ABCURED
Connecting the genetic dots with AI

LABORATORY OF PHARMACOLOGY

HPC in biomarker and drug target discovery

Nikolas Dovrolis, MSc, PhD

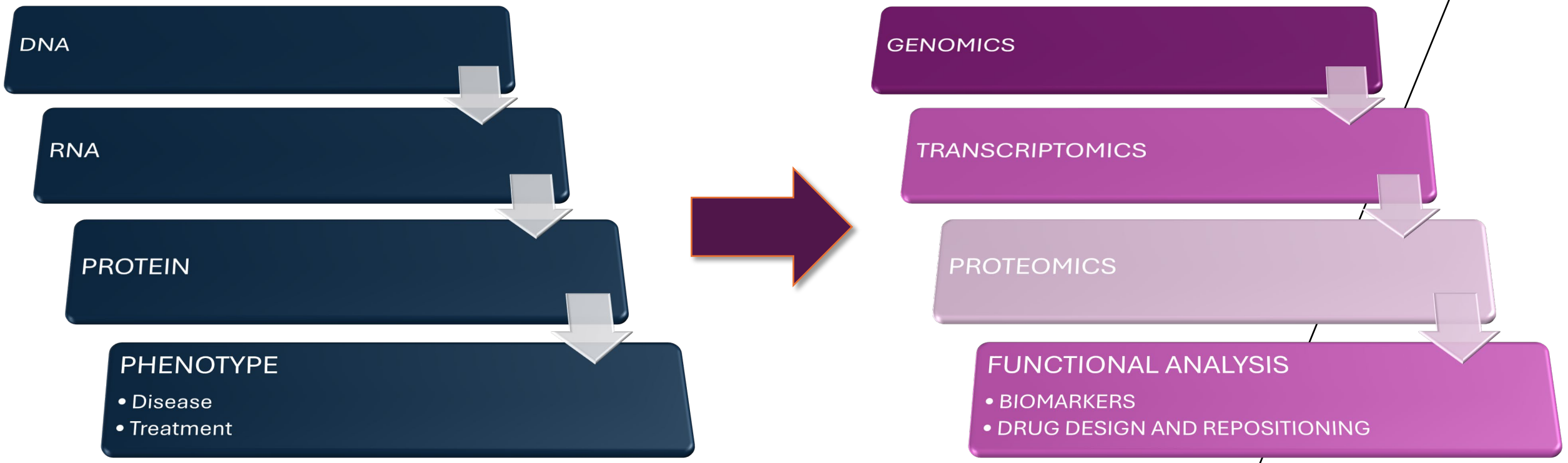
-Omics

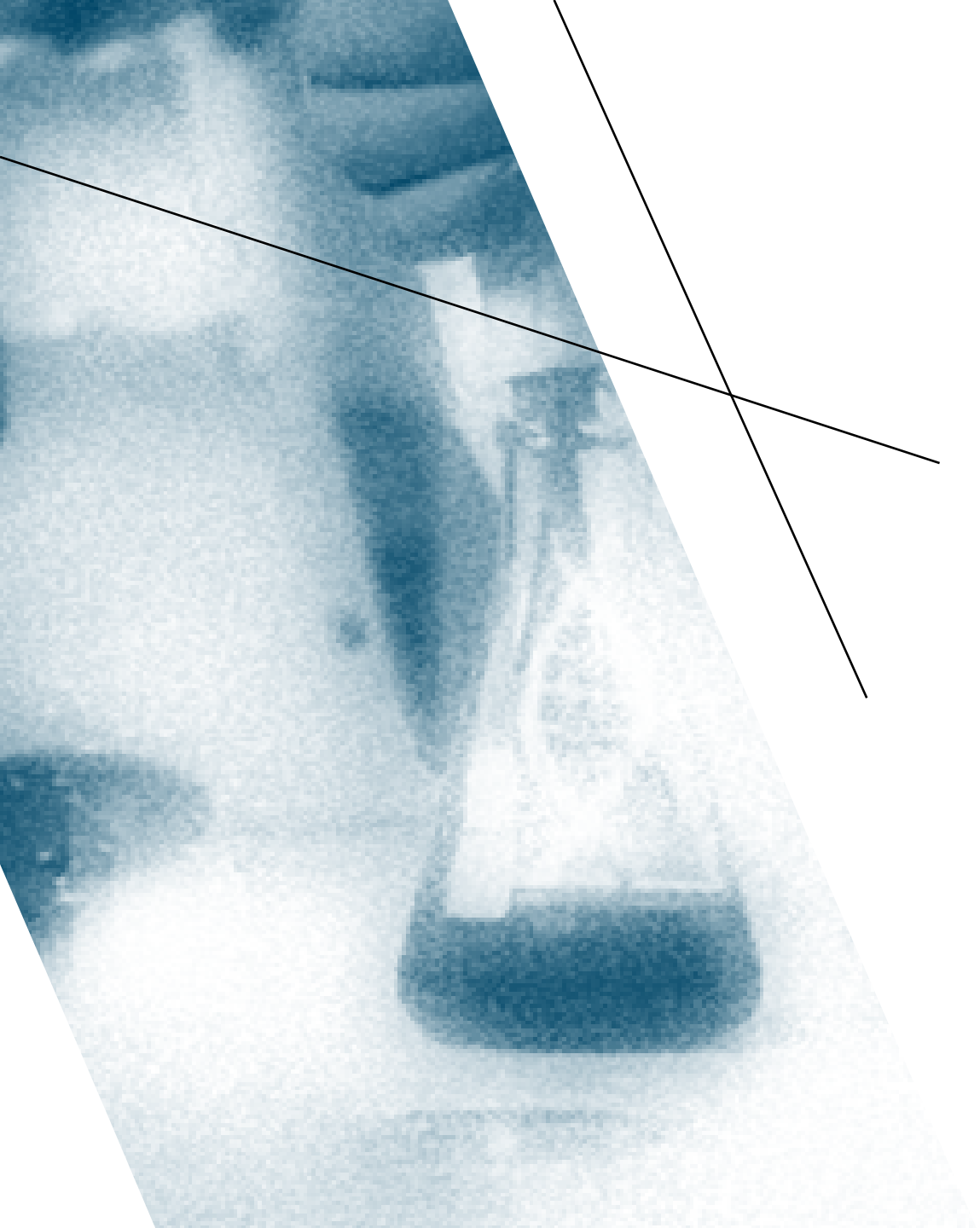
Επίθημα που καταδεικνύει την διεξοδική μελέτη βιολογικών συστημάτων και παραγόντων που μπορεί να εμπλέκονται στην παθογένεια της νόσου καθώς και στην συλλογή και ψηφιοποίηση βιολογικών δεδομένων

ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

From molecule to BIT: a systemic approach

How life can be translated into data





Biomarkers

- Biomarkers are measurable indicators of biological states or processes, including molecules, cells, or physiological signals.
- They are used to detect, diagnose, predict, and monitor diseases or treatment responses.
- Examples include genetic mutations, protein levels, imaging results, or metabolite concentrations.
- Challenges include ensuring biomarker specificity, sensitivity, reproducibility, and clinical relevance.

