IMPERIAL

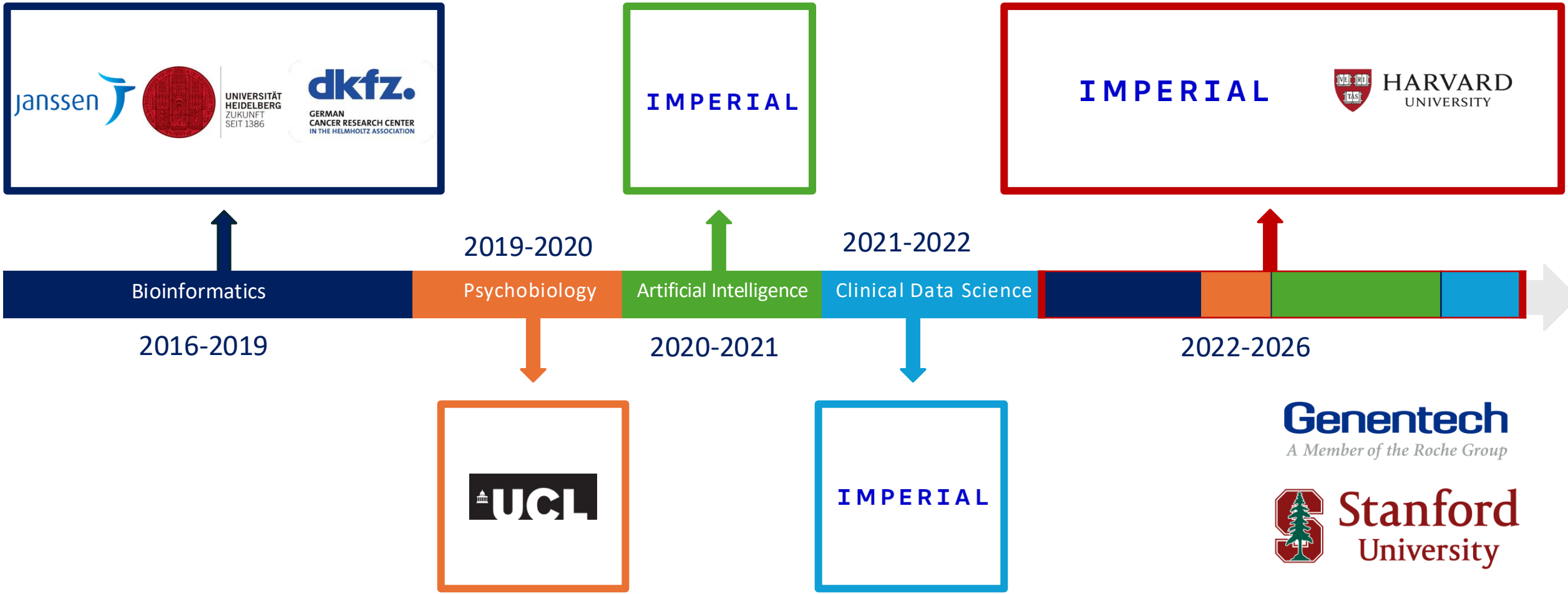
Medical Research Council

HARVARD UNIVERSITY

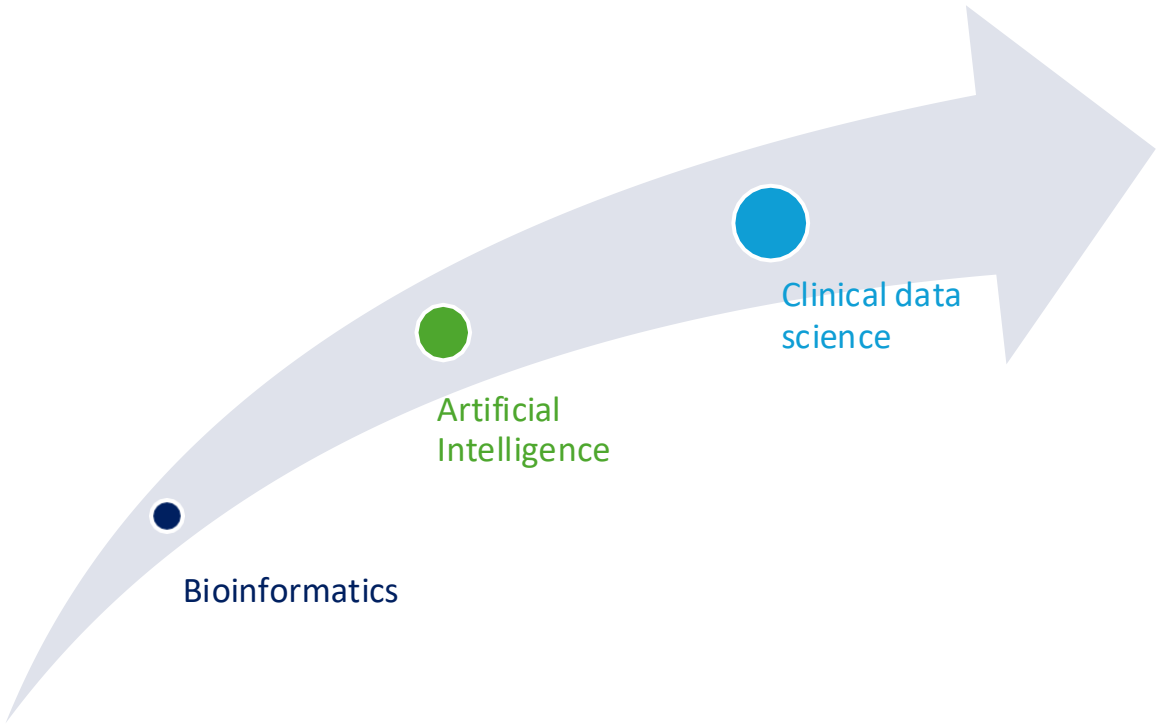
Lessons from a multiscale protein–phenotype foundation model

Valentina Giunchiglia

My background



My background



Interdisciplinarity

Research Motivation

Holistic study of diseases across **multiple scale** to drive advancements in **precision medicine**

How can we achieve this using AI?

Different approaches throughout the years….

1st paradigm shift

2nd paradigm shift

Standard ML

Models learn to complete specific, narrowly defined tasks.

