Current application of

# Artificial Intelligence

## Tools and Resources in Biomedical Research

**Alberto Berral González**
**Bioinformatics & Functional Genomics Lab**
**Centro de Investigación del Cáncer**
**(CiC, CSIC/USAL) Salamanca (Spain)**

Financiado por la Unión Europea NextGenerationEU

06/05/2026

# Table of contents


01 Why AI?

02 Paradigm change

03 The AI toolkit

04 Workload archetypes

05 Examples

06 Challenges & the road ahead

# Why AI? Why now?

Current biomedical research is a **petabyte factory**

| | | | |
|---|---|---|---|
| **30M+**<br>**PubMed articles**<br>**~4.000 new/day** | **150 GB**<br>**per WGS genome**<br>**raw sequencing** | **10 yrs**<br>**$2.6B**<br>**drug development** | **x100**<br>**Data growth in omics over the last decade** |

Biomedical research generates more data than human can process

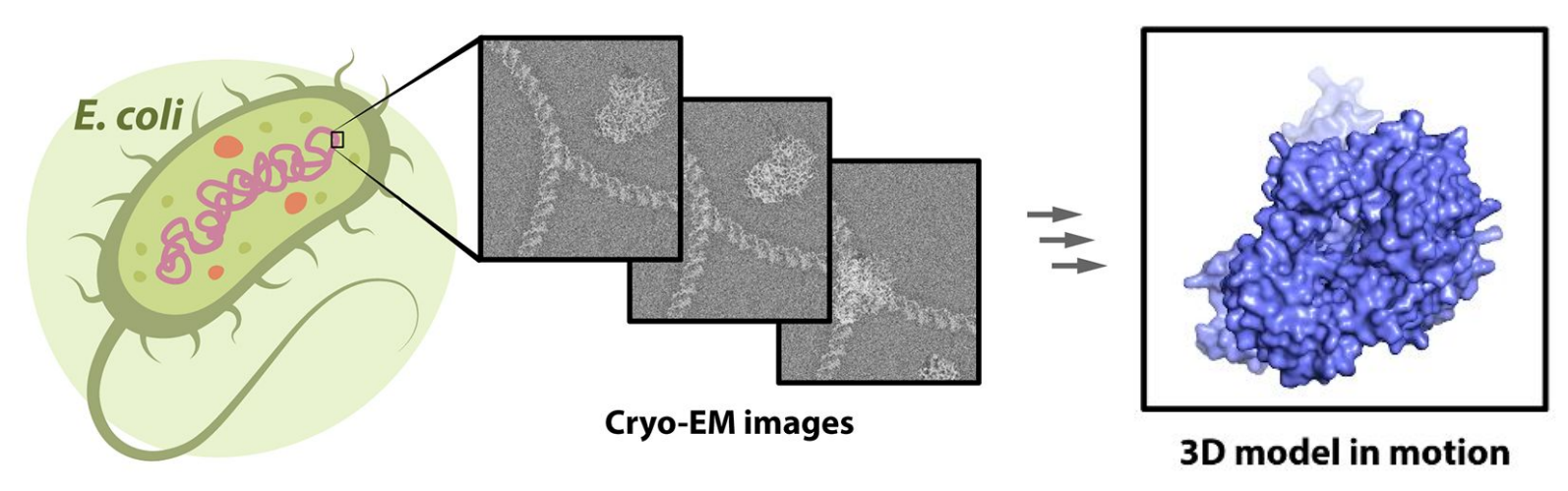
**HPC + AI bridges the gap between data and scientific insight**

# Why AI? Why now?

## Ex. Cryo-electron microscopy (cryo-EM)

**Nobel Prize 2017**

A single data acquisition session using a high-end microscope can generate **several terabytes** of raw data in a **single day**






Baldwin PR *et al*. Big data in cryoEM: automated collection, processing and accessibility of EM data (2018) doi: 10.1016/j.mib.2017.10.005.
https://morgridge.org/feature/cryo-em-primer/

