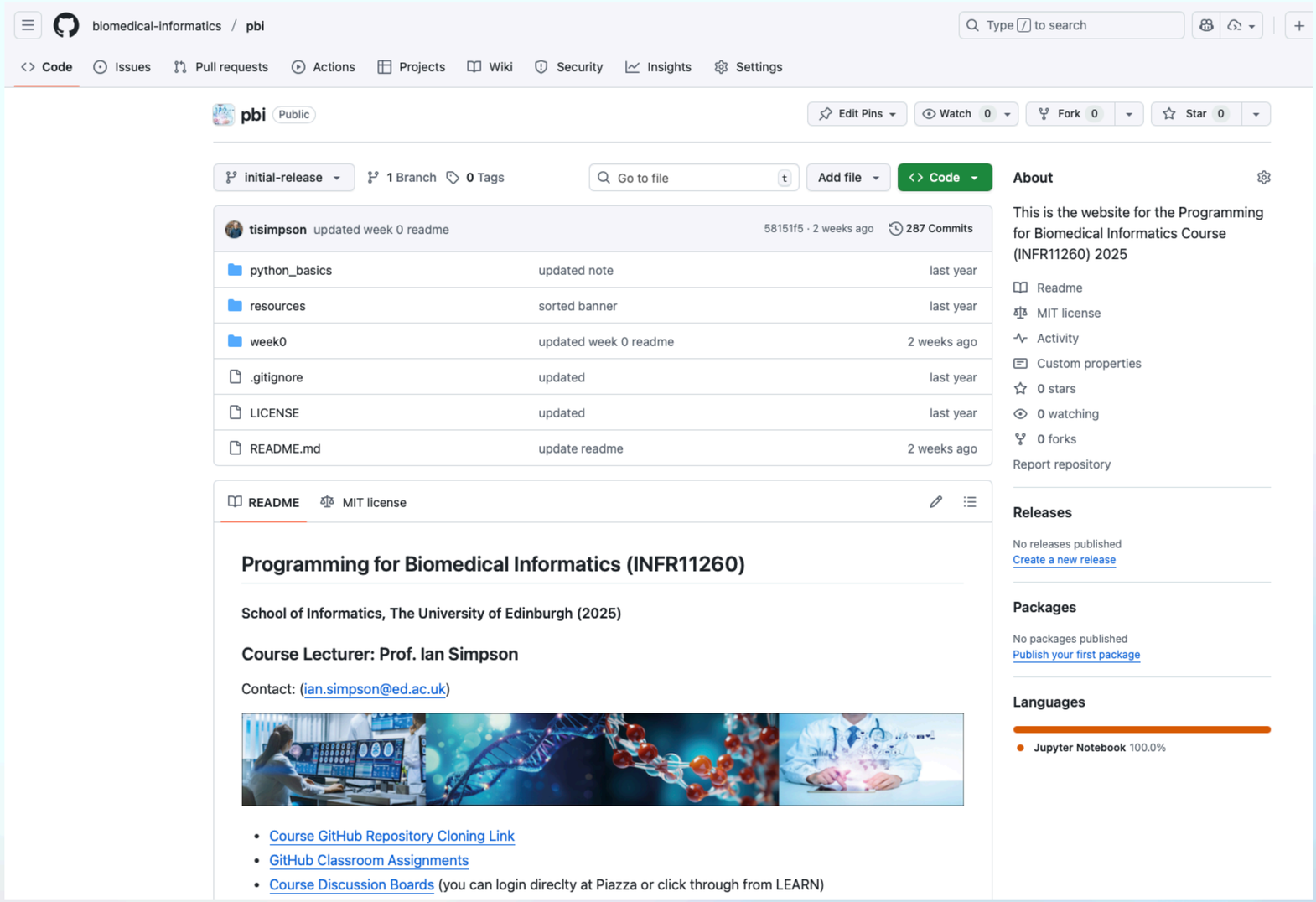
Week	Week Commencing	Weekly Topics (background, application)
1	15th September	(L1) Course Introduction & Setup, (L2) Working with Notebooks & Git
2	22rd September	(L3) Introduction to the Biomedical Dataverse, (L4) Finding & Fetching Data
3	29th September	(L5) Mapping and Harmonisation, (L6) Data Integration & 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