Τύπος Omics	Ορισμός	Αναμενόμενο αποτέλεσμα	Biomarkers
Γονιδιωματική	Πολυεπιστημονική προσέγγιση της μελέτης, του ποσοτικού προσδιορισμού και του χαρακτηρισμού όλων των γονιδίων και των μεταλλάξεων τους σε ένα δείγμα	Αλληλούχιση DNA, ταυτοποίηση SNP	**Προδιάθεση, Διάγνωση, Πρόγνωση, Σταδιοποίηση, Ταξινόμηση υπο-φαινοτύπων ασθενειών, Ανταπόκριση στη θεραπεία**
Επιγονιδιωματική	Το επιστημονικό πεδίο που μελετά τις αλληλεπιδράσεις του περιβάλλοντος με το γονιδίωμα και τον τρόπο με τον οποίο αυτές μπορούν να ρυθμίζουν τη γονιδιακή έκφραση	Διαφορικά μοτίβα τροποποιήσεων DNA και ιστονών	
Μεταγραφομική	Το πεδίο της επιστήμης που μετρά την έκφραση του RNA και ανιχνεύει ποσοτικές διαφορές και τους τρόπους αλληλεπιδράσεων των γονιδίων	Διαφορική έκφραση RNA (mRNA, rRNA, miRNA, lncRNA, tRNA, snRNA)	
Πρωτεομική	Δίνει την ευκαιρία ανίχνευσης, ταυτοποίησης και χαρακτηρισμού ολόκληρης της πρωτεϊνικής έκφρασης ενός δεδομένου κυττάρου ή ιστού σε ευρεία κλίμακα	Ανίχνευση, ταυτοποίηση και χαρακτηρισμός των επιπέδων έκφρασης του πρωτεόματος	
Μεταβολομική	Η μελέτη των μεταβολικών διεργασιών και των αλλαγών στην παραγωγή μεταβολιτών σε έναν οργανισμό	Ανίχνευση και ταυτοποίηση της σύνθεσης μεταβολιτών	
Μικροβίωμα (μετα-γονιδιωματική, μετα-μεταβολομική, κλπ)	Η συνολική γενετική σύνθεση της μικροχλωρίδας σε ένα συγκεκριμένο όργανο	Ανίχνευση, ταυτοποίηση, χαρακτηρισμός και ποσοτικοποίηση της μικροβιακής σύνθεσης και του μεταβολικού της προφίλ	

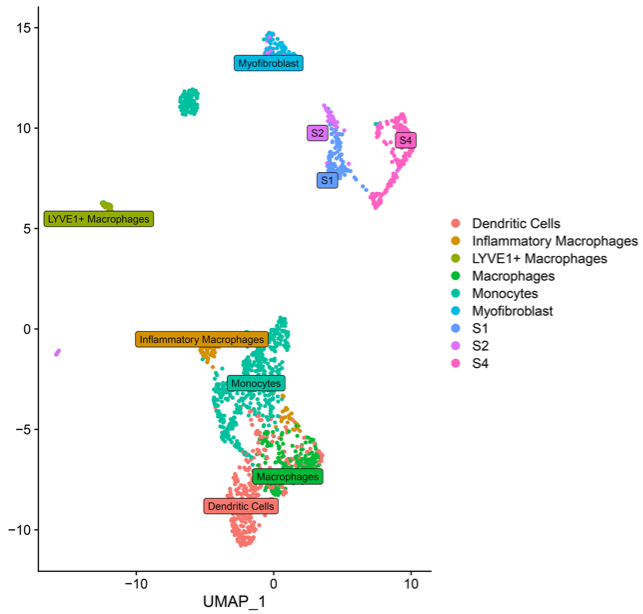
biomedicines

Single cell Transcriptomics

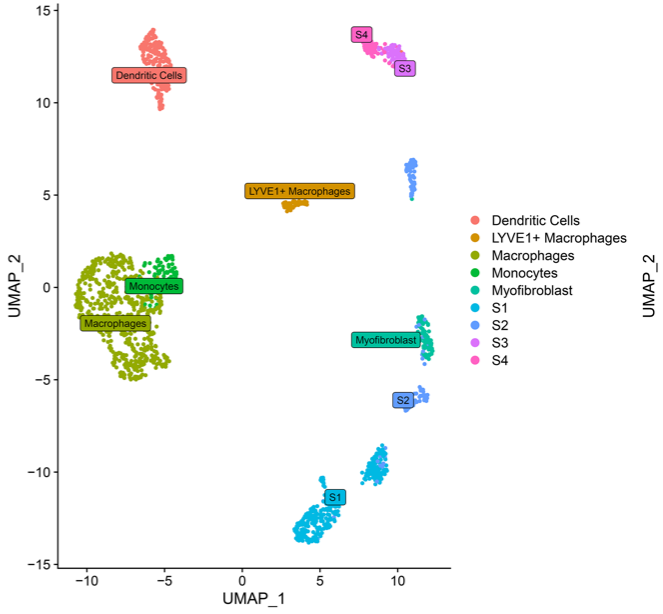
Landscape of Interactions between Stromal and Myeloid Cells in Ileal Crohn's disease