# Paradigm change

## Shifting from Deterministic Analysis to AI-Driven Insights

**Classical bioinformatics**

- **Deterministic** pipelines
- **Structured** data
- **CPU-**intensive

# Paradigm change

## Shifting from Deterministic Analysis to AI-Driven Insights

**Classical bioinformatics**

- **Deterministic** pipelines
- **Structured** data
- **CPU-**intensive

**AI + Bioinformatics**

- **Stochastic** pipelines
- **Unstructured** data
- **GPU-**intensive

# Paradigm change

## Shifting from Deterministic Analysis to AI-Driven Insights

**Classical bioinformatics**

- **Deterministic** pipelines
- **Structured** data
- **CPU-**intensive

**AI + Bioinformatics**

- **Stochastic** pipelines
- **Unstructured** data
- **GPU-**intensive

**Adapt** & **Re-think** our **tools**

# The AI Toolkit

## A methodology-first view of current algorithmic families

CPU CLUSTER

### Classical ML

RF, SVM, XGBoost

Biomarker discovery, clinical risk scores and feature selection

GPU

### Deep Learning

CNN, RNN, Transformers

Medical imaging analysis, basic sequence modeling and EHR structured data

GPU + Large RAM

### Graph Neural Nets

GNNs, DeepGraph

Drug-target interaction, knowledge graphs and protein-protein networks

High VRAM

### Generative AI

GAN, VAE, Diffusion

Drug molecule design, protein engineering and synthetic data generation

Multi-GPU

### Large Language Models

Bio-LLMs, scGPT, Geneformer

Literature mining, clinical NLP and foundational models

HYBRID CPU + GPU

### Reinforcement Learning

DQN, PPO, AlphaZero

Experimental design automation, adaptive clinical trials and robotics

# Workload archetypes

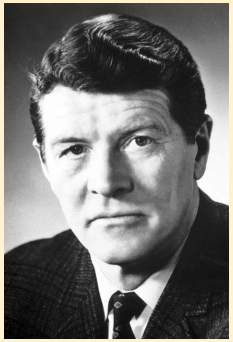
**Hardware-centric classification for high-performance biomedical workloads**

| | Type | Example Models | Critical Hardware | Execution pattern |
|---|---|---|---|---|
|  | **Bio sequences** | **AlphaFold / DNABERT**<br>Proteins & Genomics | GPU<br>RAM<br>Flash Storage | CPU heavy prep +<br>GPU modeling |
|  | **Computer vision** | **MONAI**<br>Tumor Segmentation | GPU<br>High Speed NVME | Cold training /<br>Massive inference |
|  | **NLP (Bio-LLMs)** | **scGPT / PubMedBERT**<br>EHRs & Literature | GPU<br>Text I/O | Frequent, dynamic<br>fine-tuning |
|  | **Digital Chemistry** | **GNNs / DiffDock**<br>Drug discovery | GPU<br>RDMA networking | Parallel screening /<br>Job arrays |

# AlphaFold

Bio sequences

## Solving the 50-year Protein Folding Challenge using AI



**Chemistry Nobel Prize**

**1972**

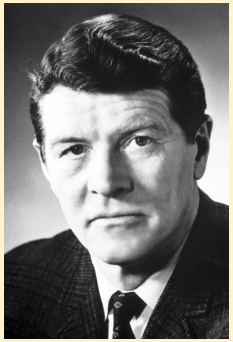
**Christian Anfinsen**

The **sequence** of amino acids, in **itself**, **determines** the way **the chain folds** itself and that **no additional** genetic **information** is **required** in this process.

# AlphaFold

Bio sequences

## Solving the 50-year Protein Folding Challenge using AI

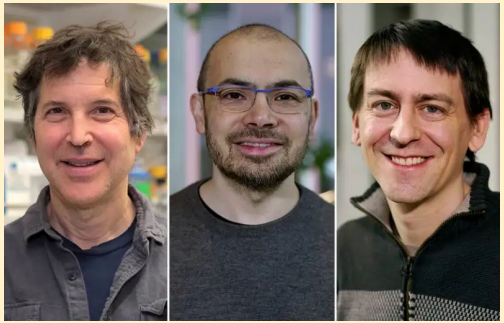


**Chemistry Nobel Prize**

**1972**

**Christian Anfinsen**

The **sequence** of amino acids, in **itself**, **determines** the way **the chain folds** itself and that **no additional** genetic **information** is **required** in this process.