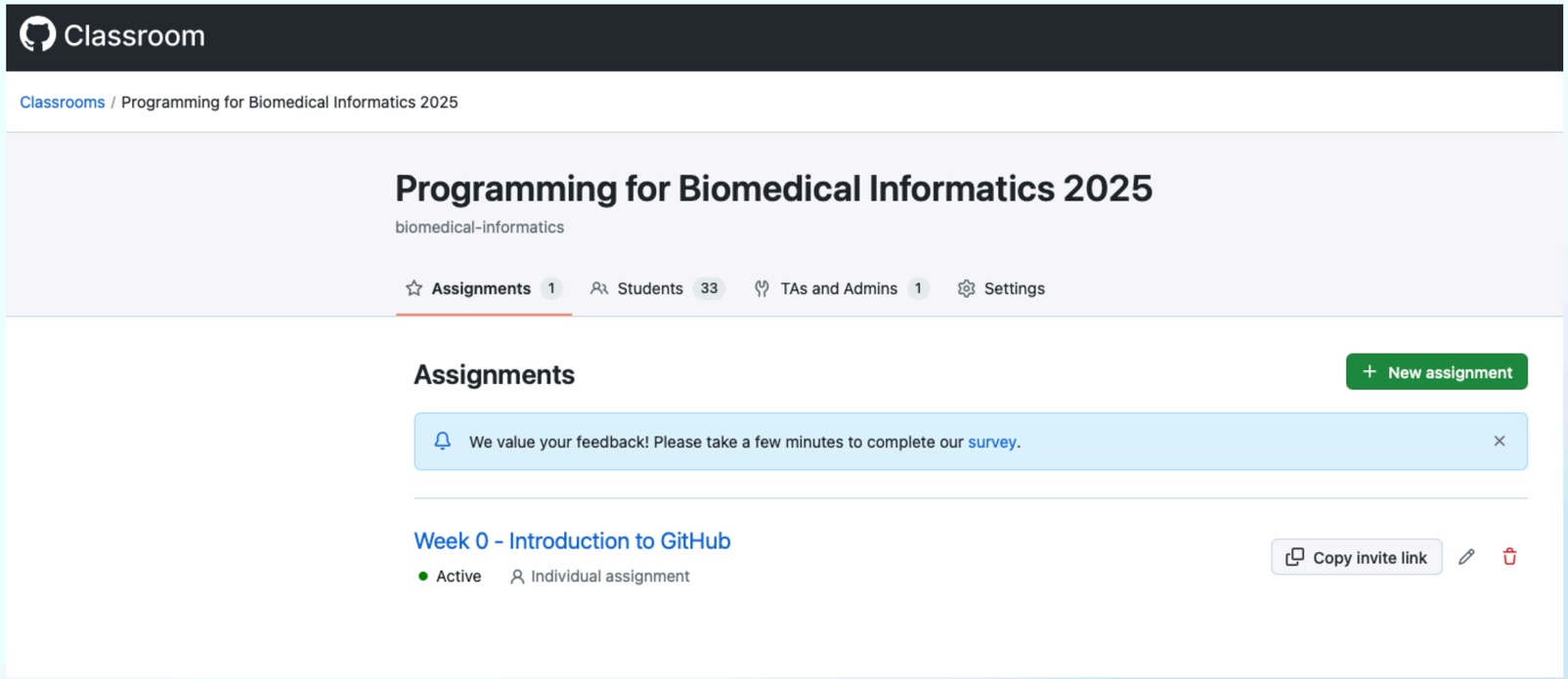
GitHub



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

Course Setup

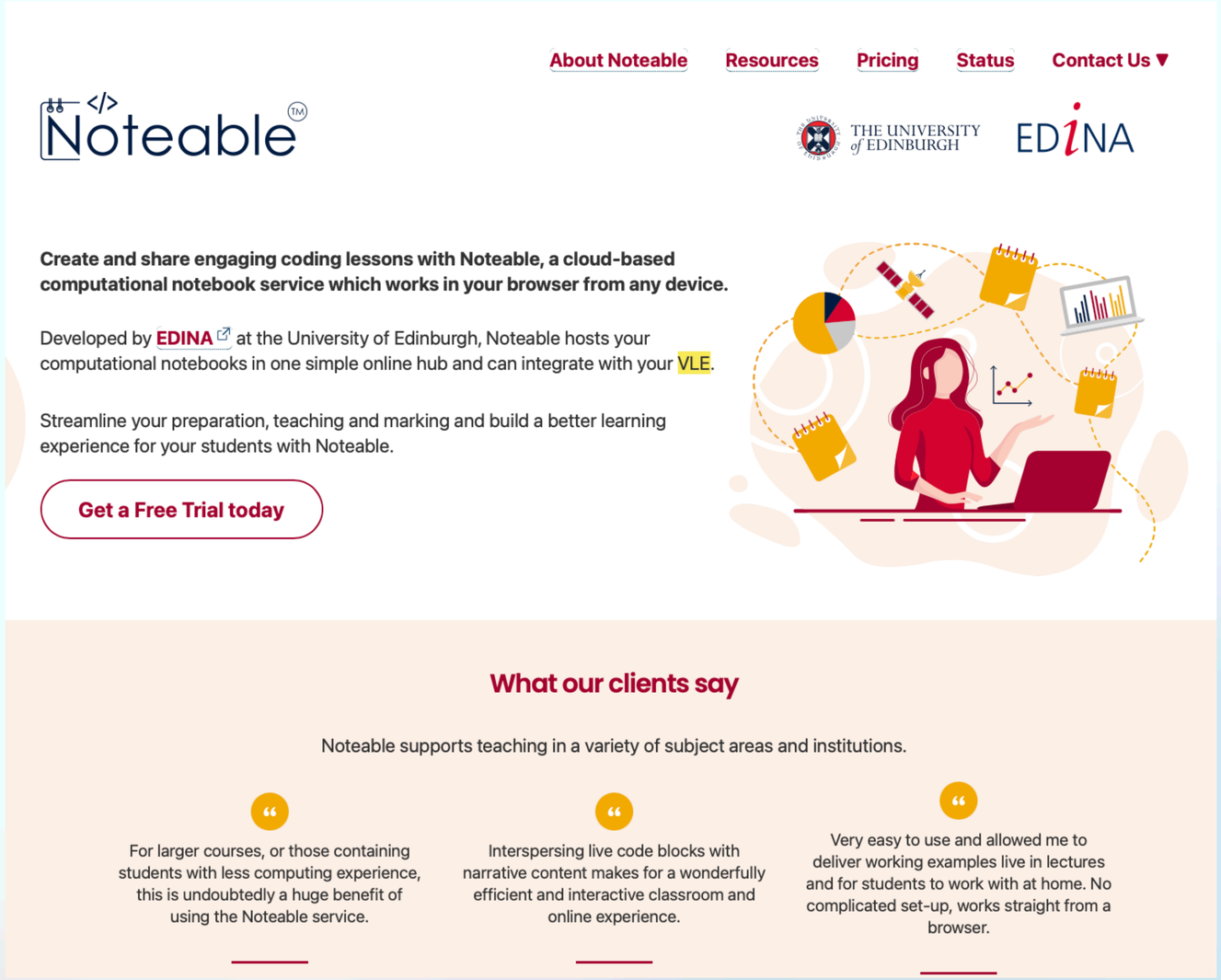
GitHub Classroom



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

Course Setup

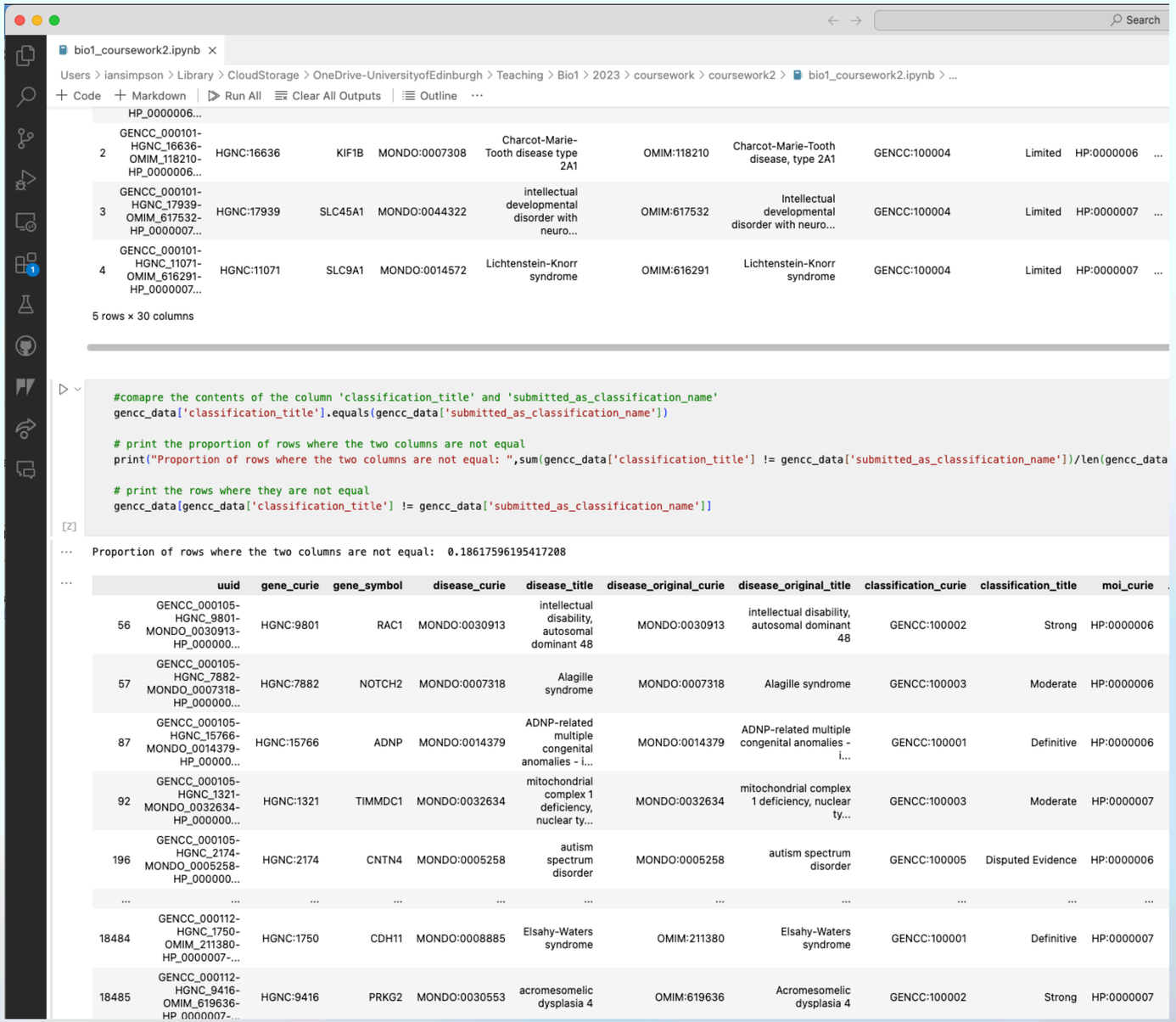
Noteable



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

Course Setup

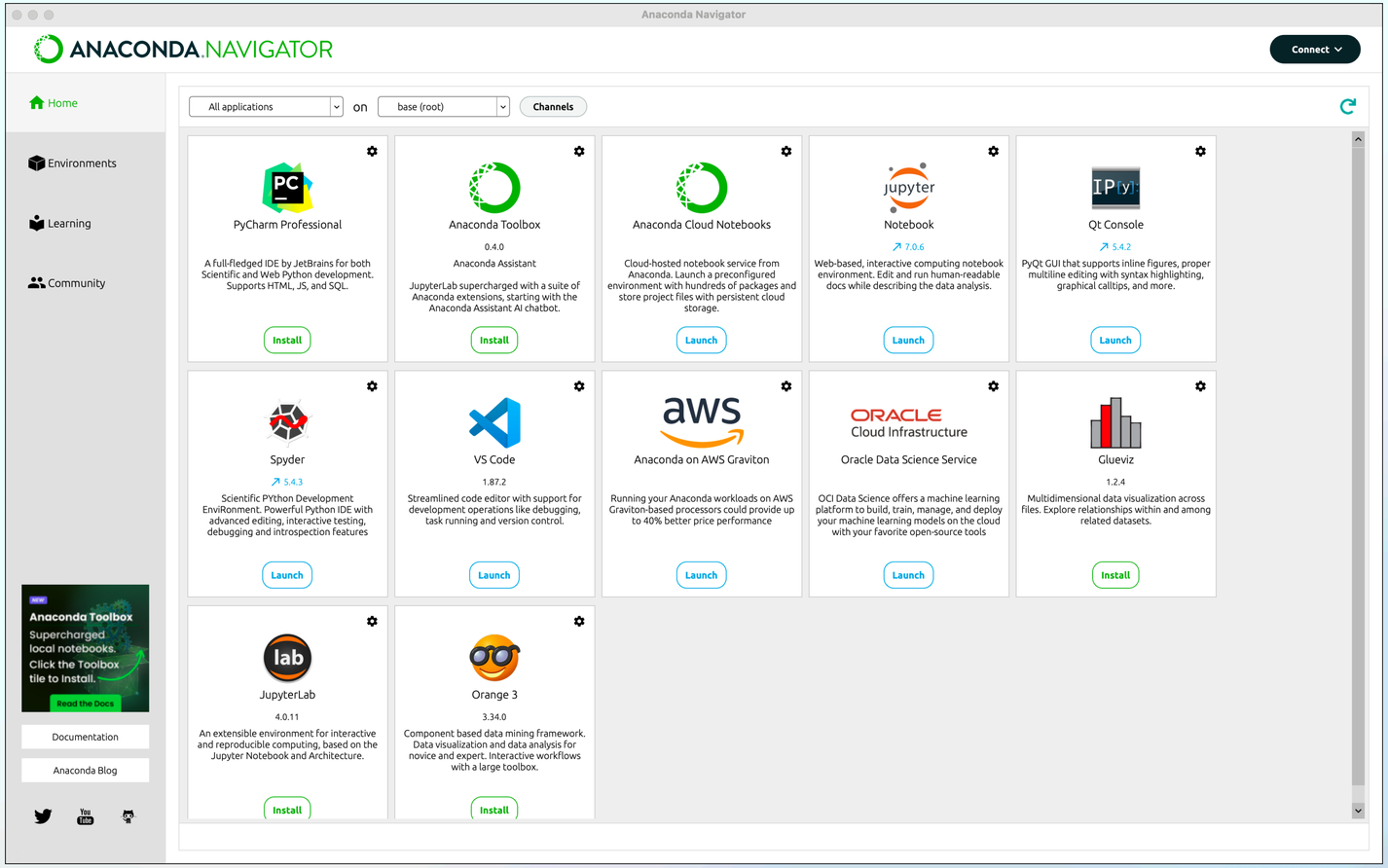
Visual Studio Code



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

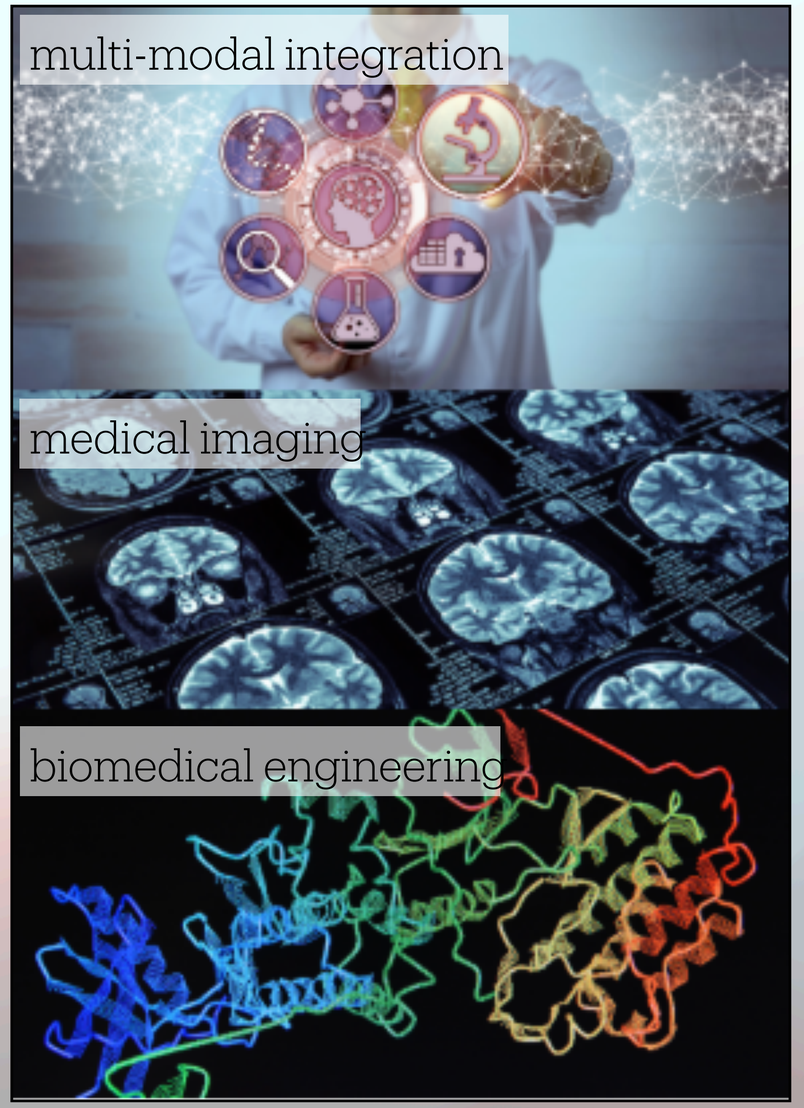
Clinical & Health