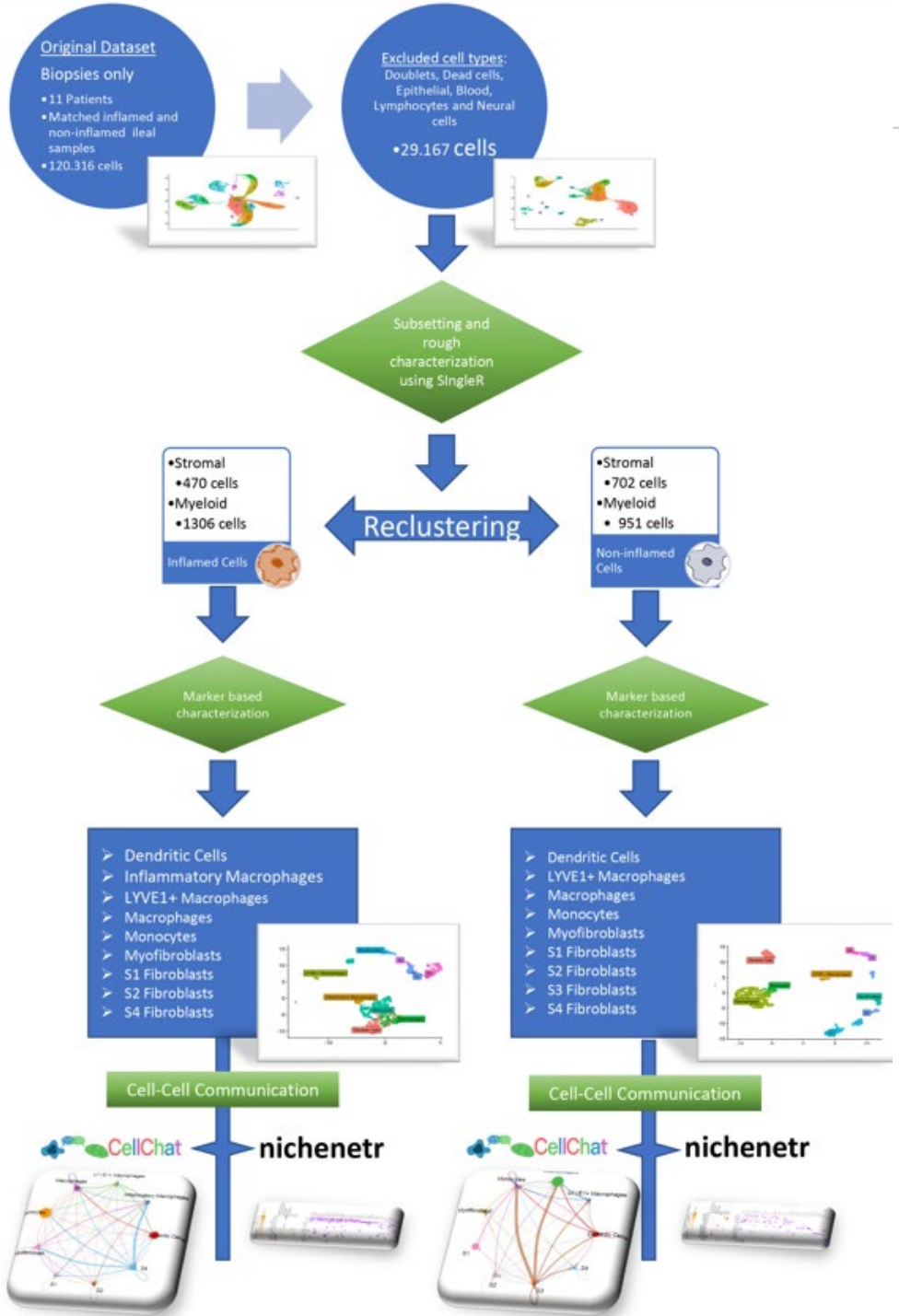
Nikolas Dovrolis[1,2*], Vassilis Valatas[1,3*], Ioannis Drygiannakis[3], Eirini Filidou[1,2], Michail Spathakis[1,2], Leonidas Kandilogiannakis[1,2], Gesthimani Tarapatzi[1,2], Konstantinos Arvanitidis[1,2], Giorgos Bamias[4], Stergios Vradelis[5], Vangelis G. Manolopoulos[1,2], Vasilis Paspaliaris[6], George Kolios[1,2]