**Chemistry Nobel Prize**

**2024**

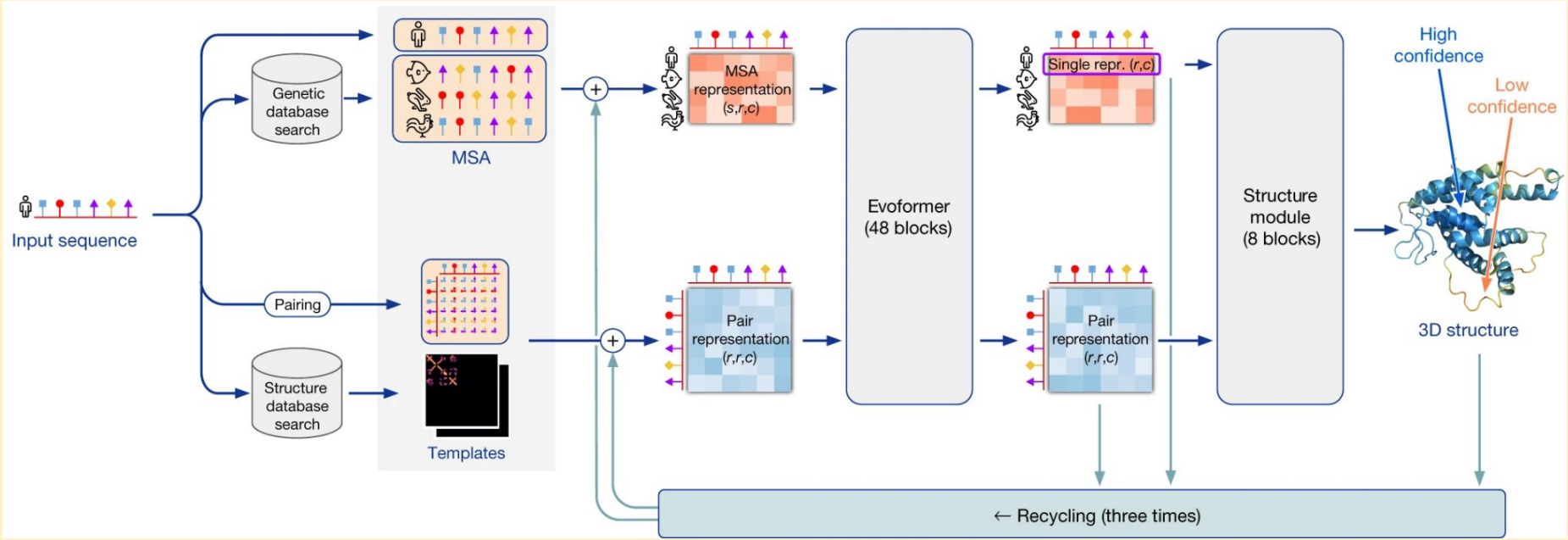
**Baker, Hassabis, Jumper**

AlphaFold2 **predicted** the 3D structure of **+200 million proteins** —virtually all known to science— **achieving in years** what would have taken **centuries of** manual lab **work**.