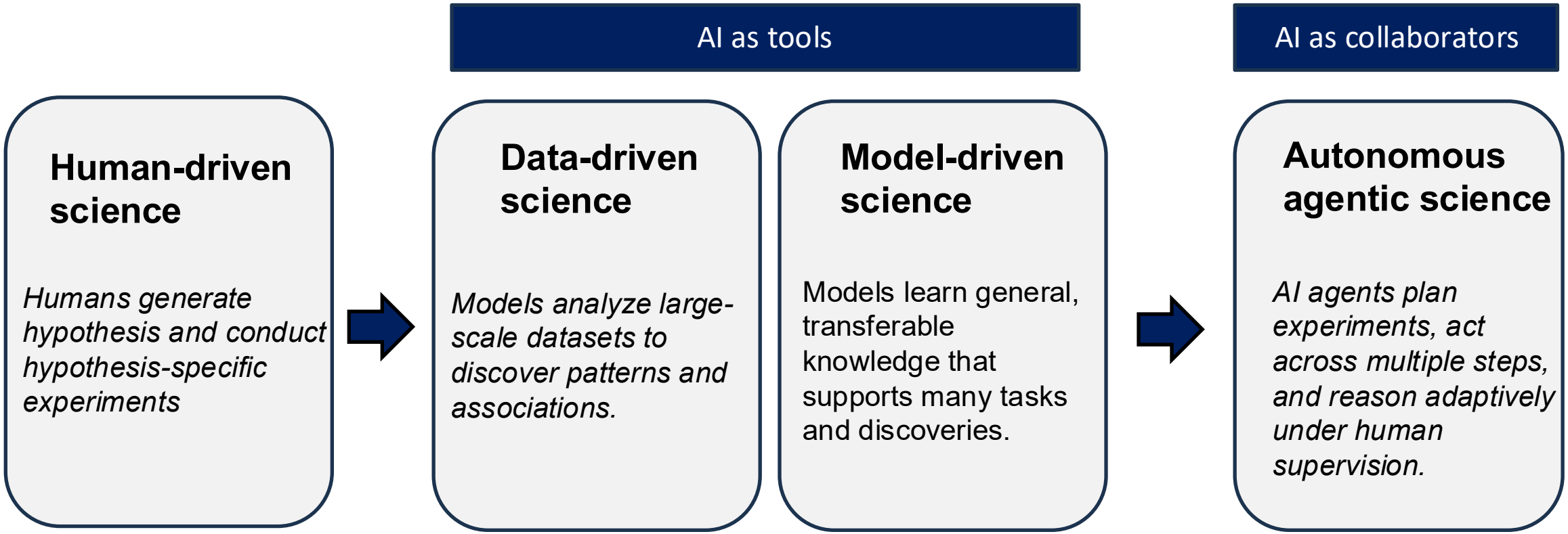
Foundation models

General-purpose models that learn transferable data representations across tasks and domains.

AI agents

Autonomous systems that plan, act, and reason across multi-step scientific workflows.

How have AI paradigm shifts changed scientific research?



Which AI approach should we use? Different AI paradigms solve different problems

- Well-defined, narrow prediction task
- Sufficient labeled, task-specific data
- Structured inputs and outputs
- Interpretability is high priority

PREDICT a defined outcome

- Complex, high-dimensional or multimodal data
- Reusable representations across multiple tasks
- Limited labels but substantial pretraining data
- Transfer learning, generation, and generalization are important

LEARN or TRANSFER knowledge

- Multi-step tasks requiring planning and adaptation
- Interaction with tools, APIs, or external environments
- Decisions depend on intermediate results and feedback
- Workflow automation with appropriate oversight and guardrails

EXECUTE a multi step goal

Regardless of the approach, choosing a model is only one part of the challenge

MODEL

How do we incorporate biological knowledge into model design?

Existing models focus primarily on scale and often lack a biologically grounded design.

VALIDATE

How do we know an AI model has learned meaningful biology?

Benchmark performance ≠ biological validity or generalisation

EXPLAIN

How do we translate model predictions into biologically meaningful insight?

Focus on benchmarking

All three paradigms require appropriate validation to support scientific conclusions.

What makes an AI prediction scientifically trustworthy?

1. DATA & LEAKAGE

What has the model actually learned?

- Training-data overlap and contamination
- Memorization vs. meaningful patterns

BEYOND MEMORIZATION

2. GENERALIZATION

Does performance hold under different conditions?

- Across datasets and biological contexts
- Across different input formulations

BEYOND BENCHMARKS

3. BIOLOGICAL VALIDITY

Do predictions reflect meaningful biology?

- Independent biological evidence
- Experimental and mechanistic support

BEYOND PREDICTION

These challenges affect all AI approaches. Let's explore them through a foundation model case study.

Foundation models as a case study in biological validation



Multimodal multiscale foundation model for protein phenotypes

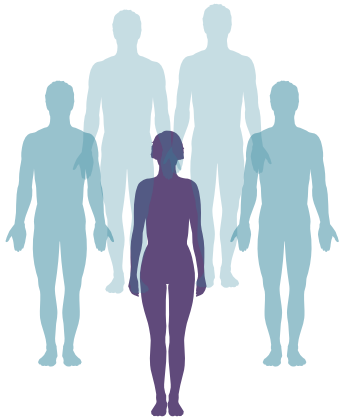
Website: https://zitniklab.hms.harvard.edu/ProCyon/
Models and Dataset: https://huggingface.co/collections/mims-harvard/procyon-6716632e60e4b3785bf8bd04
Code: https://github.com/mims-harvard/ProCyon

Most relevant publications:

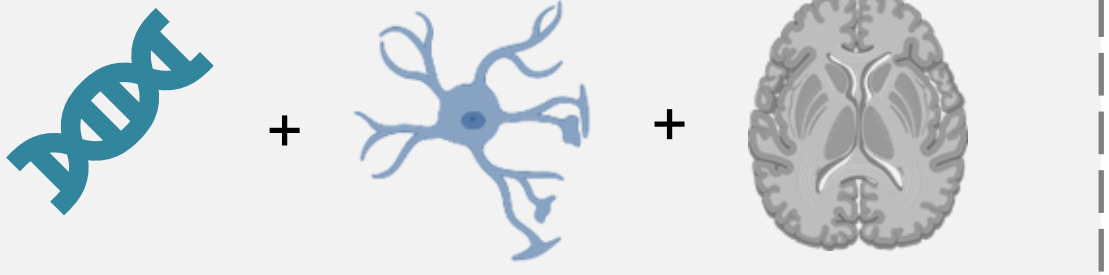
Queen, O.*, Huang, Y. *, Calef, R. *, **Giunchiglia, V.*,** et (Under Review in Nature)
Queen, O*, **Giunchiglia, V*,** Journal of Multiple Sclerosis (2025)

Complex human diseases span multiple scales and result into distinct clinical presentations

