

UMU

Disease Modeling and Prediction

Perspective based on Foundation models

Juan A. Botía

Universidad de Murcia, Spain





UNIVERSIDAD
DE MURCIA

What is a foundation mode anyway?

A foundation model (FM) is any model that

- is trained on large corpus of data,
- using **self-supervision** and
- that can be adapted to a wide range of downstream tasks once trained.

Generally speaking, they are based on deep neural networks as the way of representing and training the models.

What is actually new is the growth rate (size) they follow, with an ever growing rate year by year.

Thanks to this scale, one distinctive capability is **zero-shot learning**, i.e., showing capabilities for which they were not explicitly trained.



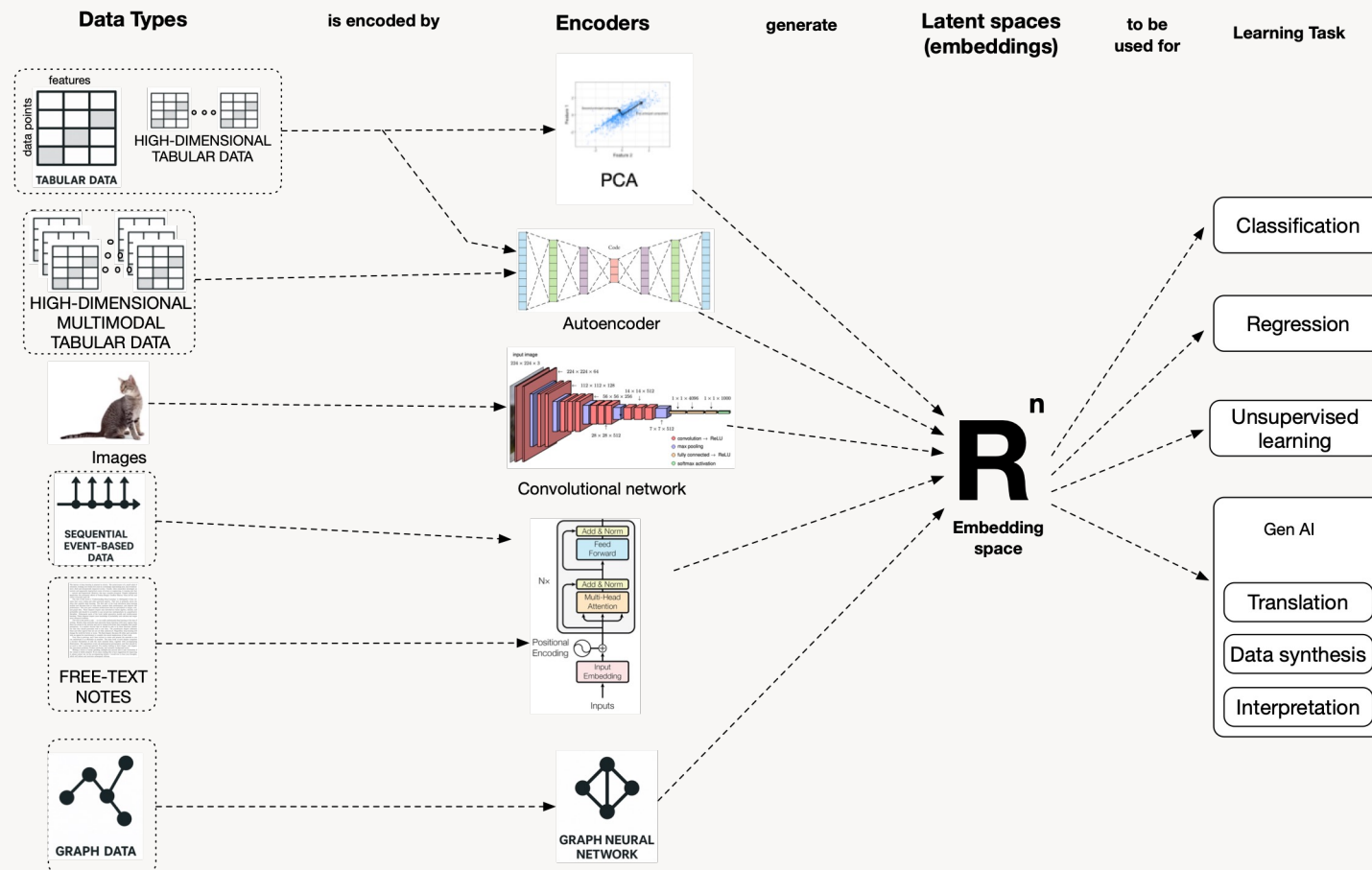
Self-supervised learning

UNIVERSIDAD
DE MURCIA

- It does not require of labelled data,
 - learns from pure raw examples before applying any constraint, as it is imposed by supervised learning when you tell the system how many classes are there and examples of them.
- Avoiding constrained learning makes it possible for these models to **learn the foundations** about the training material.
- FMs perform dimensionality reduction
 - Increases level of information representation
 - Enables discovering more accurate associations within data
 - Improve more restrictive downstream tasks.

There is a suitable technology for each training data modality

UNIVERSIDAD
DE MURCIA





UNIVERSIDAD
DE MURCIA

A practical notion of what is a successful foundation model

A FM is a non-linear and context-dependant **representation mapping** from the original space to a new space, from input examples domain X , to a vectorial space Z of d dimensions

A FM that works, for a downstream task T , and a way to measure improvement in the task, $I(T)$, we have that for the majority of x in X , it holds that

$$I(T(FM(x))) > I(T(x)),$$

i.e. there is an improvement in the task when there is this mapping.

Briefings in
Bioinformatics

[Article Navigation](#)

JOURNAL ARTICLE

Progress and opportunities of foundation models in bioinformatics 

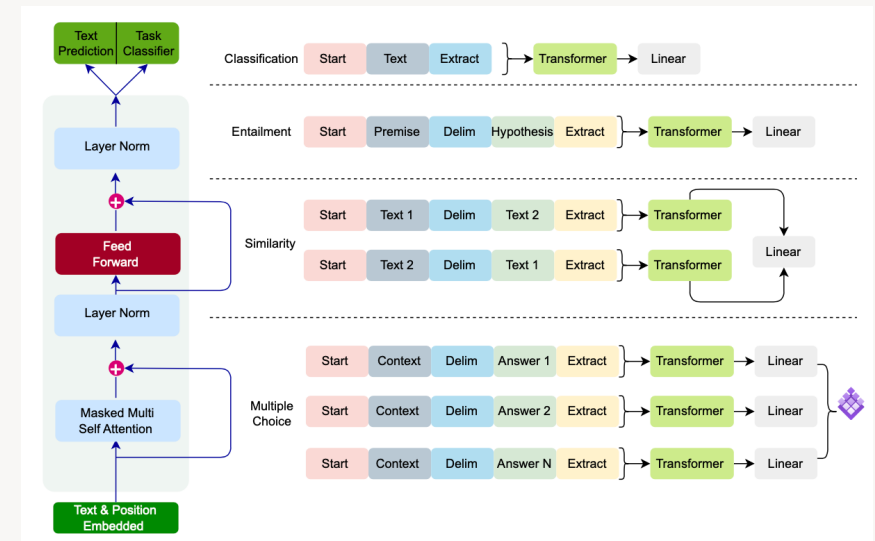
Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang , Le Song , Yu Li 



Model adaptation and downstream tasks

UNIVERSIDAD
DE VALENCIA

- FMs were created around NLP and language modelling.
- Behind any foundational model is the idea of transfer learning.
- With a foundational model we are training the model in a basic and general task, and we use this to adapt it to a slightly different task, within the same domain, e.g., language.
- Pretraining refers to the creation of the foundational model.
- Fine-tuning refers to model adaptation to a specific downstream task.



arXiv

Computer Science > Computation and Language

[Submitted on 11 May 2023 (v1), last revised 21 May 2023 (this version, v2)]

Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, Thippa Reddy Gadekallu



How did we get here?

Representing text documents/fragments

UNIVERSIDAD
DE MURCIA

- **Bag-of-words:** no explicit representation of the sequence, just tokens (i.e., words) in a bag.
 - If there are $|V|$ elements in V , then any document is represented by $|V|$ elements, and the i -th element is the count of how many times the i -th element in V appears at the document.
- **Static word embeddings:** for each word, it focuses on the word surroundings.
 - The representation is still a vector of real numbers, between 256 and 1024.
 - Creates vectors that are similar for words with similar context (i.e., the words that likely appear around them).
 - Static as there is a fixed representation for each token (see Word2vec and Glove).
- **Contextual embeddings:** We still assume that the meaning of a word is determined by its context, but its representation should be adapted to each particular occurrence.
 - In this case, a word generates an embedding that depends on the input sequence in which it appears.

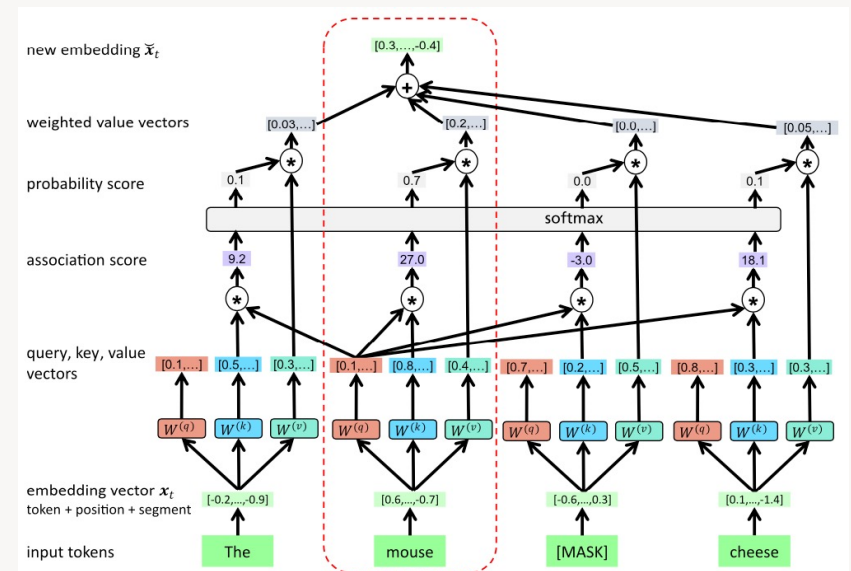


BERT (Bidirectional Encoder Representations from Transformers, 2018)

UNIVERSIDAD
DE

BERT incorporates the **attention mechanism** introduced by Vaswani and colleagues in their landmark 2017 paper, *Attention Is All You Need*. The following diagram

position embeddings	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
segment embeddings	x_A	x_A	x_A	x_A	x_A	x_A	x_B	x_B	x_B	x_B	x_B
token embeddings x_t	$x_{[CLS]}$	x_{my}	x_{dog}	x_{is}	$x_{[MASK]}$	$x_{[SEP]}$	x_{he}	x_{likes}	x_{play}	$x_{\#ing}$	$x_{[SEP]}$
input tokens v_t	[CLS]	my	dog	is	[MASK]	[SEP]	he	likes	play	##ing	[SEP]

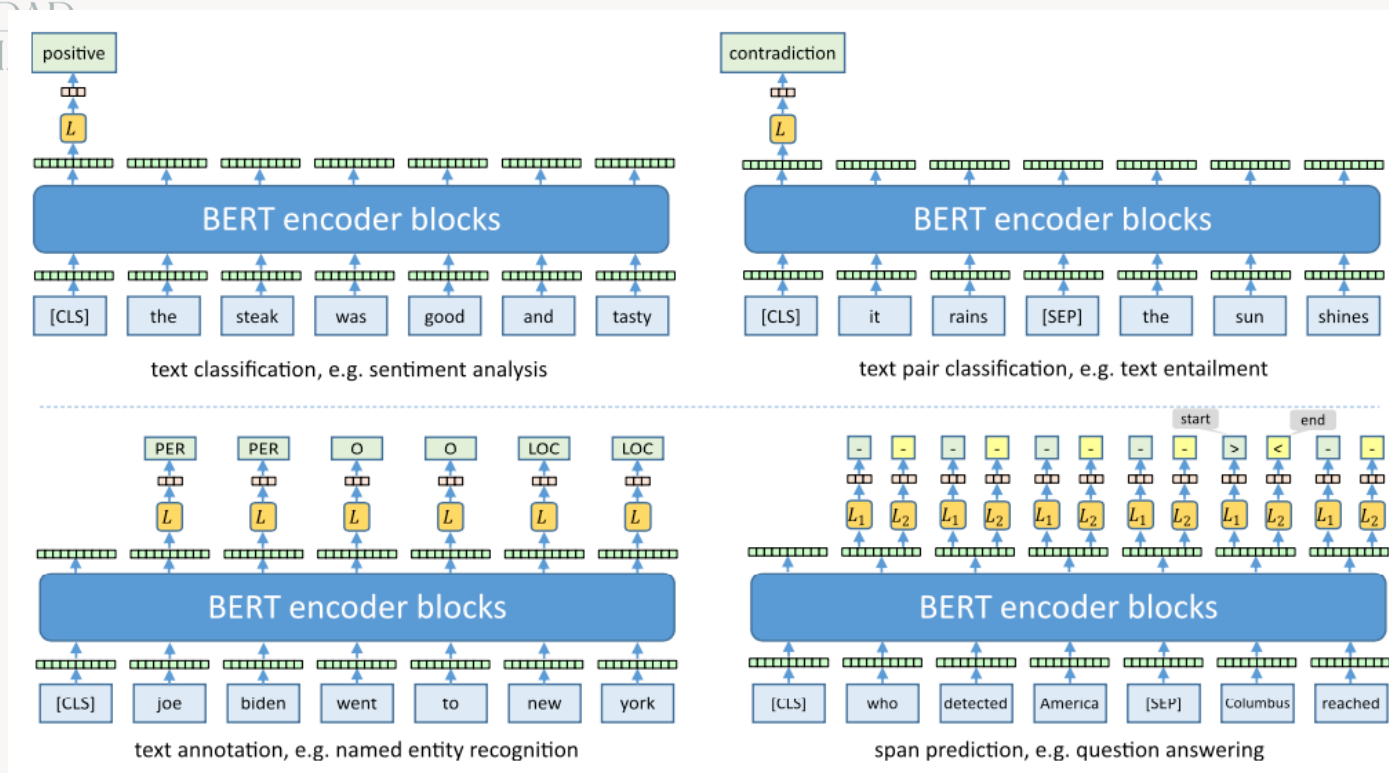


Fixed encoding for the word, a position embedding and a segment embedding (i.e. you are either on left of right of the masked word) at inference.