# AlphaFold: how does it work?

Bio sequences

## Solving the 50-year Protein Folding Challenge using AI



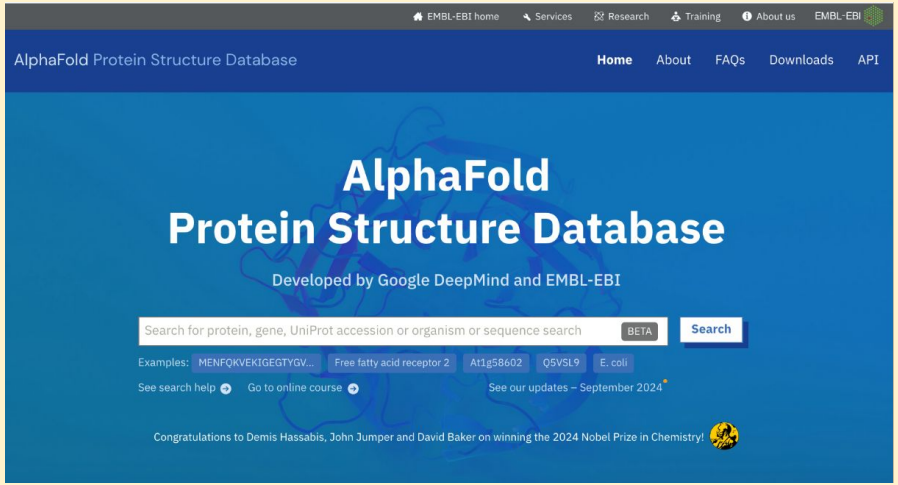


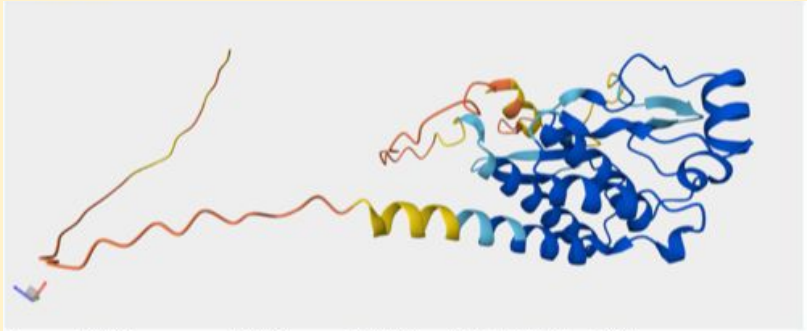

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

# AlphaFold: how does it work?

Bio sequences

## Solving the 50-year Protein Folding Challenge using AI



This work is a major step forward, but **it's not a solution**



Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

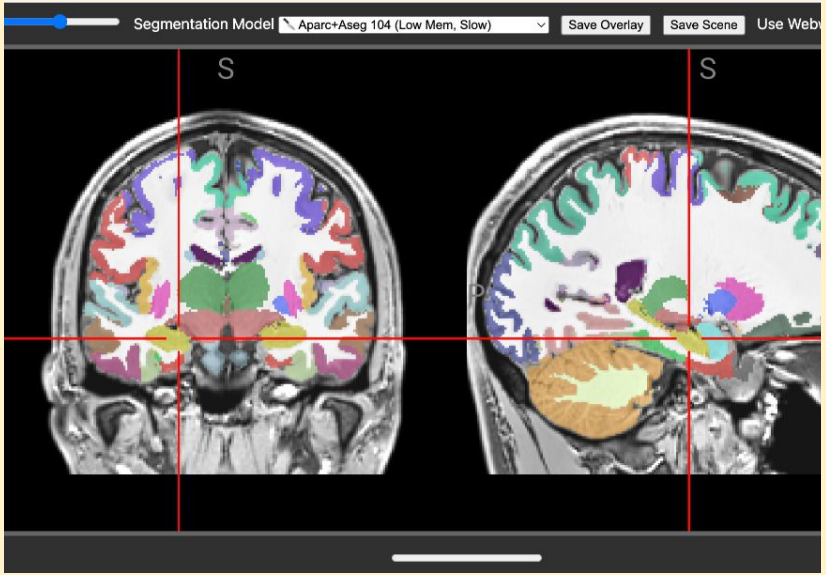
# Medical Open Network for AI

Computer vision

## MONAI: Bridging AI & Healthcare

**Based on PyTorch**

**Tools** & **libraries** for **developing**, **training** and **deploying** AI models in medical imaging applications (**segmentation**, **classification** & **registration**)






Plis, S. M., Masoud, M. et al. (2024). Brainchop: Providing an Edge Ecosystem for Deployment of Neuroimaging Artificial Intelligence Models. Aperture Neuro, 4, 2024. https://doi.org/10.52294/001C.123059

# Medical Open Network for AI

Computer vision

## MONAI: Bridging AI & Healthcare

### MONAI Label

Create **training datasets** and build AI **annotation** models

### MONAI Core

Specialized infrastructure to **build** AI-driven medical imaging **pipelines**

### MONAI Deploy

**Simplify deployment** in existing medical infrastructure



Plis, S. M., Masoud, M. et al. (2024). Brainchop: Providing an Edge Ecosystem for Deployment of Neuroimaging Artificial Intelligence Models. Aperture Neuro, 4, 2024. https://doi.org/10.52294/001C.123059

# Medical Open Network for AI

Computer vision

## MONAI: Bridging AI & Healthcare

### MONAI Label

Create **training datasets** and build AI **annotation** models

### MONAI Core

Specialized infrastructure to **build** AI-driven medical imaging **pipelines**

### MONAI Deploy

**Simplify deployment** in existing medical infrastructure

**REALITY**

Preference for **turnkey** commercial **solutions** for faster deployment.



Plis, S. M., Masoud, M. et al. (2024). Brainchop: Providing an Edge Ecosystem for Deployment of Neuroimaging Artificial Intelligence Models. Aperture Neuro, 4, 2024. https://doi.org/10.52294/001C.123059

# Medical Open Network for AI

Computer vision

## MONAI: Bridging AI & Healthcare

### MONAI Label

Create **training datasets** and build AI **annotation** models

### MONAI Core

Specialized infrastructure to **build** AI-driven medical imaging **pipelines**

### MONAI Deploy

**Simplify deployment** in existing medical infrastructure

**REALITY**

Preference for **turnkey** commercial **solutions** for faster deployment.

Using **open source frameworks** to **prevent** "reinventing the wheel" with each imaging project



Plis, S. M., Masoud, M. et al. (2024). Brainchop: Providing an Edge Ecosystem for Deployment of Neuroimaging Artificial Intelligence Models. Aperture Neuro, 4, 2024. https://doi.org/10.52294/001C.123059

# Foundational models and LLMs

NLP
(Bio-LLMs)

## Training a ‘Jack-of-all-trades’ model

**Before**

A **different dataset**
and a
**different model**
for each task

Greater **amount** of **data** required per task !!