Example



Alzheimer's patients



Molecular changes

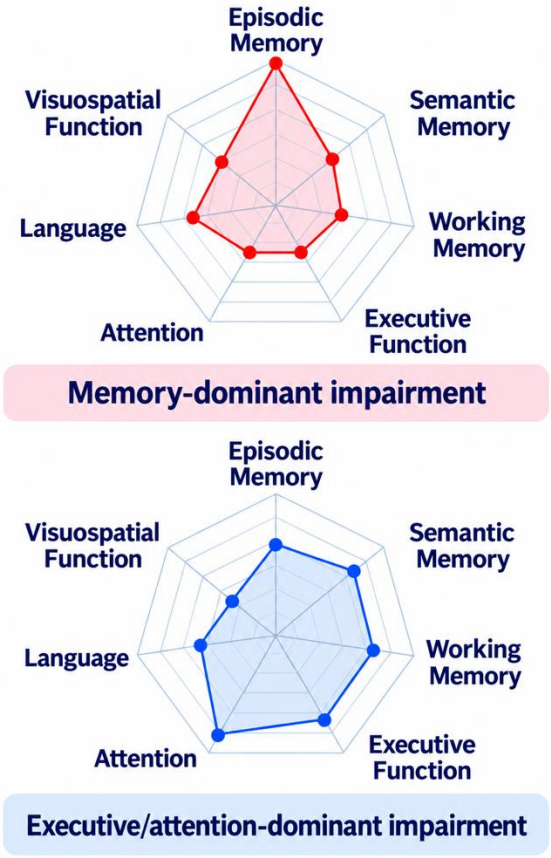
Tau, APOE4, APP

Cellular changes

Microglia activation, neuronal dysfunction

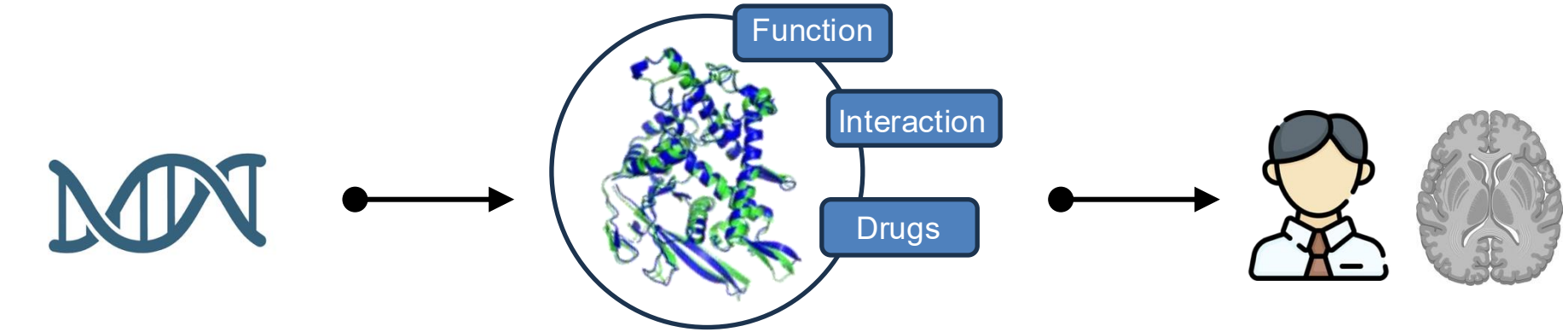
Tissue pathology

Plaques, tangles, atrophy



Clinical phenotype

Why protein phenotypes matter in this context?



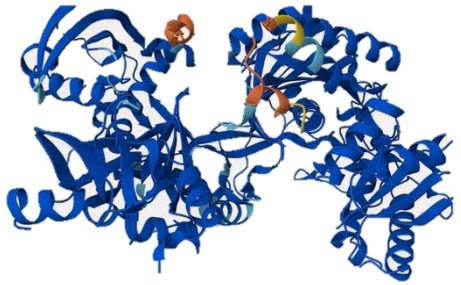
Learning protein phenotypes enables mapping from proteins to disease biology and clinical endpoints.

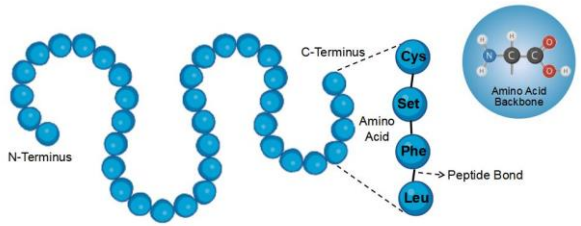
ProCyon

Foundation model trained on protein sequences, structures, and text data, leveraging natural language to **connect knowledge across biological domains**, from molecular mechanisms to therapeutics and clinical phenotypes