- Administration Support
- Decision Support
- Patient Engagement
- Synthetic Data Generation
- Clinical Trial Design & Monitoring
- Population Level Modelling
- Professional Education

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

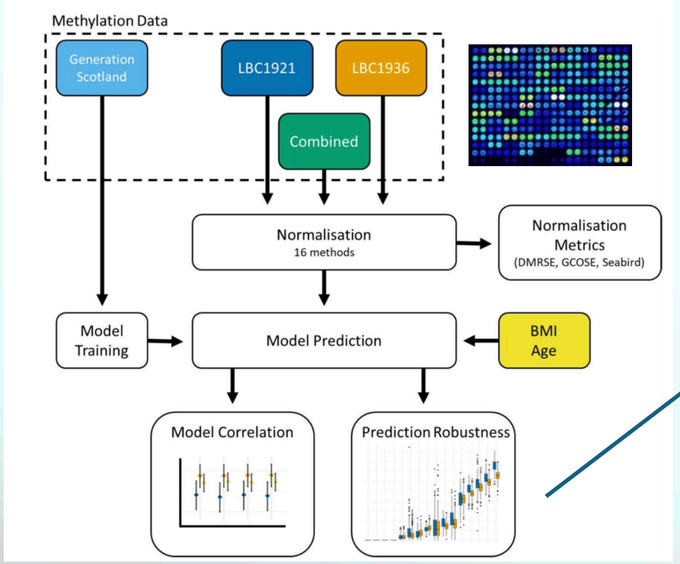
Societal & Health Systems
- Clinical safety, Efficacy, & Reliability
- Evaluation, Regulation, & Certification
- Privacy
- Copyright & Ownership
- Implementation & Adoption



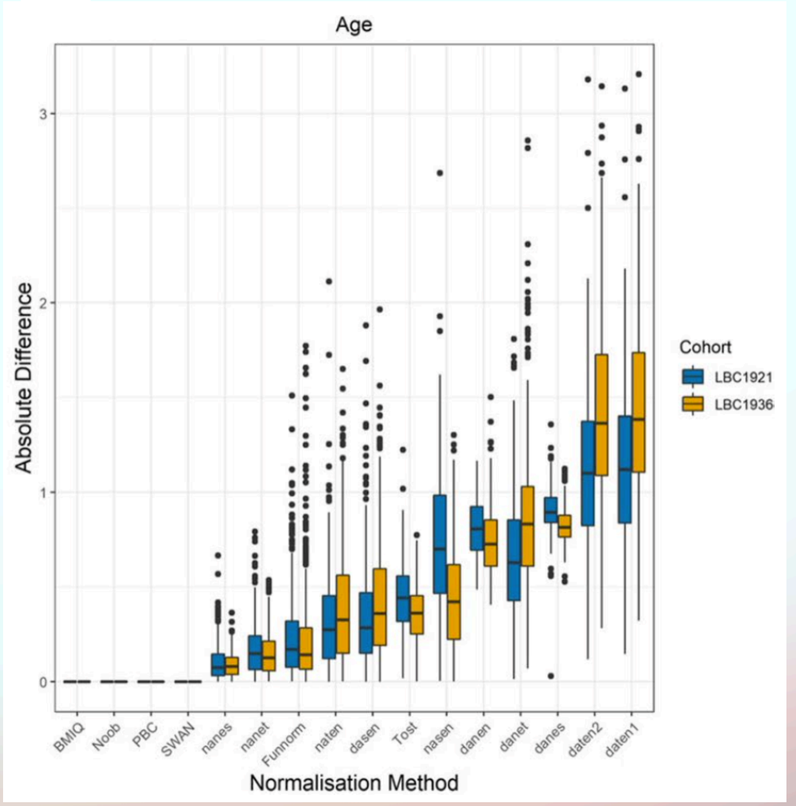
Integration of Datasets for Individual Prediction of DNA Methylation-Based Biomarkers

DNA Methylation Data

1921 Lothian Birth Cohort 1936



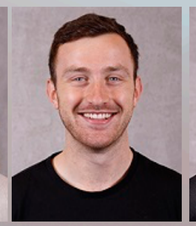
Robustness of Predictions



- Projecting data from new datsets 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

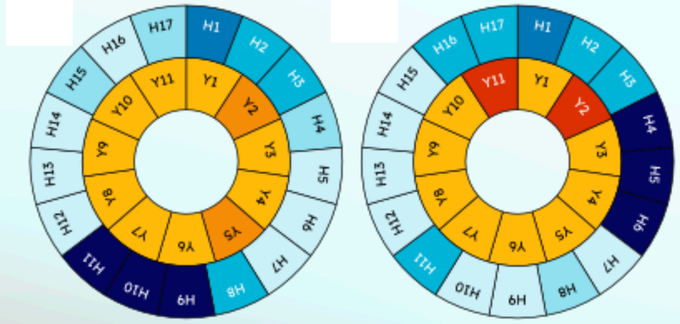
Psychiatric assessments are scheduled infrequently and rely on self-reported experiences

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

Reducing monitoring to acute episode detection misses on actionable clinical information

