

Artificial Intelligence in Biomedical R&D and Drug Discovery

mapping yourself to AI, from Algorithms and Agents to Good Practice

Cameron MacPherson, PhD
INSTITUT ROCHE, PARIS, FRANCE

Disclaimer: The views, data, and content presented in this material are either entirely my own or in the public domain and do not necessarily reflect those of my employer.

AUTHOR'S NOTE

I have made significant use of different AI tools to put this presentation together. I can honestly say that it has been a frustrating pleasure. Believe me when I say this oxymoronic statement is entirely intentional. While I have a natural inclination toward complexity, this is likely the most information dense set of slides I have ever made. I highlight this out of a sense of caution and empathy toward you, but also – mostly – to highlight that it is a lesson in itself; to showcase both the capability for large language models to encode knowledge and mirror it back in myriad ways. It is a remarkable weapon in communication but it is also, of course, a double-edged sword. As you will no-doubt see for yourself there can be such a thing as too much information and within it the certainty that some of it is wrong. I therefore encourage you not to believe anything more beyond this note and to make a study of the content yourself. I have put DOI links to all resources used on every slide. Good luck.

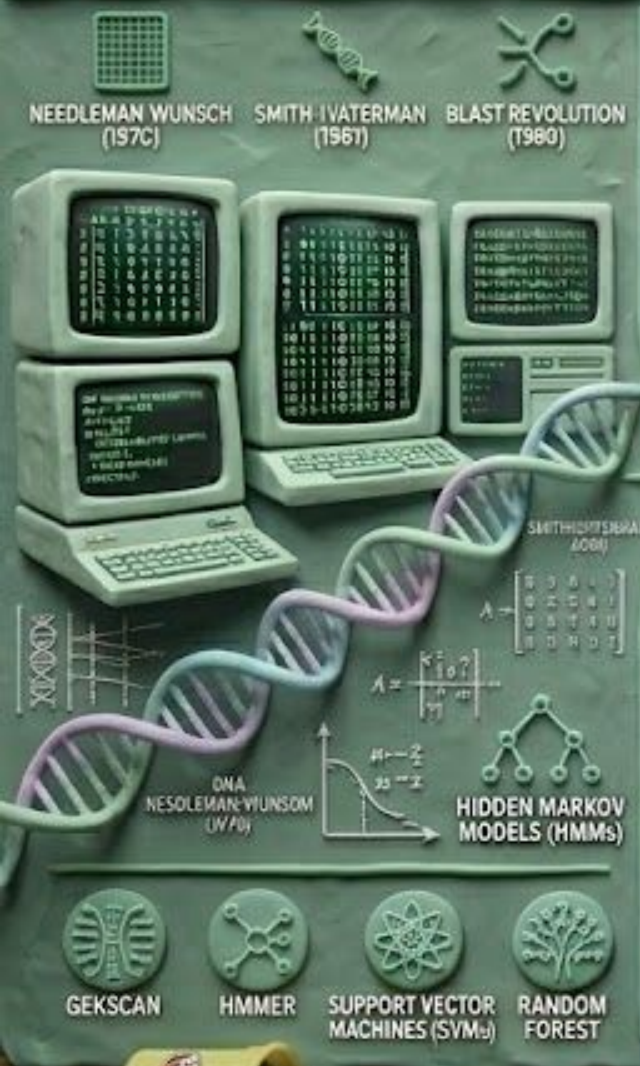
| SYSTEMS EVERYWHERE

" *Technological change is not additive; it is ecological. A new technology does not merely add something; it changes everything.*

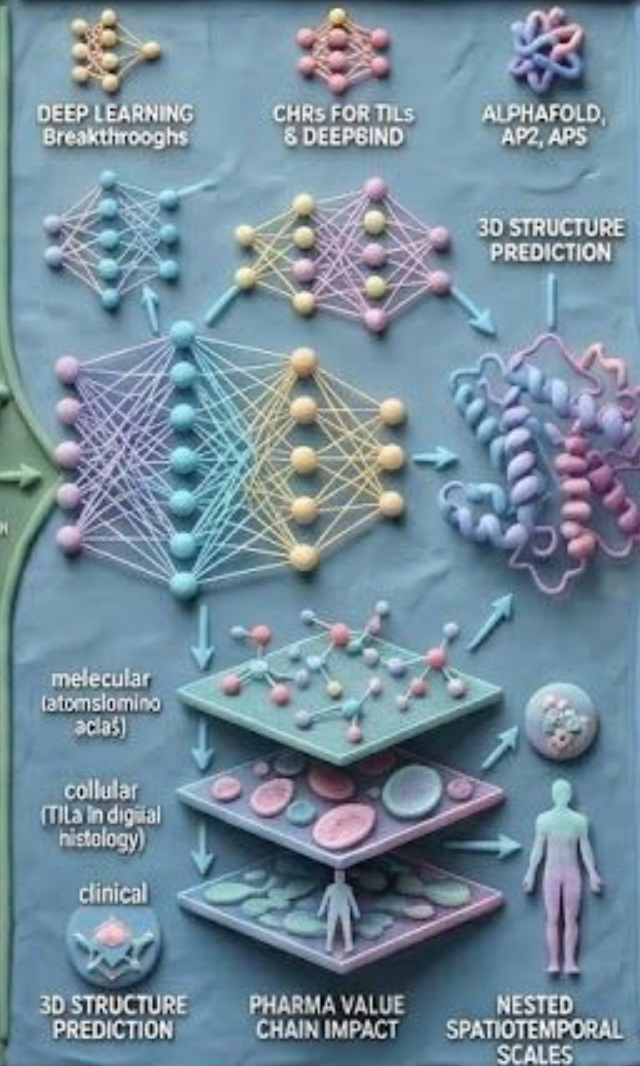
— Neil Postman

Lesson plan: The Evolution of AI in Systems Immunology - From Foundations to Agents

Phase 1: HISTORICAL FOUNDATIONS OF BIOINFORMATICS & MACHINE LEARNING



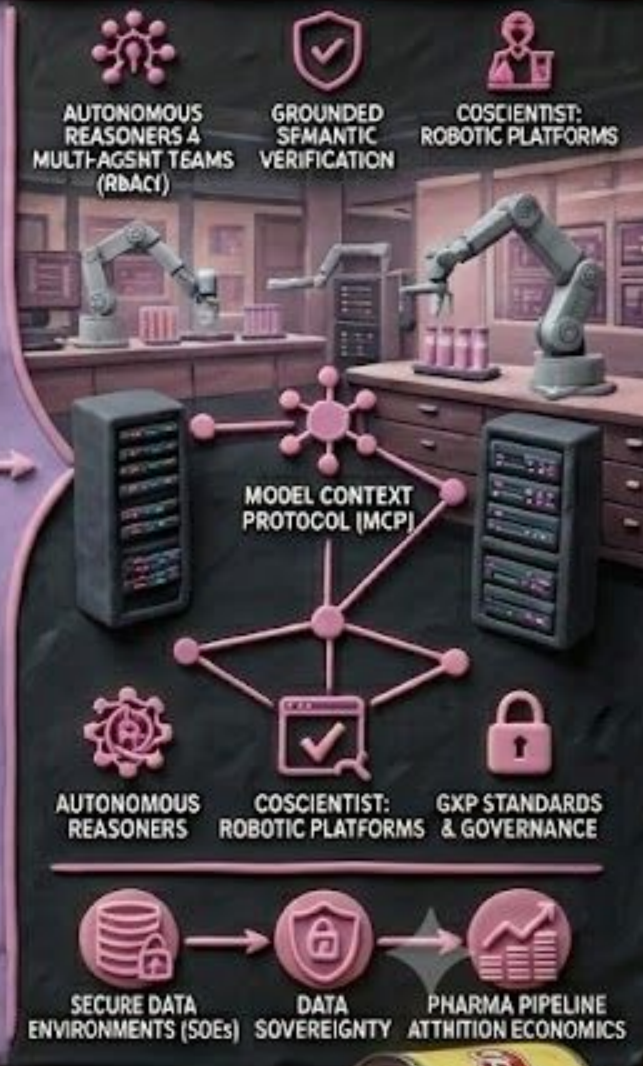
Phase 2: DEEP LEARNING, ALPHAFOLD, & SPATIOTEMPORAL SCALE



Phase 3: PREDICTIVE ML VS. GENERATIVE AI & FOUNDATION MODELS



Phase 4: AGENTIC AI LAB ARCHITECTURES & GXP BEST PRACTICES



Context

A BRIEF HISTORY

Long and large investments into machine learning

Machine learning & AI

1950-1960

Birth of Machine Learning

1970-1980

Expert Systems Era

1980-1990

Renewed Interest in Neural Networks

1990-2000

The Age of Data

2000-2010

Deep Learning Revolution

2010-2020

Early Industry Adoption and Ethical Concerns

2020-

Maturing adoption, LLMs, federated & edge computing.

Biomedical informatics

1970-1980

Basic Data Analysis

1980-1990

Molecular Modeling

1990-2000

Bioinformatics Boom

2000-2010

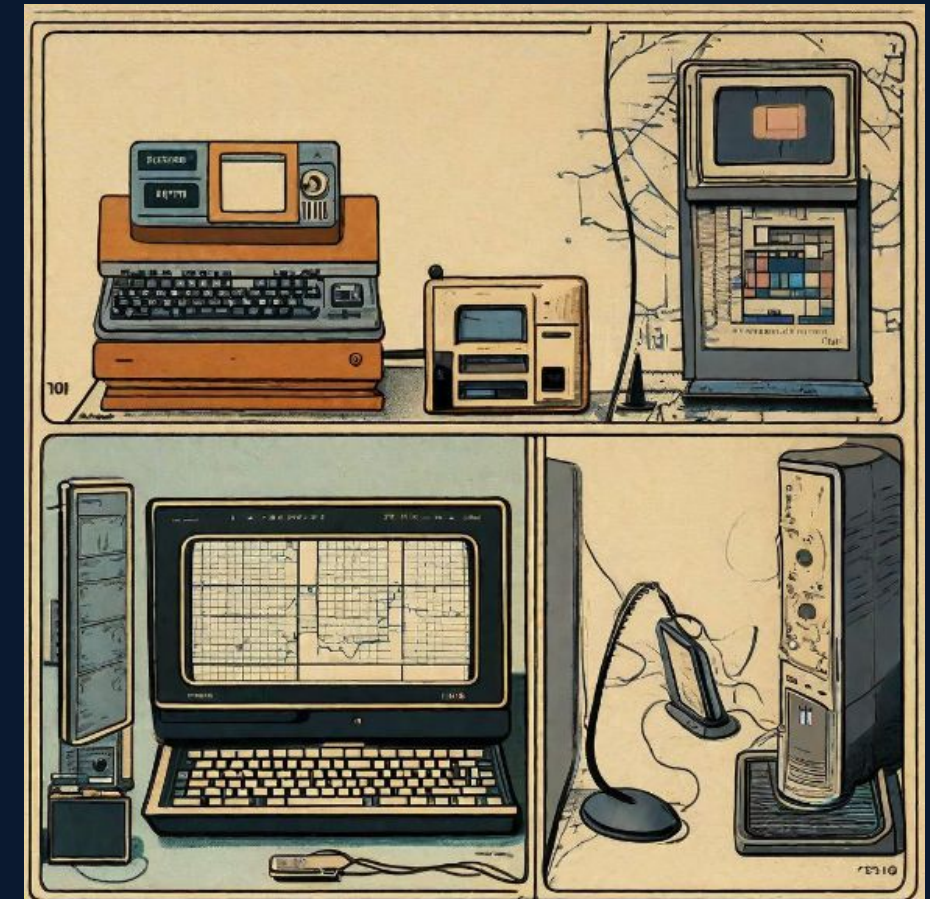
High-throughput Screening

2010-2020

Drug Repurposing and Personalized Medicine

2020-

Multomics & optimisation of biologics

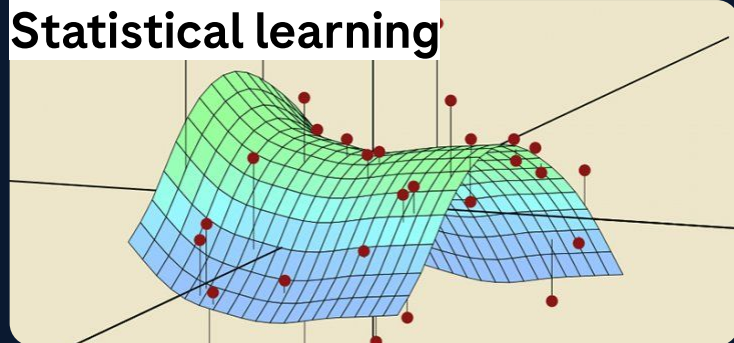


FALLING INTO LARGE LANGUAGE MODELS

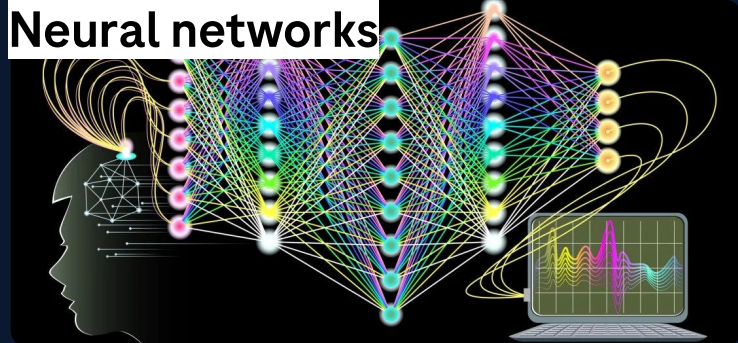
AI has been incredibly successful but LLMs have captured our imagination ... WHY?

