

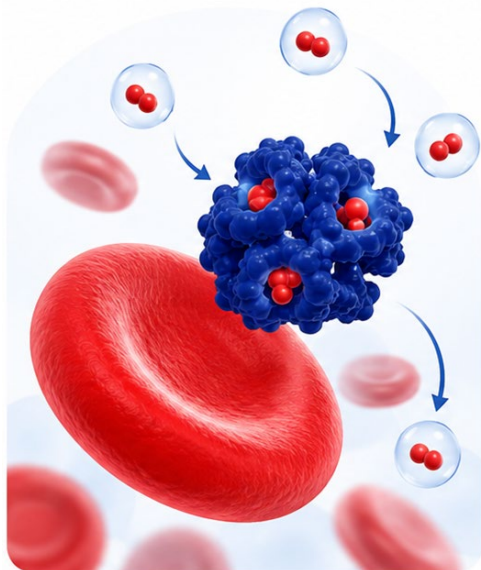
A Binder for Every Protein: Building the indome

Anna M. Díaz Rovira

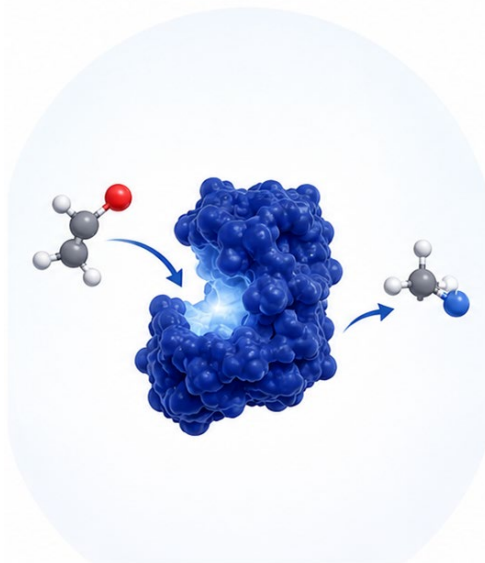
*Postdoctoral researcher in Bruno Correia's lab
École Polytechnique Fédérale de Lausanne (EPFL)*

Proteins are the molecular machines of life

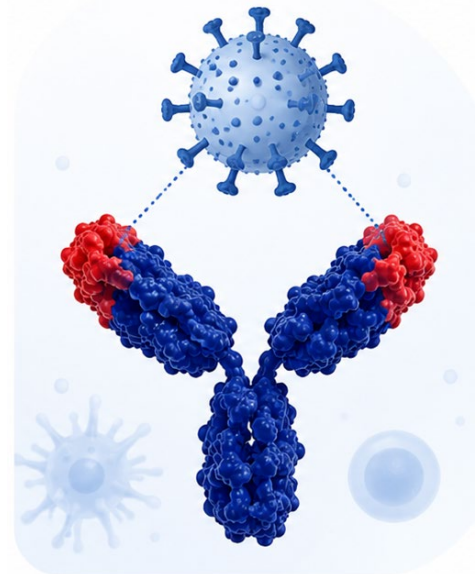
Hemoglobin carrying O₂
to our body



Catalyze chemical reactions
to digest nutrients



Molecular recognition to
prevent infections



Proteins make life possible

Proteins are the molecular machines of life

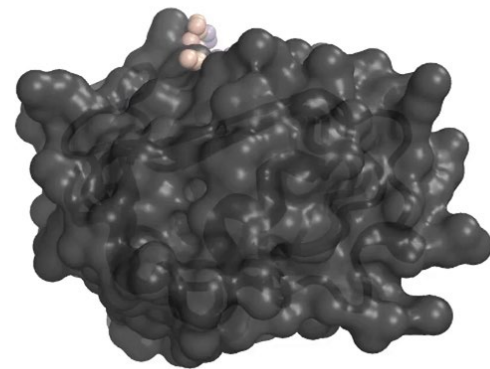
... and now we can design them on a computer using AI algorithms



Design a binder: A protein that recognizes another protein

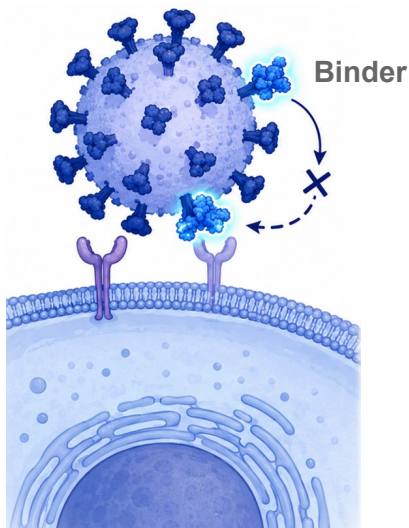


Generated with ChatGPT

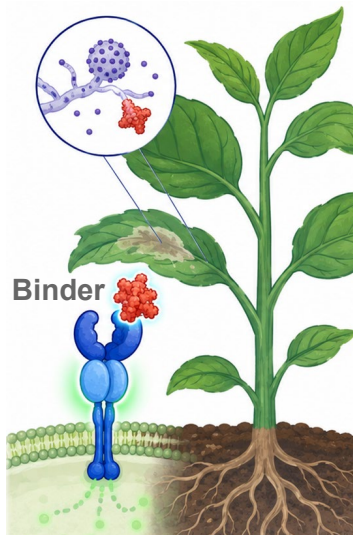


Why would we want proteins that bind proteins?

Block infections

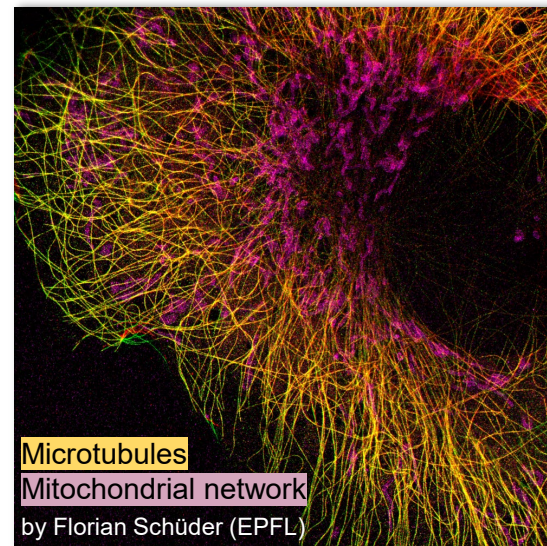


Protect crops



Zhu et al. *Science*. 2026

Image cells



Microtubules
Mitochondrial network
by Florian Schüder (EPFL)

Schnitzbauer et al. *Nat Protoc*. 2017

Because by recognizing a protein, **binders can block** it, **detect** it, or **trigger** a biological response.

But reliable binders are not available for all proteins

Antibodies are essential, but their poor characterization led to a **reproducibility problem**

US\$1 billion per year
of US taxpayers' money

Khan et al. *Nature*. 2023

NEWS | 29 May 2026

nature

Science sleuths uncover more than 100 suspicious images in Thermo Fisher antibody catalogue

Scientists have long worried about the reliability of commercial antibodies, and the latest findings have sparked fresh concerns.