INFLAMED CLUSTERS

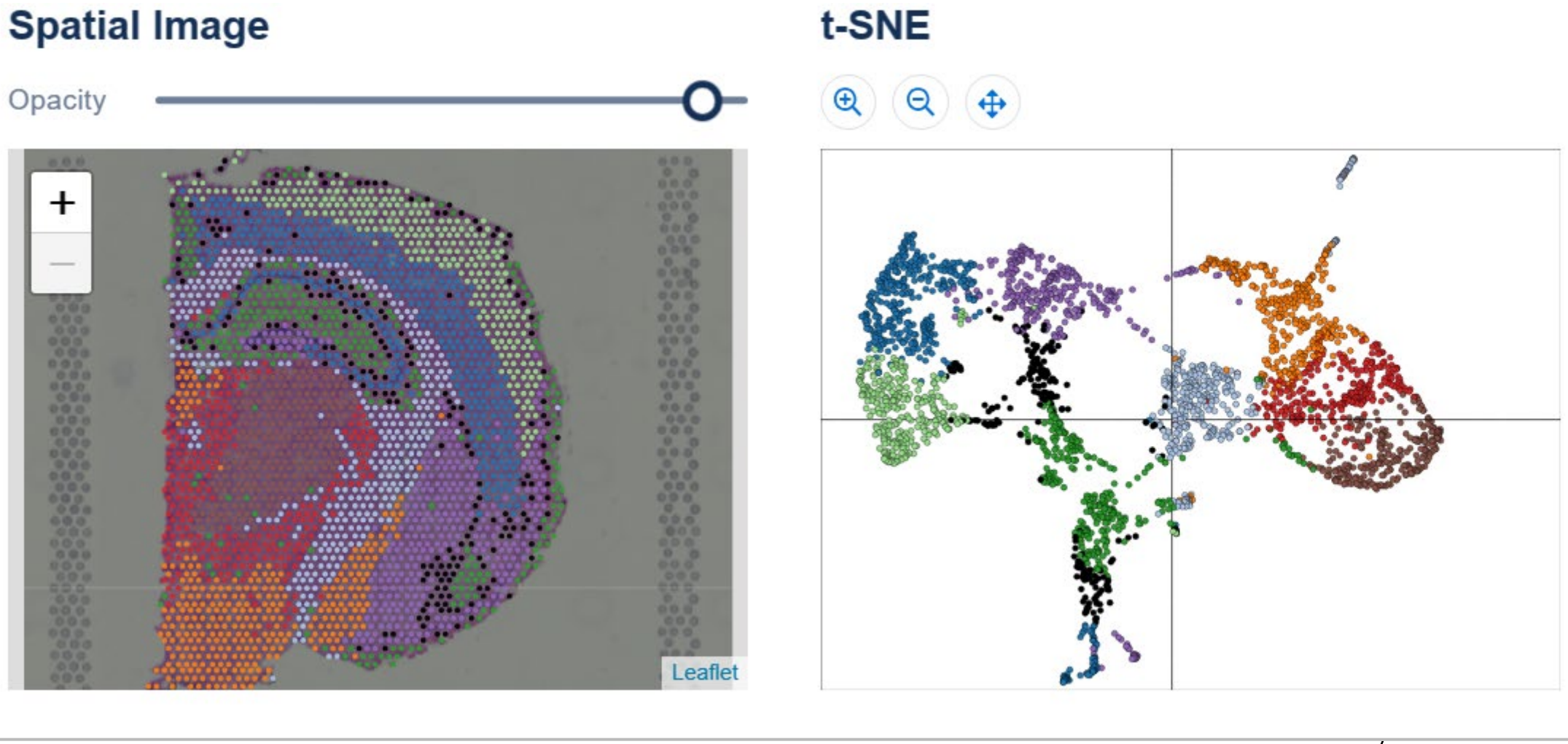


NON-INFLAMED CLUSTERS

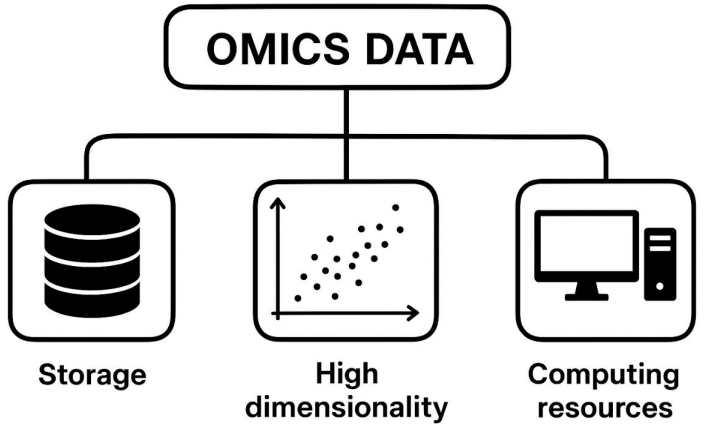




Spatial transcriptomics



[Spatial Transcriptomics - 10x Genomics]



HPC and omics

High dimensionality is a hallmark of modern biomedical and omics datasets, where the number of features (e.g., genes, CpG sites, SNPs) can vastly exceed the number of observations (samples).

Computational resource requirements are significant, as processing, analyzing, and integrating these datasets necessitate high-performance computing environments, optimized algorithms, and substantial memory and CPU/GPU resources.

High Performance Computing (HPC) becomes essential to manage, process, and analyze such data efficiently.

The Scale of Omics Data

Omics Type	File Size / Sample	# Samples / Study	# Features
Genomics (WGS)	100–200 GB (FASTQ)	100–100,000	~3–5 million variants
Transcriptomics (RNA-seq)	3–10 GB (FASTQ)	50–1,000	~20,000 genes
Epigenomics (450K/850K)	20–200 MB (IDAT/CSV)	50–10,000	450K–850K CpG sites
WGBS	100–200 GB	10–500	~28 million CpG sites
Proteomics (MS)	1–5 GB (RAW)	30–500	3,000–10,000 proteins
Metabolomics	100–500 MB (LC-MS)	30–300	100–10,000 metabolites
Metagenomics (16S)	50–150 MB (FASTQ)	50–1,000	100–1,000 taxa (OTUs/ASVs)
Metagenomics (WGS)	1–10 GB	50–5,000	10,000–1,000,000 genes

Just one repository

Category	Estimated Count
GEO Series (GSE)	~230,000
GEO Samples (GSM)	>10 million
GEO Platforms (GPL)	~5,000
Single-cell datasets	~25,000+
Studies with RNA-seq	~50,000+
Studies with array data	~100,000+
Total data volume (est.)	>2 petabytes (via SRA links)

Large Biobanks

TCGA, UK Biobank reach petabyte scale

Case Study – TCGA Pan-Cancer Atlas

Scale

Over 11,000 tumors across 33 cancer types analyzed

HPC Applications

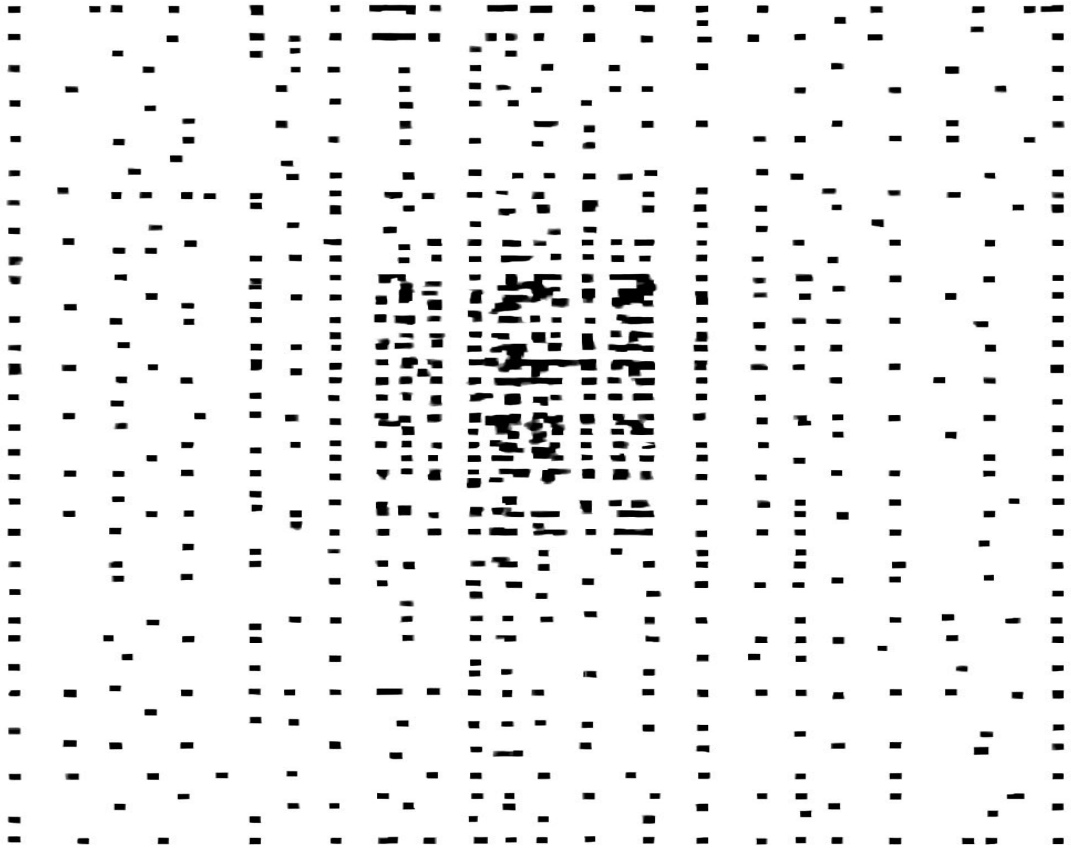
Multi-omics clustering (iCluster, MOFA), pathway enrichment, regulatory networks (ARACNe, PANDA)

Discoveries

Biomarkers like TP53 mutations and epigenetic silencing of MLH1 identified

Curse of Dimensionality

As dimensionality increases, data become sparser in the feature space, which can degrade the performance of distance-based algorithms and statistical models.