NLP

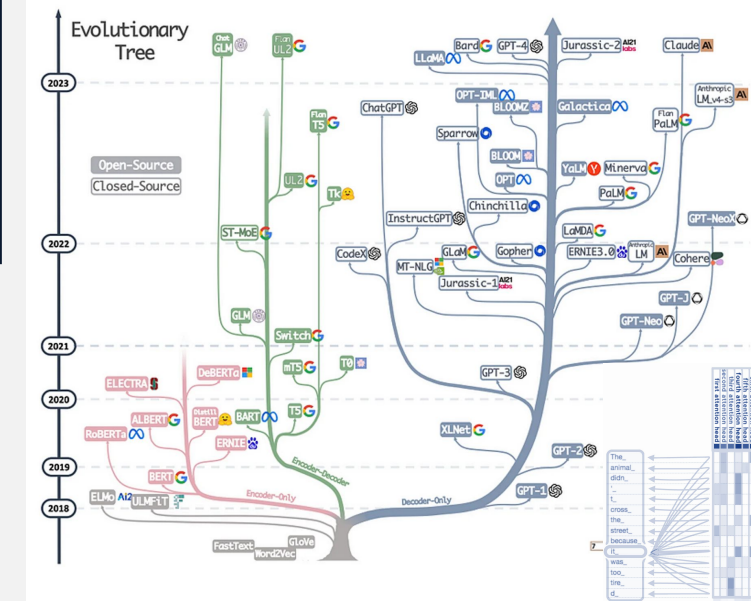
Statistical learning



Neural networks



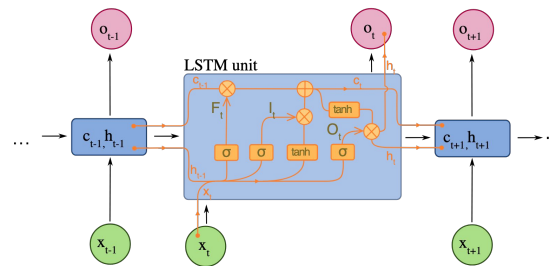
Large language models



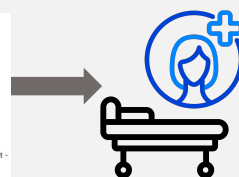
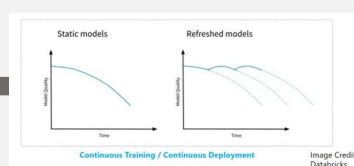
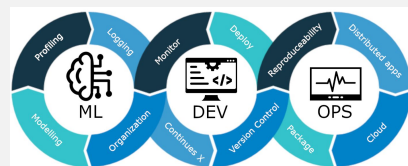
N-Gram models

N = 1 : This is a sentence *unigrams*: this, is, a, sentence
N = 2 : This is a sentence *bigrams*: this is, is a, a sentence
N = 3 : This is a sentence *trigrams*: this is a, is a sentence

Recurrent networks



Deployment models

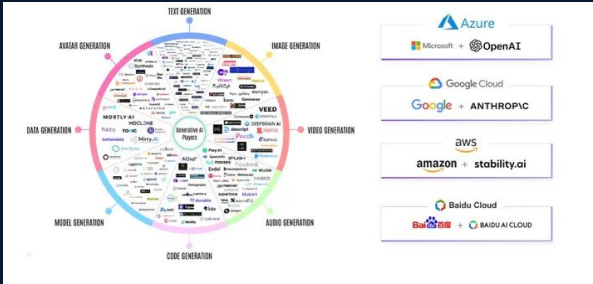
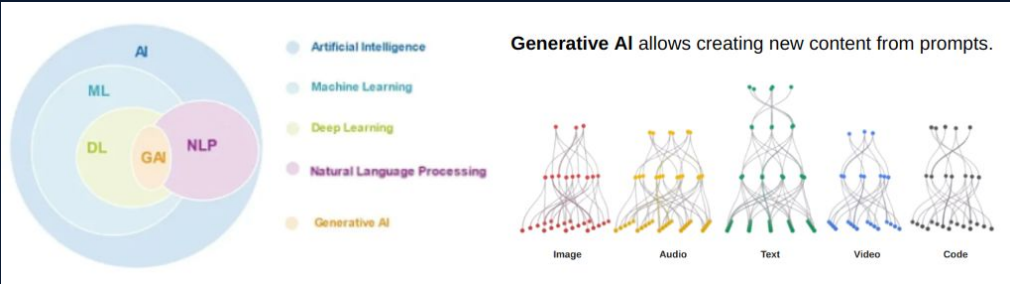
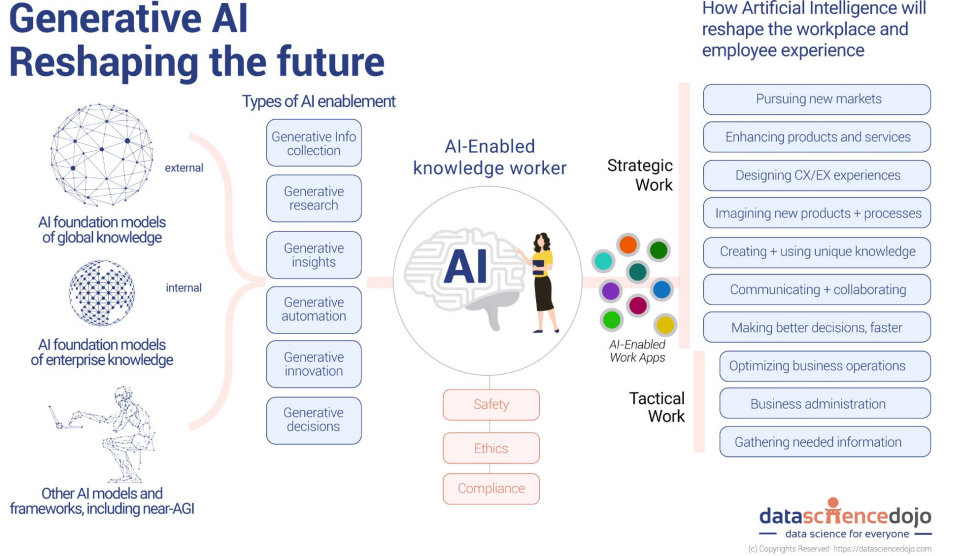
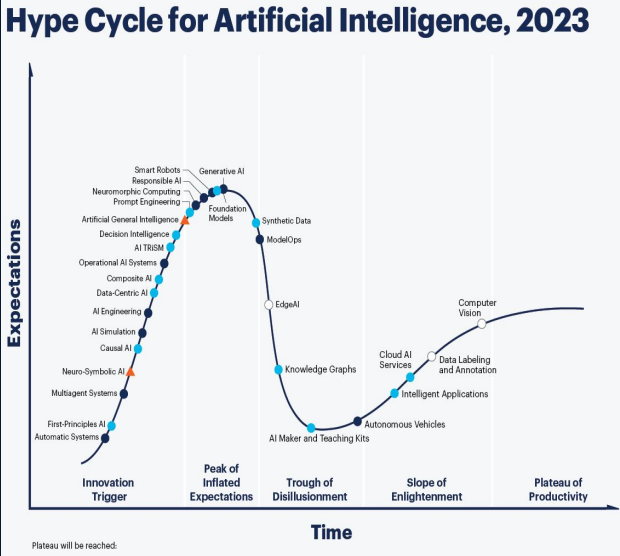
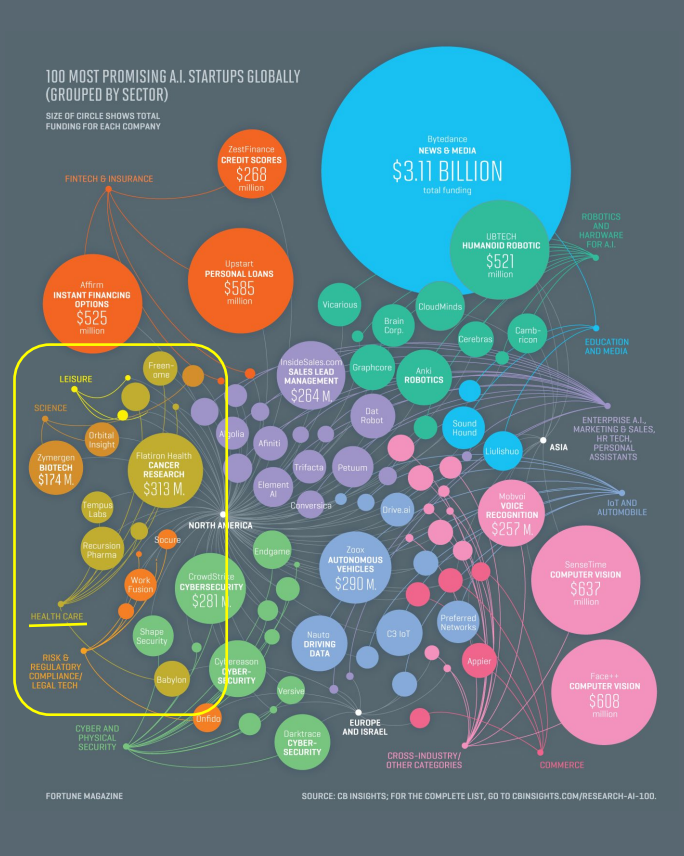


Data models (more later)



ACCELERATED PROGRESS ⇒ CHEAPER ACCESS

Each discovery prompts an even greater discovery, sooner.



THE HYPE CYCLE COMPOUNDED

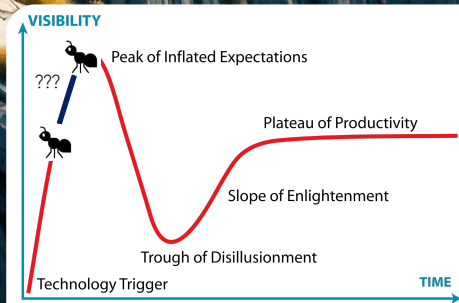
Don't focus on technology, focus on valid needs and adopt AI as necessary.
Watch, wait, learn - together.

A hype cycle depicts phases of enthusiasm for a new tech: initial excitement, over-expectation, disillusionment, and productive use.

- AI's development is marked by overlapping hype cycles, each depicting phases from excitement to productive use.
- From rule-based systems to deep learning, each era has its own cycle, with daily AI discoveries adding to the complexity.
- AI advancement is accelerating due to computational growth, data availability, and intersections with other technologies.
- This compounded and quickening trajectory makes predicting AI's future increasingly challenging.

We are in no position to accurately estimate the end-game for AI or decide on the most impactful use of AI in healthcare.

We can only assume the impact will be extreme & **we must learn**.