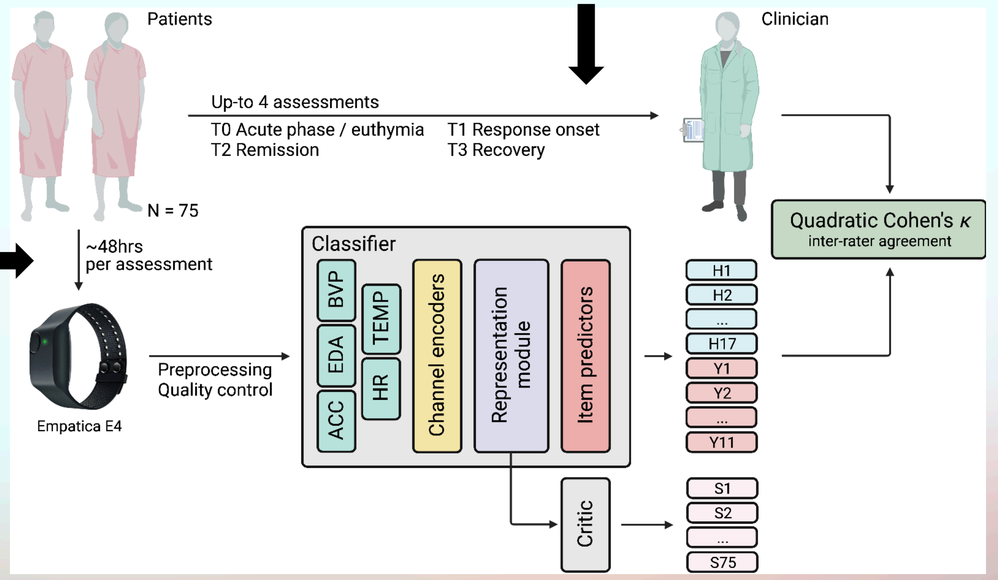
Timely Interventions
Better Outcomes

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



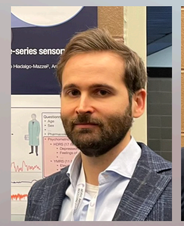
Hamilton Depression Rating Scale (HDRS)

Young Mania Rating Scale (HDRS)



Model		ACC		F_1score		
		segment	subject	segment	subject	segme
SL	ENET	66.38	71.88	66.54	70	72.
	KNN	70.37	82.81	71.34	80.6	73.
	SVM	71.25	81.25	71.63	78.81	73.
	XGBoost	72.02	82.81	71.72	82.03	72.
	E4mer	75.35	81.25	74.39	81.33	75.
SSL	MP (LR)	77.53	87.5	77.87	88.3	78.
	MP (FT)	**81.23**	**90.63**	**81.45**	**91.47**	**82.**
	TP (LR)	71.16	81.25	72.06	82.37	71.
	TP (FT)	75.69	84.38	75.1	83	75.

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

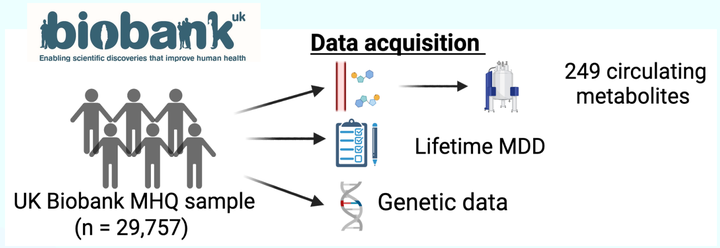


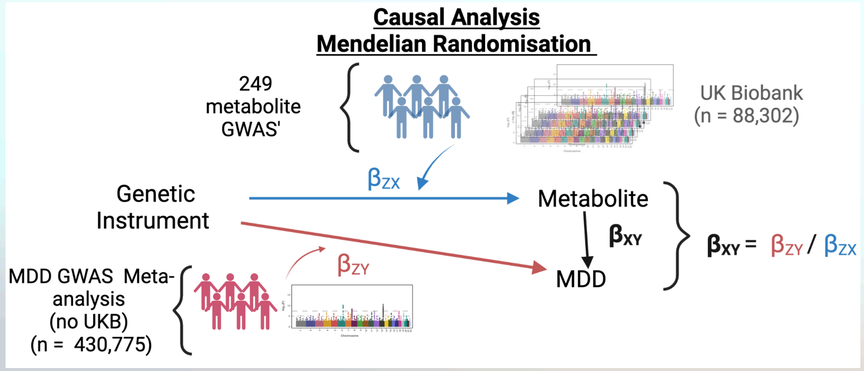
Filippo Corponi

Bryan Li

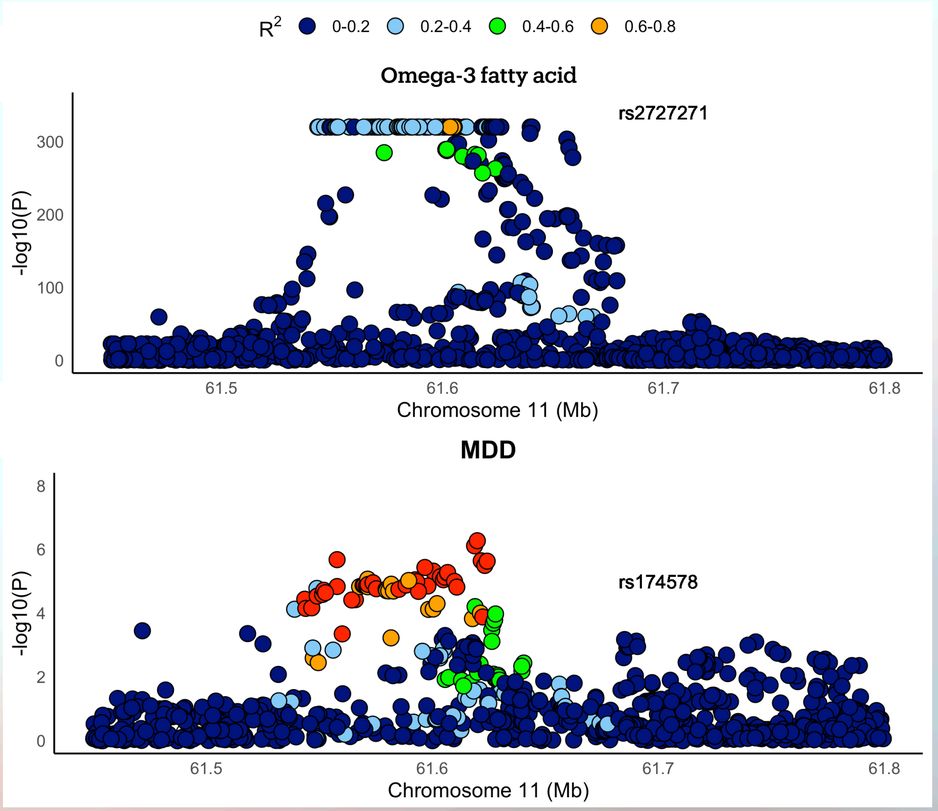
Metabolites and Major Depressive Disorder in the UK Biobank

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.