UNIVERSIDAD
DE MURCIA

BERT (& II), same encoding, different tasks





Foundation models in medicine

UNIVERSIDAD
DE MURCIA

FMs in medicine are FMs trained on data related to human disease.

- **Patient trajectory foundation models:** based on medical data, and electronic health records. Models like Delphi (Shmatko et al., 2025), Foresight 1 & 2, and the latest, Foresight England (Ellershaw et al., 2026) are trained on patient trajectories.
- **FMs on biological/experimental data:** biological data underlying the processes that may generate disease, e.g., molecular data, i.e. foundational models in biology.
- **FMs based on Knowledge:** medical ontologies like SNOMED-CT, disease ontologies like HPO, the Gene Ontology describing genes and how they organise and act within the cell.
 - And naturally we have a specific type of neural network, GNNs, to deal with this type of data. GNNs produce embeddings for graphs too.



Foundation models in medicine

UNIVERSIDAD
DE MURCIA

FMs in medicine are FMs trained on data related to human disease.

- **Patient trajectory foundation models:** based on medical data, and electronic health records. Models like Delphi (Shmatko et al., 2025), Foresight 1 & 2, and the latest, Foresight England (Ellershaw et al., 2026) are trained on patient trajectories.
- **FMs on biological/experimental data:** biological data underlying the processes that may generate disease, e.g., molecular data, i.e. foundational models in biology.
- **FMs based on Knowledge:** medical ontologies like SNOMED-CT, disease ontologies like HPO, the Gene Ontology describing genes and how they organise and act within the cell.
 - And naturally we have a specific type of neural network, GNNs, to deal with this type of data. GNNs produce embeddings for graphs too.



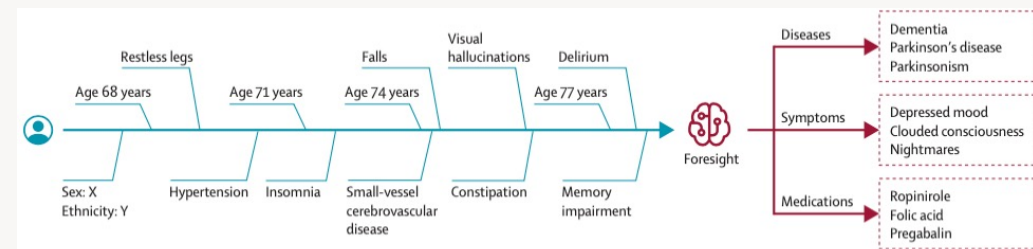
Patient trajectory foundation models

Trajectories can be formed with any information that describes a person's state of health,

- your medical history,
- the provision of health services or
- the payment for such services, created or used by
 - professionals, centres, laboratories, insurers, researchers, devices or by the patient himself.

Examples:

- Identifiers and demographics: name, date of birth, medical history number.
- Clinical record: diagnoses, procedures, constants, allergies, medication, lists of problems, clinical notes.
- Diagnostic: laboratory results, pathology reports, imaging (DICOM), signals (e.g., ECG)
- Administrative: complaints, invoicing, authorizations.
- Devices and sensors: wearables, home monitors, medical devices.
- Genetics, omics and biomarkers. Data reported by the patient.



THE LANCET
Digital Health

[This journal](#) [Journals](#) [Publish](#) [Clinical](#) [Global health](#) [Multimedia](#) [Events](#) [About](#)

ARTICLES Volume 6, Issue 4, E281-E290, April 2024 [Open Access](#) [Download Full Issue](#)

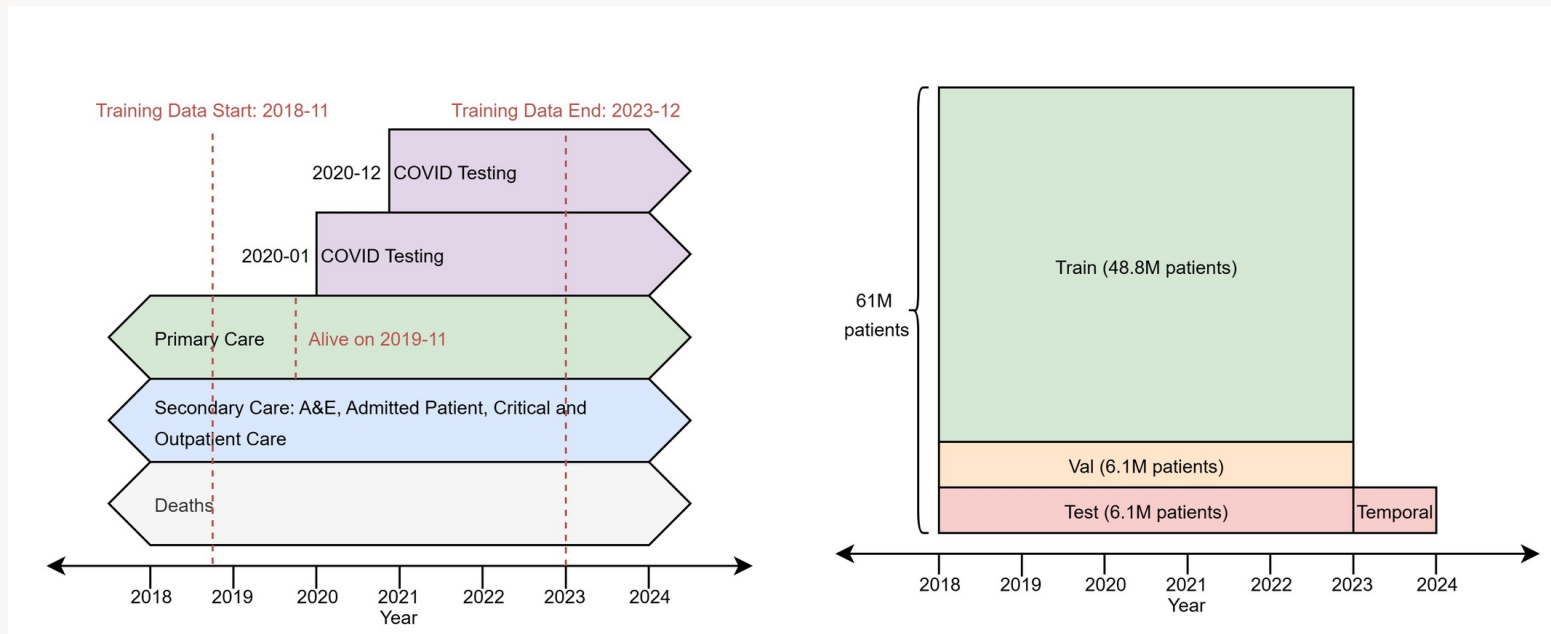
Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study

Zeljko Kraljevic, MSc ^{1,2}, Dan Bean, PhD ^{1,2}, Anthony Shek, PhD ^{1,2}, Rebecca Bendayan, PhD ^{1,2}, Prof Harry Hemingway, PhD ^{1,2}, Joshua Au Yeung, MBBS ^{1,2} et al. [Show more](#)



UNIVERSIDAD
DE MURCIA

Patient trajectory FMs: E-foresight



Foresight-England: Development of a National-Scale Generative AI Model of Electronic Health Records for Medical Event Prediction across the COVID-19 Pandemic

Simon Ellershaw^{4,2} Christopher Tomlinson^{4,5,3,3,4} Zeljko Kraljevic³
Spiros Denaxas^{1,4,5,6,7} Harry Hemingway^{1,6,7} Cathie Sudlow⁸
Angela M. Wood^{6,9,10,11,12,13,14} Anoop D. Shah^{1,4,15} Richard Dobson^{1,2,4,7,16}



Foundation models in medicine

UNIVERSIDAD
DE MURCIA

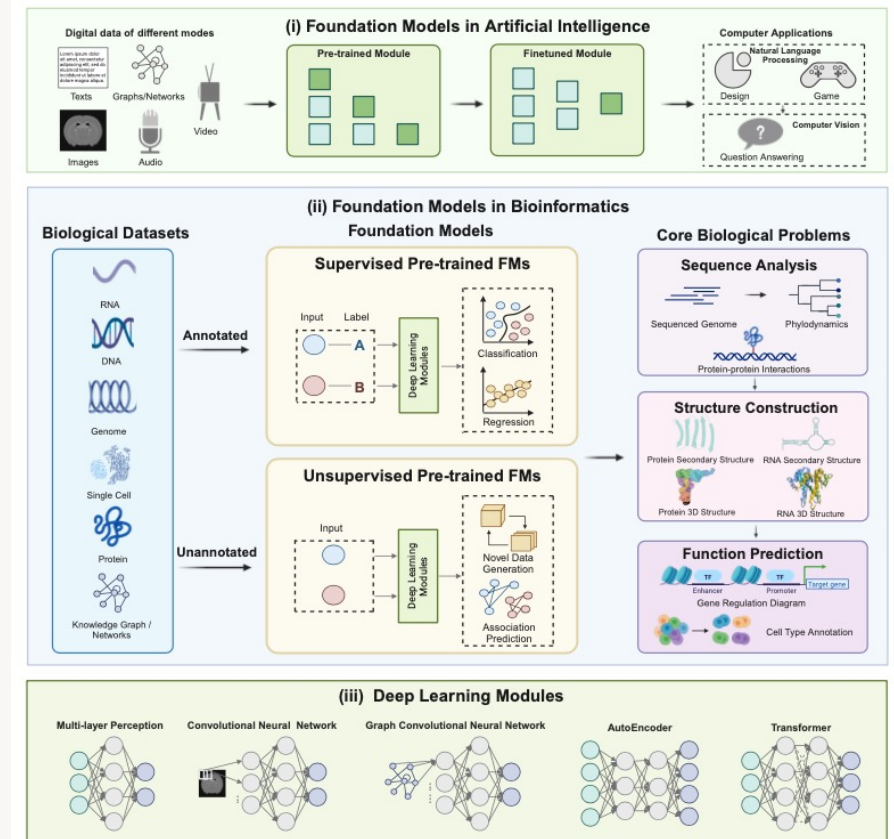
FMs in medicine are FMs trained on data related to human disease.

- **Patient trajectory foundation models:** based on medical data, and electronic health records. Models like Delphi (Shmatko et al., 2025), Foresight 1 & 2, and the latest, Foresight England (Ellershaw et al., 2026) are trained on patient trajectories.
- **FMs on biological/experimental data:** biological data underlying the processes that may generate disease, e.g., molecular data, i.e. foundational models in biology.
- **FMs based on Knowledge:** medical ontologies like SNOMED-CT, disease ontologies like HPO, the Gene Ontology describing genes and how they organise and act within the cell.
 - And naturally we have a specific type of neural network, GNNs, to deal with this type of data. GNNs produce embeddings for graphs too.



Foundation models from experimental data

- By leveraging the sequential/linear nature of information disposal biology, we can reuse with little practically no effort, all technological advances made in the NLP domain.
 - DNA is disposed in a double helix, but each thread within the helix is linear in nature. And so it is the transcription process that generates RNA from DNA. The global picture is there, in figure 1 of (Li et al., 2024)
- There are three main task groups we can focus on:
 - study of the sequence to better understand it
 - prediction of structure,
 - prediction of function.