Role of HPC in High-Dimensional Data Analysis

HPC addresses these challenges by providing parallelism, memory scaling, and I/O acceleration.

A. Parallel Processing

Parallel algorithms (e.g., in R using `BiocParallel`, `foreach`, or in Python with `Dask`, `Ray`, or native MPI/CUDA frameworks) distribute workloads such as model training or matrix operations across multiple CPU or GPU cores.

B. Distributed Memory and Storage

For large datasets that exceed RAM capacity, HPC clusters with distributed memory allow in-memory processing using Spark or Hadoop-based platforms, or distributed matrix libraries (e.g., ScaLAPACK, Elemental).

C. Machine Learning Acceleration with Optimized Backends

Omics workflows increasingly rely on machine learning models built on heavy linear algebra (e.g., PCA, NMF, LASSO, neural nets).

Optimized backends like cuBLAS, MKL, and MAGMA power tools such as TensorFlow, PyTorch, and XGBoost to enable fast, scalable training on CPUs/GPUs.

A real world example

Methylation Analysis from WGBS (Cancer dataset)

2 nodes 40 CPU Cores / 500GB RAM / NVMe storage

Samples : 96

Size of sequences: ~2 TB

Time for sequence preprocessing (quality trimming) : ~4hrs

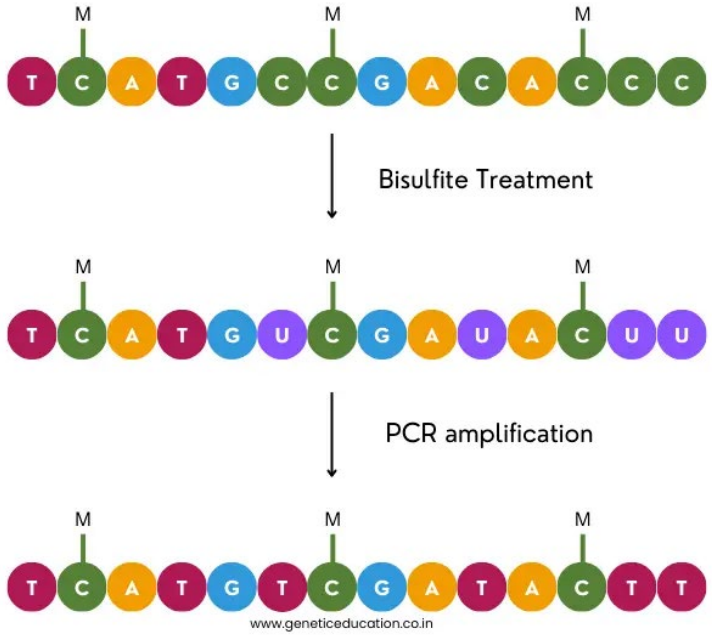
Time for sequence alignment to converted reference (with deduplication): ~72 hours, ~45min per sample

Methylation calling per cg: ~1hr

Table creation with Python: ~5hrs

Results : Sparse Table with 96 rows and 3.2 million columns . Uncompressed full matrix ~230 GB

Now imagine multiple datasets, of different omics, all combined together before ML even begins …



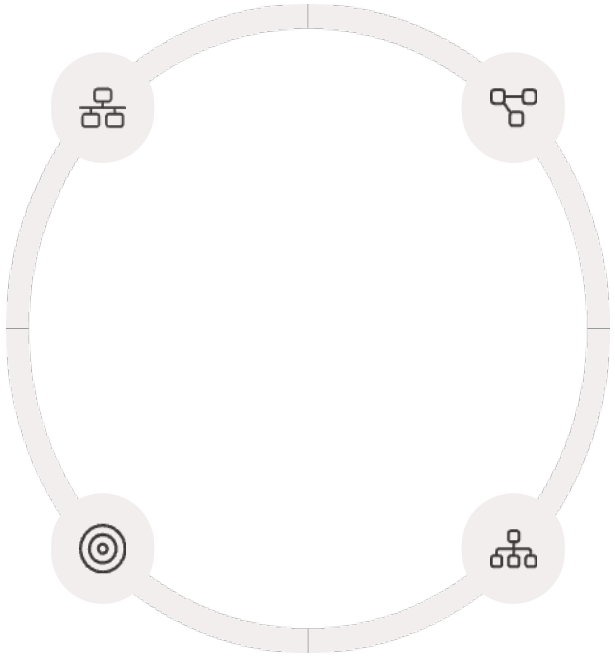
Network-Based Target Identification

Protein-Protein Interactions

Maps of physical protein interactions

Target Prioritization

Identification of hub genes and disease modules



Co-expression Networks

Genes with correlated expression patterns

HPC Algorithms

WGCNA, HotNet2, DiffusionRank

HPC, ML and OMICS

- Multi-omics Integration
  - ML: Integrative models (e.g., autoencoders, graph neural networks).
  - HPC: Manages high-dimensional matrices and iterative training.
- Drug Discovery
  - ML: Virtual screening, de novo drug design.
  - HPC: Molecular dynamics, docking simulations with ML acceleration.
- Workflow Integration: Raw data → Preprocessing → Feature selection → Model training (GPU) → Evaluation
- ML frameworks: TensorFlow, PyTorch + Snakemake/Nextflow on HPC clusters.

Molecular Docking and Virtual Screening

Structure-Based Discovery

HPC revolutionizes drug discovery through 3D molecular modeling.