Garisto. *Nature*. 2026

News Feature | Published: 19 May 2015

nature

Reproducibility crisis: Blame it on the antibodies

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

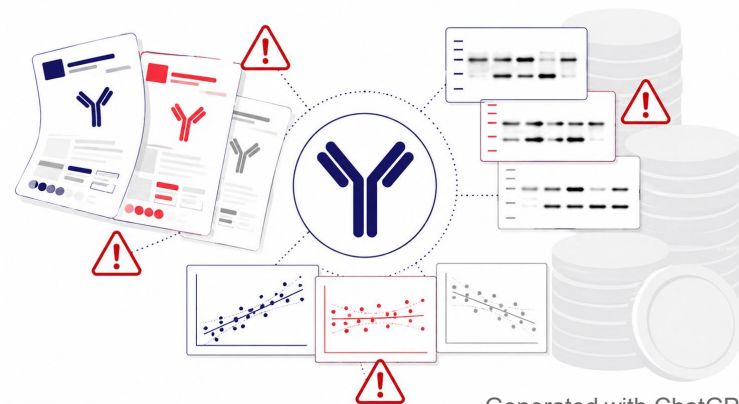
Baker. *Nature*. 2015

But reliable binders are not available for all proteins

Antibodies are essential, but their poor characterization led to a **reproducibility problem**

US\$1 billion per year
of US taxpayers' money

Khan et al. *Nature*. 2023

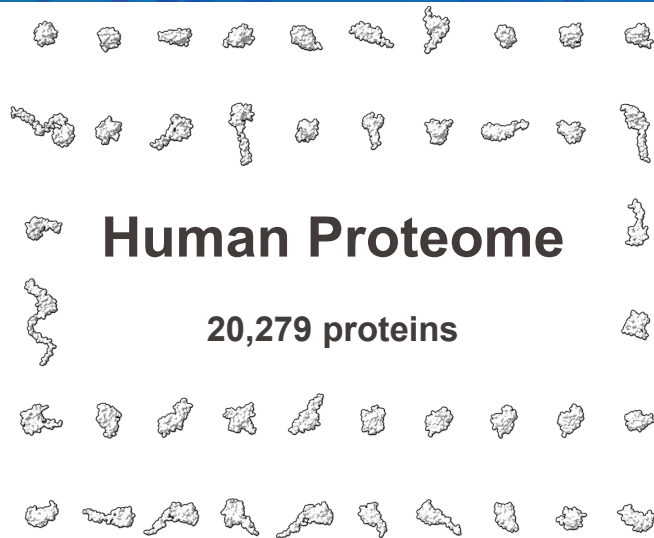


Generated with ChatGPT

We **aim** make a comprehensive **database with well characterized binders**

Could we design binders for every human protein?

AlphaFold Protein Structure Database



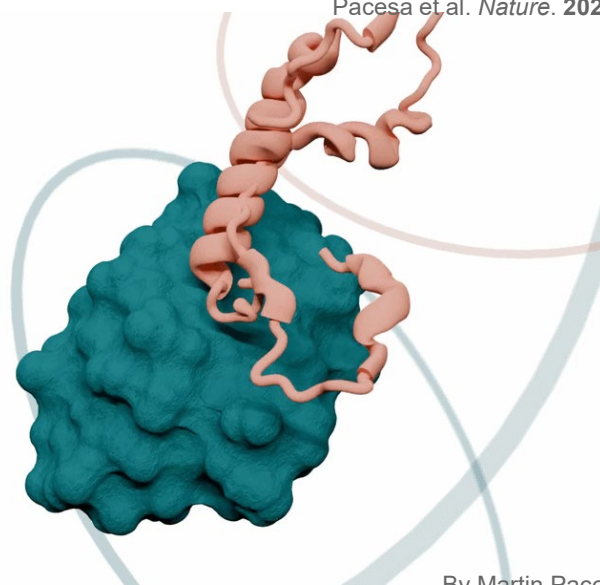
Varadi et al. *Nucleic Acids Res.* 2021

Article | [Open access](#) | Published: 27 August 2025

One-shot design of functional protein binders with BindCraft

nature

Pacesa et al. *Nature.* 2025

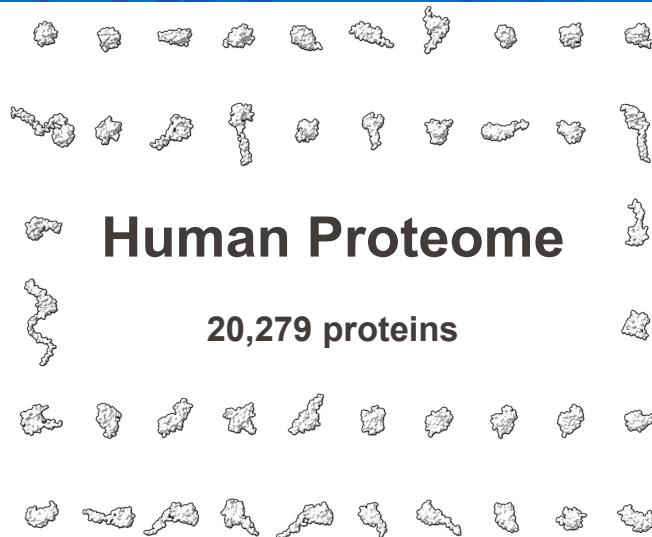


By Martin Pacesa

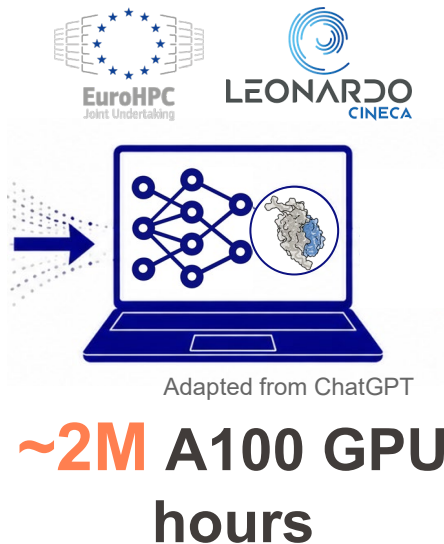
Experimental success rates between 10 –100%