UNI
DE

The link between biology and disease

Level	Variable functions and effects		Abnormalities in PAH deficiency	Treatment strategies in PAH deficiency
Genotype				
DNA	Genetic and epigenetic variants		Variants in the PAH gene	Potential gene therapy
Transcript				
mRNA	Transcript quantity and sequence		Abnormal transcript amount and/or sequence	Potential mRNA replacement therapy
Different phenotype levels				
Protein	Protein quantity, structure and function		Abnormal amount, stability, activity and/or regulation of PAH enzyme	Pharmaceutical supplementation of the cofactor tetrahydrobiopterin stabilizes mutant proteins
Multimer	Structural interactions		Some variants interfere with active tetramer formation	

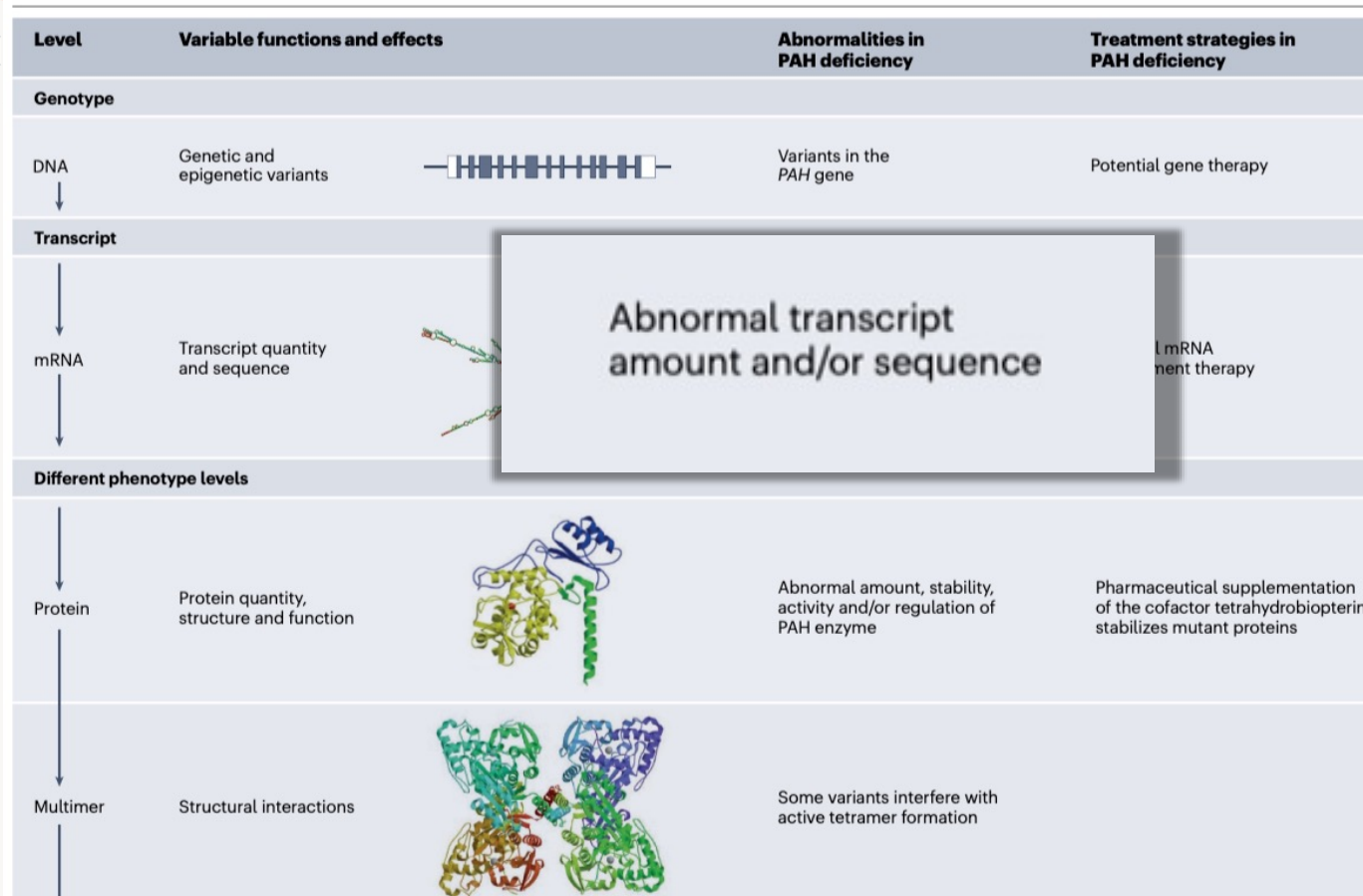
In all these information levels, we can work on useful foundation models





UNIDE

The link between biology and disease



In all these information levels, we can work on useful foundation models



Types of data from where to know about gene regulation to modulate transcript production

UNIVERSIDAD
DE MURCIA

Measuring

- Transcribed regions
- Open Chromatin
- Regions associated with transcription factors
- Regions associated with active promoters, enhancers, transcriptional repressors

While considering

- Tissue, cell and state specific measures

How is chromatin organised in 3D

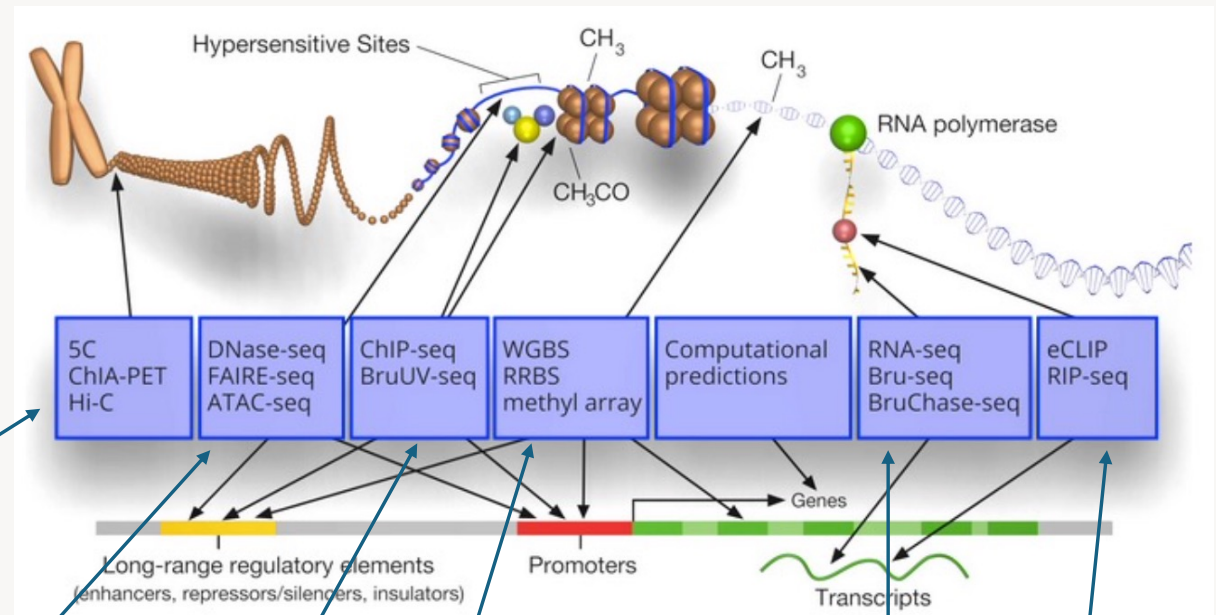
Is chromatin accessible?

Where are active promoters and enhancers? Where is transcription repressed?

Is DNA methylated?

Is the gene really active?

Which proteins are working with the gene when transcribing?



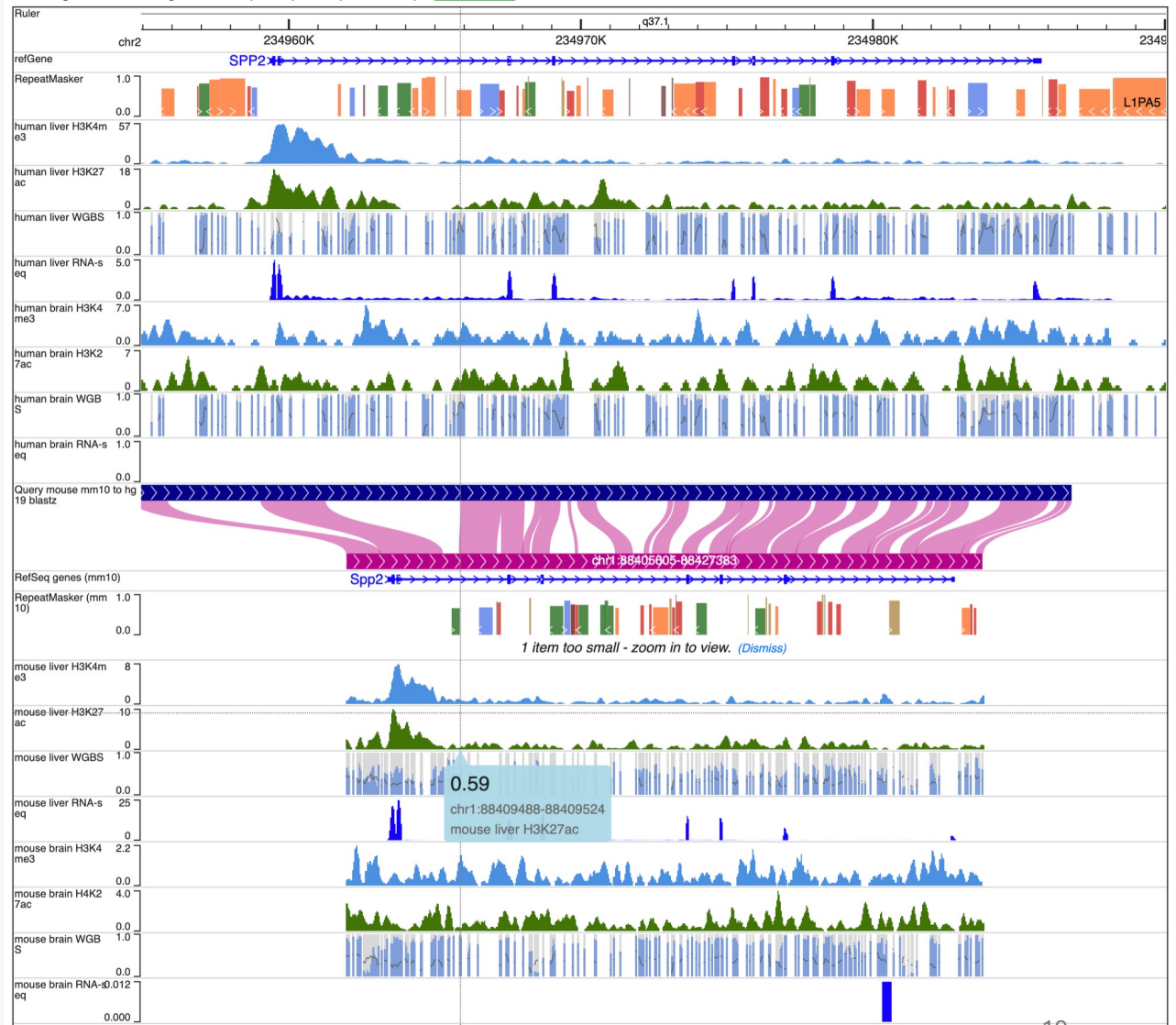
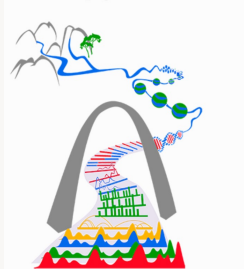


UNIVERSIDAD
DE MURCIA

How does this data look like?

- Multiple data tracks on the same genome locus
- Many species
- Many tissues
- Many experimental data types

WashU Epigenome Browser





UNI
DE

The link between biology and disease

Level	Variable functions and effects		Abnormalities in PAH deficiency	Treatment strategies in PAH deficiency
Genotype				
DNA	Genetic and epigenetic variants		Variants in the PAH gene	Potential gene therapy
Transcript				
mRNA	Transcript quantity and sequence		Abnormal transcript amount and/or sequence	Potential mRNA replacement therapy
Different phenotype levels				
Protein	Protein quantity, structure and function		Abnormal amount, stability, activity and/or regulation of PAH enzyme	Pharmaceutical supplementation of the cofactor tetrahydrobiopterin stabilizes mutant proteins
Multimer	Structural interactions		Some variants interfere with active tetramer formation	

In all these information levels, we can work on useful foundation models





UNI
DE

The link between biology and disease

Level	Variable functions and effects		Abnormalities in PAH deficiency	Treatment strategies in PAH deficiency
Genotype				
DNA	Genetic and epigenetic variants		Variants in the PAH gene	Potential gene therapy
Transcript				
mRNA	Transcript quantity and sequence		Abnormal transcript amount and/or sequence	Potential mRNA replacement therapy
Different				
Protein	Protein quantity, structure and function		Abnormal amount, stability, activity and/or regulation of PAH enzyme	Pharmaceutical supplementation of the cofactor tetrahydrobiopterin stabilizes mutant proteins
Multimer	Structural interactions		Some variants interfere with active tetramer formation	