Molecular Information

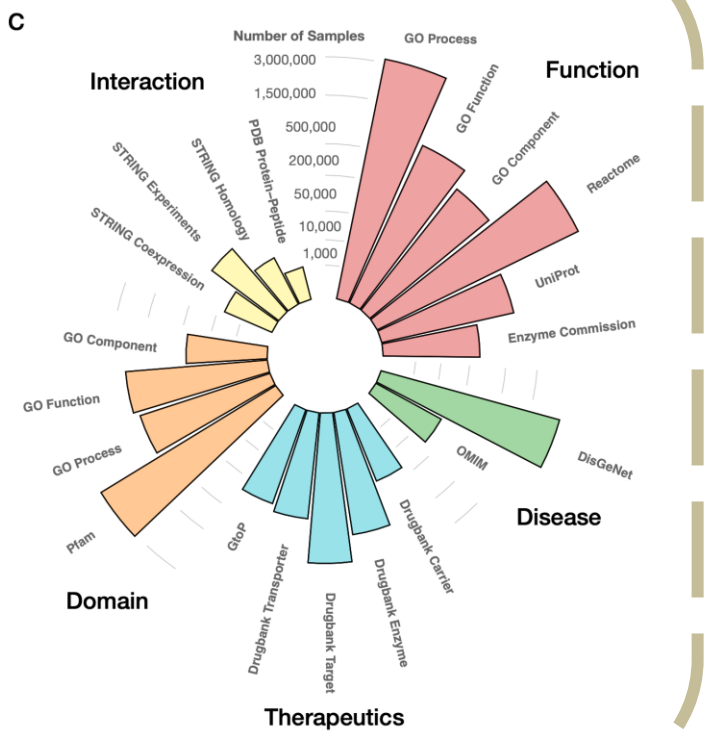
Protein Structure



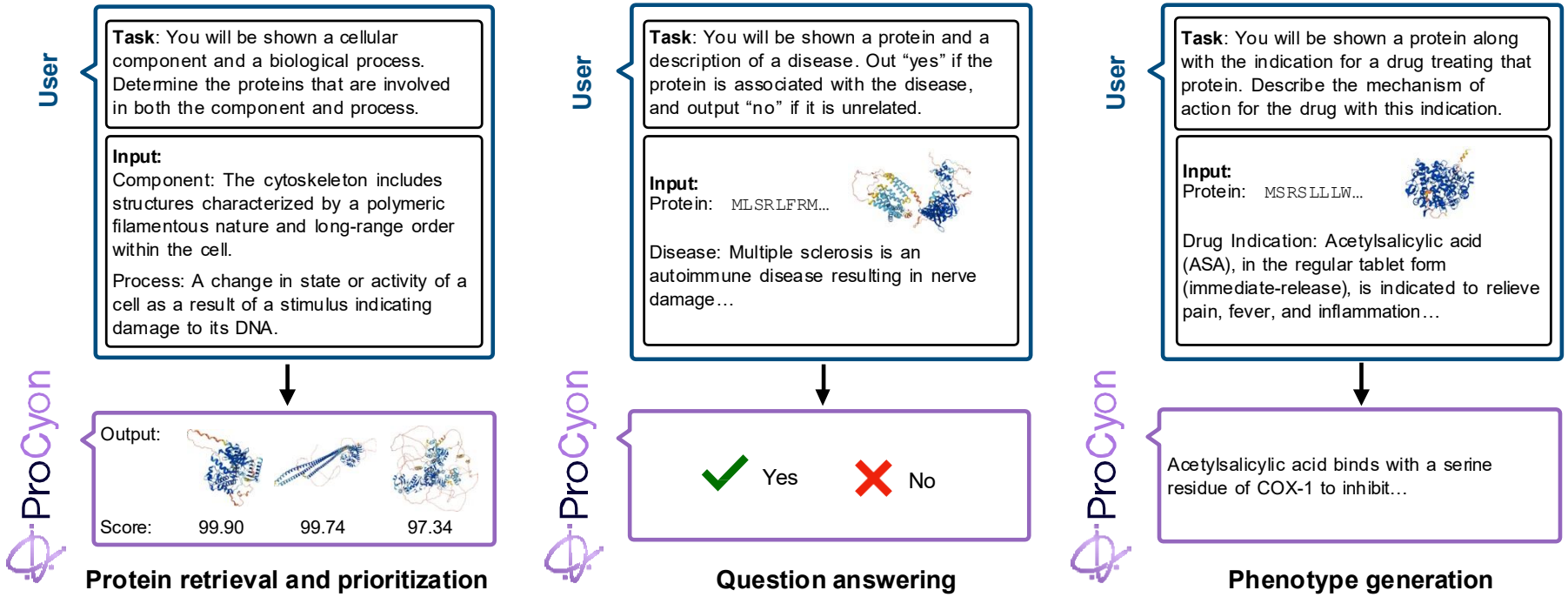


Protein Sequence

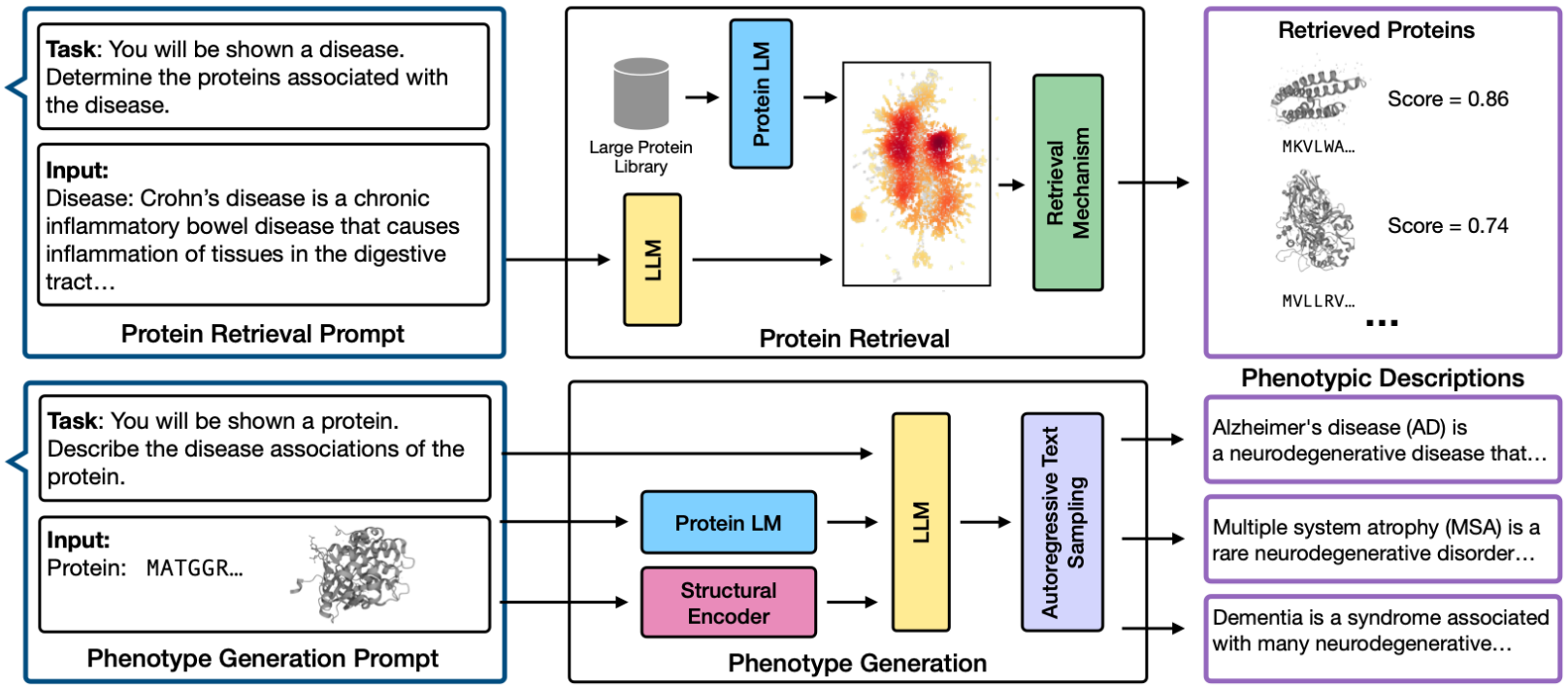
Multiscale Phenotypic information



ProCyon Capabilities



ProCyon Architecture



Co-training of a LLM with dedicated encoders for protein sequences, structures and ligand structures

Use of projectors to project modality-specific embeddings in the LLM space

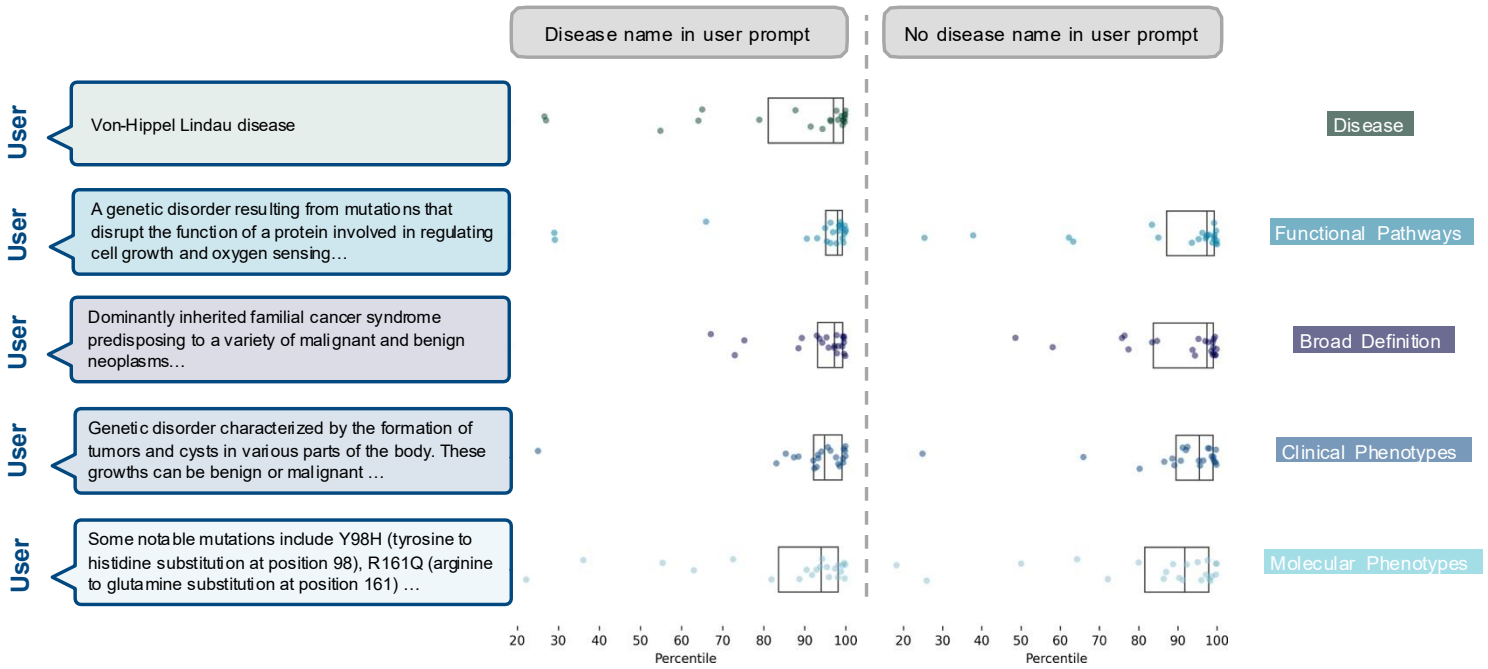
Contrastive learning for retrieval and causal language modeling for phenotype generation

How do we validate if ProCyon learnt molecular to phenotype associations?