EuroHPC enabled proteome-scale binder design

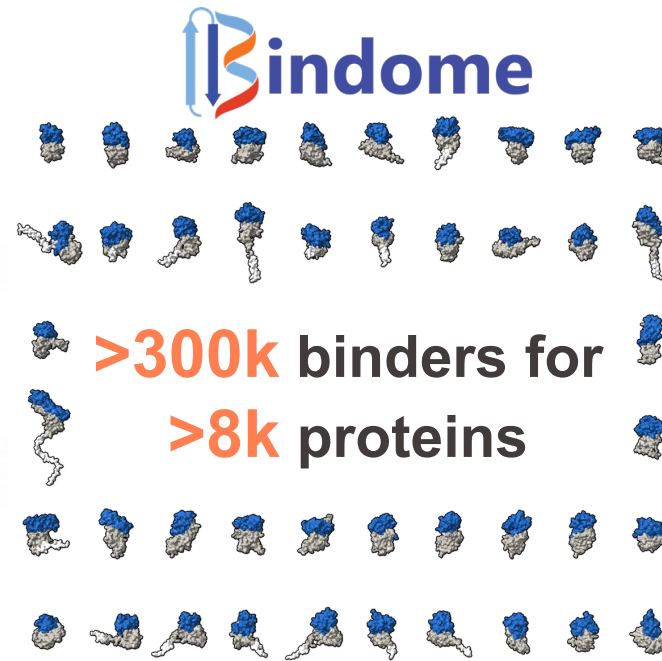
AlphaFold Protein Structure Database



Varadi et al. *Nucleic Acids Res.* 2021



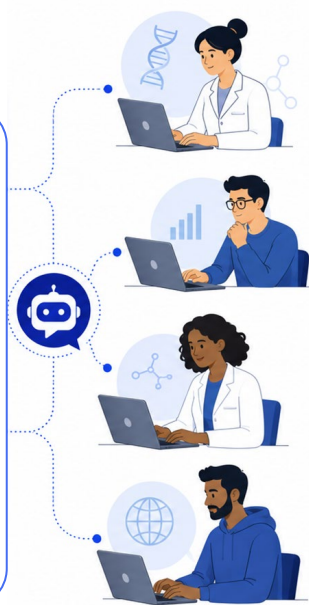
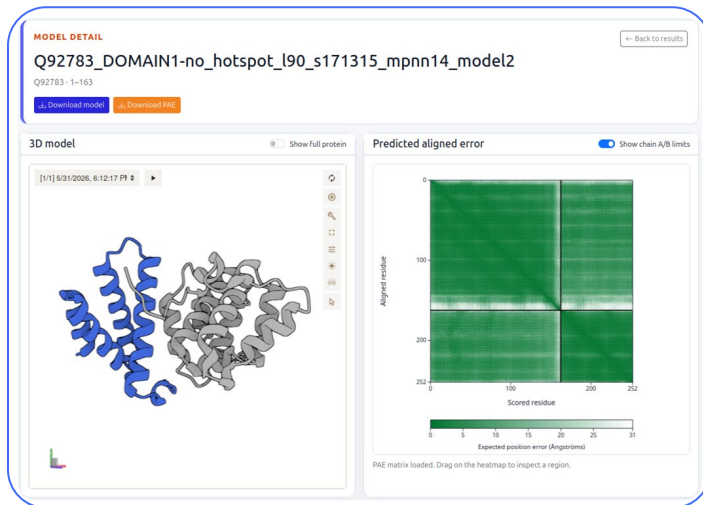
Pacesa et al. *Nature.* 2025



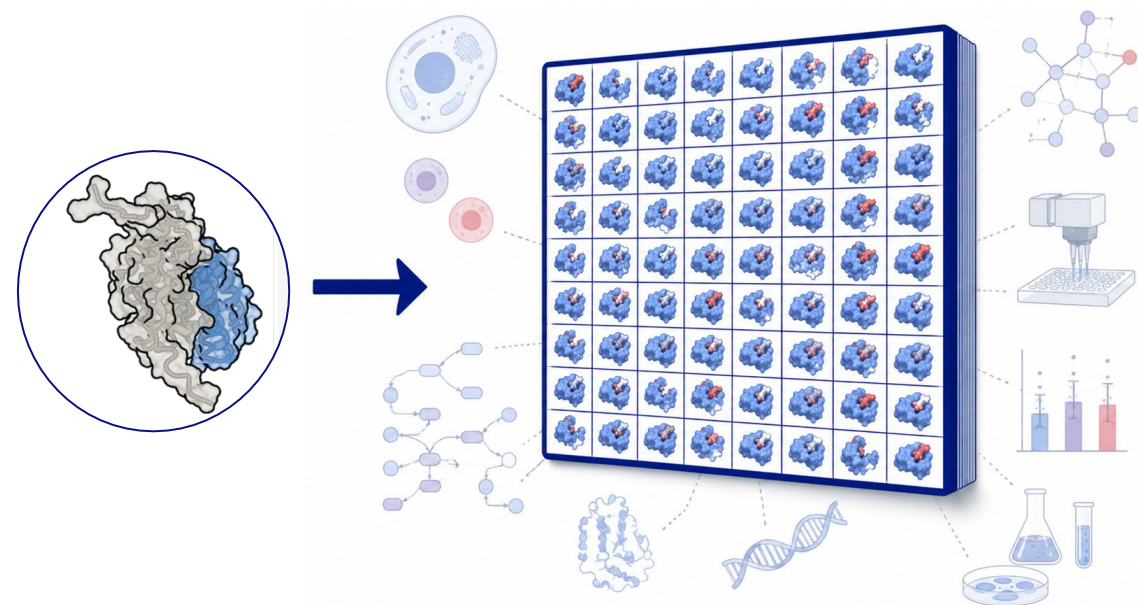
Wenckstern & Diaz-Rovira et al. *bioRxiv.* 2026

2M GPU hours are only useful if people can access the results

Bindome
+40k queries so far



From studying proteins one at a time...
to asking questions across the proteome.