In all these information levels, we can work on useful foundation models





Predicting protein sequence

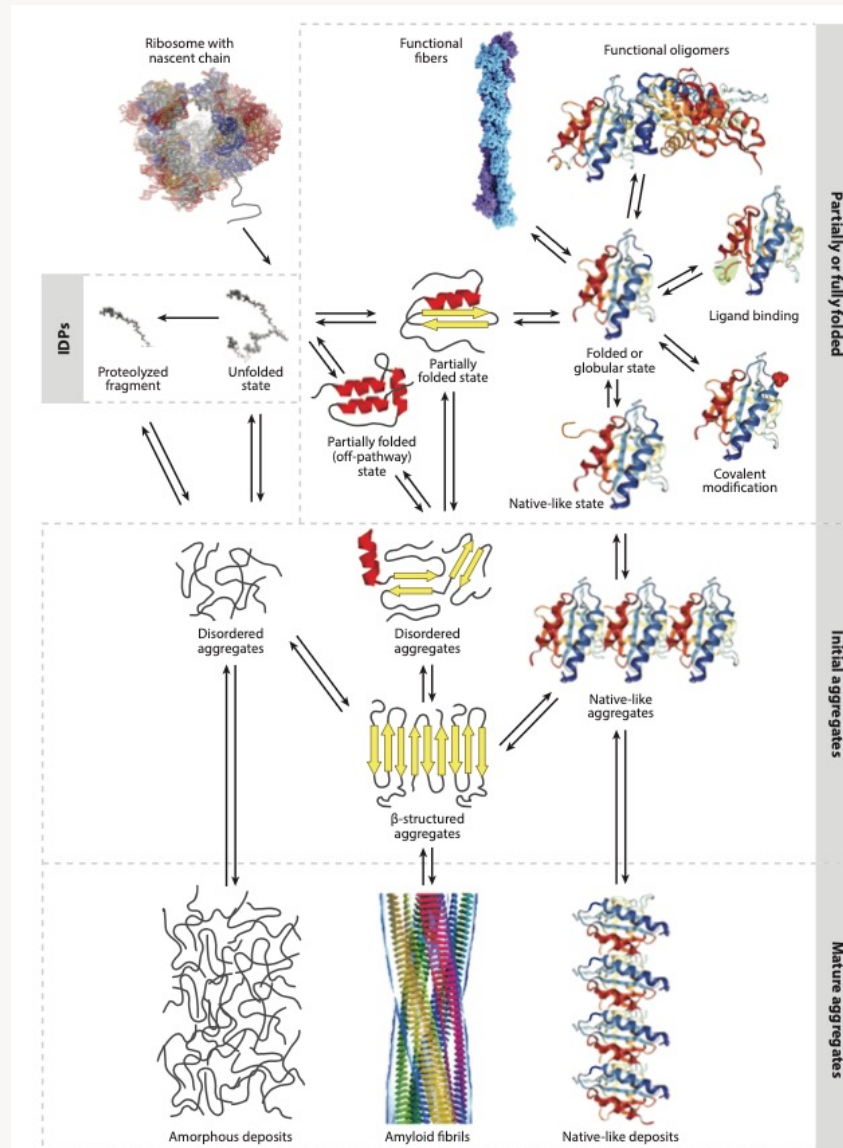
UNIVERSIDAD
DE MURCIA

- **Protein conformational state:** 3D arrangement adopted by a protein molecule.
 - Same sequence can generate different conformations depending on molecules binding to the protein and chemical environment.
- **Protein aggregation state:** describes how the molecules of the same protein associated with one another, e.g.,
 - as monomers (single protein molecule),
 - oligomer (complex formed by a small number of molecules/subunits, as dimers, trimers, tetramers, pentamers).
 - In oligomers, when the number of molecules is larger, we talk about fibrils (larger repetitive structure aggregates of molecules), including amyloid fibrils.
- **Amyloid fibrils:** long structures characterised by beta-sheets as the basic 3D structure
 - Very stable, elongated structures.
 - Examples of proteins leading to this include Tau, APP (amyloid-beta), alpha-syn (SNCA). All of them have very different sequences and functions, yet they convert to cross-beta amyloid architecture (Chiti & Dobson, 2017).

Protein states and spatial organization

UNIVERSIDAD
DE MURCIA

- Normally, protein sequence and biological environment contribute to their soluble state
- Sometimes they convert into nonfunctional and potentially damaging states
- When leading to amyloid fibrils within the cell it can cause cellular dysfunction and disease states like in Alzheimer's, Parkinson's and Type II diabetes



Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade

Fabrizio Chiti¹ and Christopher M. Dobson²

¹Department of Experimental and Clinical Biomedical Sciences "Mario Serio," Section of Biochemistry, Università di Firenze, Italy; email: fabrizio.chiti@unifi.it
²Department of Chemistry, Cambridge Amyloid Diseases, University of Cambridge, Cambridge CB2 1EW, United Kingdom; email: cm214@cam.ac.uk

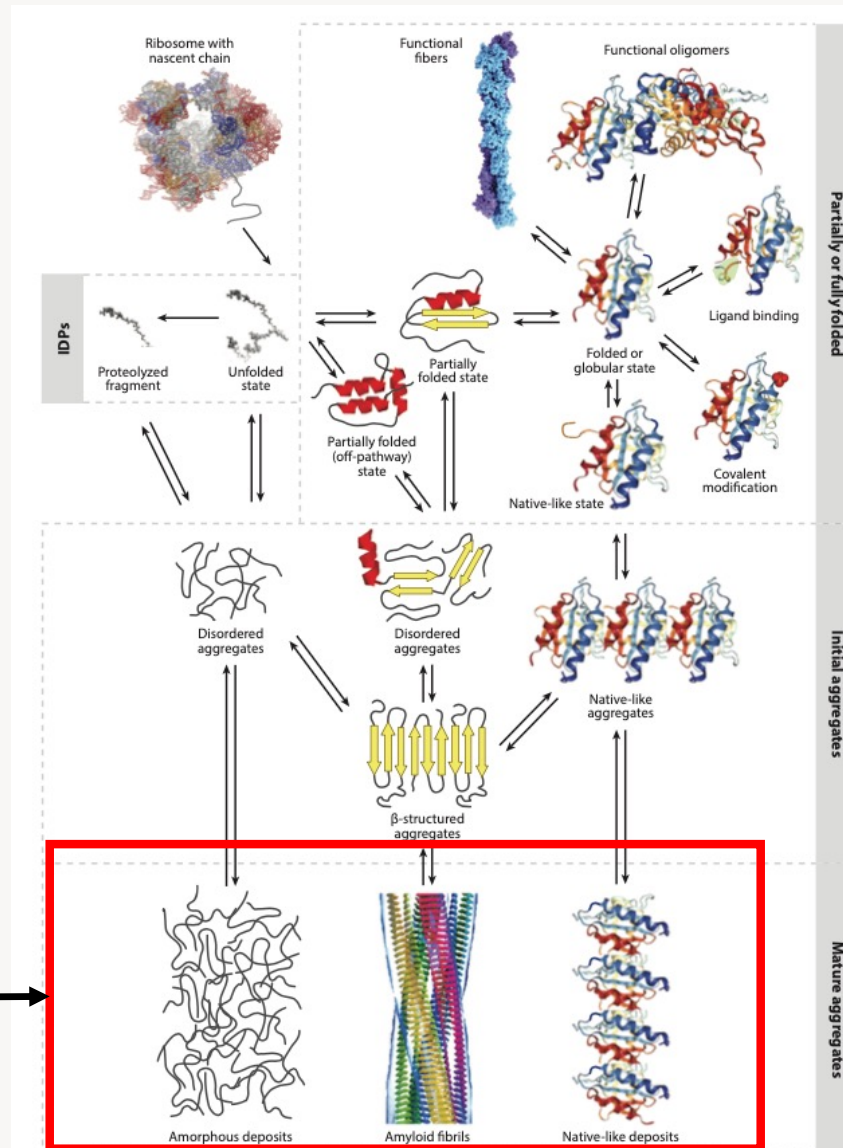


UNIVERSIDAD
DE MURCIA

Protein states and spatial organization

- Normally, protein sequence and biological environment contribute to their soluble state
- Sometimes they convert into nonfunctional and potentially damaging states
- When leading to amyloid fibrils within the cell it can cause cellular dysfunction and disease states like in Alzheimer's, Parkinson's and Type II diabetes

States potentially
associated to
disease



Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade

Fabrizio Chiti¹ and Christopher M. Dobson²

¹Department of Experimental and Clinical Biomedical Sciences "Mario Serio," Section of Biochemistry, Università di Firenze, Italy; email: fabrizio.chiti@unifi.it
²Department of Chemistry, Cambridge Amyloid Disease, University of Cambridge, Cambridge CB2 1EW, United Kingdom; email: cm214@cam.ac.uk

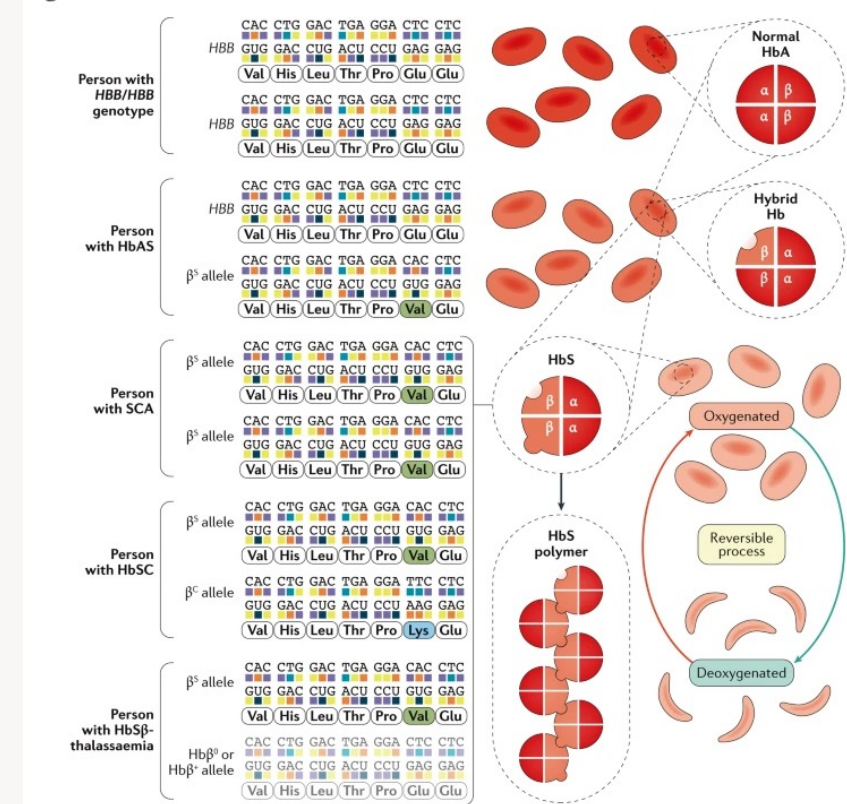
Example A: one mutation leads to sickle cell disease

It is a disease that changes red blood cells to become rigid and sticky, due to a single mutation at the haemoglobin gene HBB.

A mutation leads to the formation of

- polymer fibers,
- red blood cell deformation,
- This bad shape leads to vascular obstruction and disease
- The normal aggregation state of haemoglobin is a tetramer, therefore is an oligomeric protein composed of four globin units.

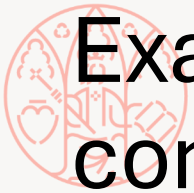
Figure 1: Genetic alterations in *HBB*.



PRIMER

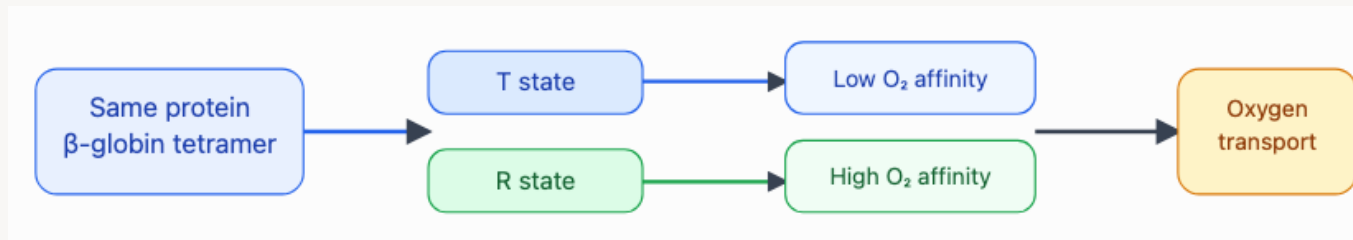
Sickle cell disease

Gregory J. Nairn, Frederick B. Paul, Clotilde D. Rosta, Marilys H. Gaudier, Keshu Chandra Freemont, Lakshmanan Krishnamurti, Wang H. Smith, Julie A. Pampeter, David J. Weatherall, Fernando F. Costa and Elliott P. Vichinsky



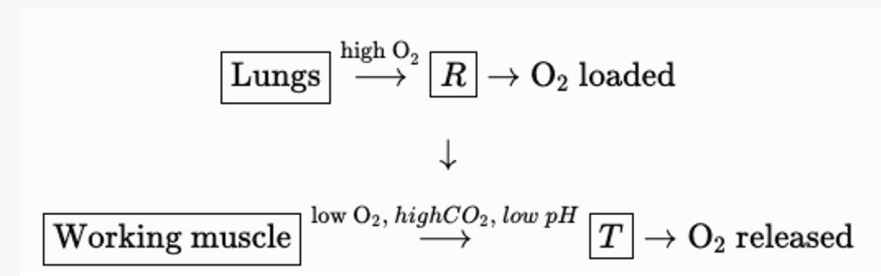
UNIVERSIDAD
DE MURCIA

Example B: no mutation, different conformations



We should account for many factors, not only sequence.

- Erythrocytes, the cells containing haemoglobin, pass through different places, lungs, blood, muscle,, and then lungs again.
- In lungs, the environment favours an R state due to oxygen abundance.
- If the muscle is exercising, there is a less O₂ abundance, making haemoglobin switching from R to T, to release more O₂.





Same protein leading to different folding can lead to different...

UNIVERSIDAD
DE MURCIA

- Protein-protein interactions
- Membrane interactions
- Stability
- Cellular uptake
- Seeding
- Propagation
- Toxicity
- Antibody recognition



The protein folding problem

UNIVERSIDAD
DE CHILE

Defined as the question of how a protein's amino acid sequence dictates its three-dimensional atomic structure (Dill et al., 2008). This can be decomposed into three subproblems

- The folding code: what balance of interatomic forces in thermodynamics lead to the structure of the protein for a given amino acid sequence
- Protein structure prediction: the prediction of the protein's native structure from the sequence
- Folding: from the perspective of kinetics, what routes or pathways some proteins use to fold so quickly

ANNUAL REVIEW OF BIOPHYSICS Volume 37, 2008

Review Article

The Protein Folding Problem

Ken A. Dill^{1,2}, S. Banu Ozkan³, M. Scott Shell⁴ and Thomas R. Weikel⁵

 View Affiliations and Author Notes

28

Vol. 37:289-316 (Volume publication date June 2008) | <https://doi.org/10.1146/annurev.biophys.37.092707.153558>



How does this data look like?