1

Recovering known biology is necessary, but not sufficient

First line of validation: monogenic disorders with established ground truth



Input: rephrased description of *monogenic diseases* that capture disease changes at different scales

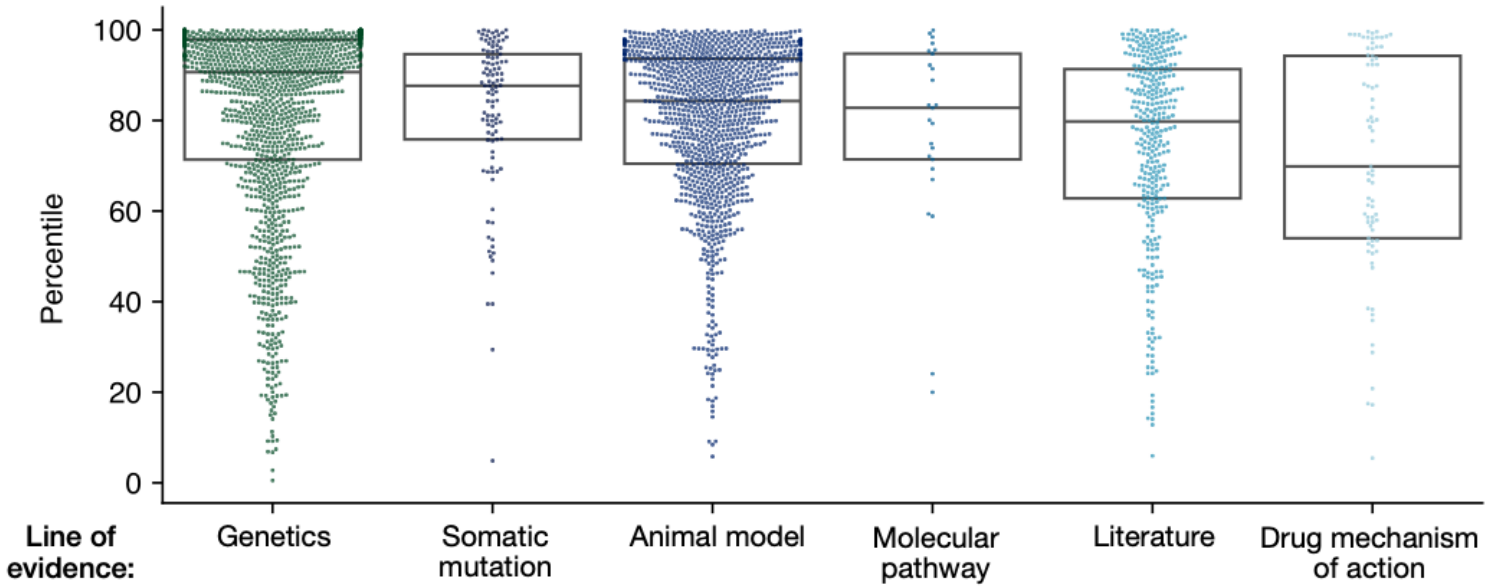
Output: ranked proteins associated to disease

ProCyon can retrieve the causal gene from multiscale phenotype descriptions

Monogenic disorders are ‘easier’ to model, no generalisation

2 Biological ground truth is not always ground truth

Second line of validation: polygenic disorders



Input: Disease description of *polygenic diseases*

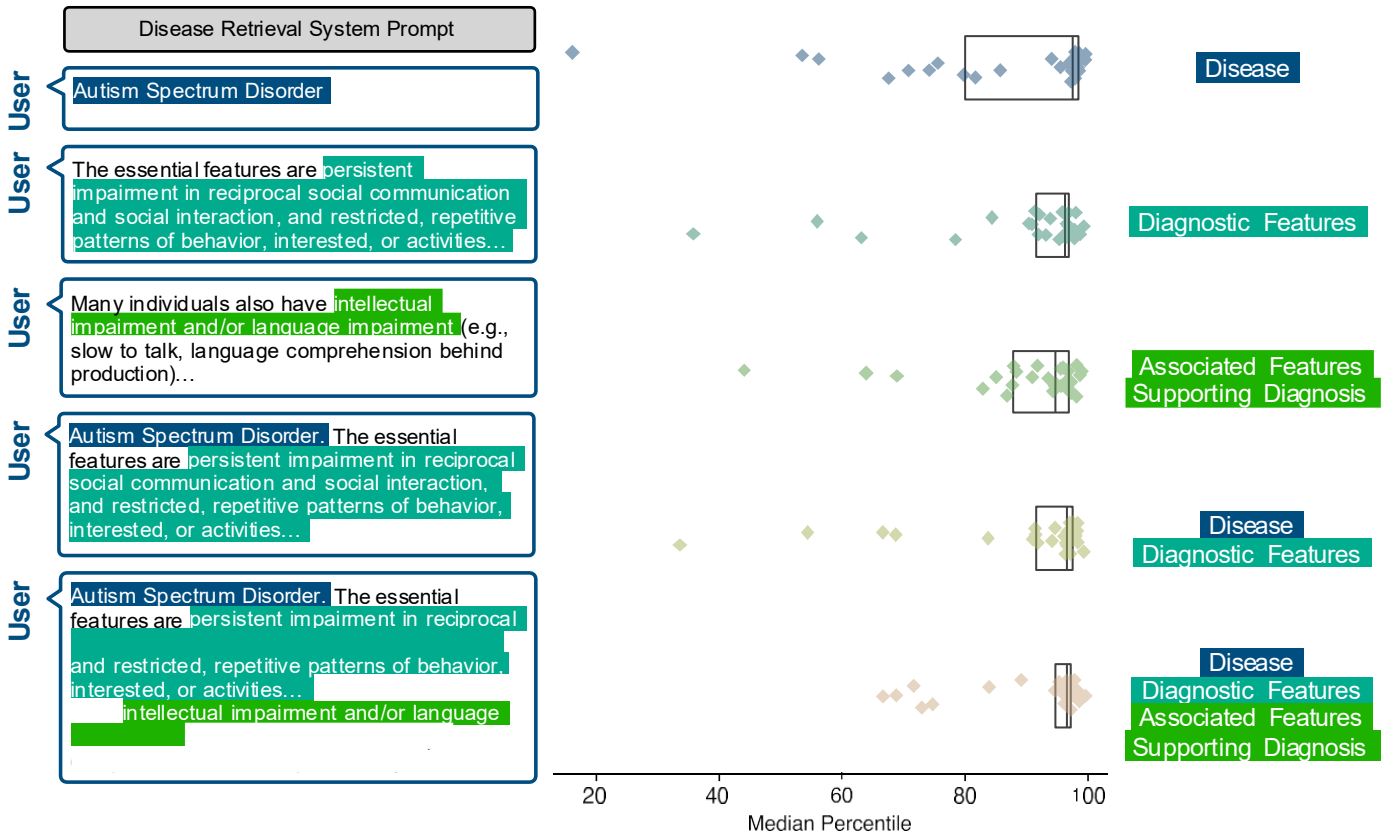
Output: ranked proteins associated to disease