Discover protein
function

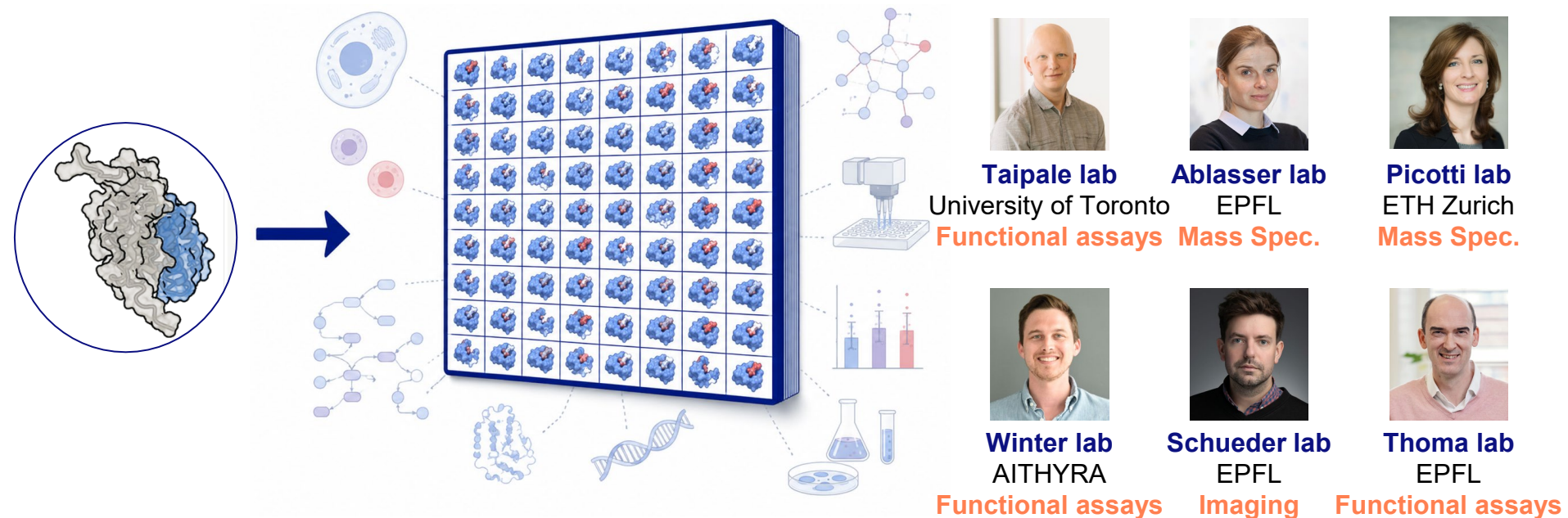
Map protein interactions

Build new AI tools

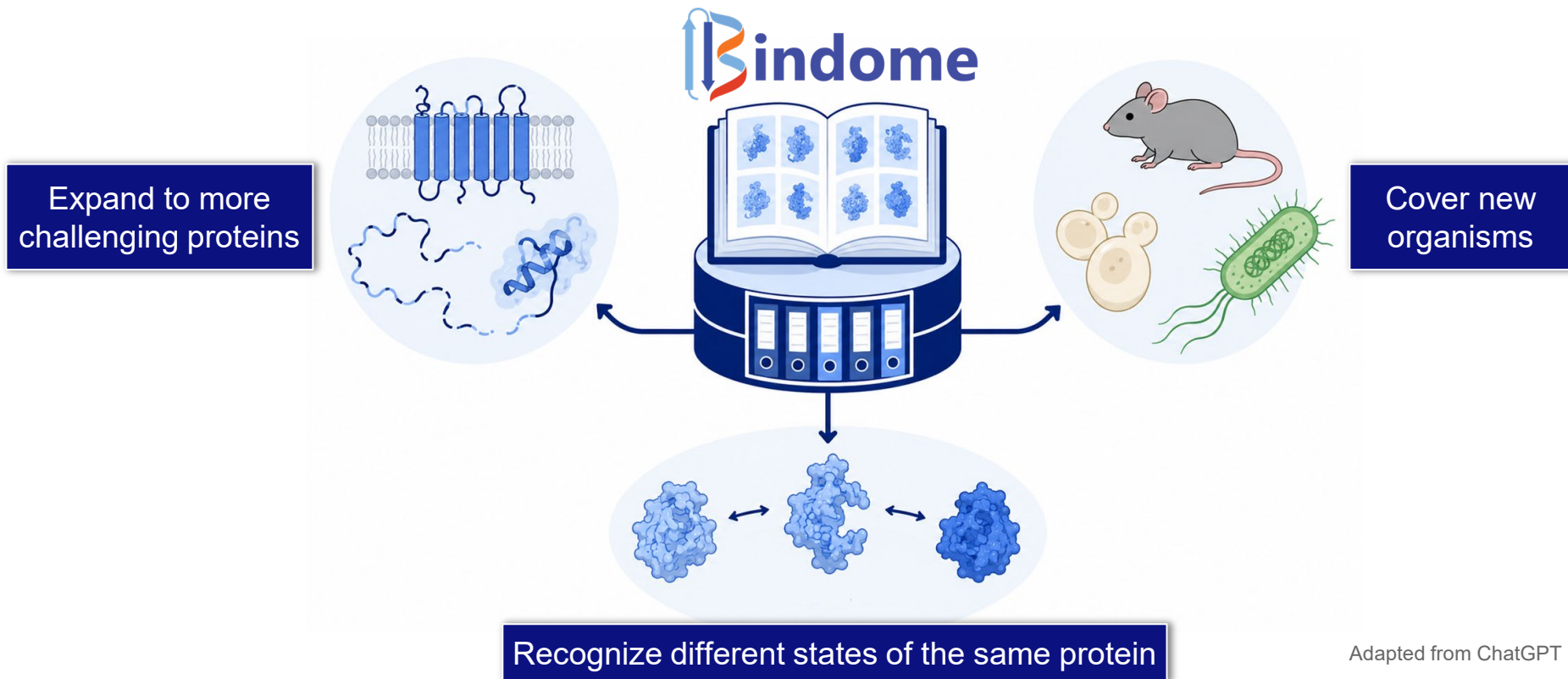


Unveil new therapeutic
targets, faster

From studying proteins one at a time... to asking questions across the proteome.

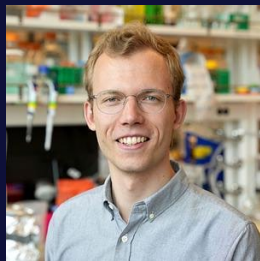


What's next?



Thanks!

EPFL  indome team:



Julius
Wenckstern



Anna M.
Díaz Rovira



Julia Aiko Kuhn



Arvid Ban



Rahma
Hamdani



Bruno Correia

<https://bindome.epfl.ch>



EuroHPC
User Days

Dublin Ireland
22→24
September 2026



PRedictive Exploration of Correlation In Specific Interactions and Optimization Networks