Fei Guo *et al.*, Foundation models in bioinformatics, National Science Review, Volume 12, Issue 4, April 2025, nwaf028, https://doi.org/10.1093/nsr/nwaf028

# Foundational models and LLMs

NLP (Bio-LLMs)

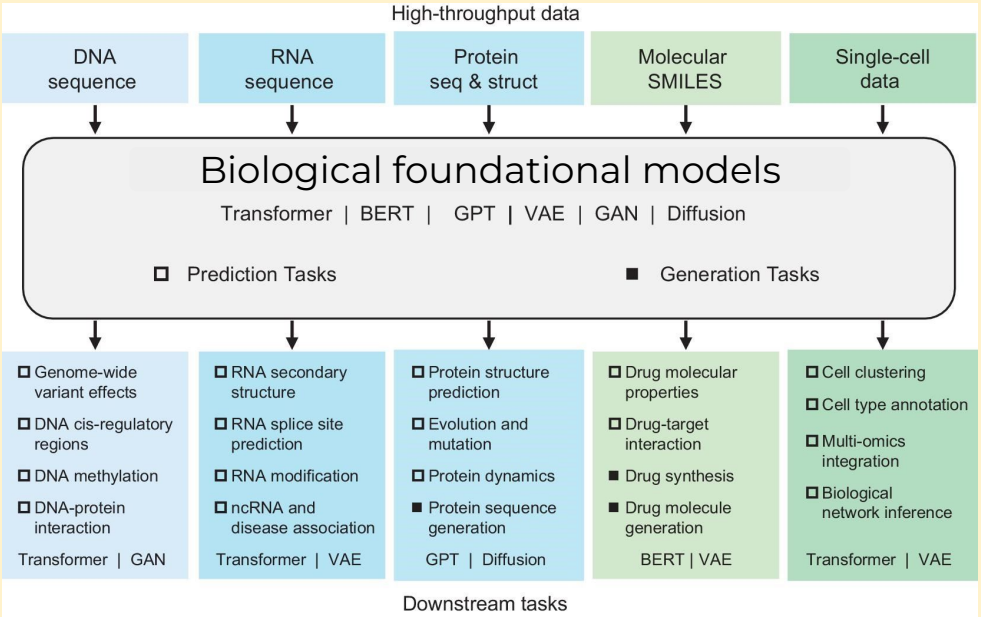
## Training a ‘Jack-of-all-trades’ model

**Before**

A **different dataset** and a **different model** for each task

Greater **amount** of **data** required per task !!

**Now**





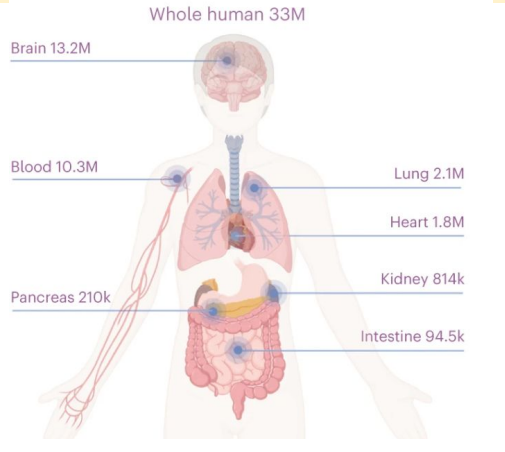

Fei Guo *et al.*, Foundation models in bioinformatics, National Science Review, Volume 12, Issue 4, April 2025, nwaf028, https://doi.org/10.1093/nsr/nwaf028