UNIVERSIDAD
DE MURCIA

- The SNCA gene leads to α -synuclein, 140 aa
- Aminoacid chain:
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTK
EGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAA
ATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGY
QDYEPEA
- In the Protein Data Bank, the experimental data about it would include (Nitrogen, alpha Carbon and Carbon atoms)

Residue	Atom	x	y	z
MET 1	N	10.42	14.73	8.21
MET 1	CA	11.31	15.48	7.35
MET 1	C	12.62	14.77	7.12

<https://www.rcsb.org/structure/9A1A>



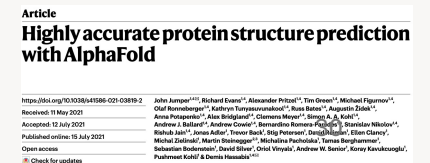
RCSB PDB



UNIVERSIDAD
DE MURCIA

AlphaFold foundation model

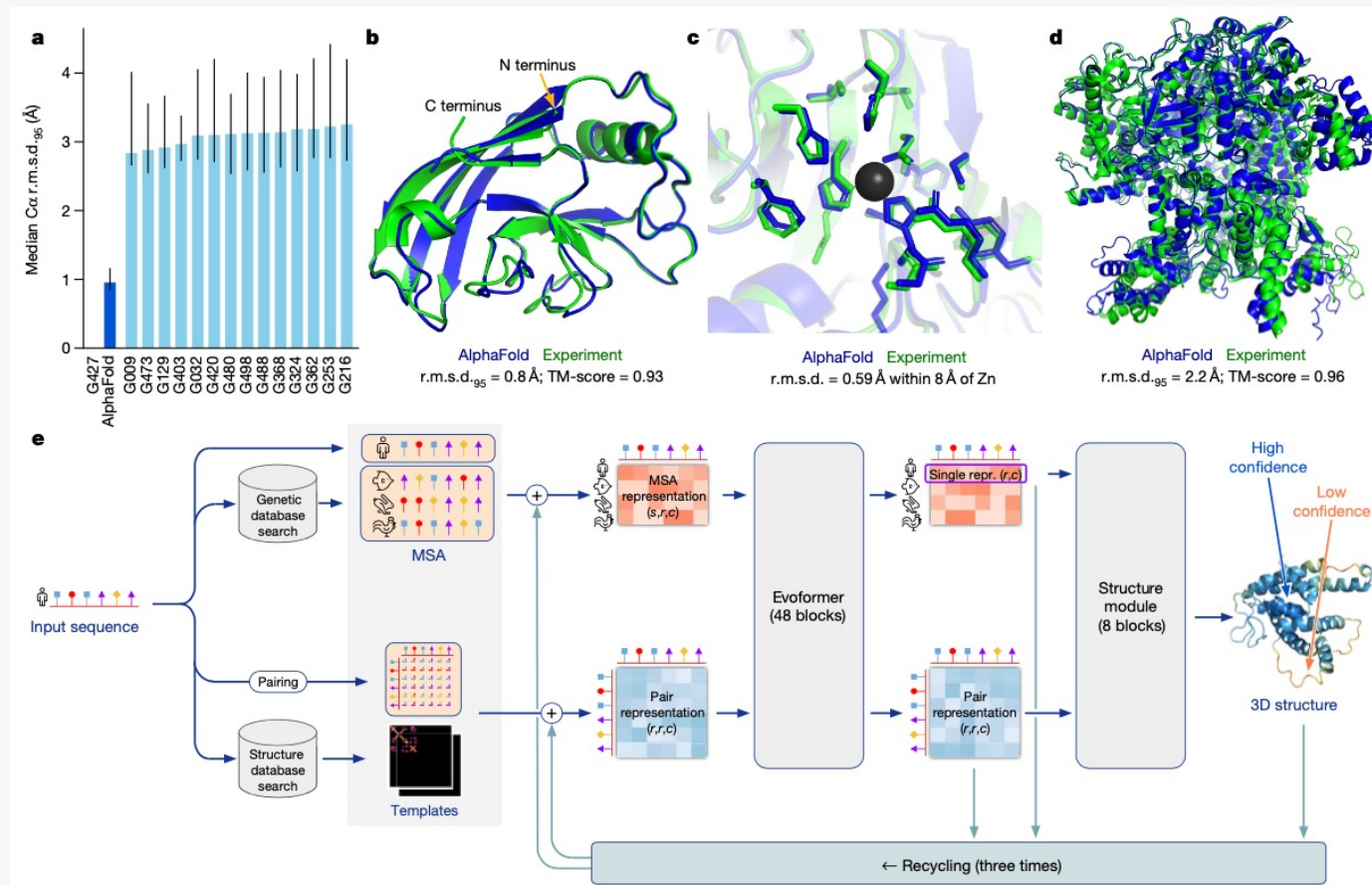
- AlphaFold predicts the 3D coordinates for all the heavy atoms for a given protein, based on
 - the primary amino acid sequence and aligned sequences of homologues
 - Working at atom level, it looks at pairs of residues.
- If two residues A and B, are closed together in the folded protein, and we had a mutation in A, then evolution would favour sequences with a compensatory mutation in B
 - Two close residues should co-vary. Therefore, this might help predicting structure as A and B interact.
- AlphaFold uses such evolutionary information together with templates (i.e., known 3D structures) from homologous proteins to infer structure. And its output is, for a sequence of L residues, a tensor of $L \times N \text{ atoms per residue} \times 3$, being the 3 dimensions x, y and z coordinates.





UNIVERSIDAD
DE MURCIA

High prediction accuracy for folded proteins



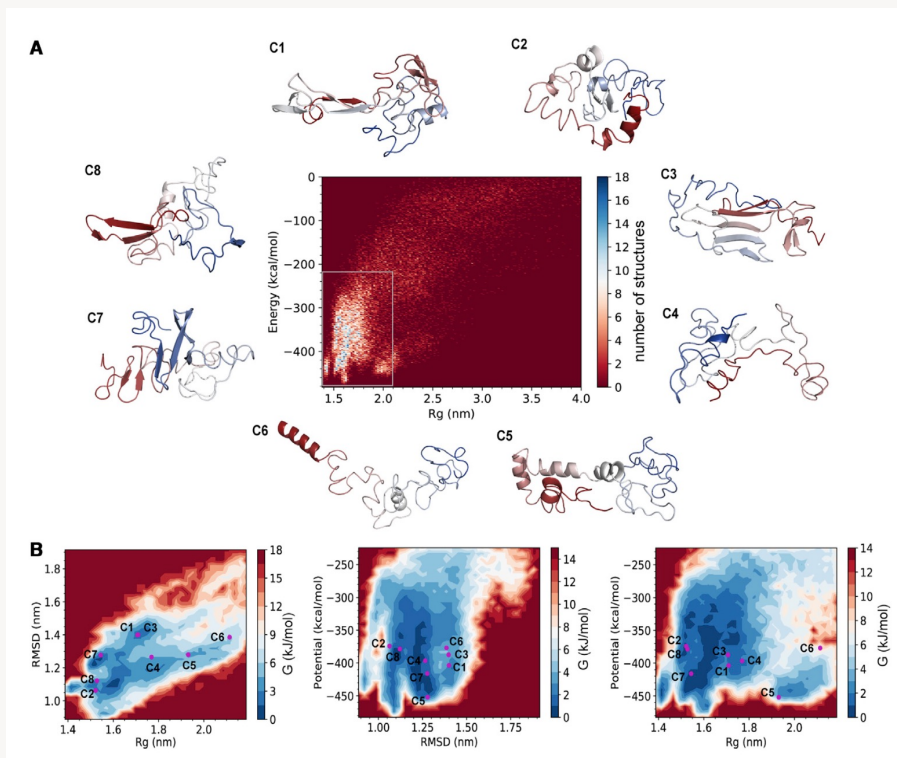


Why Parkinson's disease is so hard to investigate with foundation models

UNIVERSIDAD
DE MURCIA

- Parkinson is a disease that belongs to a family of diseases called **alphasynucleopathies**
 - characterized by the pathological accumulation of misfolded and aggregated alpha-synuclein in the nervous system.
 - DLB and MSA are the other.
- This SNCA protein, **140 amino acids, is not associated with a single 3D structure**
- It is an IDP (intrinsically disordered protein), which means that in the cytosol it does not show a stable tertiary structure, instead it can adopt many interconverting conformations.
- States influence behaviour.
- When bind to cell membranes, half of the disordered monomer transforms into an alpha-helix-like protein
- When interacting with other alpha-syn molecules, it forms oligomers (multimers): dimers, trimers and larger assemblies, this generates a variety of structures and, hence, functions
- When compounding a fibril, i.e., many, many molecules at the same time, there is also fibril polymorphism.

The experimental evidence



CellPress

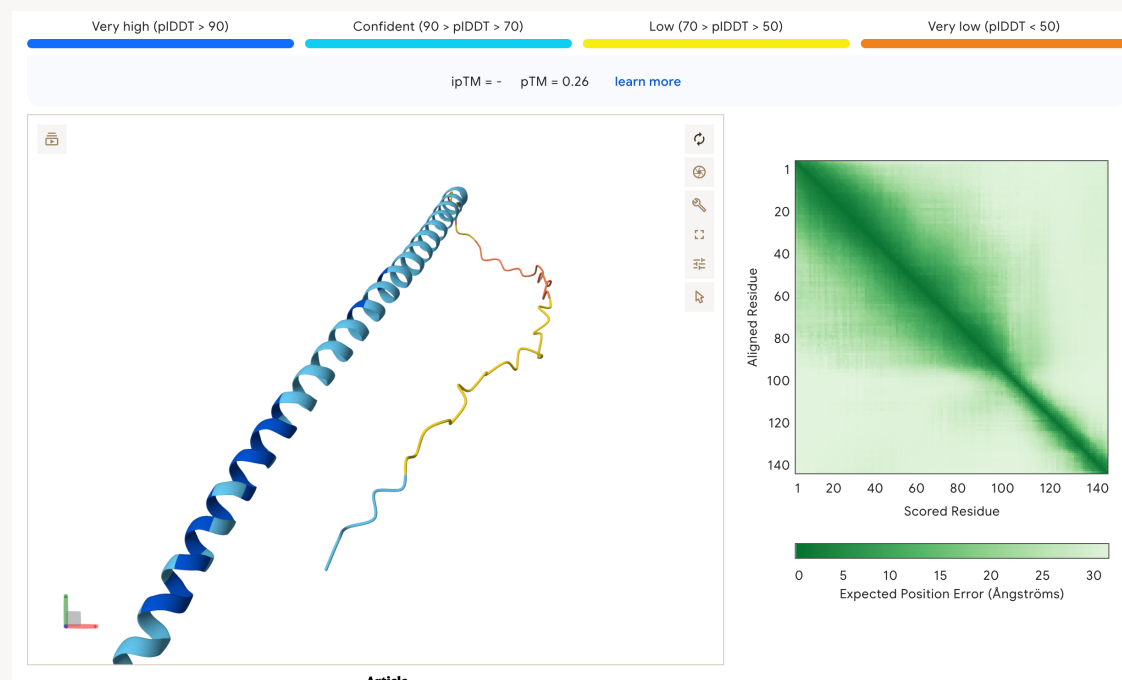
Structure

Article

The structural heterogeneity of α -synuclein is governed by several distinct subpopulations with interconversion times slower than milliseconds

Jiaxing Chen,^{1,2} Sofia Zaer,^{1,2} Paz Droni,² Joanna Zamel,² Khalil Joroni,² Nir Kallisman,² Eitan Lerner,^{1,2,3,4} and Nikolay V. Dokholyan^{1,2,3,4}