Optimization of the PRECISION workflow for Computational DrugDesign

<https://doi.org/10.1016/j.procs.2026.08.026>

Giorgia Frumenzio
CINECA- HPC SPECIALIST
g.frumenzio@cineca.it



CINECA



Computational Drug Discovery Workflow

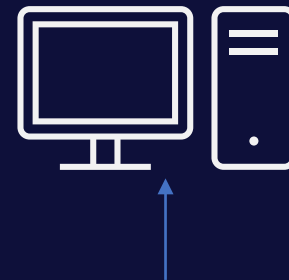
1. Molecular Docking

2. Molecular dynamics

3. Processing MD data

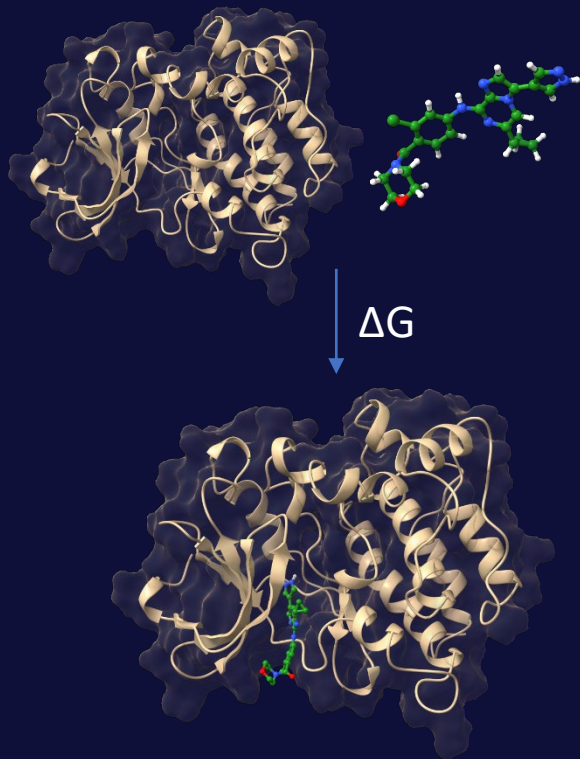


Around 16 000 simulations from LIGATE project



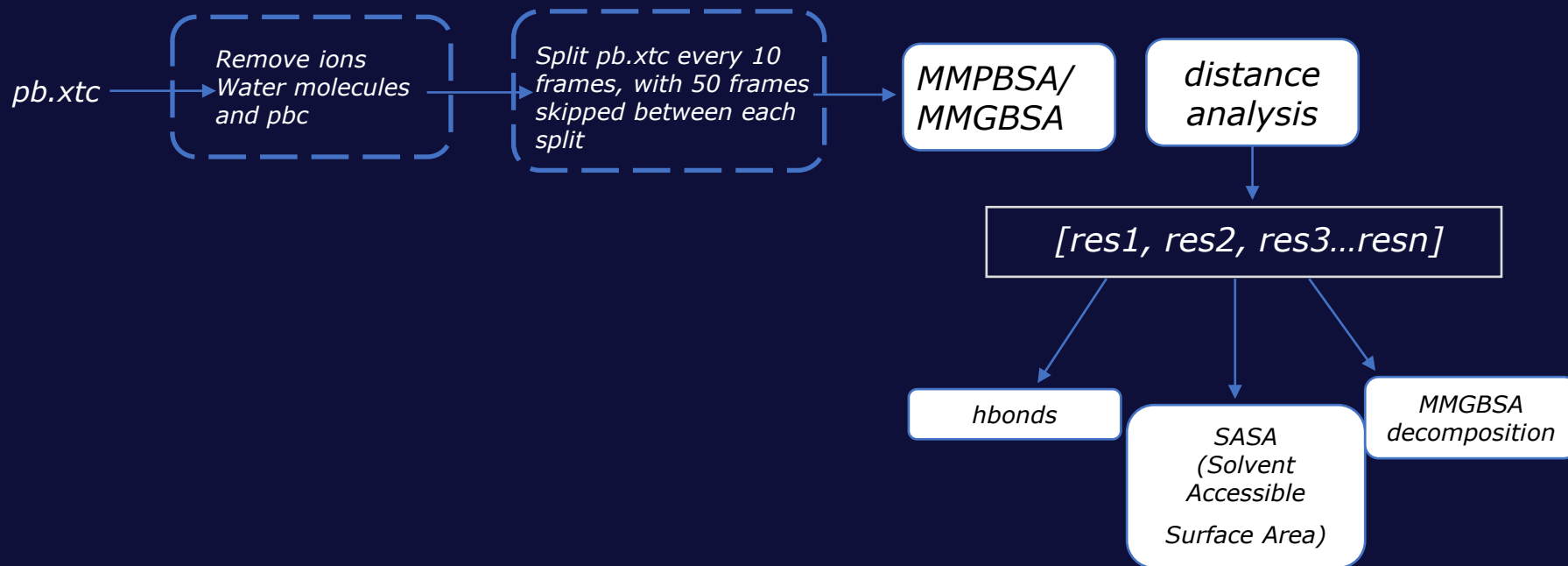
Precision workflow

Purpose of the Analyses

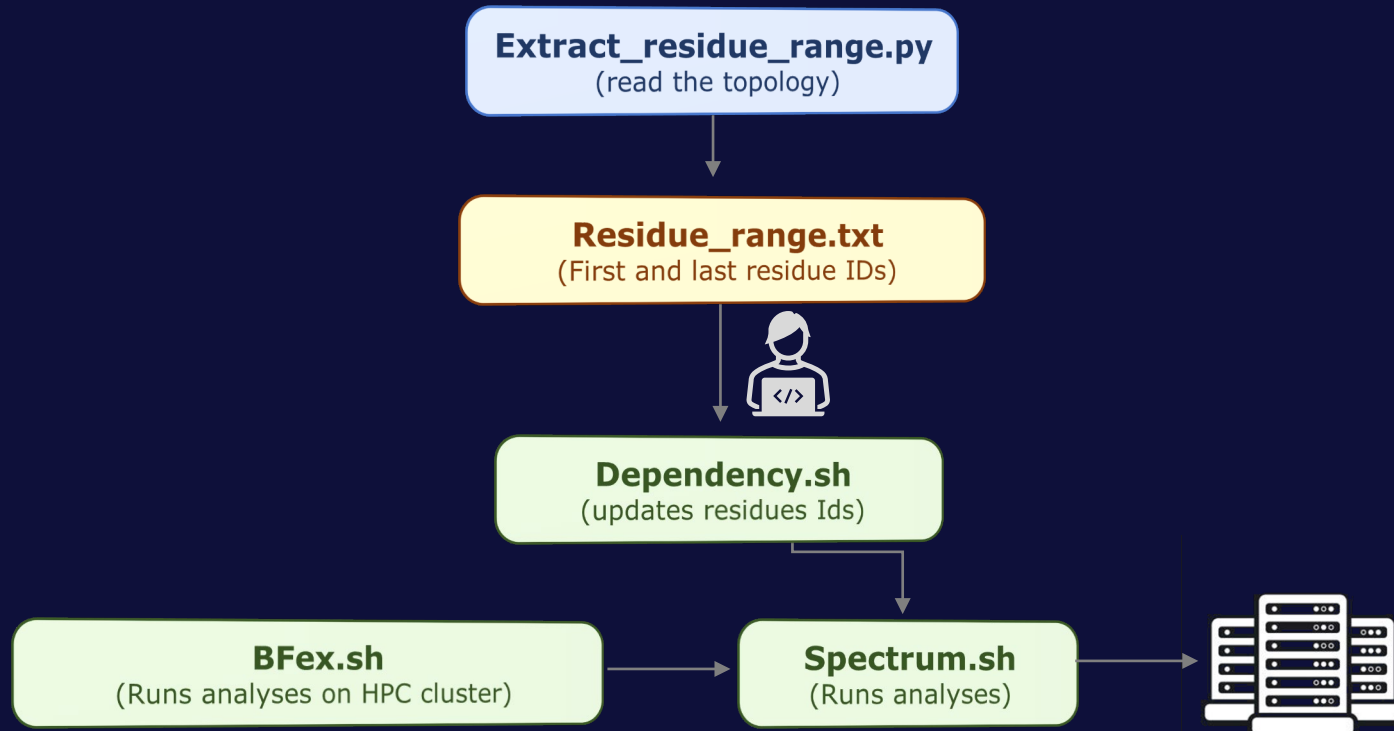


1. Estimate protein-ligand binding affinity using MMGBSA and MMPBSA analyses
2. Identify the key molecular determinants of binding through per-residue MM-GBSA decomposition, hydrogen-bond analysis, and distance-based interaction profiling.
3. Characterize interaction stability and persistence over time by integrating contact occupancy, hydrogen-bond lifetime, SASA, and structural descriptors.

Automated Post-Processing of Molecular Dynamics Simulations



Original Workflow Design and Execution Strategy



Computational Bottlenecks



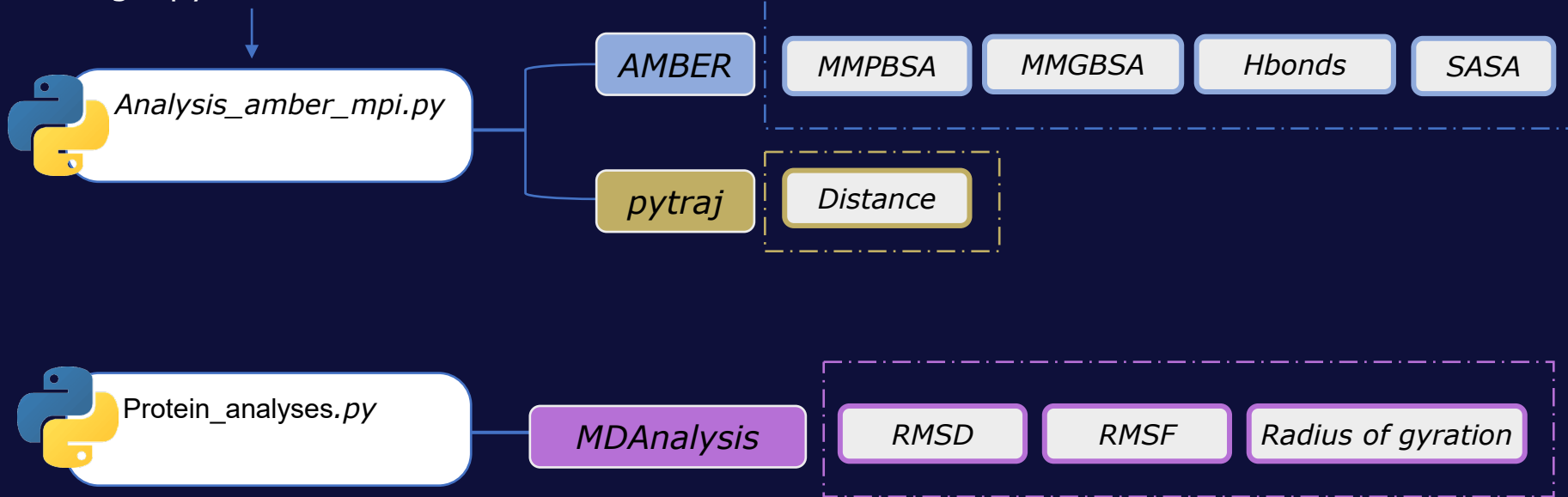
Time consuming MMPBSA and MMGBSA decomposition analyses are computationally expensive, especially for large protein-ligand systems and long trajectories.