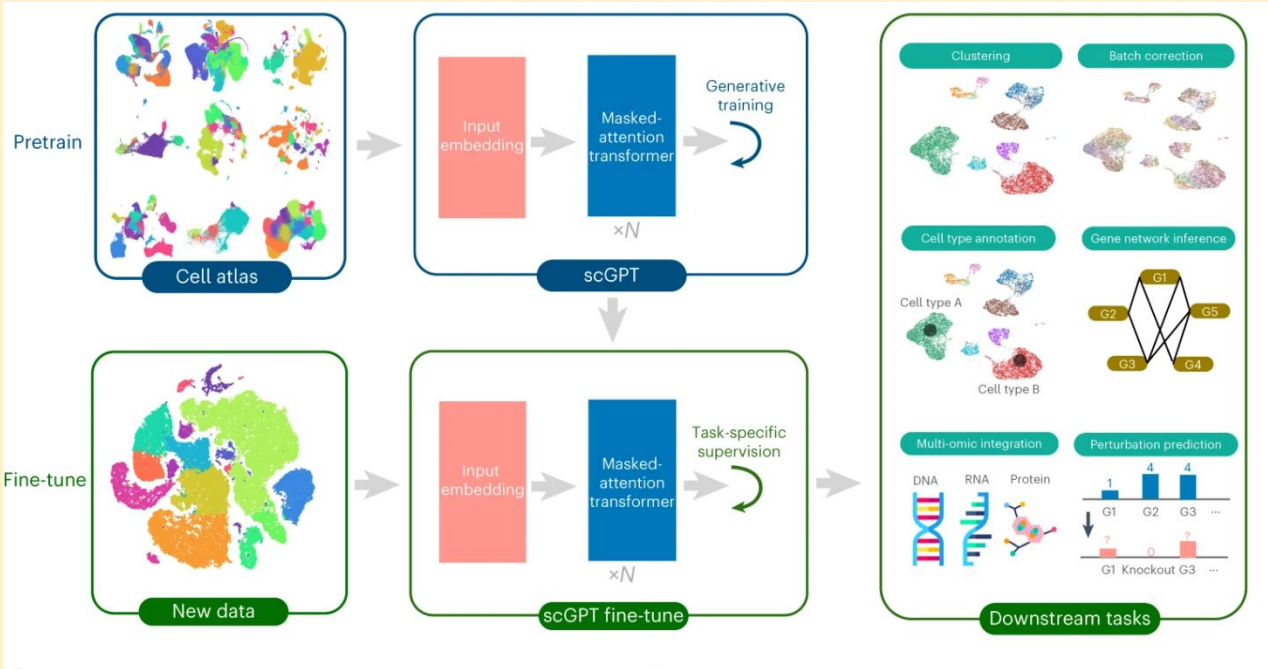
# Foundational models and LLMs

NLP
(Bio-LLMs)

## scGPT - Reading cell grammar

Trained with an **atlas** of **more** than **33M normal** cells






Cui, H., Wang, C., Maan, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 21, 1470–1480 (2024). https://doi.org/10.1038/s41592-024-02201-0

# Foundational models and LLMs

NLP (Bio-LLMs)

## LLMs: beyond software

**DNA/RNA sequence analysis, functional prediction and structure prediction**

- HyenaDNA
- DNAGPT
- CombSAFE
- BERT-Enhancer
- DNABERT
- RoBERTa
- RNABERT
- GeoBoost2

**Protein sequence analysis, functional prediction and structure prediction**

- PTMGPT
- ProGen
- ESM
- ProtGPT
- EpiBERTope
- ELMo
- BepiPred-3.0
- TransPPMP
- ProteinBERT
- MSA Transformer
- BERT4Bitter

**Multi-Omics Data Analysis**

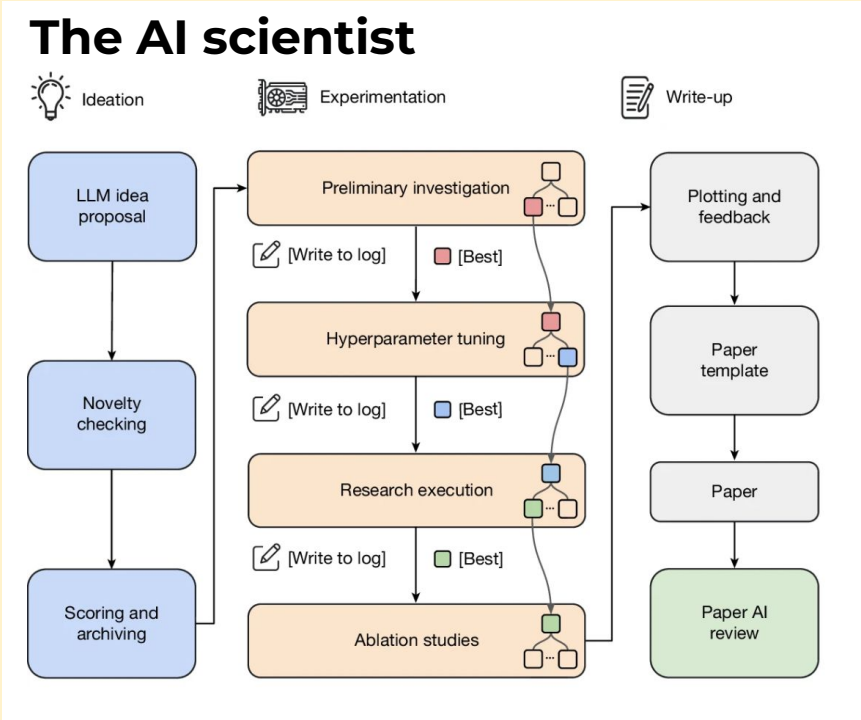
- BERT-TFBS
- scGPT
- GP-GPT
- pLM4Alg
- scELMo
- MuLan-Methyl
- iEnhancer-BERT
- scBERT
- TooT-PLM-ionCT
- UESolDS
- iDNA-ABT
- POOE
- mtx-COBRA
- DeepGene Transformer

**Computational Approaches in Drug Discovery and Design**

- MolGPT
- ChemBERTa
- SMILES-BERT
- PISTON
- MRCF
- K-BERT
- PharmBERT
- DTI-BERT
- TransDTI