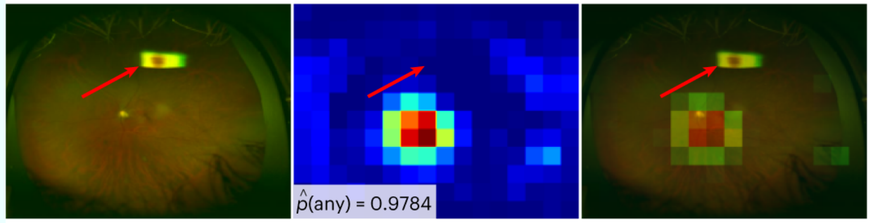


Colocalisation of MDD and Omega-3 fatty acid loci in the FADs cluster (PP.H4 > 0.90)

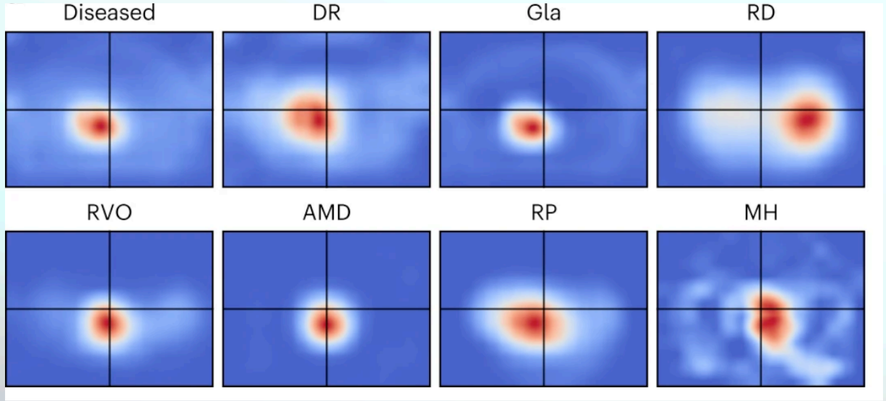
Eleanor Davyson, Xueyi Shen, Danni A. Gadd, Elena Bernabeu, Robert F. Hillary, Daniel L. McCartney, Mark Adams, Riccardo Marioni, and Andrew M. McIntosh (2023) Metabolomic Investigation of Major Depressive Disorder Identifies a Potentially Causal Association With Polyunsaturated Fatty Acids. Society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

Ella Davyson

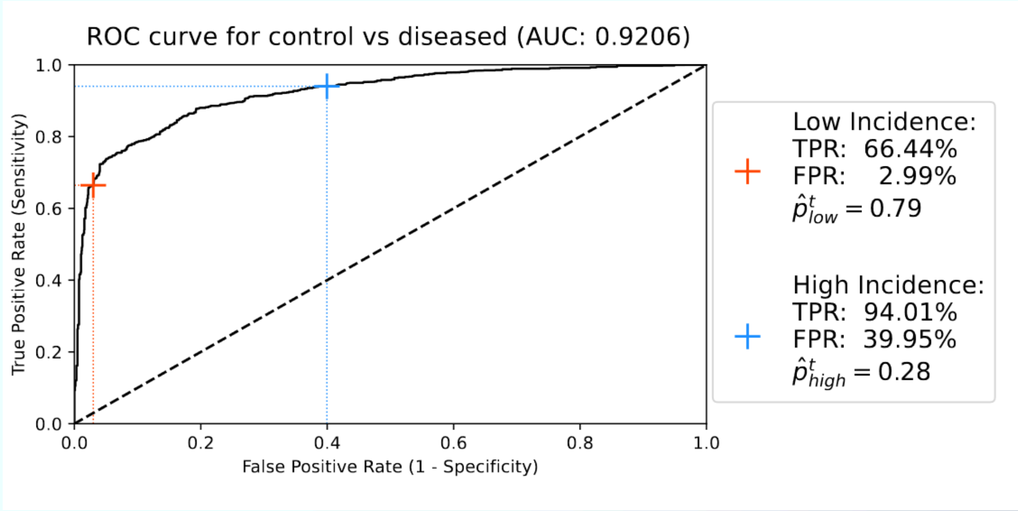
Disease Detection in Ultra-Wide-Field Retinal Images



Age: 57 | Sex: M | Dataset labels: Gla
MH: 0.055(0)| RP: 0.048(0)| AMD: 0.039(0)| RVO: 0.103(0)| RD: 0.053(0)| Gla: 0.968(1)| DR: 0.079(0)| any: 0.978(1)