The AlphaFold prediction



Article

Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

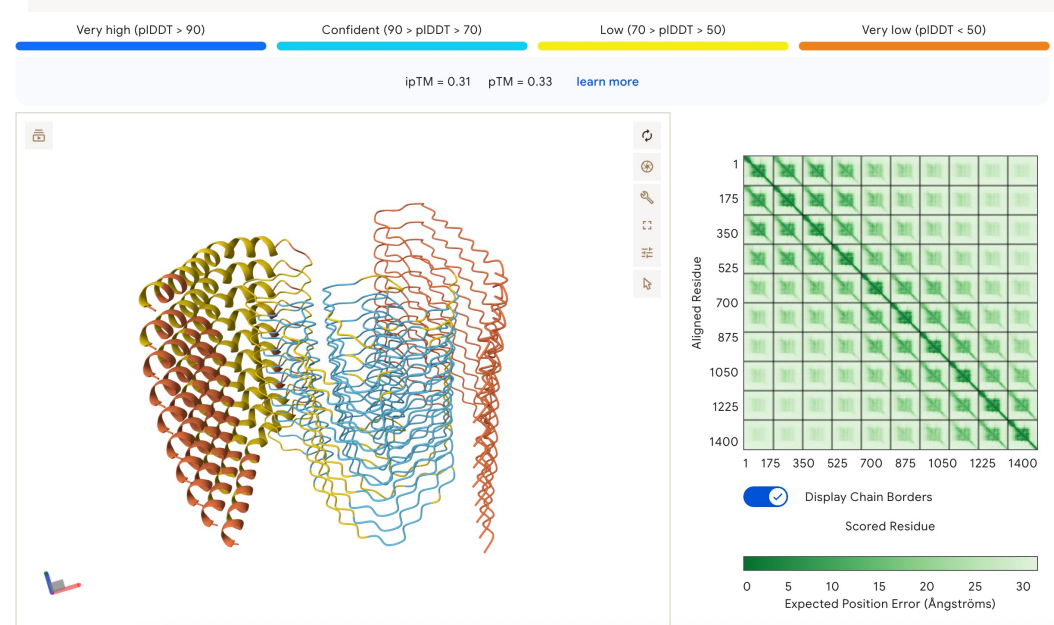
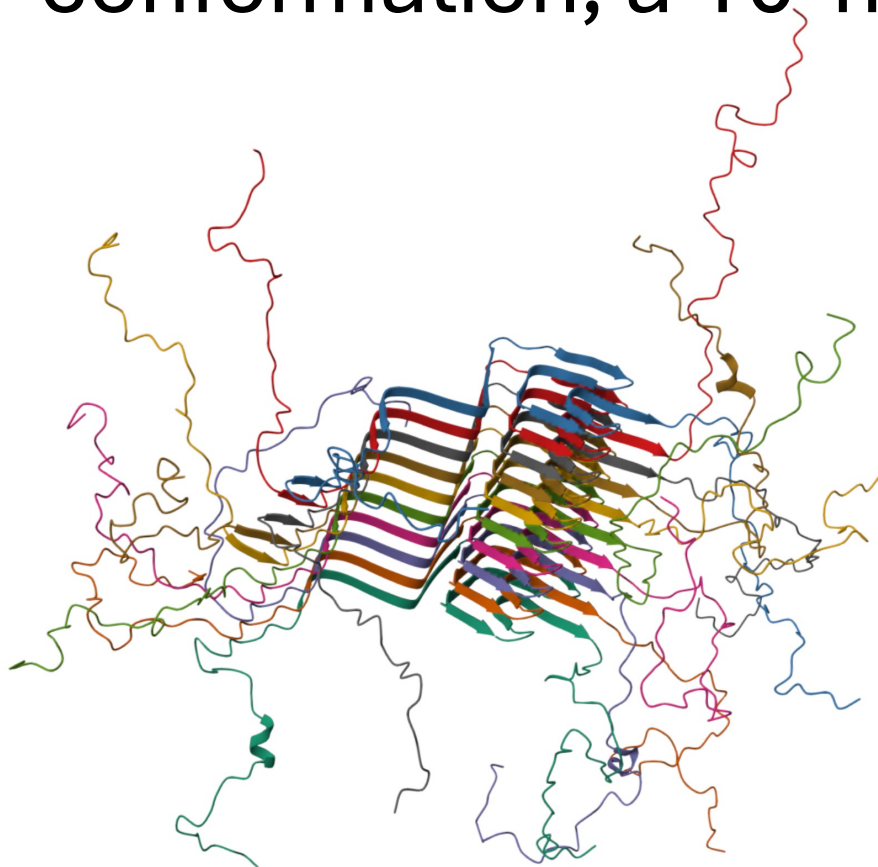
Open access

[Check for updates](#)

John Jumper^{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,135,136,137,138,139,140,141,142,143,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159,160,161,162,163,164,165,166,167,168,169,170,171,172,173,174,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,203,204,205,206,207,208,209,210,211,212,213,214,215,216,217,218,219,220,221,222,223,224,225,226,227,228,229,230,231,232,233,234,235,236,237,238,239,240,241,242,243,244,245,246,247,248,249,250,251,252,253,254,255,256,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,273,274,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,296,297,298,299,300,301,302,303,304,305,306,307,308,309,310,311,312,313,314,315,316,317,318,319,320,321,322,323,324,325,326,327,328,329,330,331,332,333,334,335,336,337,338,339,340,341,342,343,344,345,346,347,348,349,350,351,352,353,354,355,356,357,358,359,360,361,362,363,364,365,366,367,368,369,370,371,372,373,374,375,376,377,378,379,380,381,382,383,384,385,386,387,388,389,390,391,392,393,394,395,396,397,398,399,400,401,402,403,404,405,406,407,408,409,410,411,412,413,414,415,416,417,418,419,420,421,422,423,424,425,426,427,428,429,430,431,432,433,434,435,436,437,438,439,440,441,442,443,444,445,446,447,448,449,450,451,452,453,454,455,456,457,458,459,460,461,462,463,464,465,466,467,468,469,470,471,472,473,474,475,476,477,478,479,480,481,482,483,484,485,486,487,488,489,490,491,492,493,494,495,496,497,498,499,500,501,502,503,504,505,506,507,508,509,510,511,512,513,514,515,516,517,518,519,520,521,522,523,524,525,526,527,528,529,530,531,532,533,534,535,536,537,538,539,540,541,542,543,544,545,546,547,548,549,550,551,552,553,554,555,556,557,558,559,560,561,562,563,564,565,566,567,568,569,570,571,572,573,574,575,576,577,578,579,580,581,582,583,584,585,586,587,588,589,590,591,592,593,594,595,596,597,598,599,600,601,602,603,604,605,606,607,608,609,610,611,612,613,614,615,616,617,618,619,620,621,622,623,624,625,626,627,628,629,630,631,632,633,634,635,636,637,638,639,640,641,642,643,644,645,646,647,648,649,650,651,652,653,654,655,656,657,658,659,660,661,662,663,664,665,666,667,668,669,670,671,672,673,674,675,676,677,678,679,680,681,682,683,684,685,686,687,688,689,690,691,692,693,694,695,696,697,698,699,700,701,702,703,704,705,706,707,708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725,726,727,728,729,730,731,732,733,734,735,736,737,738,739,740,741,742,743,744,745,746,747,748,749,750,751,752,753,754,755,756,757,758,759,760,761,762,763,764,765,766,767,768,769,770,771,772,773,774,775,776,777,778,779,780,781,782,783,784,785,786,787,788,789,790,791,792,793,794,795,796,797,798,799,800,801,802,803,804,805,806,807,808,809,810,811,812,813,814,815,816,817,818,819,820,821,822,823,824,825,826,827,828,829,830,831,832,833,834,835,836,837,838,839,840,841,842,843,844,845,846,847,848,849,850,851,852,853,854,855,856,857,858,859,860,861,862,863,864,865,866,867,868,869,870,871,872,873,874,875,876,877,878,879,880,881,882,883,884,885,886,887,888,889,890,891,892,893,894,895,896,897,898,899,900,901,902,903,904,905,906,907,908,909,910,911,912,913,914,915,916,917,918,919,920,921,922,923,924,925,926,927,928,929,930,931,932,933,934,935,936,937,938,939,940,941,942,943,944,945,946,947,948,949,950,951,952,953,954,955,956,957,958,959,960,961,962,963,964,965,966,967,968,969,970,971,972,973,974,975,976,977,978,979,980,981,982,983,984,985,986,987,988,989,990,991,992,993,994,995,996,997,998,999,1000,1001,1002,1003,1004,1005,1006,1007,1008,1009,1010,1011,1012,1013,1014,1015,1016,1017,1018,1019,1020,1021,1022,1023,1024,1025,1026,1027,1028,1029,1030,1031,1032,1033,1034,1035,1036,1037,1038,1039,1040,1041,1042,1043,1044,1045,1046,1047,1048,1049,1050,1051,1052,1053,1054,1055,1056,1057,1058,1059,1060,1061,1062,1063,1064,1065,1066,1067,1068,1069,1070,1071,1072,1073,1074,1075,1076,1077,1078,1079,1080,1081,1082,1083,1084,1085,1086,1087,1088,1089,1090,1091,1092,1093,1094,1095,1096,1097,1098,1099,1100,1101,1102,1103,1104,1105,1106,1107,1108,1109,1110,1111,1112,1113,1114,1115,1116,1117,1118,1119,1120,1121,1122,1123,1124,1125,1126,1127,1128,1129,1130,1131,1132,1133,1134,1135,1136,1137,1138,1139,1140,1141,1142,1143,1144,1145,1146,1147,1148,1149,1150,1151,1152,1153,1154,1155,1156,1157,1158,1159,1160,1161,1162,1163,1164,1165,1166,1167,1168,1169,1170,1171,1172,1173,1174,1175,1176,1177,1178,1179,1180,1181,1182,1183,1184,1185,1186,1187,1188,1189,1190,1191,1192,1193,1194,1195,1196,1197,1198,1199,1200,1201,1202,1203,1204,1205,1206,1207,1208,1209,1210,1211,1212,1213,1214,1215,1216,1217,1218,1219,1220,1221,1222,1223,1224,1225,1226,1227,1228,1229,1230,1231,1232,1233,1234,1235,1236,1237,1238,1239,1240,1241,1242,1243,1244,1245,1246,1247,1248,1249,1250,1251,1252,1253,1254,1255,1256,1257,1258,1259,1260,1261,1262,1263,1264,1265,1266,1267,1268,1269,1270,1271,1272,1273,1274,1275,1276,1277,1278,1279,1280,1281,1282,1283,1284,1285,1286,1287,1288,1289,1290,1291,1292,1293,1294,1295,1296,1297,1298,1299,1300,1301,1302,1303,1304,1305,1306,1307,1308,1309,1310,1311,1312,1313,1314,1315,1316,1317,1318,1319,1320,1321,1322,1323,1324,1325,1326,1327,1328,1329,1330,1331,1332,1333,1334,1335,1336,1337,1338,1339,1340,1341,1342,1343,1344,1345,1346,1347,1348,1349,1350,1351,1352,1353,1354,1355,1356,1357,1358,1359,1360,1361,1362,1363,1364,1365,1366,1367,1368,1369,1370,1371,1372,1373,1374,1375,1376,1377,1378,1379,1380,1381,1382,1383,1384,1385,1386,1387,1388,1389,1390,1391,1392,1393,1394,1395,1396,1397,1398,1399,1400,1401,1402,1403,1404,1405,1406,1407,1408,1409,1410,1411,1412,1413,1414,1415,1416,1417,1418,1419,1420,1421,1422,1423,1424,1425,1426,1427,1428,1429,1430,1431,1432,1433,1434,1435,1436,1437,1438,1439,1440,1441,1442,1443,1444,1445,1446,1447,1448,1449,1450,1451,1452,1453,1454,1455,1456,1457,1458,1459,1460,1461,1462,1463,1464,1465,1466,1467,1468,1469,1470,1471,1472,1473,1474,1475,1476,1477,1478,1479,1480,1481,1482,1483,1484,1485,1486,1487,1488,1489,1490,1491,1492,1493,1494,1495,1496,1497,1498,1499,1500,1501,1502,1503,1504,1505,1506,1507,1508,1509,1510,1511,1512,1513,1514,1515,1516,1517,1518,1519,1520,1521,1522,1523,1524,1525,1526,1527,1528,1529,1530,1531,1532,1533,1534,1535,1536,1537,1538,1539,1540,1541,1542,1543,1544,1545,1546,1547,1548,1549,1550,1551,1552,1553,1554,1555,1556,1557,1558,1559,1560,1561,1562,1563,1564,1565,1566,1567,1568,1569,1570,1571,1572,1573,1574,1575,1576,1577,1578,1579,1580,1581,1582,1583,1584,1585,1586,1587,1588,1589,1590,1591,1592,1593,1594,1595,1596,1597,1598,1599,1600,1601,1602,1603,1604,1605,1606,1607,1608,1609,1610,1611,1612,1613,1614,1615,1616,1617,1618,1619,1620,1621,1622,1623,1624,1625,1626,1627,1628,1629,1630,1631,1632,1633,1634,1635,1636,1637,1638,1639,1640,1641,1642,1643,1644,1645,1646,1647,1648,1649,1650,1651,1652,1653,1654,1655,1656,1657,1658,1659,1660,1661,1662,1663,1664,1665,1666,1667,1668,1669,1670,1671,1672,1673,1674,1675,1676,1677,1678,1679,1680,1681,1682,1683,1684,1685,1686,1687,1688,1689,1690,1691,1692,1693,1694,1695,1696,1697,1698,1699,1700,1701,1702,1703,1704,1705,1706,1707,1708,1709,1710,1711,1712,1713,1714,1715,1716,1717,1718,1719,1720,1721,1722,1723,1724,1725,1726,1727,1728,1729,1730,1731,1732,1733,1734,1735,1736,1737,1738,1739,1740,1741,1742,1743,1744,1745,1746,1747,1748,1749,1750,1751,1752,1753,1754,1755,1756,1757,1758,1759,1760,1761,1762,1763,1764,1765,1766,1767,1768,1769,1770,1771,1772,1773,1774,1775,1776,1777,1778,1779,1780,1781,1782,1783,1784,1785,1786,1787,1788,1789,1790,1791,1792,1793,1794,1795,1796,1797,1798,1799,1800,1801,1802,1803,1804,1805,1806,1807,1808,1809,1810,1811,1812,1813,1814,1815,1816,1817,1818,1819,1820,1821,1822,1823,1824,1825,1826,1827,1828,1829,1830,1831,1832,1833,1834,1835,1836,1837,1838,1839,1840,1841,1842,1843,1844,1845,1846,1847,1848,1849,1850,1851,1852,1853,1854,1855,1856,1857,1858,1859,1860,1861,1862,1863,1864,1865,1866,1867,1868,1869,1870,1871,1872,1873,1874,1875,1876,1877,1878,1879,1880,1881,1882,1883,1884,1885,1886,1887,1888,1889,1890,1891,1892,1893,1894,1895,1896,1897,1898,1899,1900,1901,1902,1903,1904,1905,1906,1907,1908,1909,1910,1911,1912,1913,1914,1915,1916,1917,1918,1919,1920,1921,1922,1923,1924,1925,1926,1927,1928,1929,1930,1931,1932,1933,1934,1935,1936,1937,1938,1939,1940,1941,1942,1943,1944,1945,1946,1947,1948,1949,1950,1951,1952,1953,1954,1955,1956,1957,1958,1959,1960,1961,1962,1963,1964,1965,1966,1967,1968,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020,2021,2022,2023,2024,2025,2026,2027,2028,2029,2030,2031,2032,2033,2034,2035,2036,2037,2038,2039,2040,2041,2042,2043,2044,2045,2046,2047,2048,2049,2050,2051,2052,2053,2054,2055,2056,2057,2058,2059,2060,2061,2062,2063,2064,2065,2066,2067,2068,2069,2070,2071,2072,2073,2074,2075,2076,2077,2078,2079,2080,2081,2082,2083,2084,2085,2086,2087,2088,2089,2090,2091,2092,2093,2094,2095,2096,2097,2098,2099,2100,2101,2102,2103,2104,2105,2106,2107,2108,2109,2110,2111,2112,2113,2114,2115,2116,2117,2118,2119,2120,2121,2122,2123,2124,2125,2126,2127,2128,2129,2130,2131,2132,2133,2134,2135,2136,2137,2138,2139,2140,2141,2142,2143,2144,2145,2146,2147,2148,2149,2150,2151,2152,2153,2154,2155,2156,2157,2158,2159,2160,2161,2162,2163,2164,2165,2166,2167,2168,2169,2170,2171,2172,2173,2174,2175,2176,2177,2178,2179,2180,2181,2182,2183,2184,2185,2186,2187,2188,2189,2190,2191,2192,2193,2194,2195,2196,2197,2198,2199,2200,2201,2202,2203,2204,2205,2206,2207,2208,2209,2210,2211,2212,2213,2214,2215,2216,2217,2218,2219,2220,2221,2222,2223,2224,2225,2226,2227,2228,2229,2230,2231,2232,2233,2234,2235,2236,2237,2238,2239,2240,2241,2242,2243,2244,2245,2246,2247,2248,2249,2250,2251,2252,2253,2254,2255,2256,2257,2258,2259,2260,2261,2262,2263,2264,2265,2266,2267,2268,2269,2270,2271,2272,2273,2274,2275,2276,2277,2278,2279,2280,2281,2282,2283,2284,2285,2286,2287,2288,2289,2290,2291,2292,2293,2294,2295,2296,2297,2298,2299,2300,2301,2302,2303,2304,2305,2306,2307,2308,2309,2310,2311,2312,2313,2314,2315,2316,2317,2318,2319,2320,2321,2322,2323,2324,2325,2326,2327,2328,2329,2330,2331,2332,2333,2334,2335,2336,2337,2338,2339,2340,2341,2342,2343,2344,2345,2346,2347,2348,2349,2350,2351,2352,2353,2354,2355,2356,2357,2358,2359,2360,2361,2362,2363,2364,2365,2366,2367,2368,2369,2370,2371,2372,2373,2374,2375,2376,2377,2378,2379,2380,2381,2382,2383,2384,2385,2386,2387,2388,2389,2390,2391,2392,2393,2394,2395,2396,2397,2398,2399,2400,2401,2402,2403,2404,2405,2406,2407,2408,2409,2410,2411,2412,2413,2414,2415,2416,2417,2418,2419,2420,2421,2422,2423,2424,2425,2426,2427,2428,2429,2430,2431,2432,2}