**Biomedical Literature Mining**

- Galactica
- BioBERT
- ClinicalBERT
- GENEVIC
- MarkerGenie
- LBERT
- GenCLiP 3
- BioBERT




Lin A, *et al*. Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. Brief Bioinform. 2025 Jul 2;26(4):bbaf357. doi: 10.1093/bib/bbaf357.
Lu, C., Lu, C., Lange, R.T. et al. Towards end-to-end automation of AI research. Nature 651, 914–919 (2026). https://doi.org/10.1038/s41586-026-10265-5

# AI in drug discovery

Digital Chemistry

## From a Tortuous Path to a Streamlined Pipeline


### Target structure prediction

**AlphaFold**: Predicted 3D structure >200M proteins

### Virtual screening

ML models screens billions of compounds in silico, reducing wet lab screening

### Generative chemistry

Models in silico generate novel drug-like molecules with desired properties

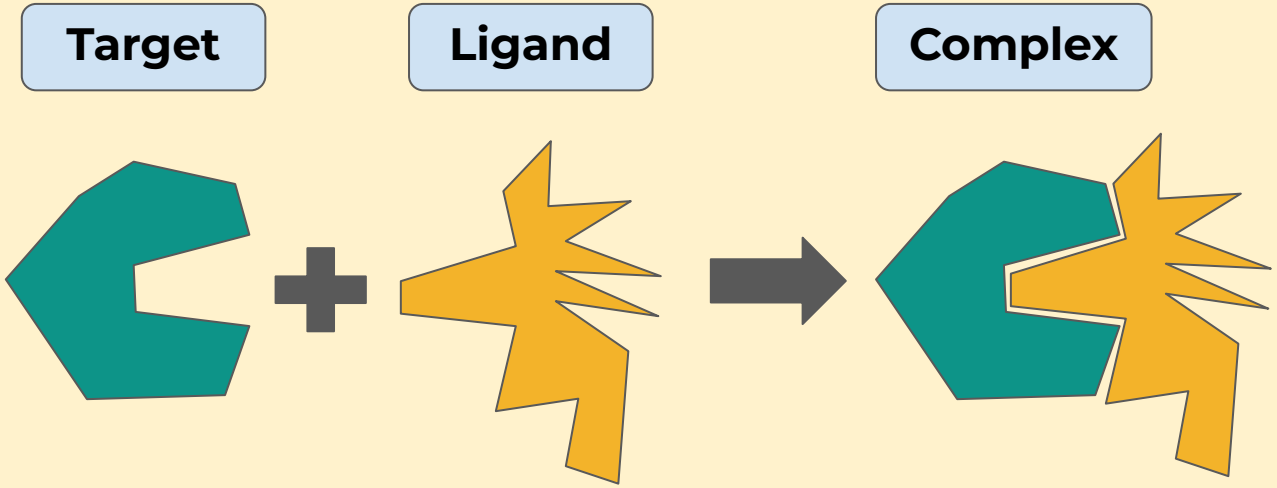
### ADMET prediction

AI predicts absorption, distribution, metabolism, excretion and toxicity early

# DiffDock: difussion and GNNs

Digital Chemistry

## Generative AI for molecular docking

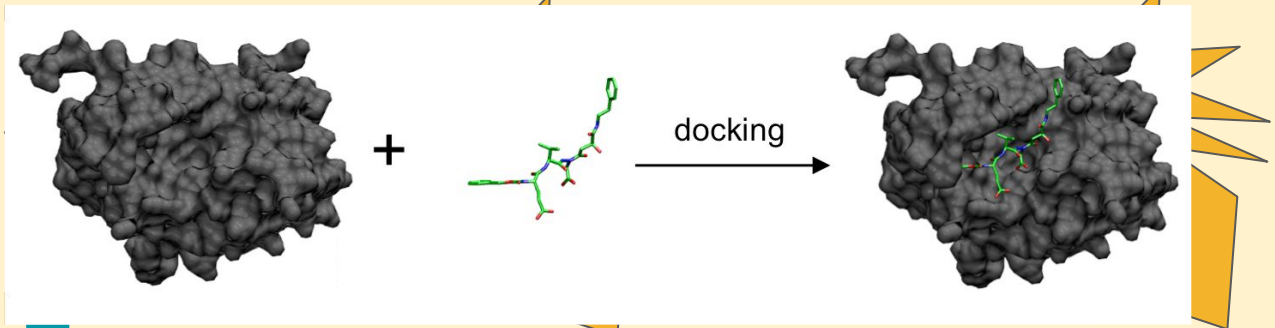


# DiffDock: difussion and GNNs

Digital Chemistry

## Generative AI for molecular docking

**Target** + **Ligand** → **Complex**




**Traditional**

**Brute force search**: **Test thousands** of **random** rotations and positions **until** an **energy minimum** is **found** (minutes to hours)