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

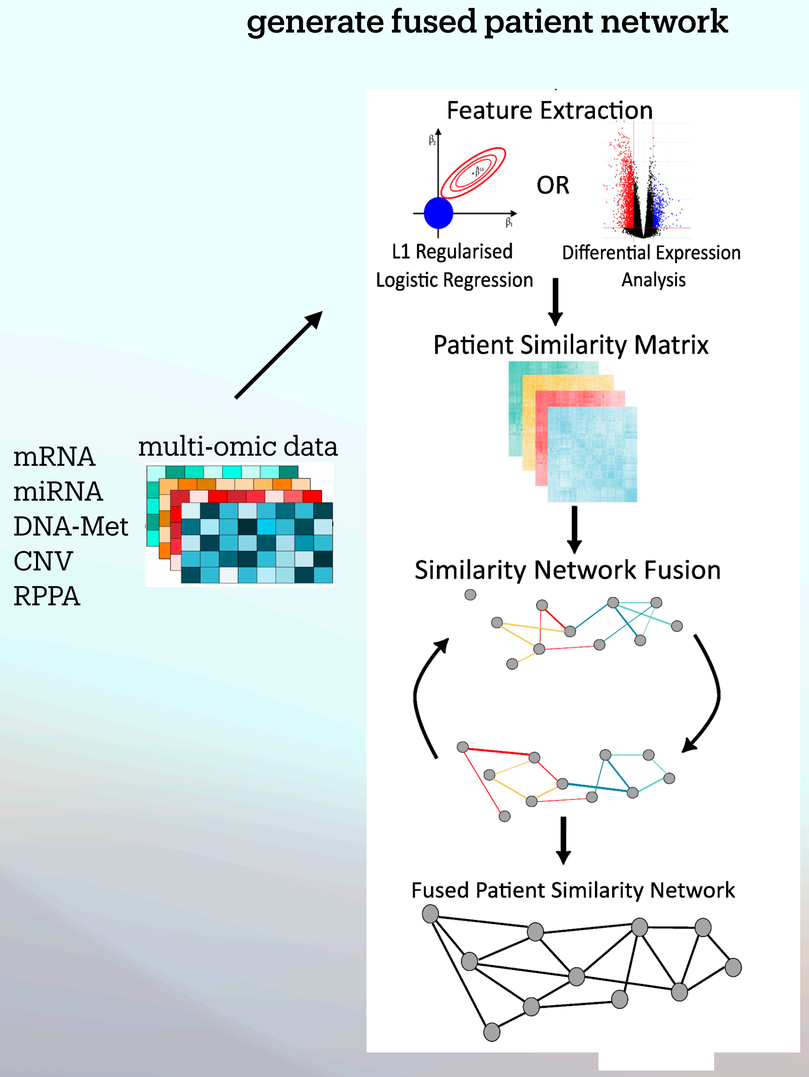


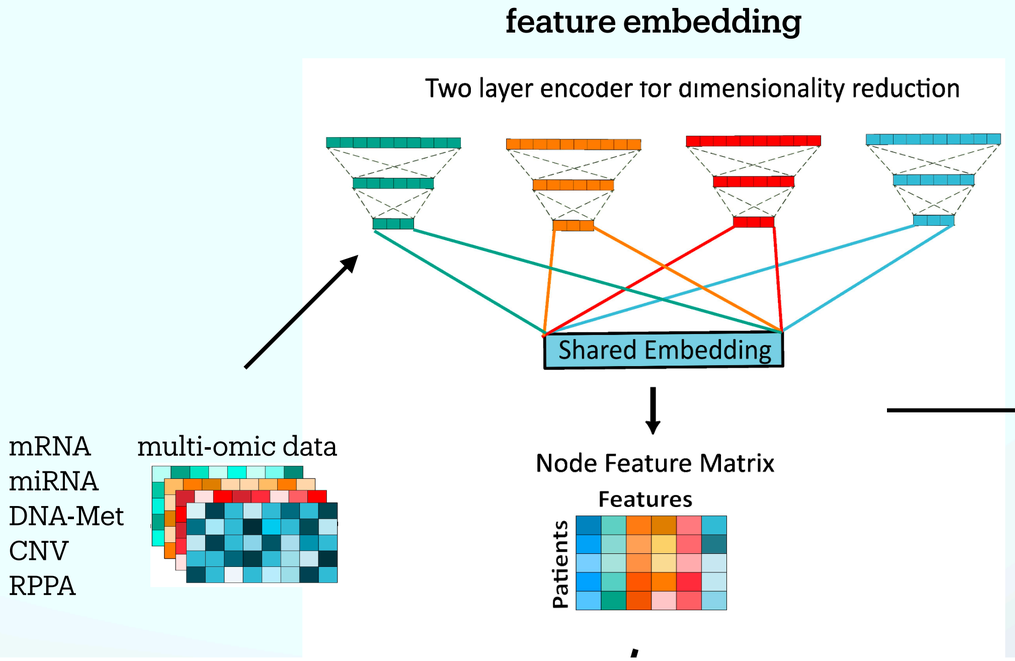
	Diseased	DR	Gla	RD	RVO	AMD	RP	MII
Logistic Regression with Age + Sex	0.5964	0.5988	0.5155	0.7676	0.4892	0.8021	0.6776	0.5625
Ensemble of Experts (binary DL models + balanced data)	0.8318 *	0.8432	0.9141	0.9217	0.8996	0.7113	**0.9490**	0.6454
Ours (Single multi-label DL model + realistic data)	**0.9206**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

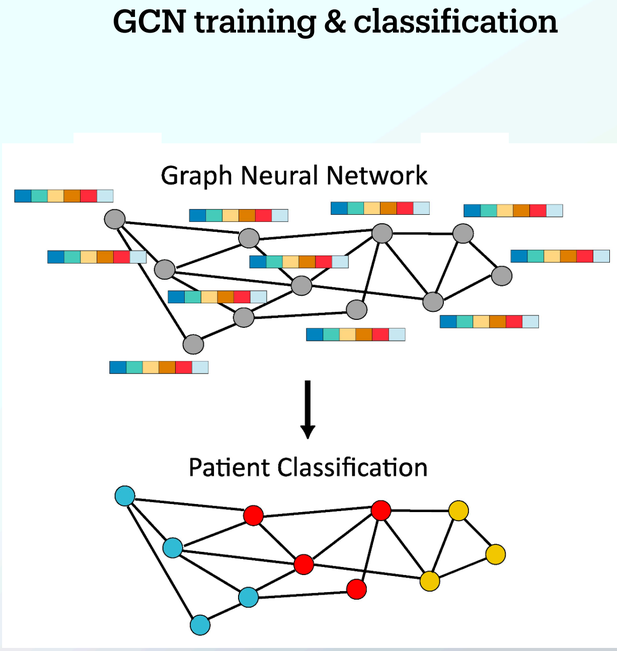
Justin Engelmann, Alice D. McTrusty, Ian J. C. MacCormick, Emma Pead, Amos Storkey & Miguel O. Bernabeu (2022) Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. Nature machine intelligence. https://doi.org/10.1038/s42256-022-00566-5

Justin Engelmann

Cancer Classification From Multi-Modal Fused Patient Networks







Method	Dataset	Number of Modalities	Number of Samples	Number of Classes	Accuracy	F1
MOGDx	BRCA	4	1083	5	0.893 ± 0.014	0.874 ± 0.012
	BRCA	4	1043	4	0.904 ± 0.014	0.887 ± 0.016
	LGG	1	457	2	0.899 ± 0.016	0.881 ± 0.019
	KIPAN	4	888	3	0.958 ± 0.003	0.948 ± 0.004
MOGONET	BRCA	3	875	5	0.829 ± 0.018	0.825 ± 0.016
	LGG	3	510	2	0.816 ± 0.016	0.814 ± 0.014
	KIPAN	3	658	3	0.999 ± 0.002	0.999 ± 0.002
MoGCN	BRCA	3	511	4	0.898 ± 0.025	0.902 ± 0.024
	KIPAN	3	698	3	0.977 ± 0.017	0.977 ± 0.017
SVM	BRCA	1	869	5	0.782 ± 0.033	0.721 ± 0.030
Lasso	BRCA	1	1047	5	0.829 ± 0.014	0.771 ± 0.012
XGBoost	BRCA	1	1047	5	0.762 ± 0.036	0.692 ± 0.033