A History of Bioinformatics

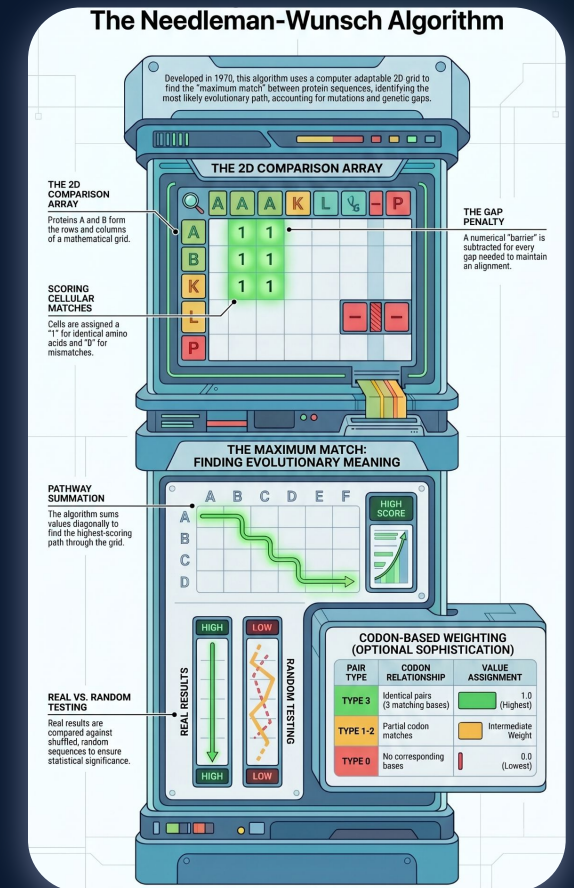
Tracing 50 Years of Molecular Logic

| 1970: The Global Alignment Milestone

" *A general method applicable to the search for similarities in the amino acid sequence of two proteins.*

— Needleman, S. B., & Wunsch, C. D. (1970)

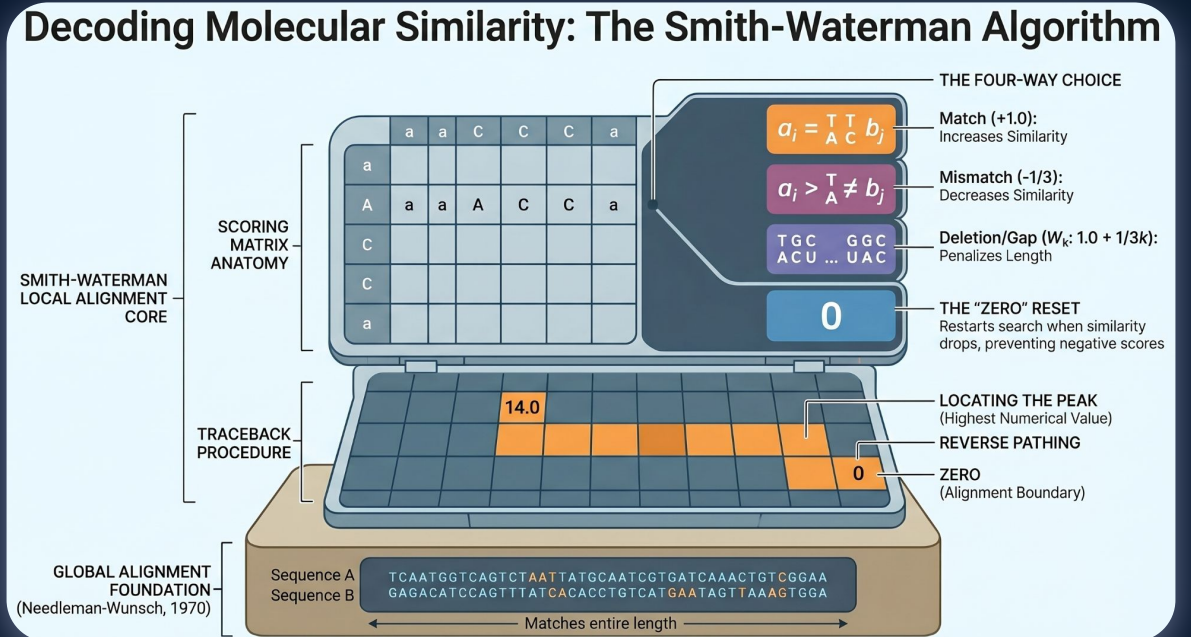
Established the dynamic programming framework that remains the bedrock of molecular biology.



[https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)

1981: Refining the Local Search

- ✓ Smith-Waterman Algorithm: Optimized for sensitive **local alignment**.
- ✓ Enabled identification of common molecular subsequences between divergent species.
- ✓ Foundation for domain discovery and evolutionary conservation mapping.



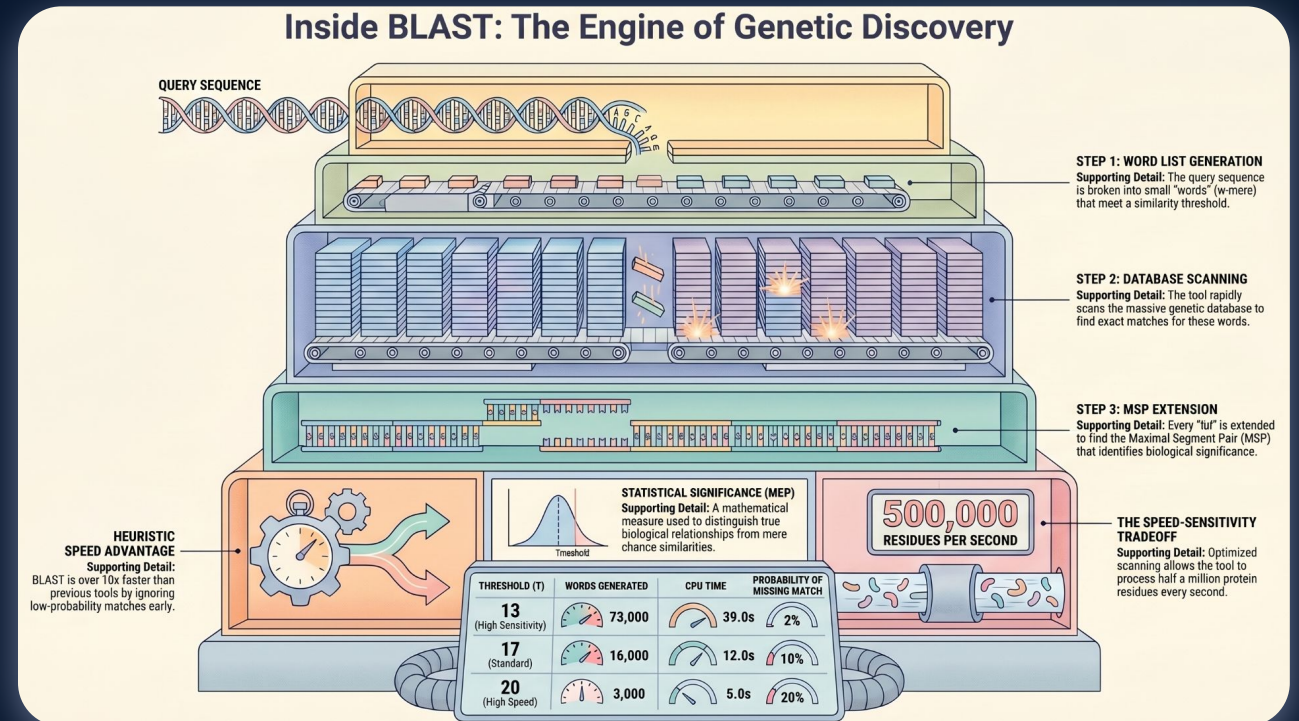
[https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)

1990: The BLAST Heuristic Revolution

- ⚡ Speed over Exhaustiveness: Heuristic search allowed indexing massive databases.
- ⚡ Proved that scaling biological data requires smart algorithmic "shortcuts."
- ⚡ Became the most cited work in bioinformatics history (Altschul et al. 1990).

Citations

Citation Indexes	84207
Patent Family Citations	878
Policy Citations	126



[https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Probabilistic Models & Hidden Markov Models

- Transition from rigid rules to probabilistic state modeling.
- Perfect for sequential data like DNA, where states (exon/intron) are "hidden."
- Allowed the community to handle biological noise and uncertainty for the first time.

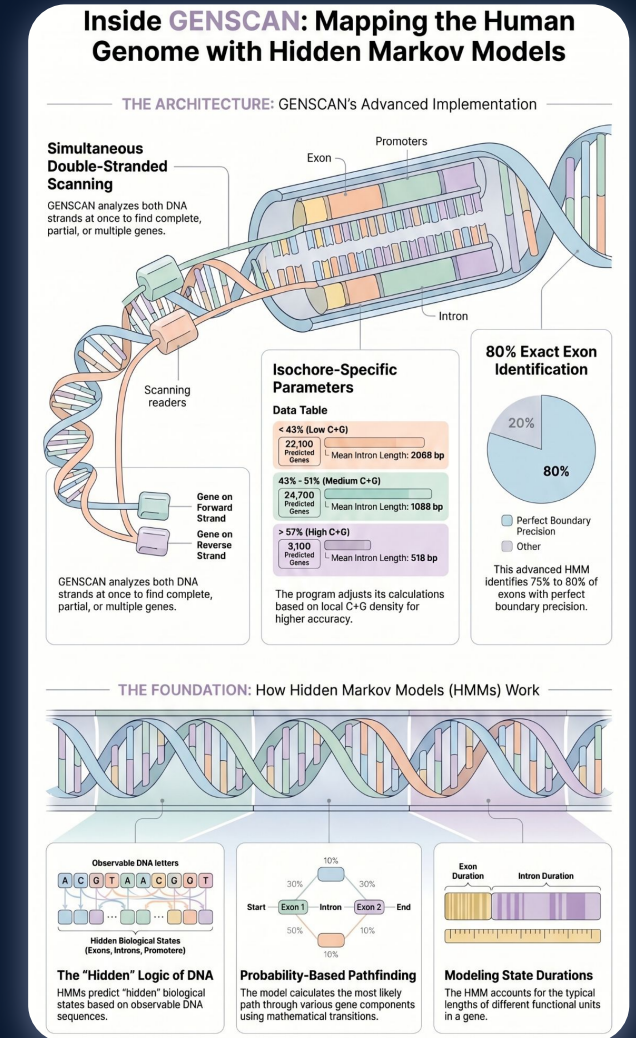
1997: Predicting the Human Genome

"

Prediction of complete gene structures in human genomic DNA.

— Burge, C., & Karlin, S. (1997)

GENSCAN used HMMs to map the human genome's functional architecture before the first draft was even finished.



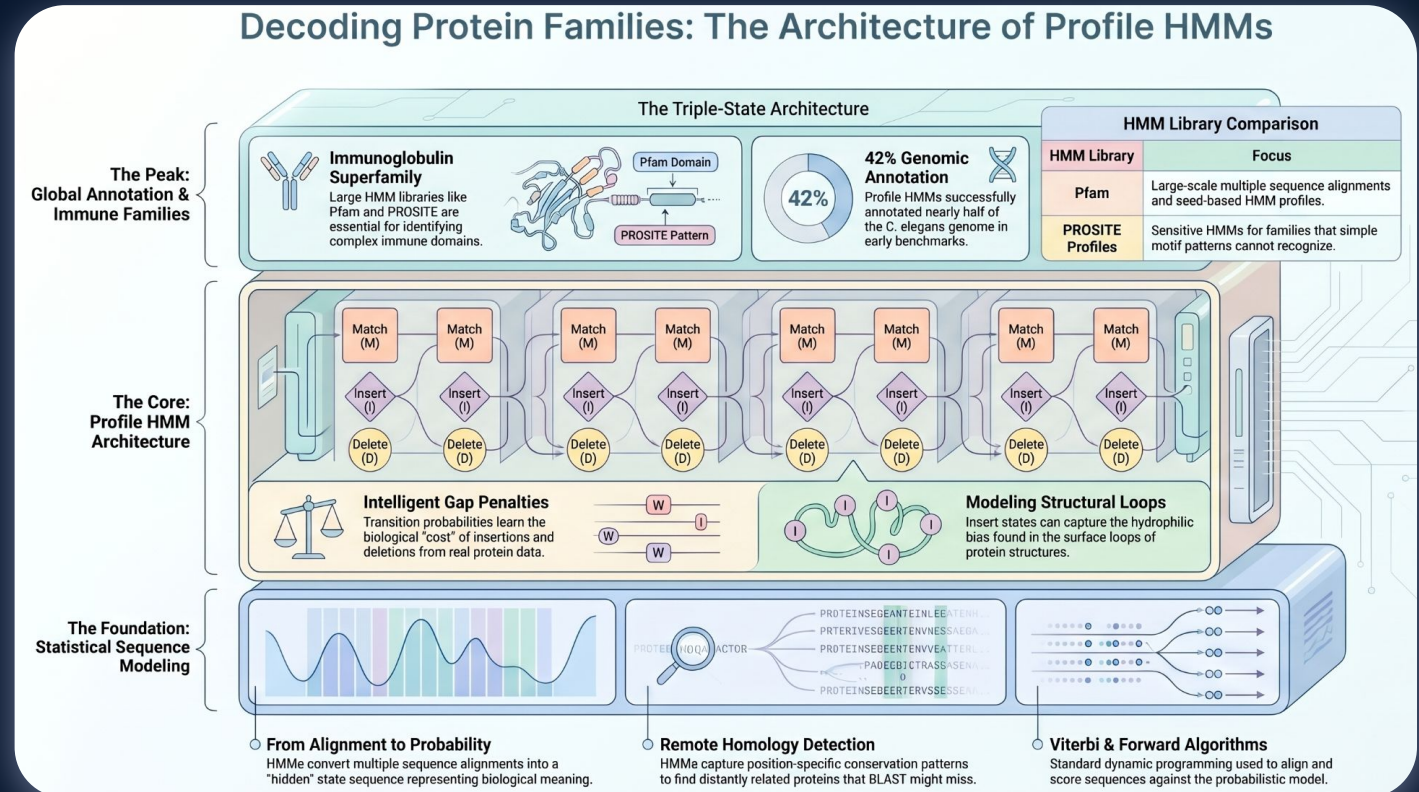
<https://doi.org/10.1006/jmbi.1997.0951>

1998: HMMER & Protein Domain Families

👉 Sean Eddy (1998): Established profile HMMs for protein signatures.