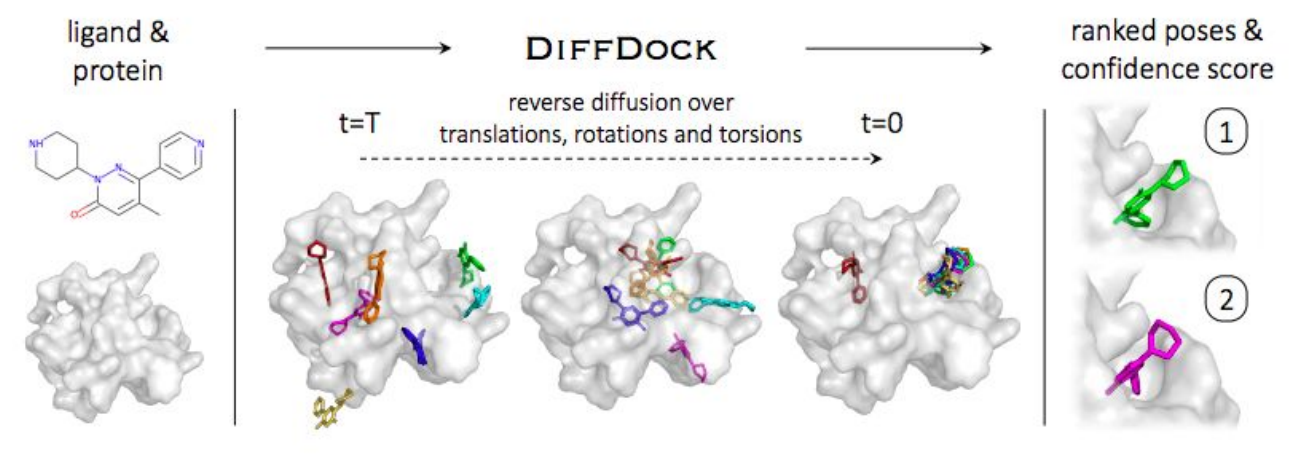
Estimating this binding site is a **computationally intensive search problem**.

# DiffDock: difussion and GNNs

Digital Chemistry

## Generative AI for molecular docking

**Initial noise** → **Reverse diffusion** → **Final pose**




**DiffDock**

**Diffusion models** to 'learn' the optimal position, achieving **unprecedented accuracy** in a matter of seconds.

# Challenges & Ethical considerations

**Data quality & bias**

Models trained on biased data can perpetuate health disparities.

**Interpretability (XAI)**

Black box models hinder clinical trust. Explainable AI techniques are essential.

**Regulatory approval**

FDA, EMA frameworks for AI as software as a Medical device (SaMD) are still evolving rapidly.

**Data privacy**

Patient data governance, federated learning, and differential privacy are critical for compliant AI development.

**Reproducibility**

Many published AI models lack code/data sharing - reproducibility crisis mirrors classical biomedical research.

**Clinical validation**

Prospective clinical trials needed to prove real-world impact beyond in silico tests.

# The road ahead

| | |
|---|---|
| **Now** | AI-assisted imaging, genomics,<br>drug screening validated tools in research pipelines |
| **2026-28** | AI co-pilots in clinical settings<br>FDA/EMA-approved AI diagnostics mainstream |
| **2029-32** | Multimodal foundation models integrating imaging +<br>omics + EHR for holistic patient modeling |
| **2033+** | Autonomous AI-driven hypothesis generation and<br>experimental design closing the R&D loop |

**The question is no longer IF AI will transform biomedicine, but HOW FAST and with what safeguards.**

# Take home message

a. AI is not a single tool, it's a **family of techniques** each **suited** to **different biomedical problems**.

b. **Proven impact areas**: protein structure prediction, medical imaging, drug virtual screening.

c. **Challenges remain**: data bias, interpretability, regulatory pathways, and rigorous clinical validation.

d. Biomedical **researchers benefit from AI literacy**, not necessarily coding expertise, but conceptual understanding.

Thanks!

# ANY QUESTIONS?

You can also find me at:

aberralgonzalez@usal.es

CIC Centro de Investigación del Cáncer
CSIC Consejo Superior de Investigaciones Científicas
Universidad de Salamanca
Momentum CSIC Desarrolla tu talento digital
Financiado por la Unión Europea NextGenerationEU
Gobierno de España Ministerio para la Transformación Digital y de la Función Pública
red.es
Plan de Recuperación, Transformación y Resiliencia