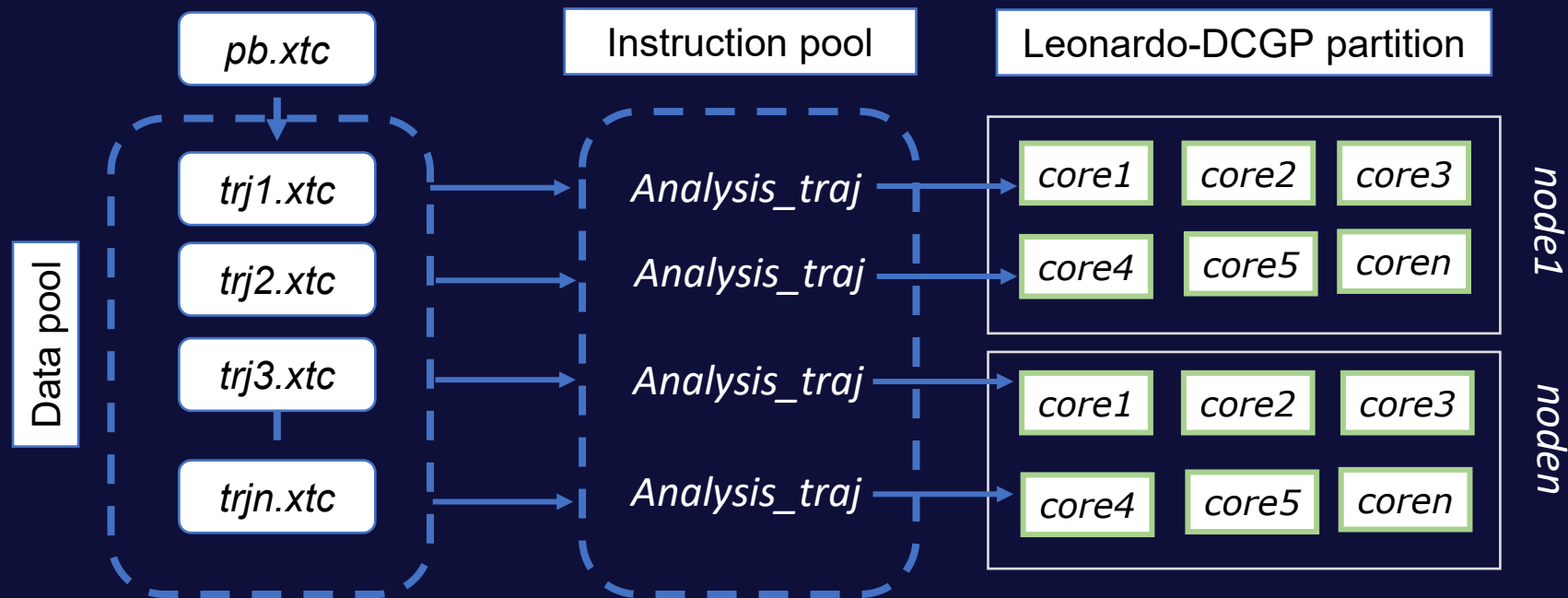
Serial execution Most of analyses lack native parallel implementation. Only for MMPBSA and MMGBSA a MPI implementation exists, allowing a parallelization only over simulation frames, poorly efficient for long simulation.

Workflow Refactoring and Python-Based Implementation

All bash script condensed in a
single python workflow



MPI-Based Parallelization Strategy



MPI Workflow Execution Model

```
for batch_start in range(0, n_tasks, size):  
    idx = batch_start + rank  
    if idx >= n_tasks:  
        continue  
    block = frames_blocks[idx]  
    xtc_out = process_traj(block, U, paths)  
    local_trj_paths.append(xtc_out)
```

Numpy array

[[1..10], [51..60], [111..120]...[171..n]]

1

2

3

MPI
processes

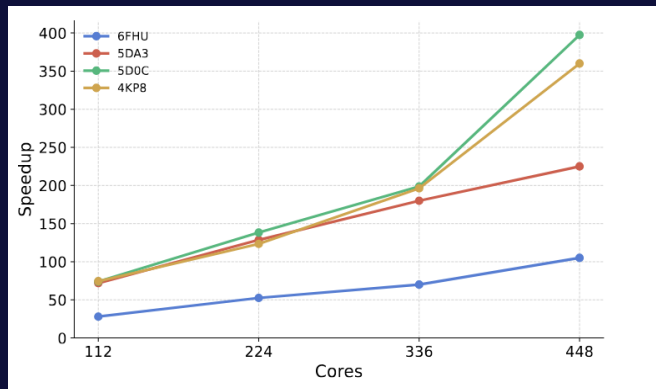
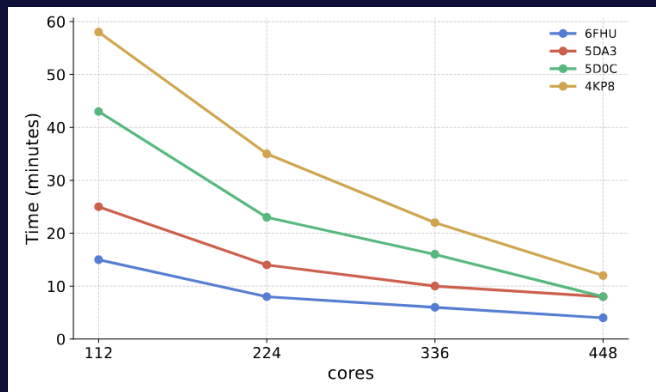
Block_1_10.xtc Block_51_60.xtc Block_111_120.xtc

[Block_1_10.xtc, Block_51_60.xtc, blockn..blockn+10]

```
for xtc in local_trj_paths:  
    analysis_function(xtc, first_prot_res, end_prot_res, first_lig_res, paths, ref_file)
```

Scaling Performance on Leonardo DCGP

PDB ID	Num. atoms	Serial time (hours)
6FHU	836	7
5DA3	4319	30
5D0C	6364	53
4KP8	8172	72



Key Takeaways and Future Directions

MPI-based parallelization achieved a **74x speedup** for the largest system tested, using a single Leonardo DCGP node

Accelerates traditionally slow analyses such as MMPBSA and MMGBSA decomposition

Generates large-scale simulation datasets for future machine learning applications