👉 Enabled homology detection beyond simple sequence similarity.

👉 Standardized how we define and annotate immune protein families today.



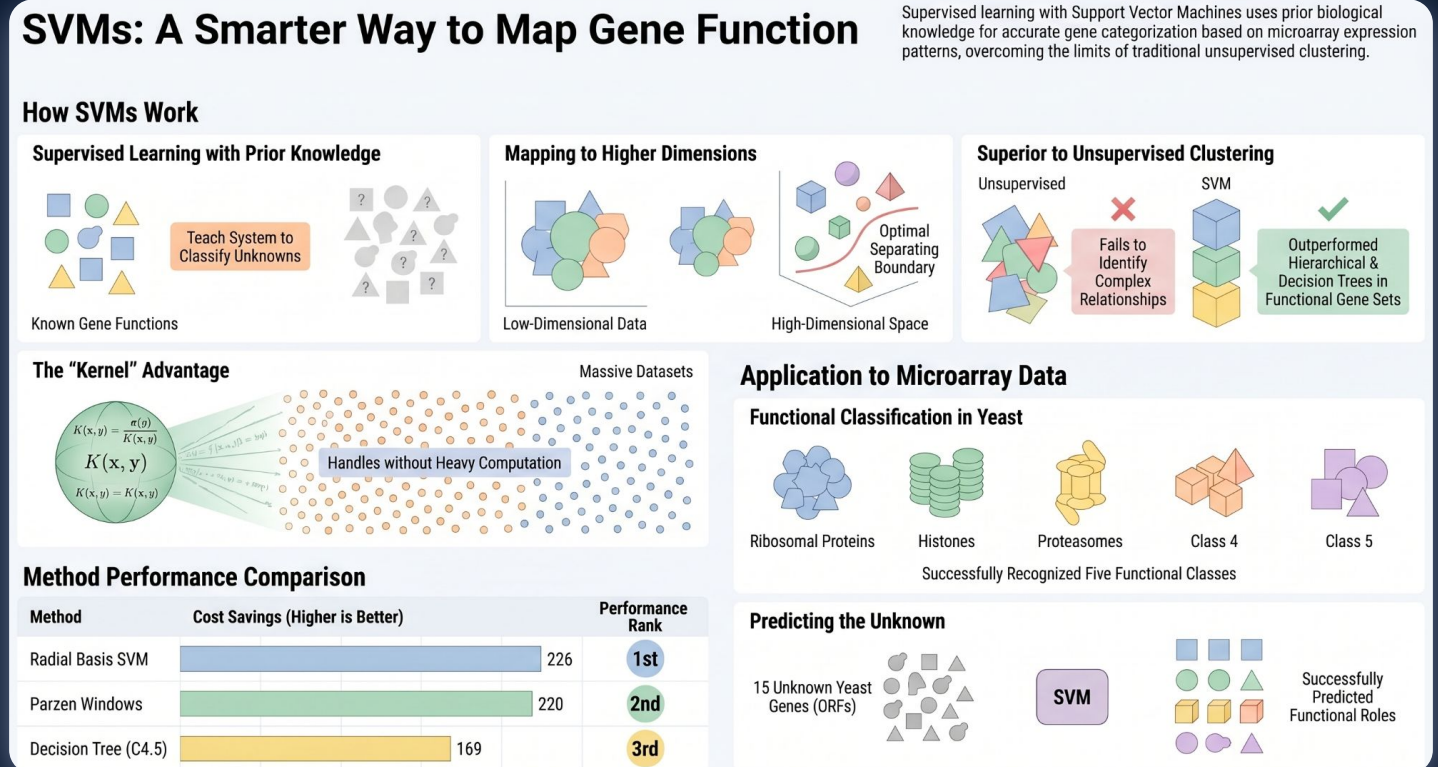
<https://doi.org/10.1093/bioinformatics/14.9.755>

The Rise of Support Vector Machines

Classification in the High-Dimensional Era

2000: SVMs & Gene Expression Matrices

- ✖ Brown et al. (2000): Applied SVMs to microarray data.
- ✖ Solved the "large p, small n" problem in early genomics.
- ✖ Allowed classification of functional gene expression profiles in oncology.

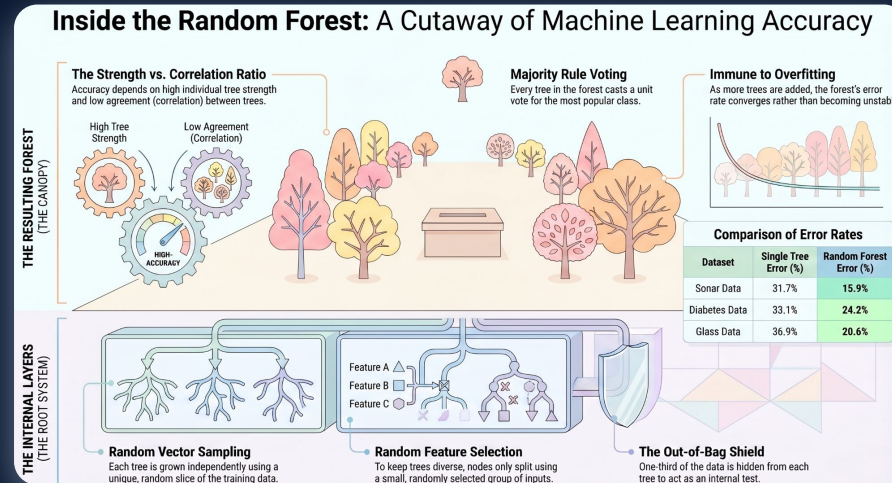


<https://doi.org/10.1073/pnas.97.1.262>

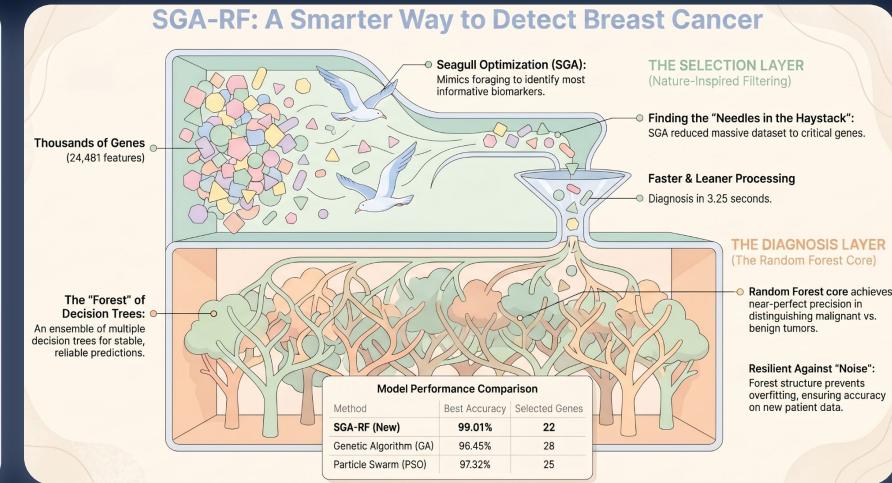
Random Forests in Clinical Omics

99%

Accuracy in Early Diagnostic Signatures



<https://doi.org/10.1023/A:1010933404324> (Breiman)



<https://doi.org/10.1038/s41598-025-95786-1> (Yaqoob)

Ensemble methods like Random Forests brought robust, interpretable "feature importance" to R&D, helping scientists identify the specific cytokines driving patient response.

- Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001)
- Yaqoob, A., et al. ... Random forest classification for enhanced breast cancer ... Sci Rep 15, 10944 (2025).

2010s: The Deep Learning Revolution

From Engineered Features to Representations

20??: Convolutional Neural Networks

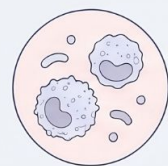
//
Spatial Organization and
Molecular Correlation of
Tumor-Infiltrating
Lymphocytes Using Deep
Learning on Pathology Images.

— Saltz et al. (2018)

Proved that Convolutional Neural
Networks could automatically map the
spatial organization of immune cells
directly from standard digital
pathology images.

Mapping the Body's Defense: How AI Visualizes Cancer Immunity

THE TECHNOLOGY: HOW AI "SEES" IMMUNITY



What are TILs?

Tumor-Infiltrating Lymphocytes are white blood cells that move from the blood into a tumor to attack it.



Deep Learning (CNN) Mapping



Neural networks analyze tiny image patches to distinguish between immune cells and dead tumor tissue.



The Human-AI Feedback Loop

Pathologists review AI predictions and provide "corrections" to fine-tune the model's accuracy.

BIOLOGICAL SIGNIFICANCE: TIL PATTERNS & SURVIVAL

Predicting Patient Outcomes

High densities and specific spatial patterns of TILs strongly correlate with improved survival rates.

VISUALIZING THE "BRISK" RESPONSE (Brisk, Diffuse)

A "Brisk" pattern indicates a strong, dense immune response spreading through the tumor. Dense infiltrate across >30% of the tumor.

BRISK, BAND-LIKE

Organized Border. Infiltrate forming a boundary around the tumor edge.

NON-BRISK, FOCAL

Very Weak. Minimal immune presence in one small, localized area.

MAPPING 13 CANCER TYPES

The AI identified unique immune structures across 5,202 slides, linking spatial patterns to genetic data.