ProCyon can retrieve the genes related to different polygenic disorders

The ground truth definition can affect performance

3 The way we ask a question changes what we measure

Third line of validation: polygenic disorders with complex etiology



Input: clinical phenotypes of 63 mental disorders derived from DSM-5 (without disease name)

Output: ranked proteins associated to disease

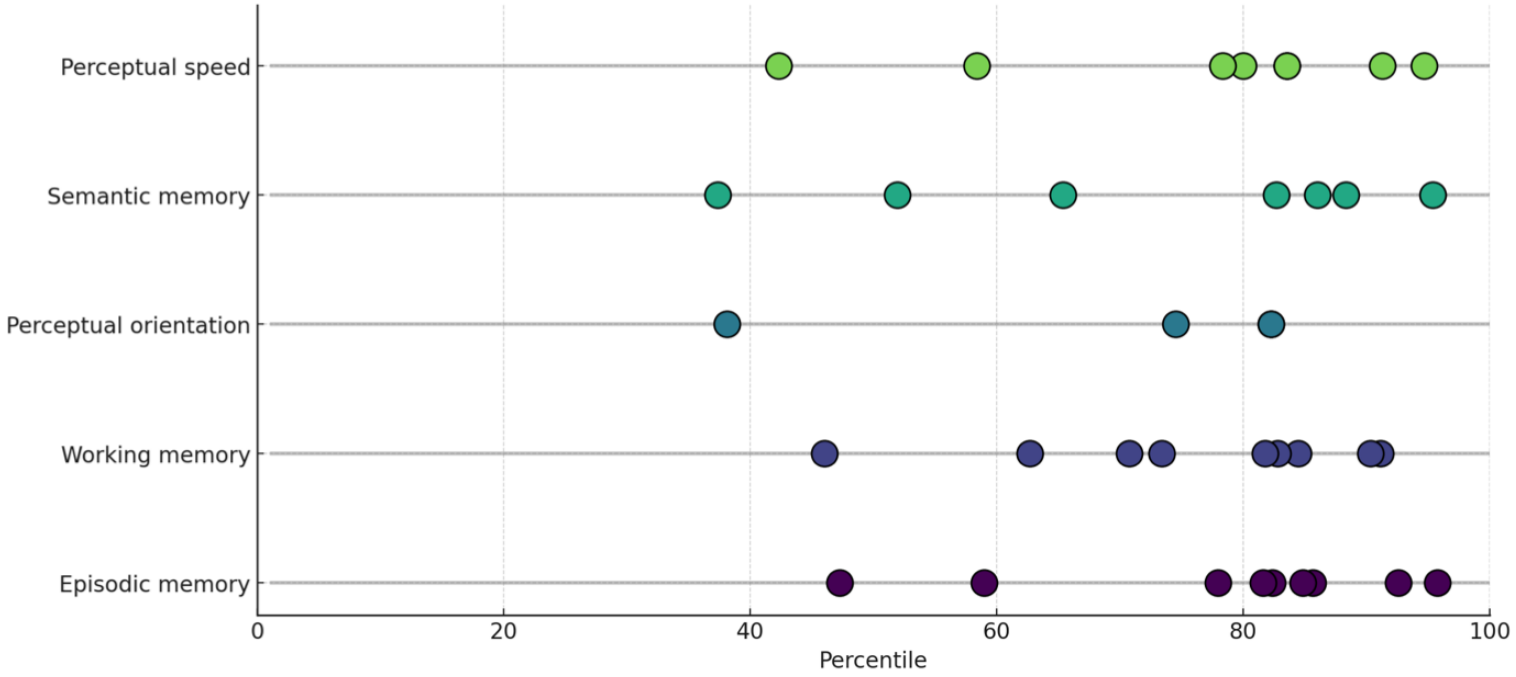
ProCyon can retrieve from clinical phenotypes genes known to be related to mental disorders

Performance is sensitive to input formulation

4

How do we validate knowledge without a definitive reference?

Fourth line of validation: polygenic disorders with complex etiology



Input: cognitive neuroscientist approved descriptions of cognitive domains

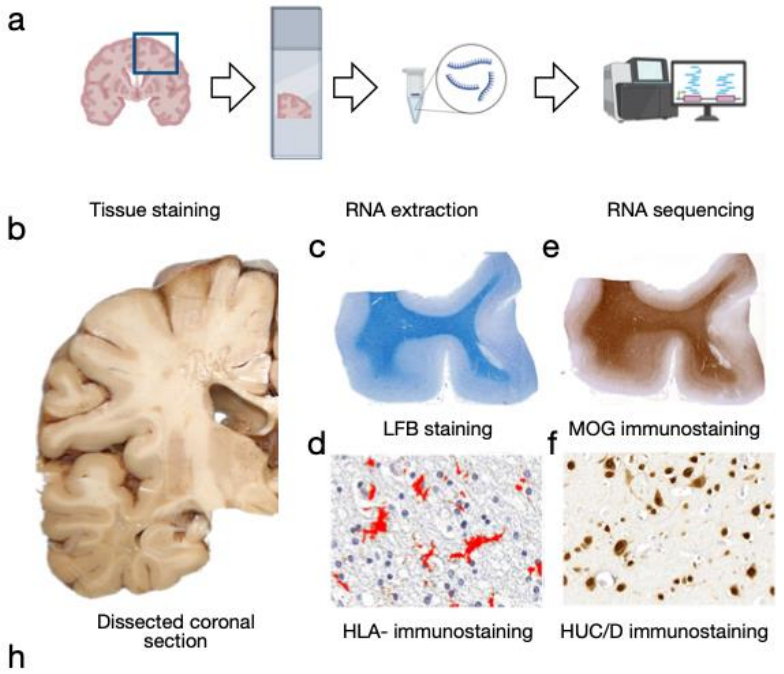
Output: ranked proteins associated to disease

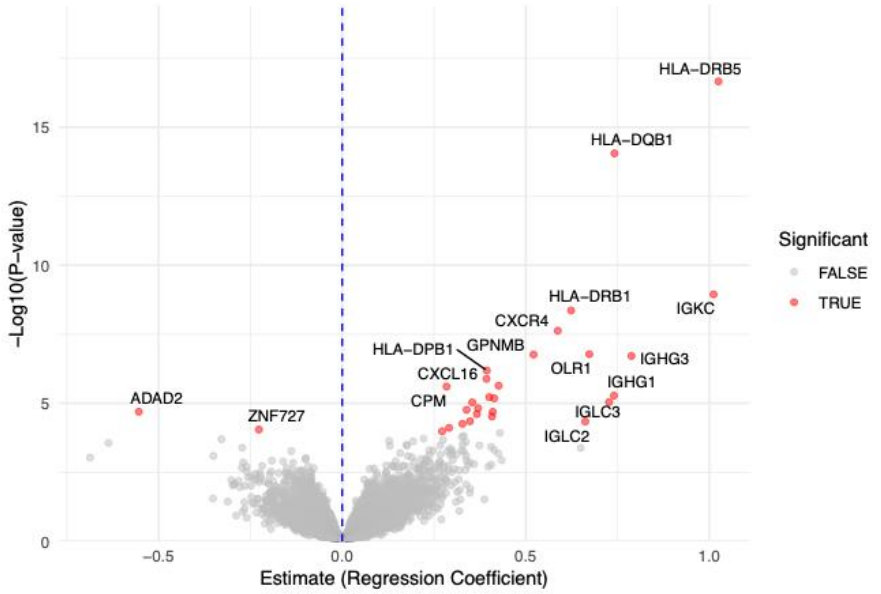
ProCyon can retrieve genes related to domains unrelated to the ones used during training