Parallel Processing

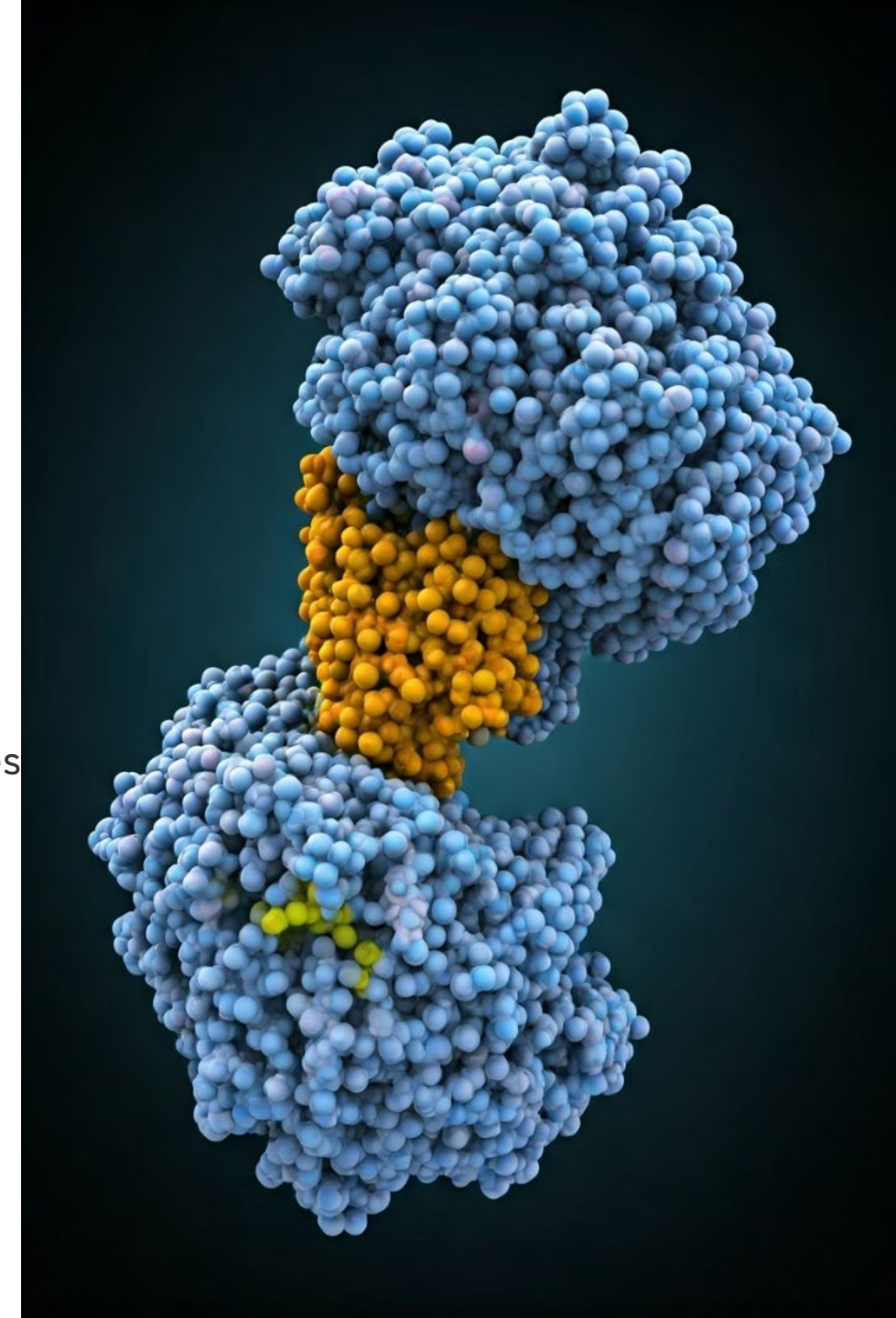
GROMACS and AutoDock-GPU leverage hundreds of CPU/GPU cores

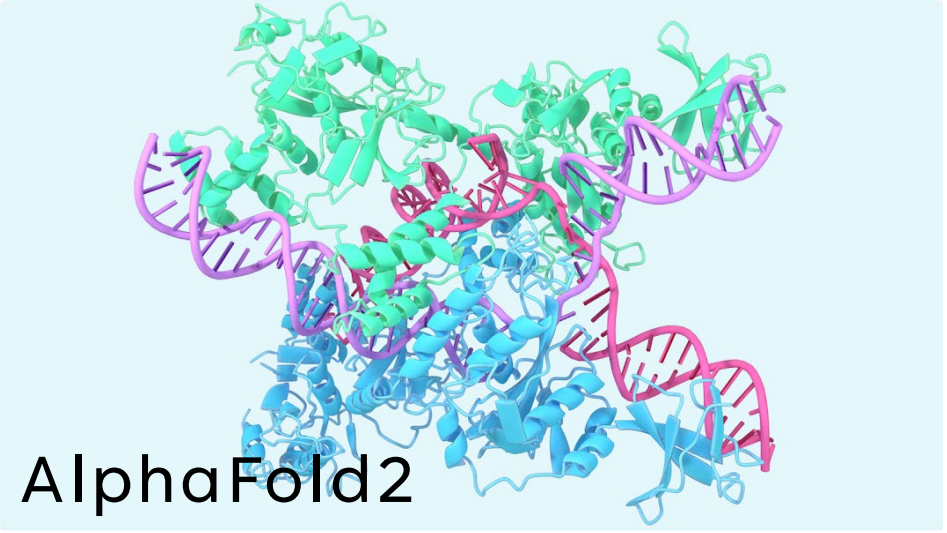
Accelerated Screening

Millions of compounds screened in silico within days.