Thank you

We acknowledge EPICURE, a EuroHPC Joint Undertaking initiative for supporting this project on Leonardo DCGP through the EuroHPC development call

g.frumenzio@cineca.it-CINECA

a.emerson@cineca.it-CINECA

akashdeep.biswas@dompe.com-DOMPE

CINECA



EPICURE
Unlocking European-level HPC Support



Dompe



EuroHPC
User Days

Dublin Ireland
22→24
September 2026



Funded by
the European Union



EuroHPC
Joint Undertaking



An Roinn Breitheamhais agus Ardoidreachais,
Eighe, Neilataichta agus Iolaithe
Department of Further and Higher Education,
Research, Innovation and Science

ICHEC
Irish Centre for High-End Computing



ÉIRE2026.EU
IRELAND2026.EU

OPADE on LUMI-G

Multimodal modelling for antidepressant-response prediction

Alexandru Floares, Lampros Androutsos, Adrian Zety, Sheraz Naseer, Maaz Mohsin

Artificial Intelligence Expert (AIE), Romania · EHPC-AIF-2025SC04-075 · OPADE (GA 101095436)



OPADE: predicting antidepressant response

Clinical data and omics measured before treatment

350

patients in the target cohort

Four recruiting countries

- Outcome: $\geq 50\%$ HDRS reduction at week 8 (T0→T1)
- Inputs: clinical data plus six omics families
- Labelled cohorts vary by modality and recruiting centre
- Challenge: many features and uneven centre coverage
- Goal: accurate predictions with interpretable outputs



LUMI supports independent modelling at scale

Combined project activity through 3 September 2026

3,860

jobs submitted

two active users

2,653

GPU-hours billed

project accounting total

384

jobs over 1 hour

97 exceeded 4 hours

15.6 h

longest job

elapsed walltime

Concurrent experiments across modalities, sites and model configurations

Why LUMI-G fits

Independent fits can run concurrently across model families and data subsets. GPU-capable models use MI250X accelerators; CPU-only methods require a different resource mix.



Model diversity drives the compute requirements

OPADE modelling and public-data benchmarks on LUMI-G

WORKLOAD	ENGINE	SCOPE	COMPUTE	STATUS
Tabular foundation models	TabPFN · TabICL · TabFM	tabular data	transformer inference	active
Microbiome foundation model	MGM · GPT-2, 9M	15 cohorts	pretrained embeddings	benchmarked
AutoML ensembles	AutoGluon extreme	15+ families	GPU-capable learners	active
Multimodal fusion sweep	pairwise + tri-modal	71 models	varies by model	validating

Match resources to the model

The mixed GPU research campaign comprised 1,251 jobs. Accelerator use depends on the learner and preset; verify execution before launching a large job array.



OPADE results: preliminary single-modality models

Preliminary results from earlier data snapshots · modality-specific labelled cohorts

LAYER	AUC	EVIDENCE	NEXT TEST	LIMITATION
Genomic / immunoprofiling	0.854 / 0.855	internal	centre-held-out	sets differ
Clinical	0.800	internal	site stability	preliminary
Methylation	0.773	internal	nested selection	preliminary
Microbiome	0.750	historical	larger cohort	two centres

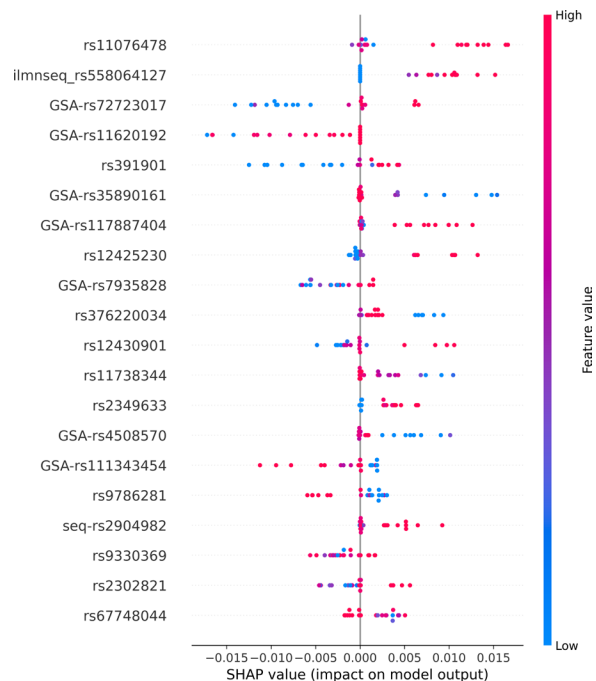
Next: test generalisation

These results used different eligible patient subsets. Updated data require re-evaluation before comparing modalities.



SHAP explains the fitted genomic model

An exploratory view of feature contributions in the current OPADE model



SHAP summary — genomic model

PLOT ELEMENT	READ AS	SCOPE	INTERPRETATION
Feature order	rank	model	Mean absolute contribution
Horizontal axis	SHAP	output	Sign and size of the contribution
Point colour	value	feature	Red = high; blue = low
Each point	sample	plotted	One sample's feature contribution
Next check	stability	pending	Repeat across folds and centres

Attributions describe this fitted model. Robust biological interpretation requires reproducible feature effects and independent validation.

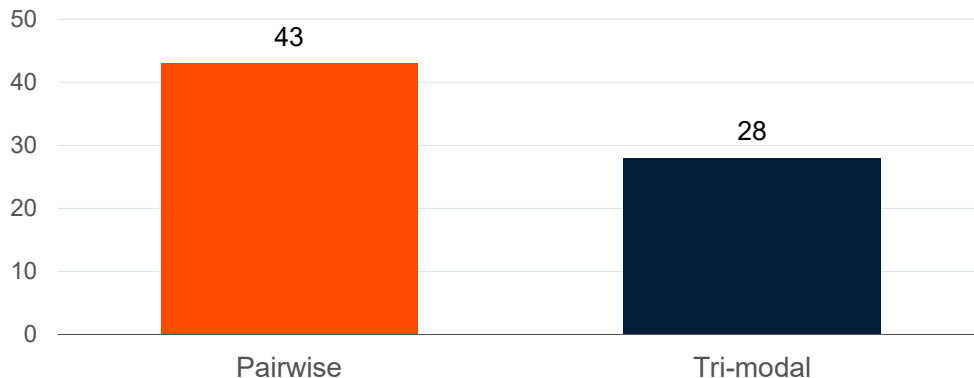


OPADE explored 71 multimodal models

Exploratory comparisons across six modality families

71

combined models



Initial comparisons: several configurations exceeded their best component baseline.

NEXT STEPS FOR OPADE

Match patients → Lock splits → Test unseen centres



Public benchmarks informed OPADE choices

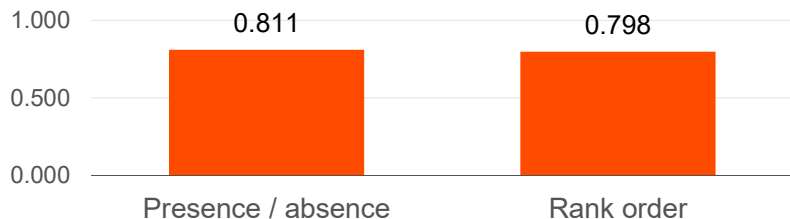
Public microbiome and cancer cohorts · separate from OPADE patient data

MICROBIOME

FM advantage depended on sample size

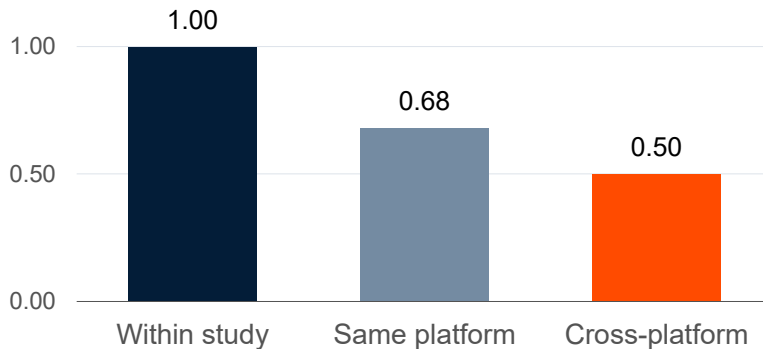
5/class: +0.048 AUC → 100/class: ≈ 0

Encoding benchmark (AUC)



SERUM miRNA

Transport benchmark (approximate AUC)



NEXT STEPS FOR OPADE

Test FM benefit and encoding.