The best experimental conformation, a 10-mer

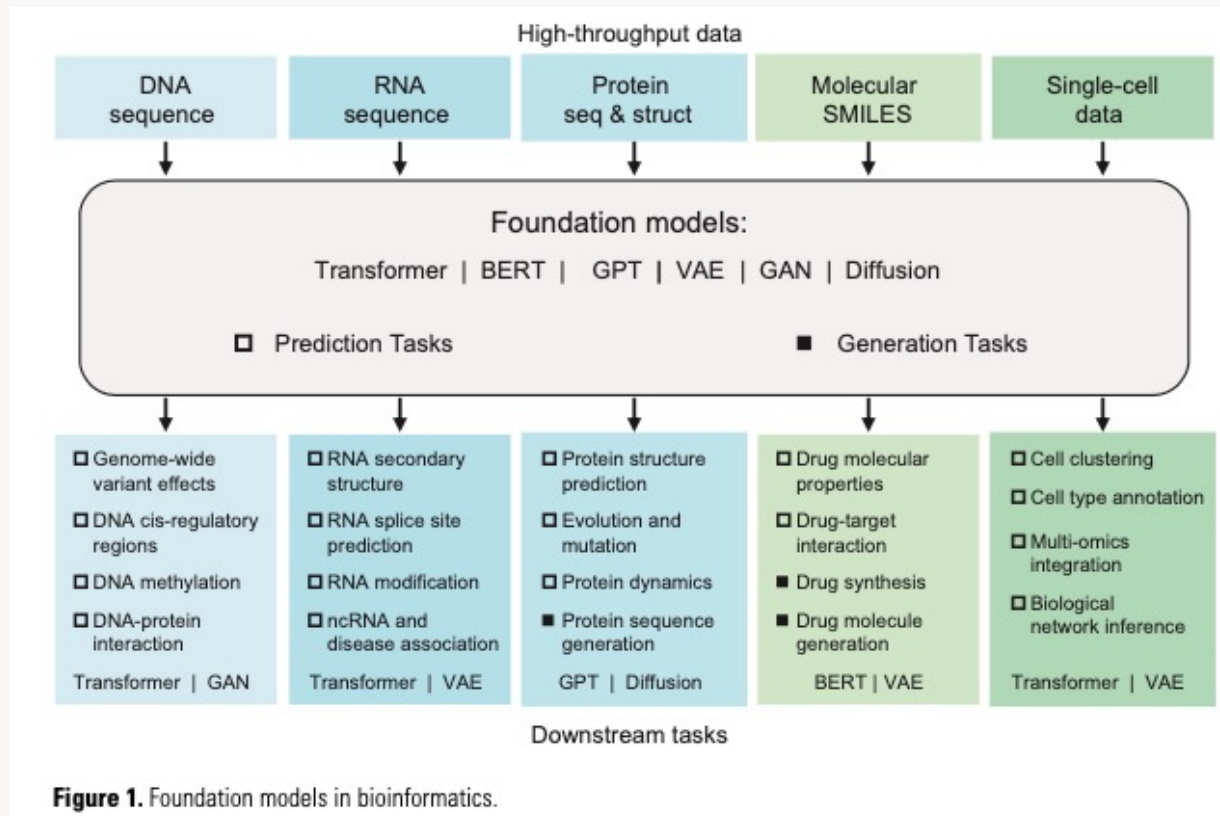
The AlphaFold prediction





UNIVERSIDAD
DE MURCIA

But there are many others





UNIVERSIDAD
DE MURCIA

Conclusions to the first part

- Foundation models are notoriously data-hungry, FMs in medicine and biology are no exception
- The nature of this data is heterogeneous
 - May include sensible data like EHRs and genetic data

But, are foundation models compromising our privacy?

- The context of data governance
 - Data governance defines **who** can use data, for **what purposes**, under **what conditions**, and with **what safeguards**.



UNIVERSIDAD
DE MURCIA

The data governance barrier problem, paradigm shift

- Usually data can't leave the institution where it is maintained.
 - What happen when infrastructure providers are outside the data source environment?
 - Secure data environments (SDE) become the norm.
- With FMs on medical data there is a paradigm shift
 - In conventional biomedical research, protected patient data serves an analysis and some aggregate results are published.
 - FMs imply that outcome of analysis is a model of millions to billions parameters that may contain critical data from patients
- Should we allow model parameters to leave the secure environment?
Who can use those and what type of access should be use with them?



The data governance barrier problem, examples

UNIVERSIDAD
DE ALABAMA
Foresight-E

The NHS (UK) paused data access after the governance committee questioned its use to train the model. Even though they concluded in April 2026 no GDPR breach, the project's access to the NHS England (SDE) is stopped.



DeepMind and acute kidney injury

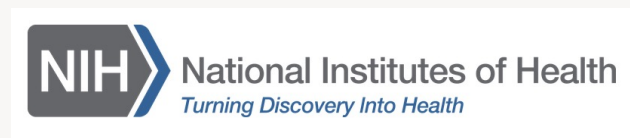
The Royal Free London NHS Foundation Trust gave DeepMind access to records from **1.6 million NHS patients** to develop software for patients at risk of developing acute kidney injury. Investigation concluded GDPR breach due to lack of transparency in informing patients on how their data is being used.



All of Us dataset restrictions

An EEUU dataset of more than 800K enrolled participants, with WGS (500K), approximately 500K EHRs, 600K of physical measurements and more than 750K surveys, RNA-seq ~ 9K and proteomics for ~ 10K with also longitudinal clinical and phenotypic information.

- Users can work with the data within their Researcher Workbench
- No ML model built there can leave the system.





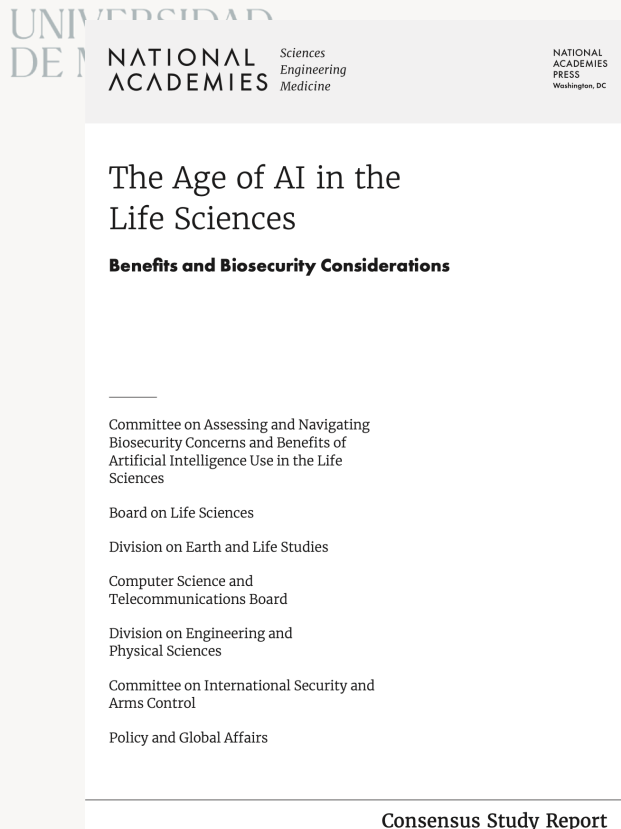
UNIVERSIDAD
DE MURCIA

The data governance barrier problem: four barriers

- Purpose compatibility: were the data collected/approved for FM training?
- **Patient transparency:** would patients reasonable expect this use?
- **Model Export:** can weights trained on protected data leave the SDE
- **Secondary uses of the model:** can others fine-tune and/or deploy it for purposes not in the original approval?



The biosecurity take on this matter



Examples of datasets of interest:

- Associations between viral sequences and virulence parameters
- viral phylogenomic sequence datasets linked to epidemiological data

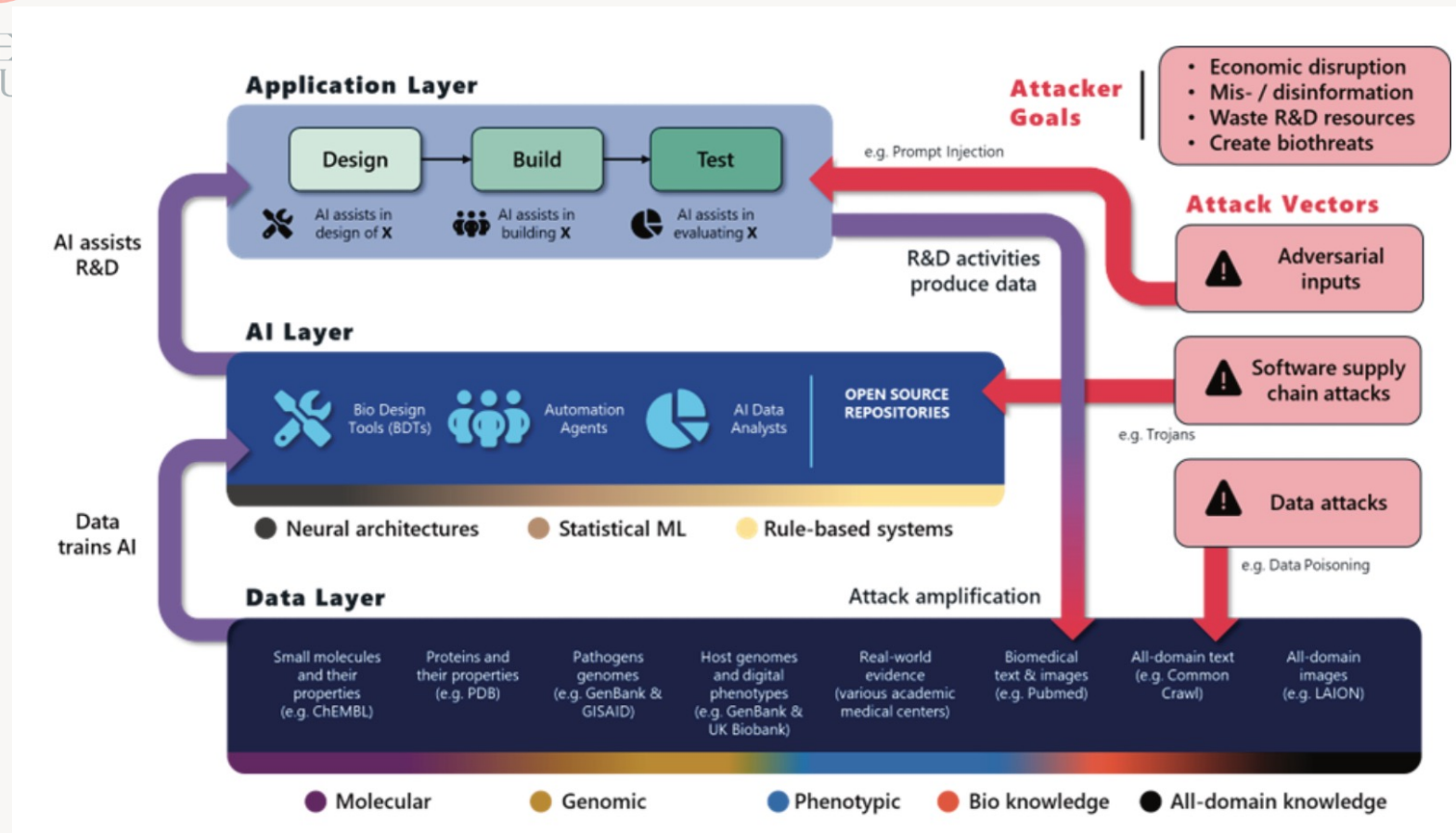
Example AI models

- FMs that infer mechanisms of pathogenicity and transmissibility from pathogen sequencing data
- FMs that could predictably generate a novel replication-competent virus