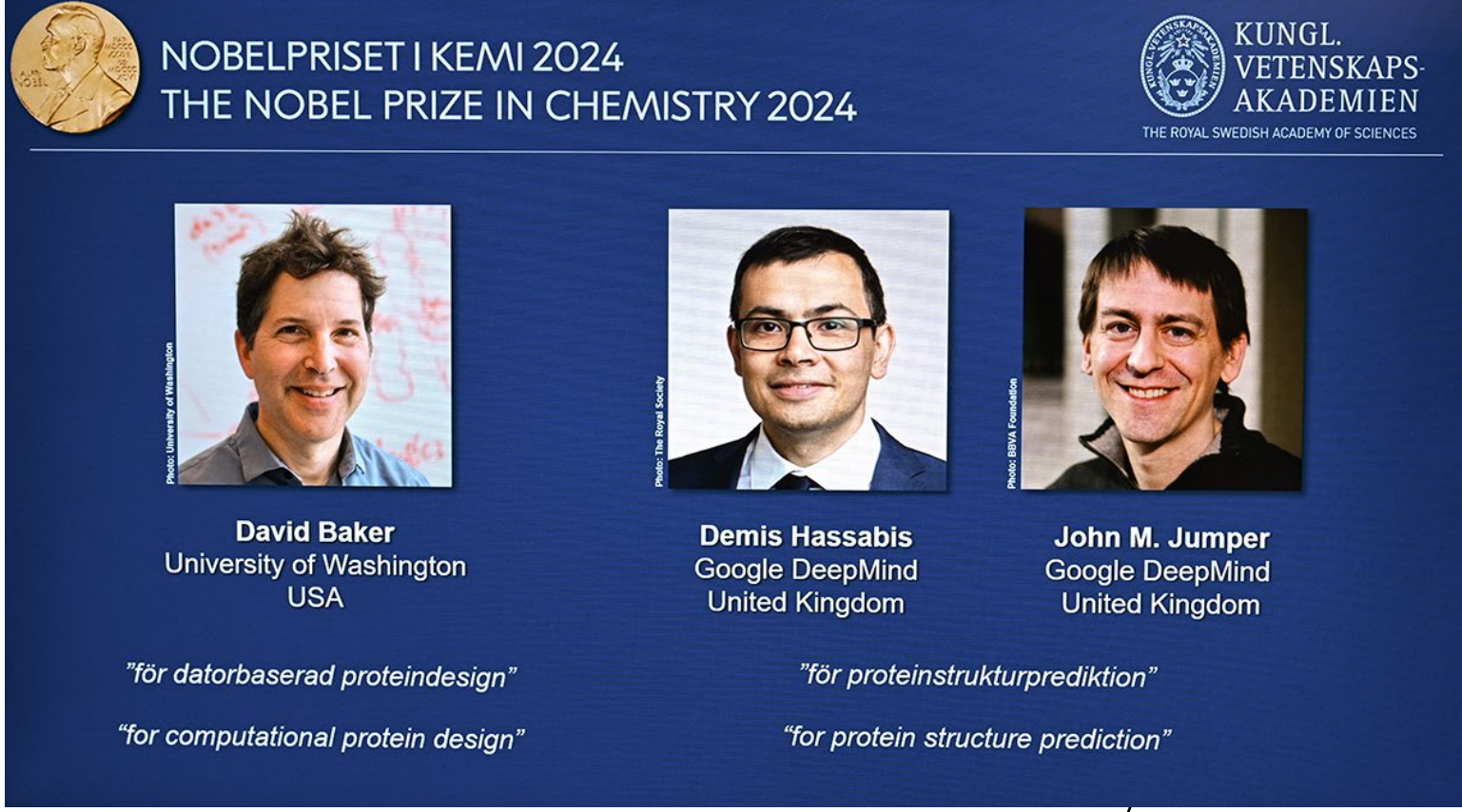
Target Binding

Quickly identifies candidate drugs that bind to high-confidence targets.





AlphaFold2



A deep learning model trained to predict protein 3D structures from amino acid sequences.

- Required thousands of GPUs and TPUs for model training.
- 200 million protein structures predicted
- Structure availability accelerates function prediction, target validation, and protein–protein interaction modeling.

- ✓ Rational Drug Design
- ✓ Virtual Screening
- ✓ De Novo Drug Generation
- ✓ Drug Repositioning
- ✓ Molecular Dynamics

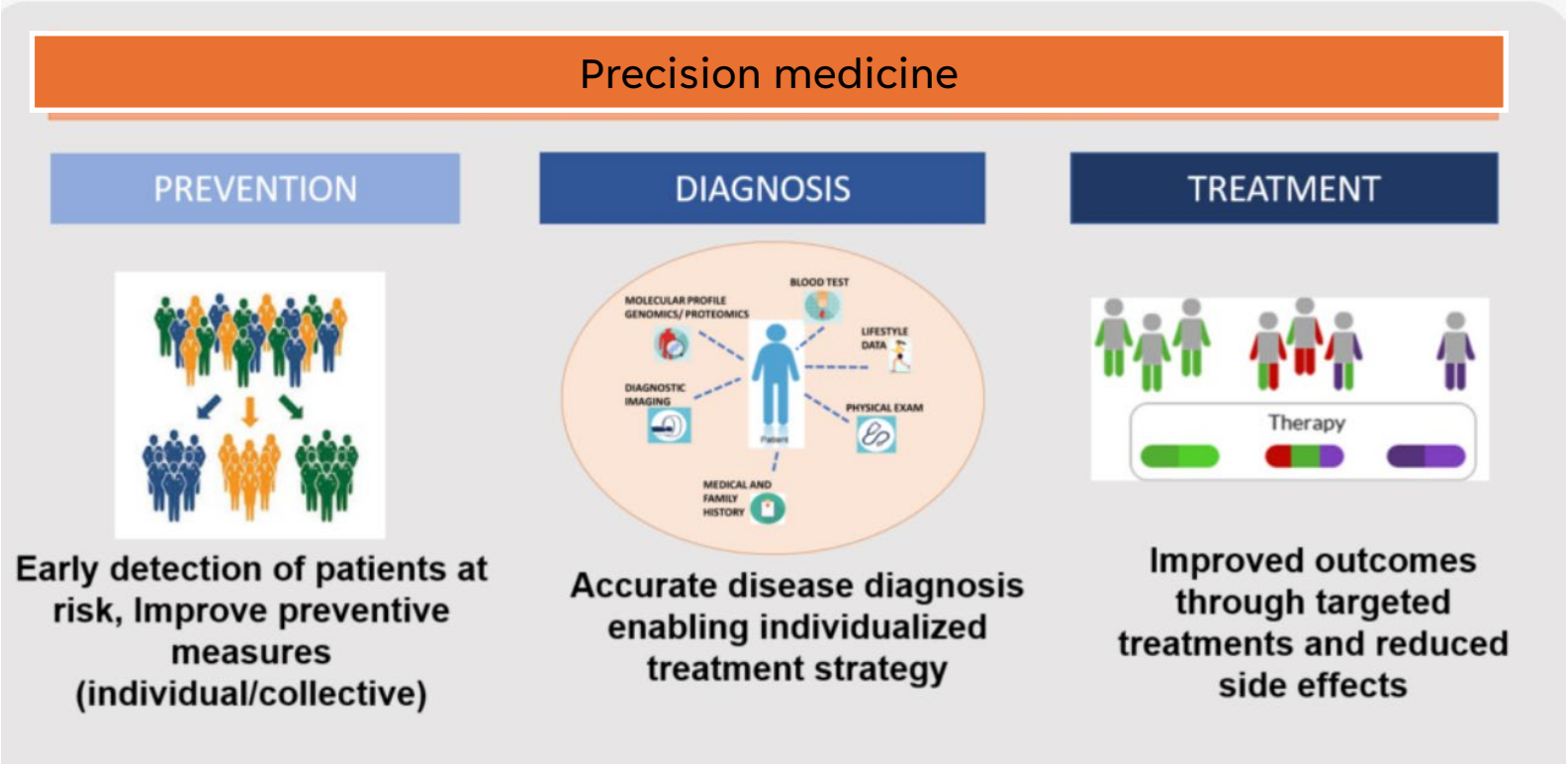
What’s next for drug discovery?