<https://doi.org/10.1016/j.celrep.2018.03.086>

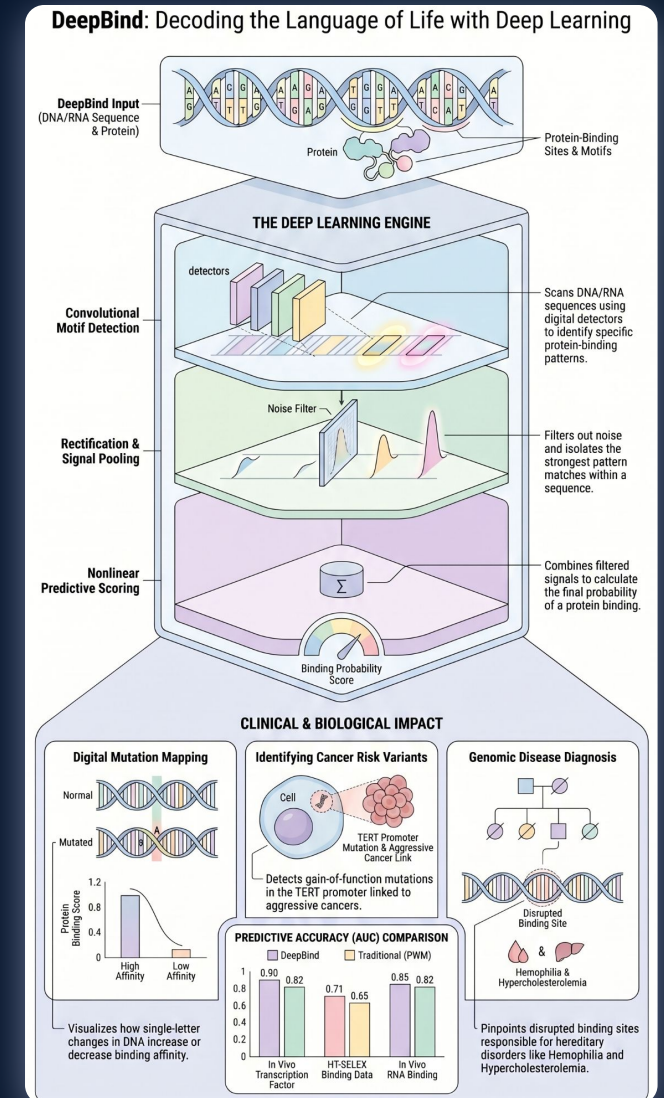
2015: DeepBind – CNNs Meet DNA

“

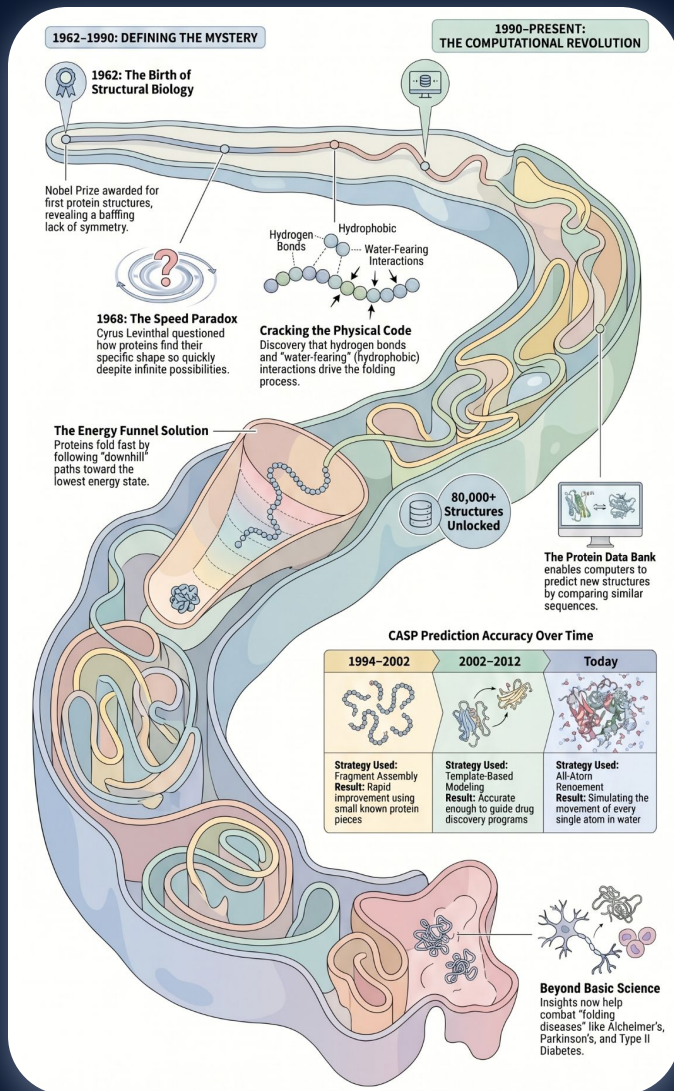
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.

— Alipanahi, B., et al. (2015)

Proved that Convolutional Neural Networks could learn "motifs" directly from raw DNA sequences.



<https://doi.org/10.1038/nbt.3300>



<https://doi.org/10.1126/science.1219021>

The Protein Folding Challenge

60-year-old challenge: Predict 3D structure from 1D amino acid strings.

Traditional costs to resolve one protein structure:

- From years and \$1m down to months and \$100k - \$300k <https://doi.org/10.1146/annurev.biophys.050708.133740>
- NIH PSI initiative optimised down to \$66k <https://doi.org/10.1016/i.str.2007.12.002>

R&D Bottleneck: Validation timelines for new therapeutic targets.

— Dill and MacCallum (2012)

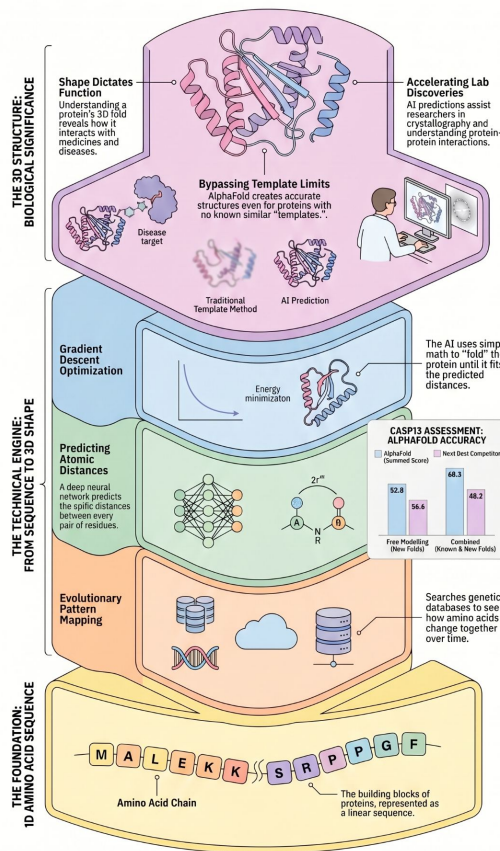
2020 – 2024: The AlphaFold Revolution

John Jumper and the DeepMind team solved the 50-year folding problem with atomic accuracy.

Reframed structural biology as a statistical learning problem on the Protein Data Bank (PDB).

2020

AlphaFold: Decoding the 3D Secrets of Life with AI

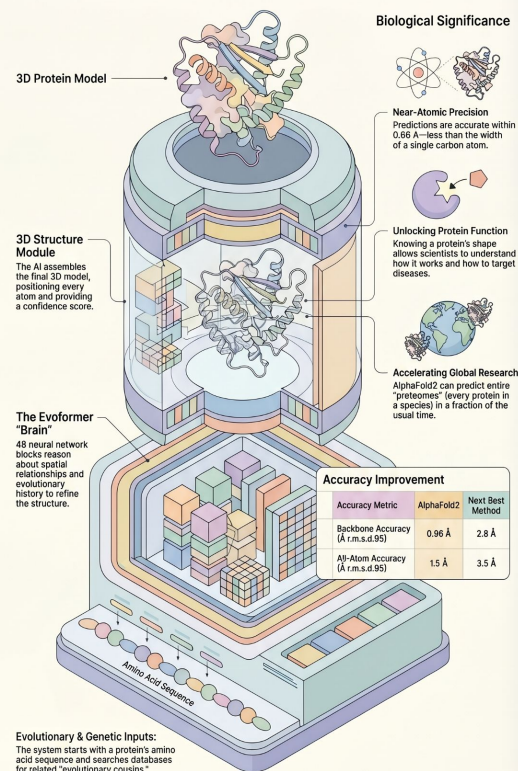


<https://doi.org/10.1038/s41586-019-1923-7>

2021

AlphaFold2: Solving Biology's 50-Year "Folding" Mystery

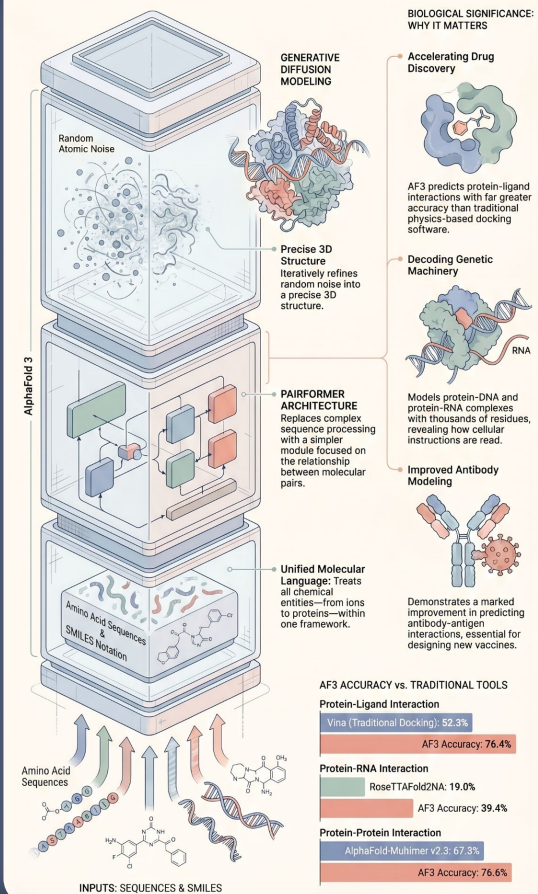
Proteins are the building blocks of life, but their functions are determined by their complex 3D shapes. AlphaFold2 is a revolutionary AI system that uses deep learning to predict these structures with atomic accuracy, turning a process that once took years of lab work into a task that takes mere minutes.



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

2024

AlphaFold 3: Mapping the Building Blocks of Life



<https://doi.org/10.1038/s41586-024-07487-w>

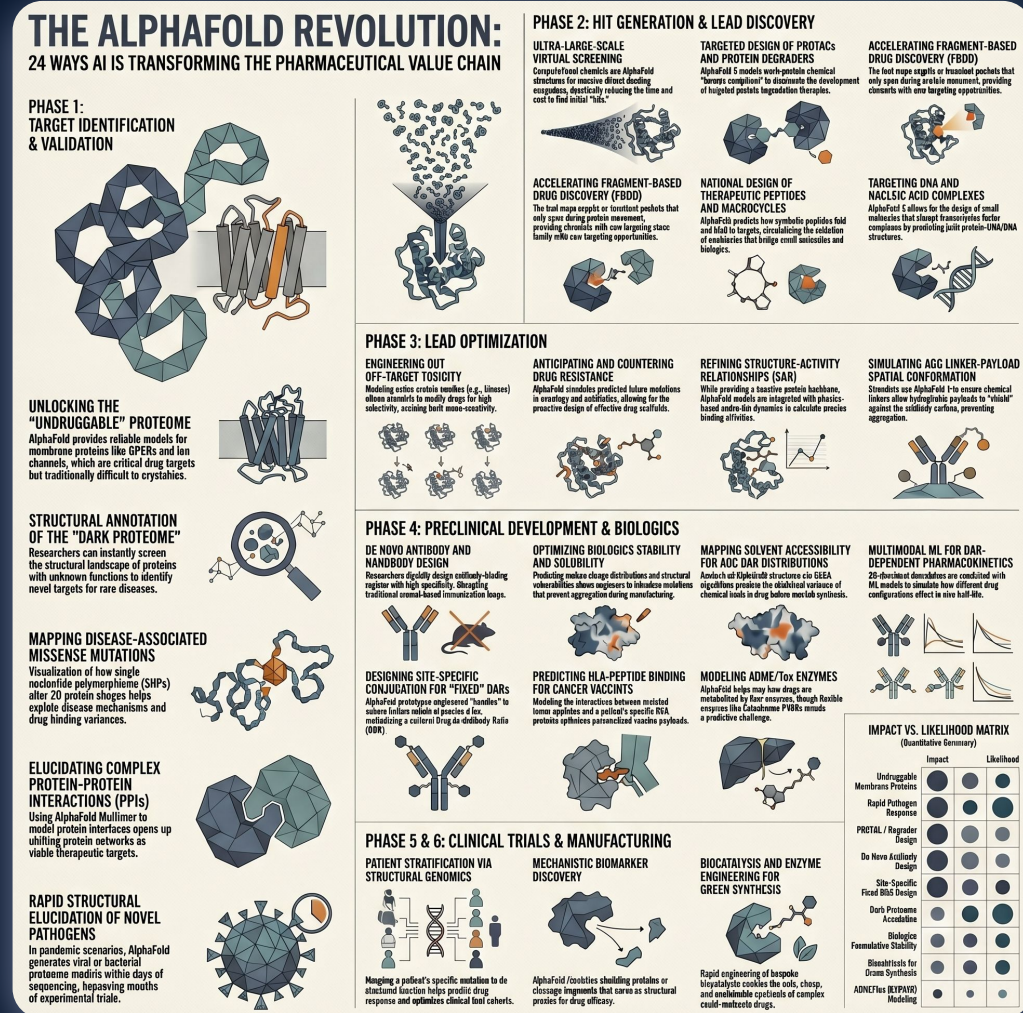
Impact: Collapsing R&D Timelines

AlphaFold enabled the mapping of the entire human proteome in a single release.