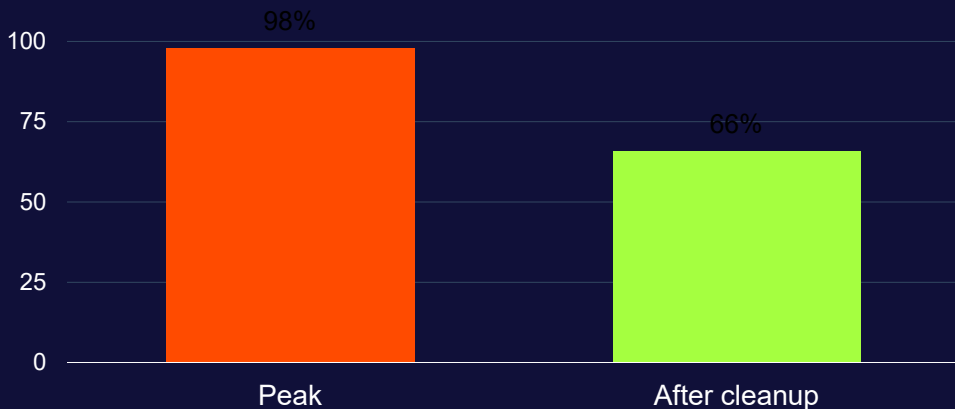
Test assay and centre effects.



File quotas constrained the workflow

Scratch file count reached 98% while byte usage remained low

Scratch file quota used



≈700k

entries removed

90%

flash file quota still used

FIX APPLIED

SquashFS packaging · Output verification · File and byte quota monitoring



Three practical lessons for GPU workflows on AMD

Recommendations from the OPADE LUMI-G campaign

1

Match model to partition

A GPU queue does not guarantee GPU execution. Route CPU-only learners to CPU nodes and choose GPU-capable presets intentionally.

2

Verify accelerator execution

Check the allocated devices and run a short GPU training step. Stop if the expected accelerator is unavailable, rather than continuing silently on CPU.

3

Control the software stack

Verify that the intended ROCm PyTorch libraries load before the package overlay. Test each learner: CatBoost GPU training was unavailable in our environment.

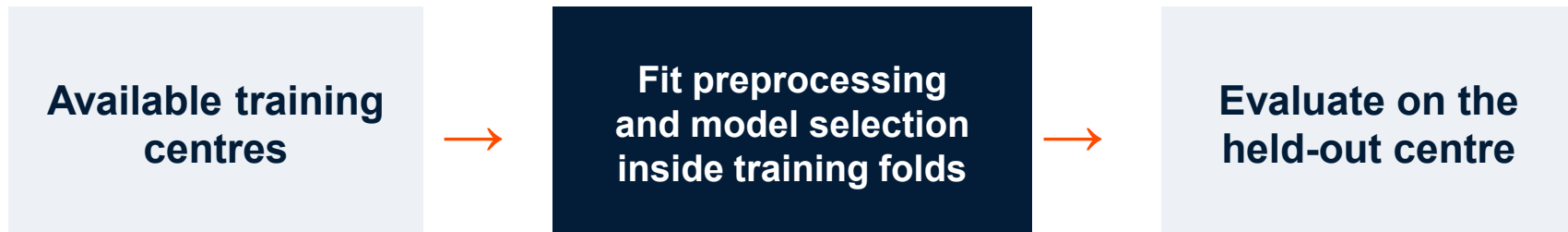
Preflight before the array

Before launching the full array, check device visibility, a short training run, saved-model reload and output paths.



Next steps: test generalisation

Planned validation by 30 November 2026 · one centre-held-out round shown



Repeat across feasible centres and outer splits.

Compare fusion fairly

Same patients and locked splits

Report performance

AUC, uncertainty and calibration

Check interpretation

Feature stability across folds

After primary validation: synthetic-data pilot, subject to utility and privacy assessment.



LUMI enabled OPADE's multimodal modelling campaign

We acknowledge EuroHPC JU for awarding the project ID EHPC-AIF-2025SC04-075 access to LUMI-G hosted by CSC, Finland.

Three takeaways

- Initial OPADE models and 71 multimodal combinations
- Public cohorts informed method and platform choices
- File quotas and GPU execution need explicit checks

The next milestone

Establish performance on unseen centres, with uncertainty and calibration, before drawing clinical conclusions.

OPADE consortium

Clinical sites: UNISI (Italy) · IDIBGI (Spain) · Medipol (Türkiye) · FUS (Colombia) · EBRIS (coordination)

Omics partners: CEINGE · Protobios · Perseus Biomics · EURECAT · EBRIS · IDIBGI

Funding

OPADE (GA 101095436) has received funding from the European Union's Horizon Europe research and innovation programme. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Backup: allocation and job accounting

Snapshot: 7 September 2026 · allocation consumption, not device utilisation

84,000

GPU-hours awarded

21,000 full-node-hours
equivalent allocation

2,653

GPU-hours billed

3.2% of the award
≈663 full-node-hours

Three phases

Data preparation

Feb–Jun

*Late omics data shifted
validation*

Model campaign

Jul–Aug

Validation extension

Sep–Nov

Deadline

30 Nov

Job status at the snapshot

3,860 submitted; 2,922 completed; 542 failed or timed out; 395 cancelled; 1 pending. These are scheduler states. A completed job is not a count of validated models.

Backup: week-8 labelled cohort availability

Assay $\cap$ response_t1 label · current counts supplied 7 September 2026

MODALITY	N	RESPONDERS	CENTRES	N BY CENTRE
Clinical reference	300	101	CO · IT · ES · TR	63 · 129 · 58 · 50
CANTAB	204	73	CO · IT · ES · TR	62 · 91 · 1 · 50
Genomic	216	67	CO · IT · ES	63 · 97 · 56
Microbiome	123	54	IT · ES	76 · 47
Immunoprofiling	95	45	IT · ES	66 · 29
miRNA	80	31	CO · IT	23 · 57
Cytokines	89	37	CO · IT · ES · TR	20 · 22 · 29 · 18
Nucleotides	89	37	CO · IT · ES · TR	20 · 22 · 29 · 18
Lipoprotein	88	36	CO · IT · ES · TR	20 · 21 · 29 · 18
Acylcarnitines	67	27	CO · ES · TR	20 · 29 · 18

Current availability; earlier results used other cohorts

Counts are not the sample sizes behind slide 5. Fusion needs patients shared across assays. CO: Colombia · IT: Italy · ES: Spain · TR: Turkey.