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

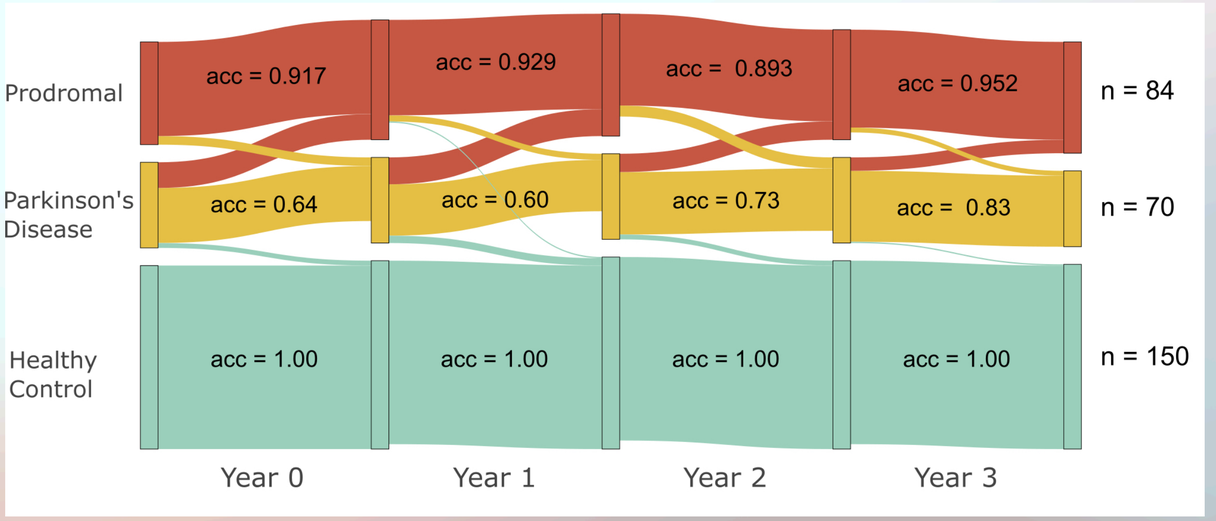
Barry Ryan

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



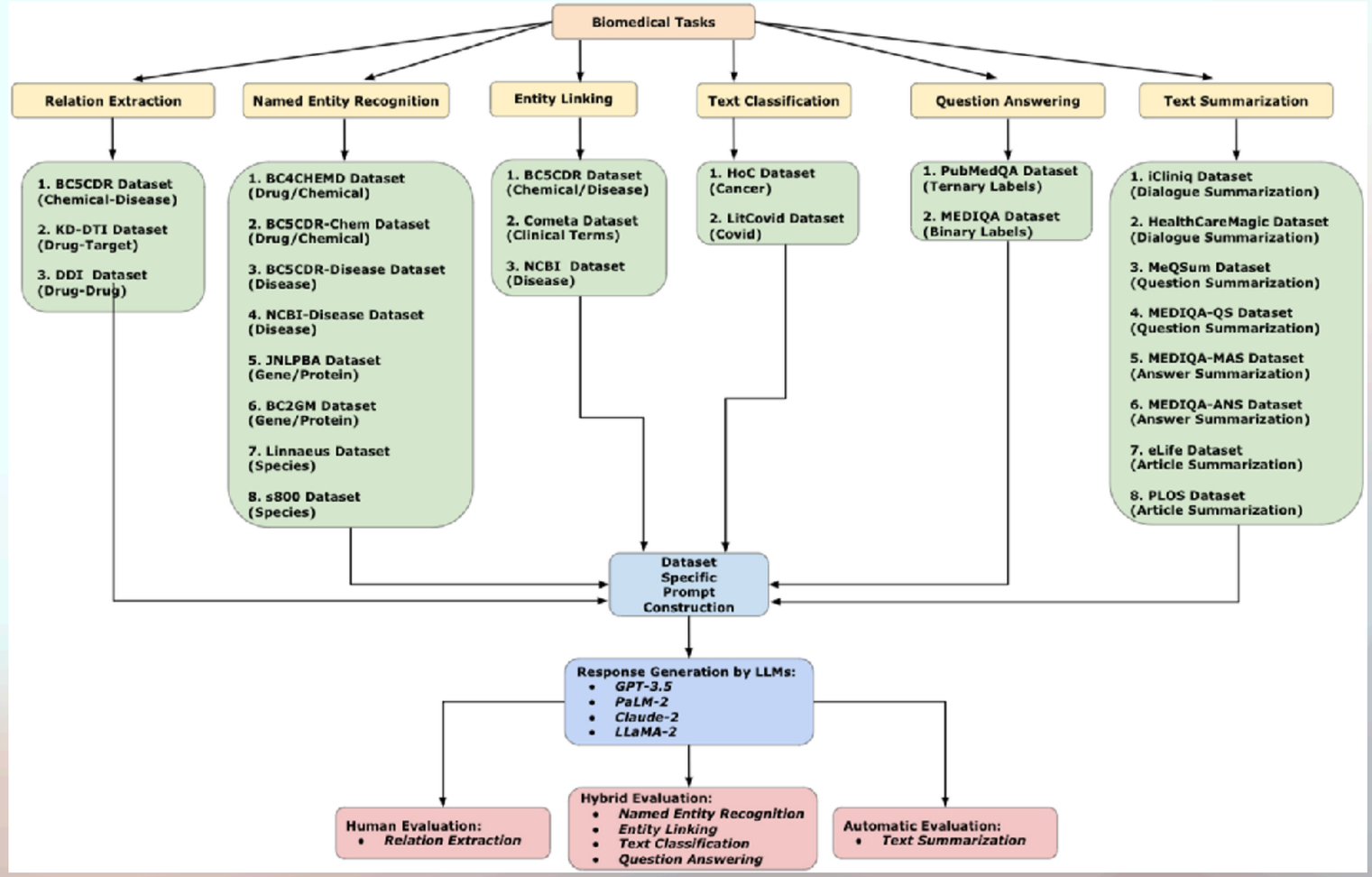
Ryan B, Marioni R, Simpson TI. An integrative network approach for longitudinal stratification in Parkinson's disease. PLOS Computational Biology 2025 21(3): e1012857. https://doi.org/10.1371/journal.pcbi.1012857

Barry Ryan

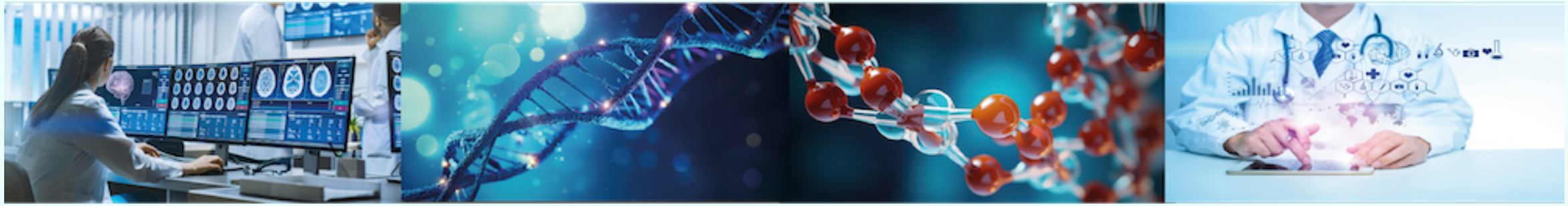
Large Language Models for Biomedical Text

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.



Jahan et al. 'A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks'. doi.org/10.1016/j.compbiomed.2024.108189



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

THE UNIVERSITY of EDINBURGH
informatics