-  Rapid Target ID.
-  Epitope Mapping.
-  Drug-Target Binding Site Analysis.

IMO. Systemic change to *status quo* on methodology & mindset was *the* most significant contribution.

AlphaFold raised the question: Can we do everything *in silico*?



What is data anyway?

Understanding systems and the beauty of scale.

BACK TO SCHOOL

Geological, Meteorological, and Ecological Systems

Earth's Nested Systems: The Science of Scale

THE GLOBAL SCALE (TOP LAYER)

Represents massive systems like global climate, tectonic shifts, and regional ecosystem health.

THE LOCAL SCALE (MIDDLE LAYER)

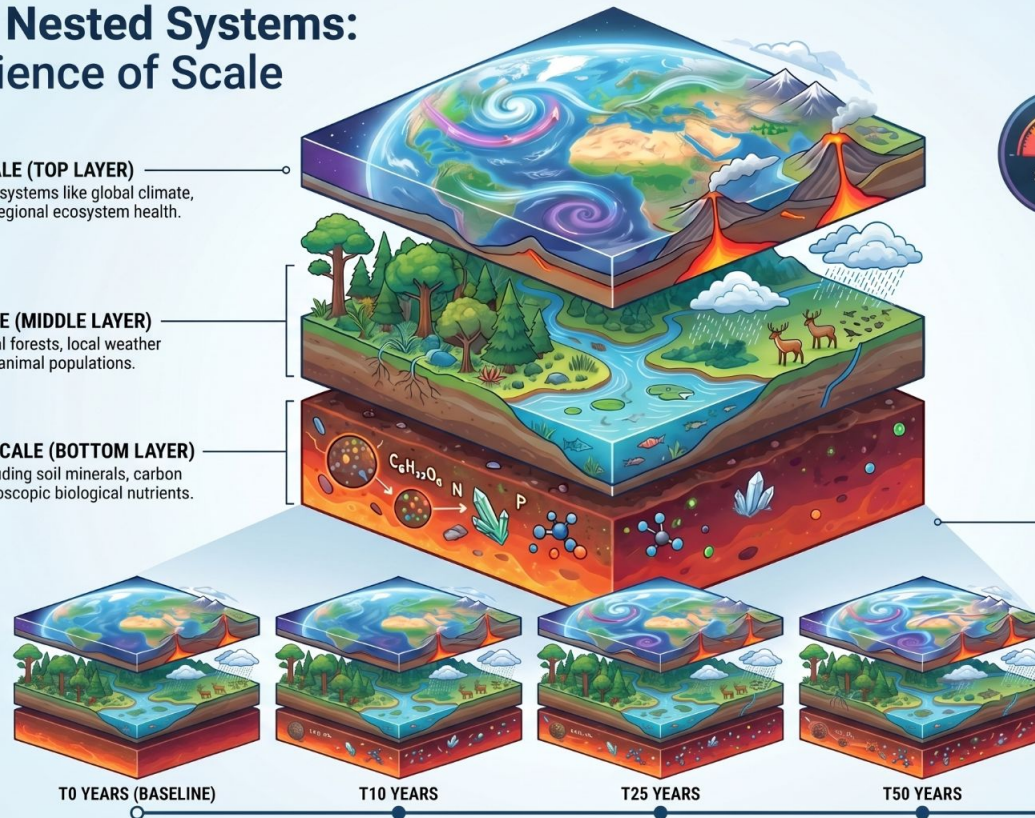
Focuses on individual forests, local weather events, and specific animal populations.

THE CHEMICAL SCALE (BOTTOM LAYER)

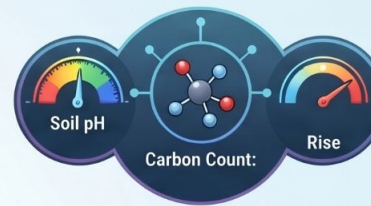
The foundation, including soil minerals, carbon molecules, and microscopic biological nutrients.

SCALE TRANSITIONS:

Small chemical changes take years to travel upward and alter the global climate.



Vector of Characteristics



The Data "Exploded Pixel":
A single measurement is a "pixel" in a larger map.

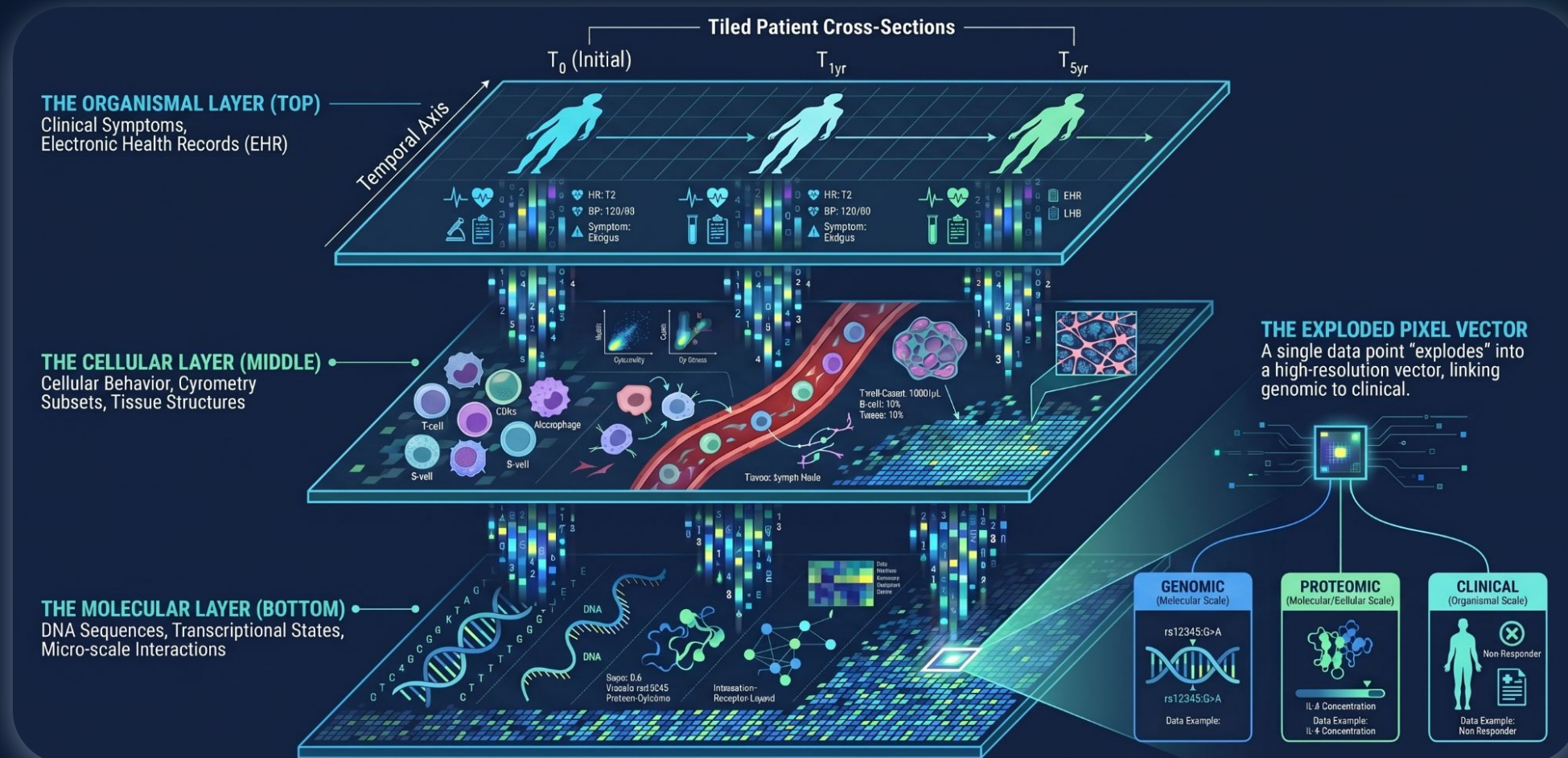
TILED SNAPSHOTS OF TIME:

We view the planet in "slices," comparing the same landscape across years or decades.

SYSTEM SCALE	TYPICAL COMPONENTS	OBSERVATION WINDOW
GLOBAL	Climate, Tectonics	Decades to Millennia
REGIONAL	Forests, Weather	Months to Years
CHEMICAL	Minerals, Carbon	Seconds to Days

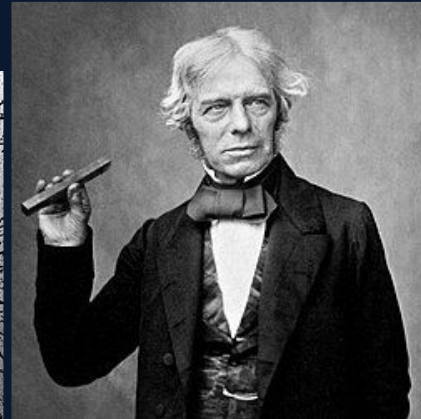
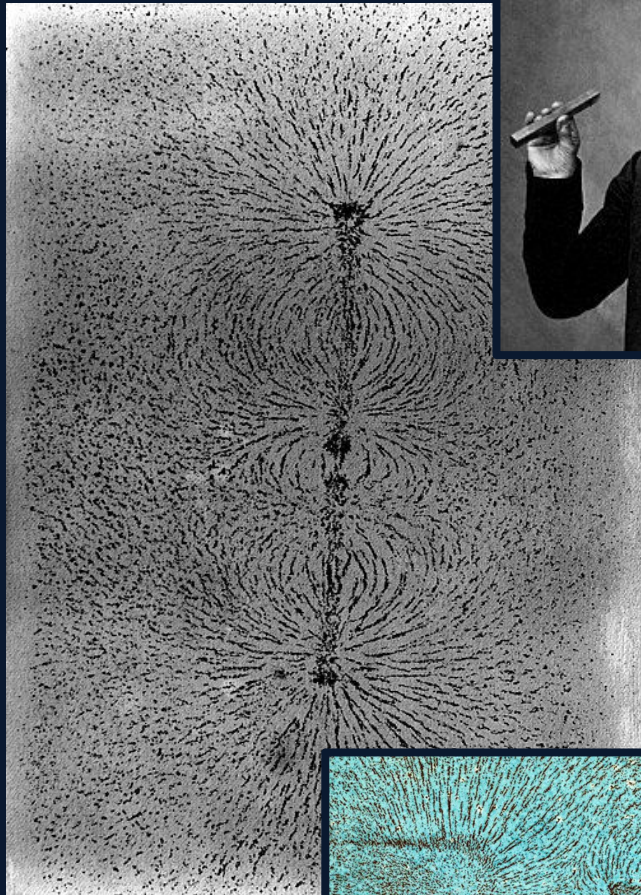
-  **Data is the recorded observation**
It isn't the system itself; it is simply a recorded observation of the system's state at a single coordinate.
-  **"Scale" Means Different Rules**
These layers don't just differ in size; they operate on entirely different rules and timelines.
-  **Time is Tiled When Observed**
We infer the dynamic health of the planet by comparing these static, cross-sectional slices.
-  **Indifferent Scales**
Scale transitions buffer upward and larger scales tend to shape choice at low scales $\Rightarrow$ constrain.