UNIVERSITY
OF MANCHESTER

The biosecurity take on this matter (cont.)





UNIVERSIDAD
DE MURCIA

Our projects using this technology

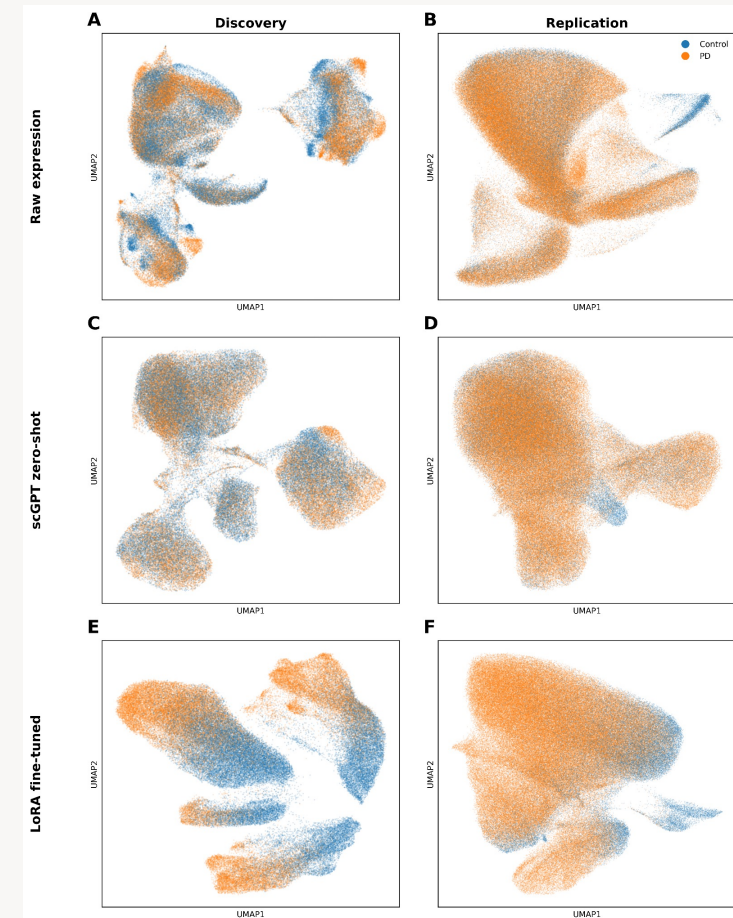


UNIVERSIDAD
DE MURCIA

Towards disease specific foundation models

Hypothesis: single-cell RNA-seq models of case/control datasets can be adapted to a specific disease to increase what we know with state-of-the-art conventional ML

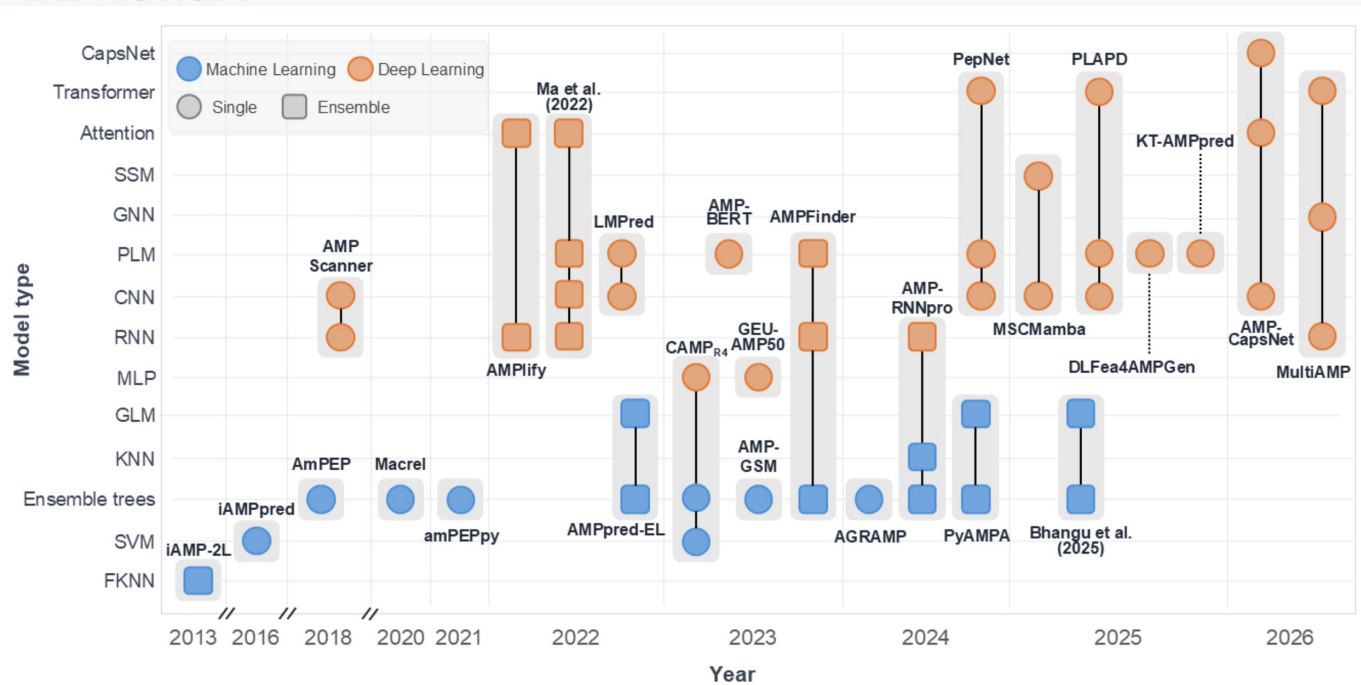
Approach: build specific pipeline to adapt, evaluate and assess from scGPT, geneFormer and others





UNIVERSIDAD
DE MURCIA

BlueAI, predicting antibacterial properties in peptides



BlueAI development platform for marine environments sustainability

START EXPLORING

COORDINATED PROJECT	BlueAI
DURATION	48 months
APPLICATION SCOPE	Mar Menor - Rias Baixas
COORDINATING INSTITUTION	University of Murcia (UMU)
LEAD PIs	Antonio F. Skarmeta Gomez - Juan A. Botia Blaya
SUBPROJECTS	BlueAI-UMU - BlueAI-UPV - BlueAI-USAL - BlueAI-UVigo

Logos: MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES, FEDERACIÓN EUROPEA, R, I+D+i, AGENCIA ESPAÑOLA DE INVESTIGACIÓN CIENTÍFICA

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT

New Results

Follow this preprint

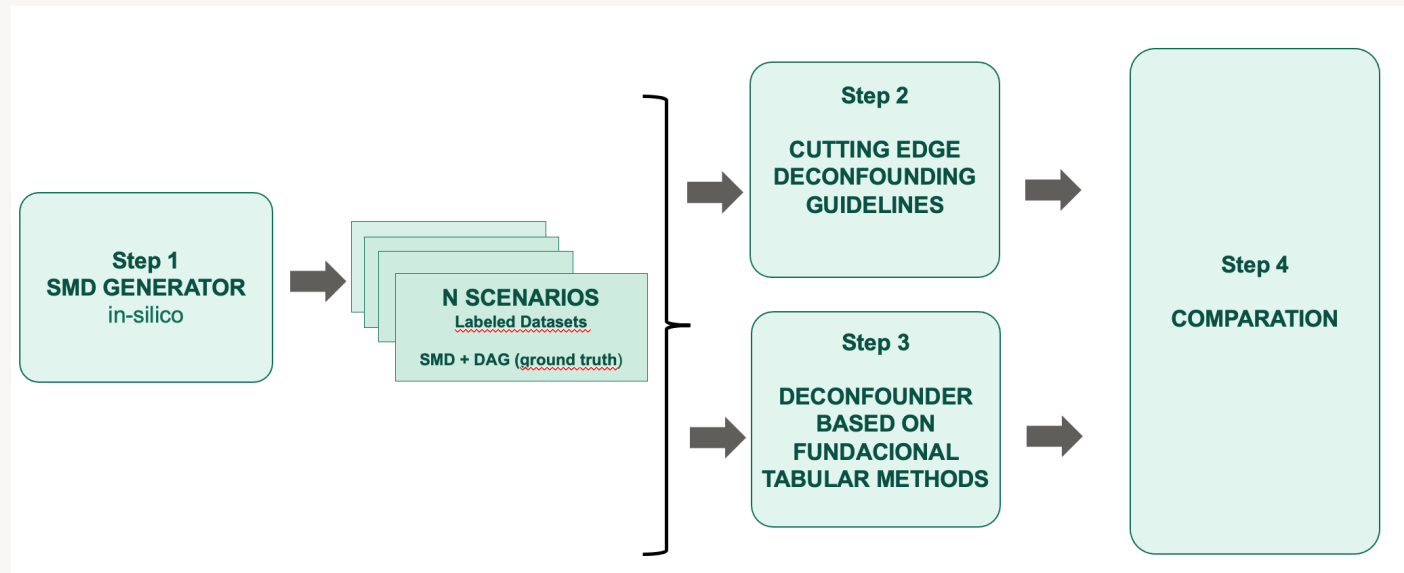
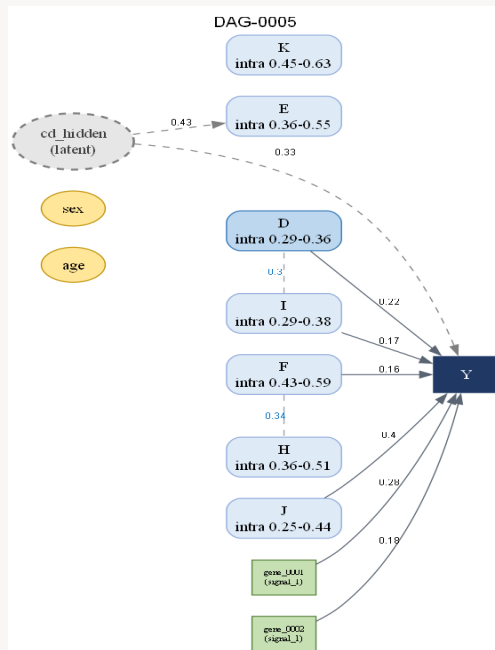
Antimicrobial peptide databases and prediction tools: Toward a standard evaluation framework

10 Alejandro Cisterna-García, 10 Alvaro M. González-López, 10 Aspasia Vozi, 10 María A. Esteban, 10 Adrian Egli, 10 Catherine Jutzeler, 10 Jose Palma, 10 Alvaro Sánchez-Ferrer, 10 Juan A. Botia
doi: <https://doi.org/10.64898/2026.05.19.726290>



UNIVERSIDAD
DE MURCIA

Tabular foundation models for causality

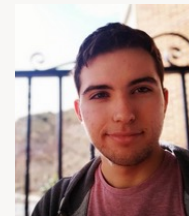
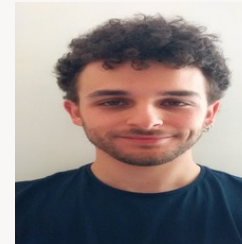
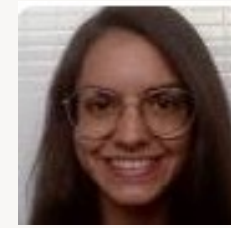




UNIVERSIDAD
DE MURCIA

Thanks

juanbot@um.es



José T. Palma
Ana Luisa Gil
Alicia Gómez-Pascual
Toni Guillén
Jaime Ñíguez
Joselu Mellina
Jose Adrian Pardo
Alejandro Belda
Alejandro Cisterna