No reliable ground truth + risk of data leakage through the LLM

5 Testing model predictions against independent disease biology

Fifth line of validation: ground truth from experimentally collected data





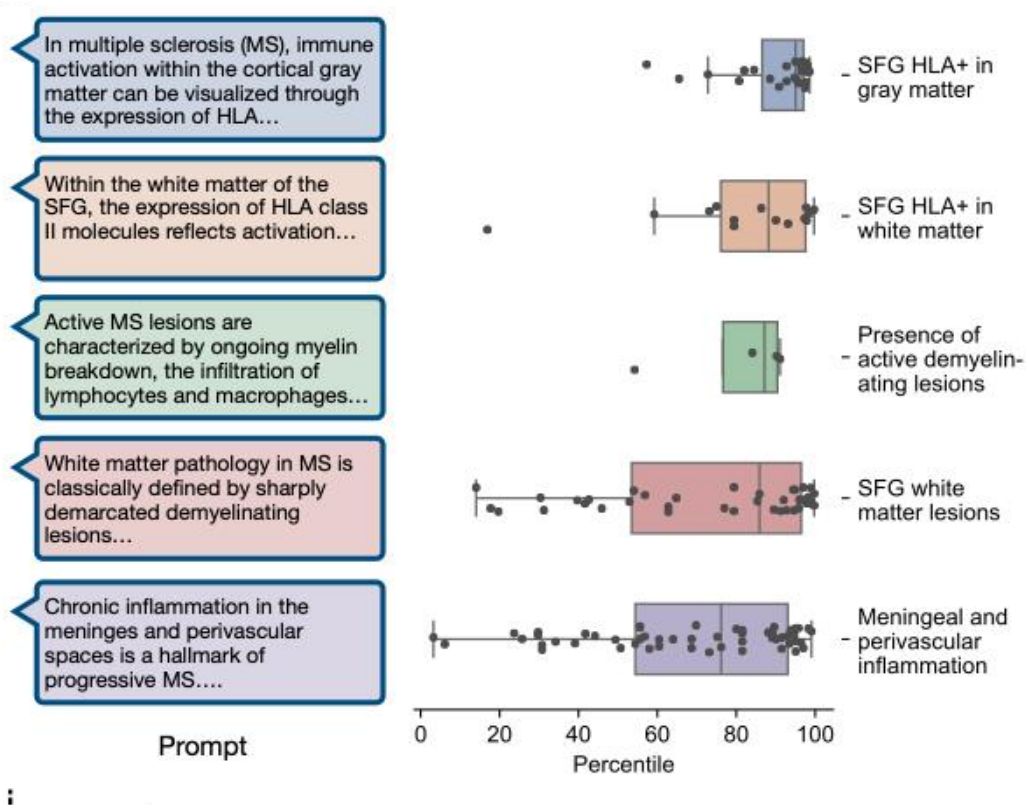
Assembly of a fully independent dataset for leakage-free experimental validation

Data collection from 174 Multiple Sclerosis (MS) brain donors and neuropathological assessment

Differential expression analysis after controlling for biological and technical cofounders

6 Experimental validation in MS using independent brain tissue data

Fifth line of validation: ground truth from experimentally collected data



Input: clinically-validated descriptions of Multiple Sclerosis (MS) tissue pathology

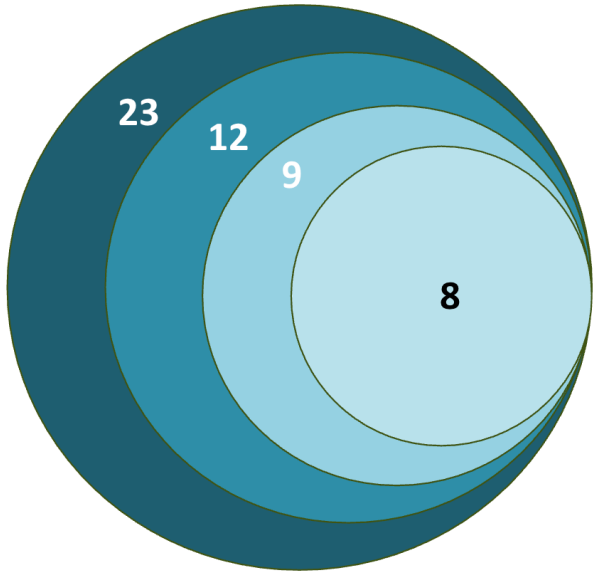
Output: ranked proteins associated to MS hallmarks

ProCyon can retrieve genes related to MS pathology

Novel discovery?

7 Can ProCyon identify novel MS biomarkers?

From validation to biomarker discovery: use ProCyon to identify potential MS markers

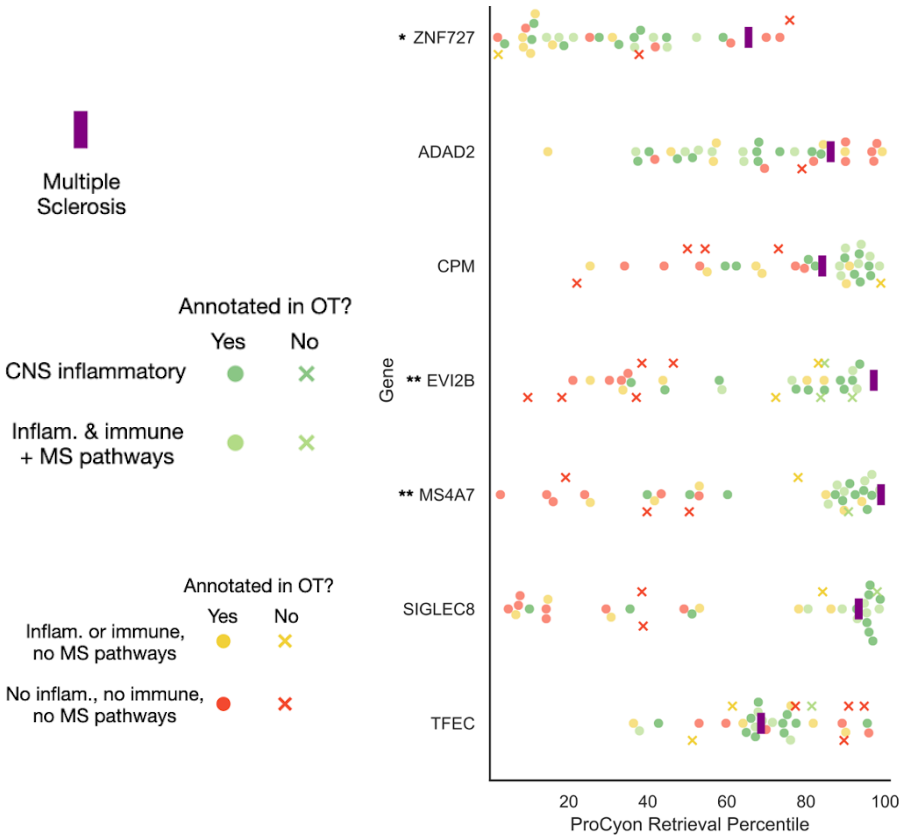


8 HLA-associated genes remained despite no established links to MS or canonical MS pathways

Can we further prioritize candidates and reduce noise?