The Spatiotemporal Landscape of Systems Immunology

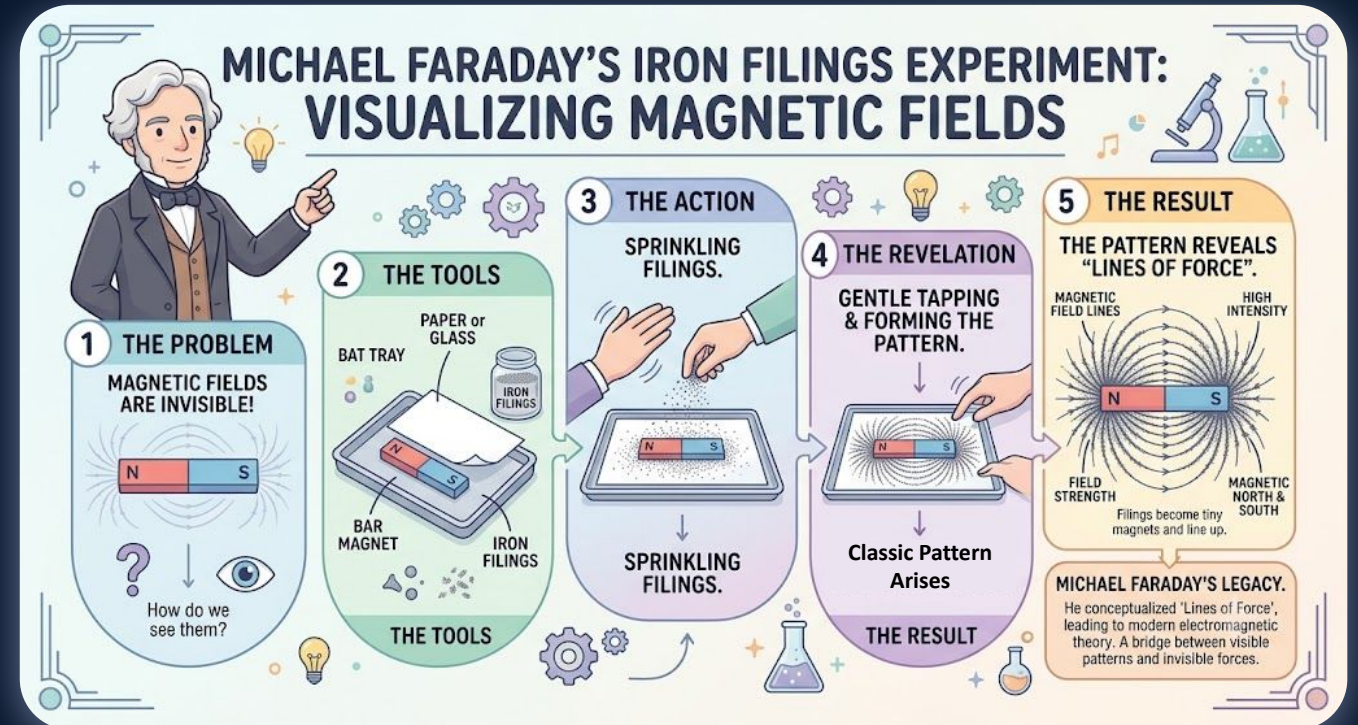
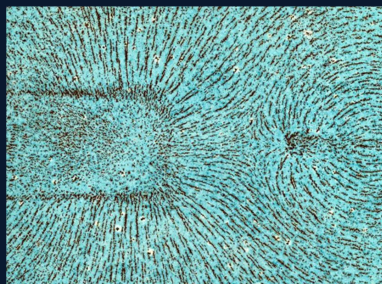


- ✓ Data is some set of observations recorded from a space generated by a system.
- ✓ Observations are often subjective ⇒ All but the simplest of systems are subjectively described.

Practical Implications When Experimenting



Michael Faraday (1851)



1 This is data, but how do you record (encode) it?



2 Observing the phenomena changes it, creates a system.

ML vs. Generative AI

A change from Classifying & Predicting to Constructing Equivalents

Predictive ML: The Discriminative Filter



The "Filter" Role

Traditional ML acts as a filter on existing space. It finds the boundary between known states.



Decision Mapping

Models learn the mapping $Y = f(X)$, where Y is a clinical outcome or label.



R&D Use Case

Virtual screening of compound libraries to filter for potency or ADMET safety.

| The Decision Boundary Concept

- ÷ Goal: Maximizing the margin between biological classes (e.g., Healthy vs. Inflamed).
- ÷ Requires "Labelled" data: We must know the ground truth before training.
- ÷ Highly efficient for narrow tasks but lacks "creative" capacity.

Feature Engineering in Systems Omics

In traditional ML, humans must decide which biomarkers to measure (e.g., GZMA, PRF1).

This "Hand-Crafting" of features limits discovery to our current biological knowledge.

New Concepts / Proxies

- Pooling Baseline Predictors Tsang et al., 2014 <https://doi.org/10.1016/j.cell.2014.03.031>
- Blood Transcription Modules Li et al., 2014 <https://doi.org/10.1038/ni.2789>
- Immune Setpoint Ramos 2025 <https://doi.org/10.1128/msphere.00502-24>

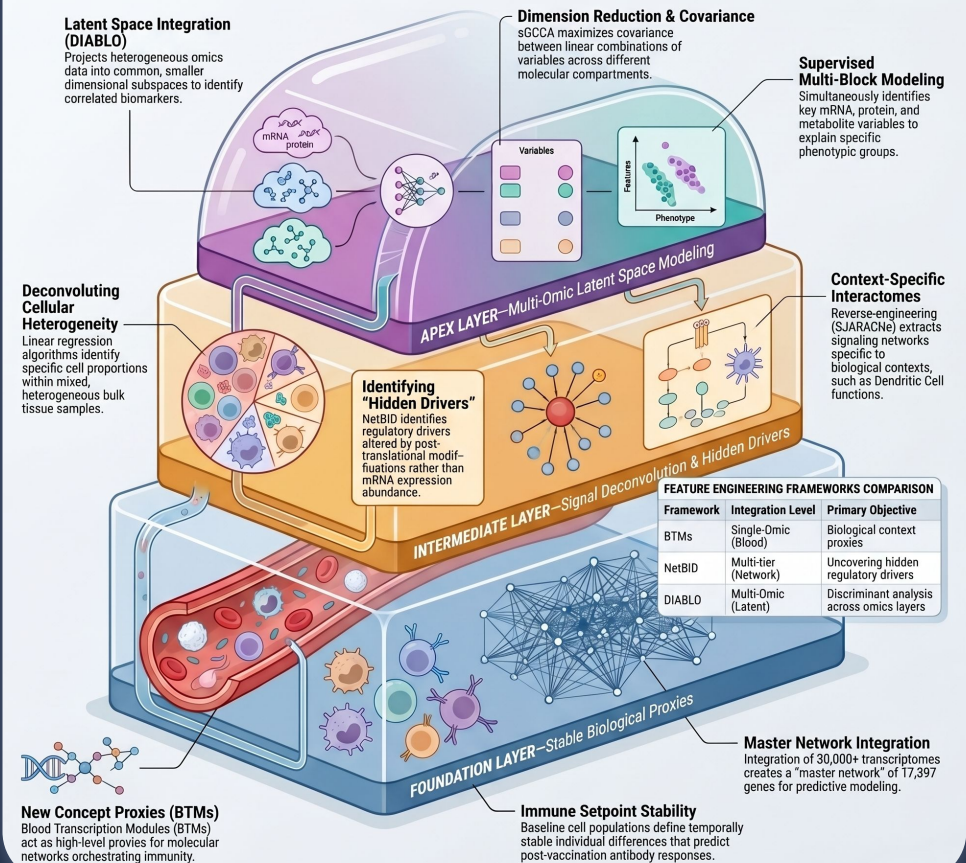
Deconvoluting Pooled Signals

- Tissue → Cell Deconvolution Yu et al., 2019 <https://doi.org/10.1016/j.coisb.2019.03.003>

Using Latent Spaces

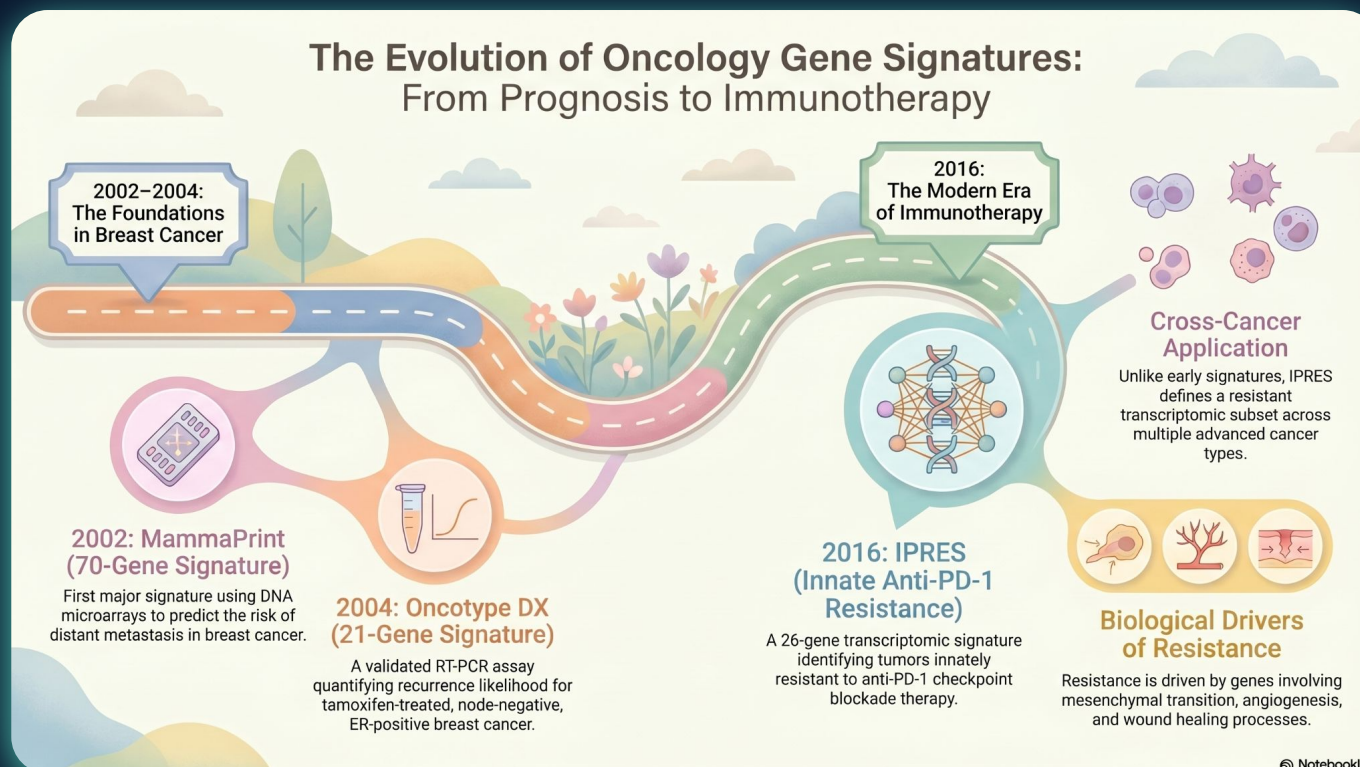
- Multi-omic biomarkers Rohart et al., 2017; Singh et al., 2019 <https://doi.org/10.1093/bioinformatics/bty1054>; Sanches et al., 2024 <https://doi.org/10.3390/biology13110848>

The Hierarchy of Feature Engineering in Systems Immunology



2016: Patient Stratification Milestones

“ Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma.
– Hugo, W., et al. (2016)



IPRES Showed that discriminative gene signatures could predict which patients fail immunotherapy.

<https://doi.org/10.1038/415530a> <https://doi.org/10.1056/NEJMoa041588> <https://doi.org/10.1038/s41586-021-03819-2>

CASE STUDIES

The Data Drift Problem

90%

INFECTIOUS DISEASE
Of Immune-Prognostic AI
Models Carried a High Risk of
Bias and Failed to Generalize

IN GENERAL | Machine learning algorithms tracking complex immune profiles—such as lymphocyte counts, C-reactive protein (CRP), and cytokine changes—frequently suffer from acute dataset shift. Because immune responses are highly dynamic and heterogeneous across different patient demographics, models trained in a single clinical environment regularly collapse when exposed to external real-world settings.

<https://doi.org/10.1136/bmj.m1328>

IMMUNOTHERAPY

A supervised deep-learning pipeline was developed to predict immunotherapy response using tumor-immune features from histologic images. While the model achieved a strong AUC of 0.75 within its US development cohort, its accuracy fell sharply to an AUC of 0.66 when deployed on an independent European validation cohort (Rakaee et al., 2025). This drop demonstrates the hidden vulnerability of clinical AI to international dataset shift.

<https://doi.org/10.1001/jamaoncol.2024.5356>

SEPSIS

EPIC's sepsis model was independently validated across multiple health systems and suffered an absolute performance drop of .1 AUC, or 13.7% (AUC fell from 0.76 to 0.63). This real-world decline led to severe alert fatigue, false positives, with only 7% performance gain over humans.