HPC Specs: 4 compute nodes (each with 32 CPU cores, 256 GB RAM, SSD/NVMe local scratch)

Just for 1 protein we have:

Step	Runtime (Wall Time)	Parallelization
Protein prep	~30 min	None
Ligand prep	~2.5 hrs	Across nodes
Docking	~20–40 hrs	Highly parallel
Scoring	~1.5 hrs	Moderate
Post-processing	~2–4 hrs	Optional parallelization

The end goal...



Precision medicine is a medical model that proposes the customization of healthcare, with medical decisions, treatments, practices, or products being tailored to a subgroup of patients, instead of a one-drug-fits-all model. In precision medicine, diagnostic testing is often employed for selecting appropriate and optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. Tools employed in precision medicine can include molecular diagnostics, imaging, and analytics.

ABCURED

Connecting the genetic dots with AI

a spin-off of Democritus University of Thrace

built to exploit results from our innovative research

BRIDGING BioMedical RESEARCH to HEALTH Solutions

- *biomarkers*
- *early detection*
- *precision medicine*
- *personalized pharmacotherapy*
- *non-invasive diagnostics*
- *novel drug targets*

ABCureD among winners of the CERPrize

Winners for the '23 CERPRIZE are:

*listed in alphabetical border

Yifat Anzelevich

Katerina Alexiou Chatzaki

Manuel Opitz

LPancreas
Liquid Pancreas Diagnosis

Karaglani, Makrina, et al. "Liquid biopsy in type 2 diabetes mellitus management: building specific biosignatures via machine learning." *Journal of Clinical Medicine* 11.4 (2022): 1045.

An Innovative Approach to Combating Diabetes

PREDIABETES

10% of Global Population has Prediabetes

Screening and Prevention

Informed timely interventions

DIABETES TYPE II

>1/2 Billion w/ Diabetes 2 (2M New Cases/Year)

Precision Treatment
Informed treatment choice

LADA

One out of Ten Diabetes Cases

Accurate LADA diagnosis
(Latent Autoimmune Diabetes in Adults)

DIABETES TYPE I

8.4 Million Individuals Suffer from Diabetes type 1

Immunotherapy
companion diagnostics

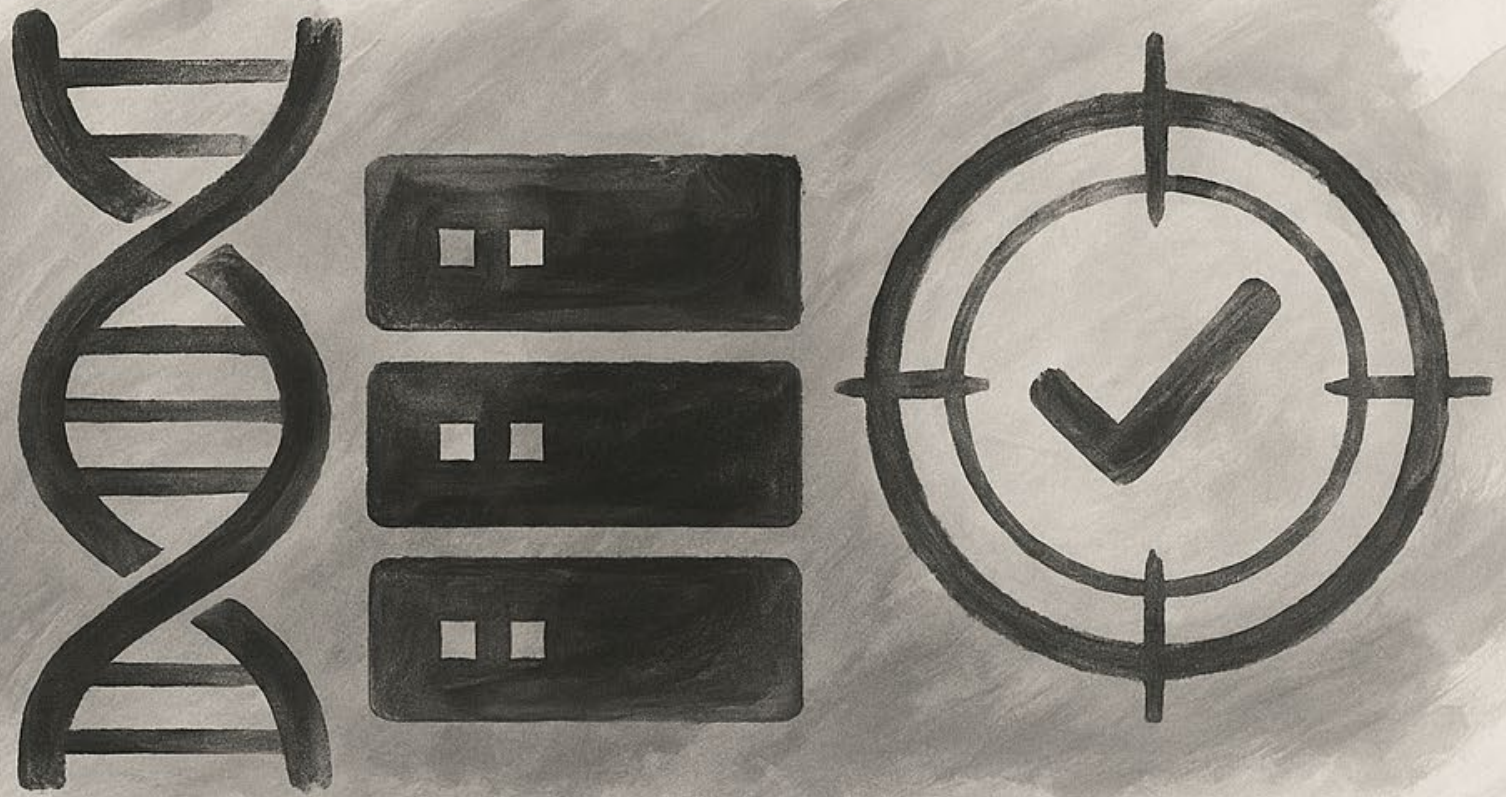
GESTATIONAL DIABETES

10% of pregnacies

Screening for after-birth
effects of gestational diabetes

LABORATORY OF PHARMACOLOGY

ABCURED
Connecting the genetic dots with AI



THANK YOU