Disease and gene enrichment analysis using ProCyon

8 ProCyon identifies 3 potential novel MS biomarkers



DISEASE ENRICHMENT – ProCyon association of **target gene** to 4 groups of diseases with different degree of similarity to MS

⟶ *Are the genes ranked higher for MS?*

GENE ENRICHMENT – ProCyon association **of control genes** to MS

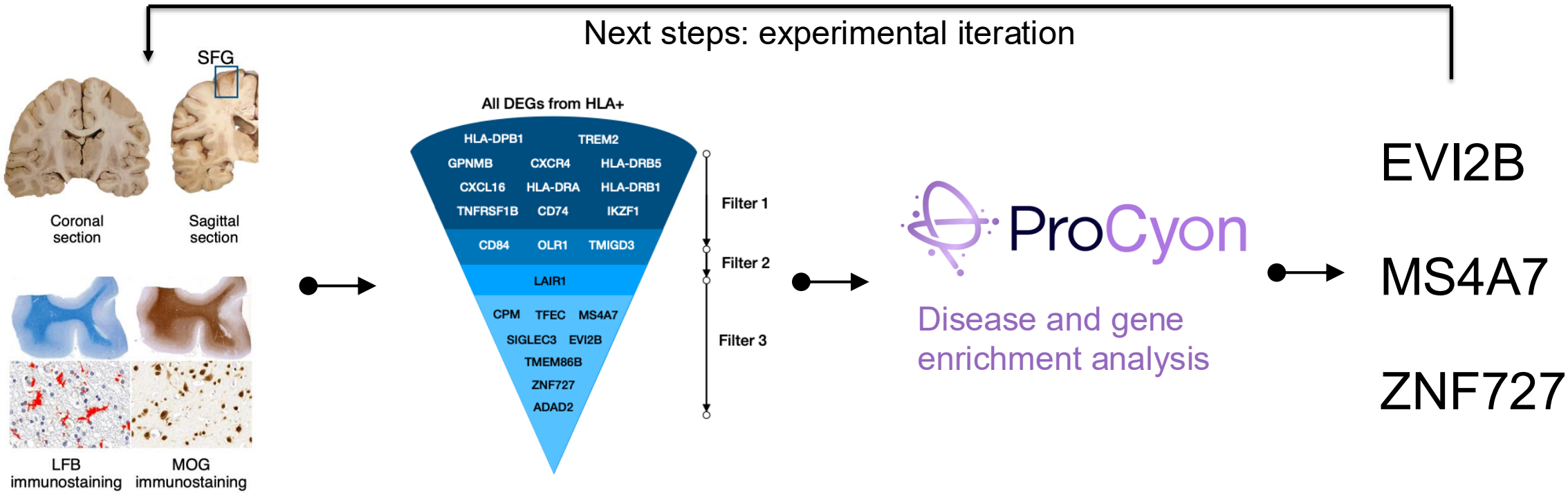
⟶ *Are the target genes ranked higher for MS compared to other control genes?*

Neuron
A Cell Press journal

NEURORESOURCE · Online now, October 10, 2025 · ***Open Access***

Integrated omics reveals disease-associated radial glia-like cells with epigenetically dysregulated interferon response in multiple sclerosis

ProCyon supports target prioritisation in an experimentally derived MS dataset



Experimental data collection

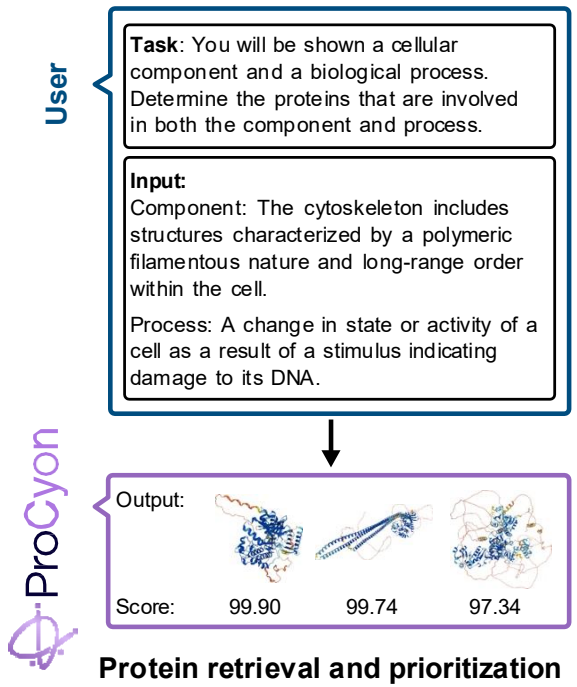
Plausible computationally derived targets

ProCyon target prioritisation

Candidate targets

Tiered validation of ProCyon across increasingly challenging settings

General-purpose foundation model trained on protein sequences, structure and text



Protein retrieval and prioritization

Validation study	Take-home message
Monogenic disorders	Establish basic capability, no generalisation
Polygenic disorders	Performance depends on ground truth definition
Complex mental disorders	Performance is sensitive to input formulation
Cognitive domain	Without ground truth, only plausibility can be tested
Multiple sclerosis	Held-out wet-lab validation provides the strongest test

Progressively harder validation: more indirect biology, less certain ground truth

From benchmark scores to biological confidence

Three lessons from validating a biological foundation model

1 QUESTION THE GROUND TRUTH

What does the benchmark actually measure?

- Recovering known links is not proof of understanding.
- Incomplete annotations change apparent performance.

ProCyon: monogenic and polygenic disorders

2 STRESS-TEST GENERALIZATION

Does the result survive a harder question?

- Vary biological contexts and input formulations.
- Check for memorization, shortcuts, and leakage.

ProCyon: mental disorders and cognition

3 SEEK INDEPENDENT EVIDENCE

Does the prediction agree with real biology?

- Evaluate with orthogonal experimental data.
- Prioritized targets still require follow-up.

ProCyon: MS brain tissue and target ranking

The scientific claim should never be stronger than the evidence used to validate it.

IMPERIAL

UKRI Medical Research Council

HARVARD UNIVERSITY

Special thanks to my main collaborators:



Robert Calef

Massachusetts Institute of Technology



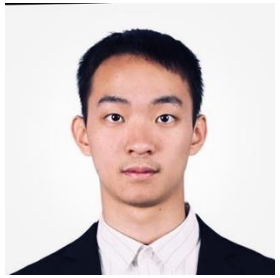
Owen Queen

Stanford University



Dragos Gruia

IMPERIAL



Yepeng Huang

HARVARD UNIVERSITY



Prof. Marinka Zitnik

HARVARD UNIVERSITY



Prof. Adam Hampshire

KING'S College LONDON IMPERIAL



Prof. Richard Nicholas

IMPERIAL

Thanks for your attention!