<https://doi.org/10.1001/jamainternmed.2021.2626>

| The Limitations of 'Static' ML

- ⚠ **Fragility:** Fails when clinical cohorts or sequencing platforms change.
- ⚠ **No Synthesis:** Cannot create a **new** drug; only screens **existing** ones.
- ⚠ **Hallucination of Correlation:** Prone to finding "lucky" signatures in small cohorts.

Generative AI: Synthesizing Space

Moving from Filters to Creators

Learning the Joint Distribution

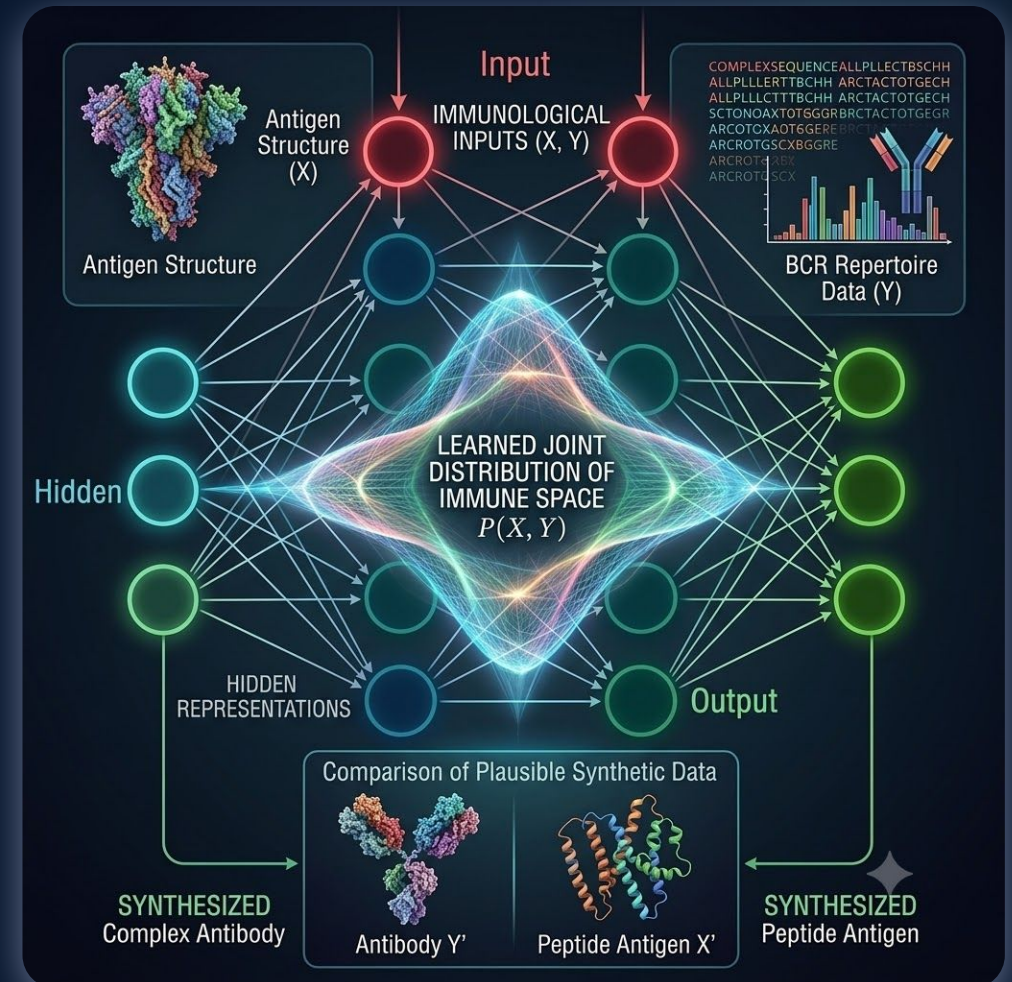
The Generative Principle

Generative AI learns how the data *is generated* by approximating the distributions. With two multi-omic inputs it is a joint-distribution $P(X, Y)$.

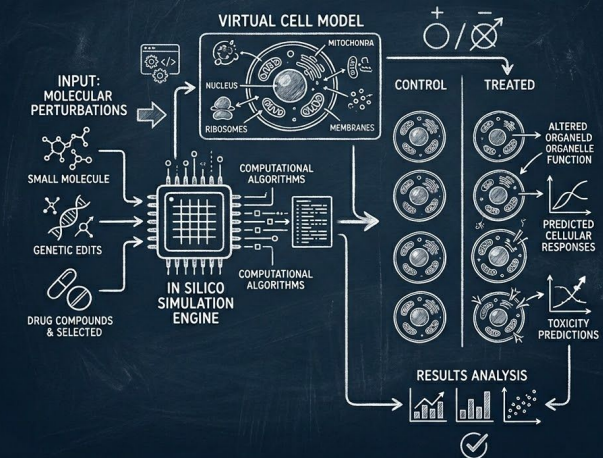
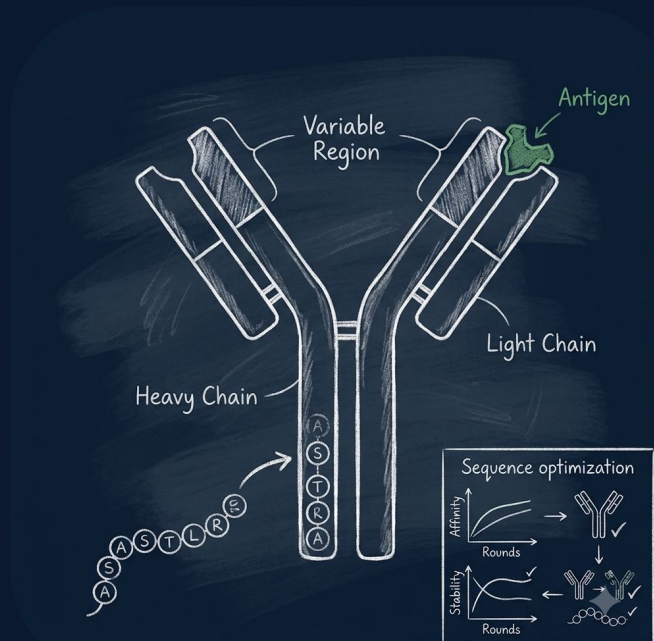
Biological Grammar

By modeling the underlying structure, it can synthesize entirely new data points that are statistically plausible and biologically relevant.

⚡ **Take home:** We don't typically care about how a system generates the data, only that the space we approximate to generate the data is correct.



Synthetic Biology & De Novo Design



Antibody Design

Generating new heavy-chain sequences optimized for binding.

Lead Gen

Synthesizing new small molecules with specific ADMET profiles.

Simulation

Generating "Virtual Cells" to test perturbations in-silico.

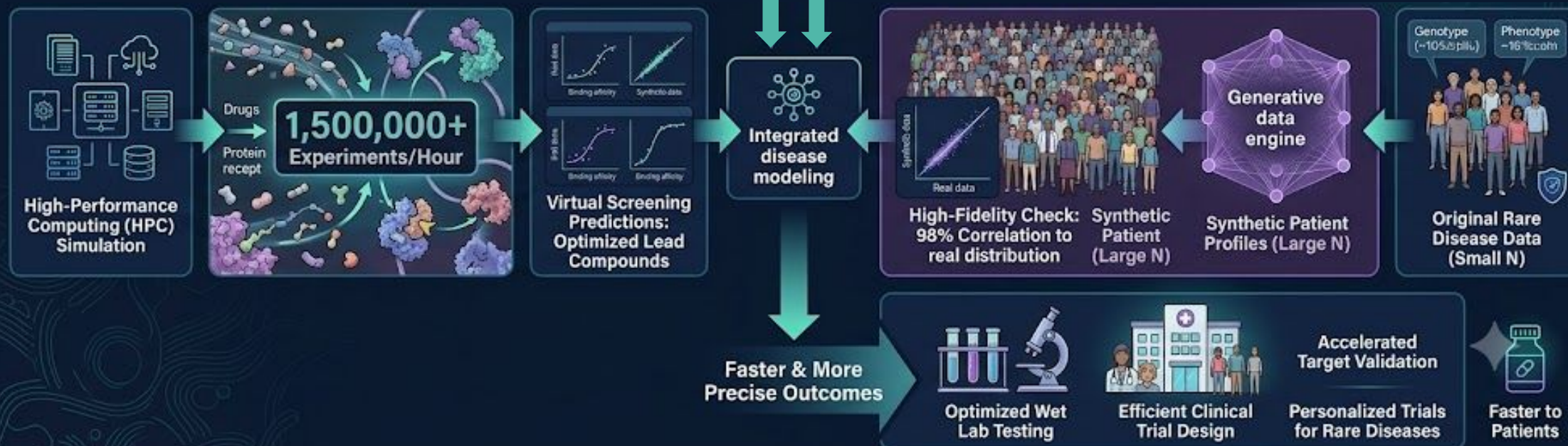
The In-Silico Lab of the Future

Digital Wet Labs

Running millions of virtual drug-receptor binding experiments per hour.

Synthetic Cohorts

Generating high-fidelity synthetic patient data for rare disease modeling where N is low.



Impact: From Filter to Creator

2X

Reduction in
Wet-Lab
Optimization
Rounds

Generative models produce "correct-by-construction" candidates, bypassing the massive screening failures of the past.

NOTE Hallucinations are a feature! See Frank et al., 2023 bioRxiv who demonstrated similar performance to RFdiffusion

<https://doi.org/10.1101/2023.02.24.529906>

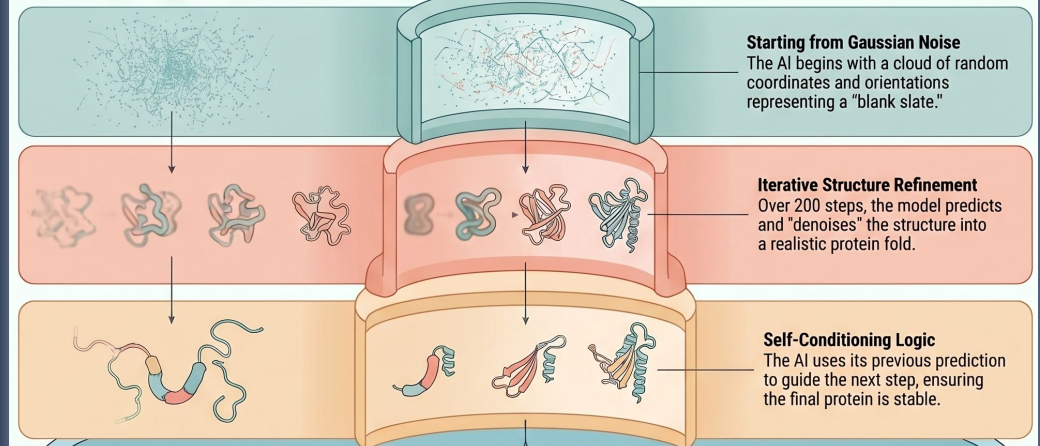
⚡ Take home: Fundamental rules of protein folding do not matter.

❓ Question: Does the model encode the system - can we study it?

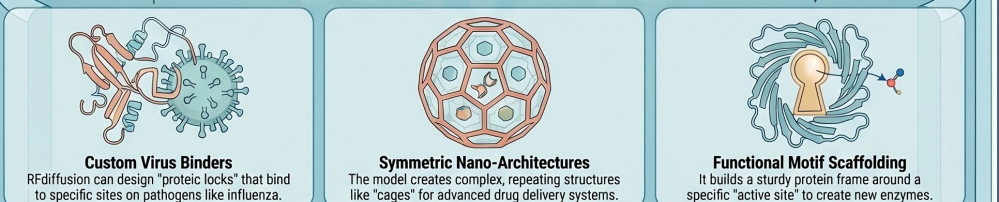
AI Protein Architect: How RFdiffusion Designs the Future of Biology

RFdiffusion is a breakthrough generative AI model that designs new protein structures from scratch by repurposing denoising logic to create custom proteins for medicine and science, from virus binders to industrial catalysts.

FROM CHAOS TO ORDER: THE DENOISING PROCESS



BIOLOGICAL SIGNIFICANCE: DESIGNING WITH PURPOSE



PERFORMANCE COMPARISON: PREVIOUS METHODS VS. RFDIFFUSION

Previous Methods (Hallucination)		RFdiffusion	
~8.5 Minutes	Generation Speed	~11 Seconds	
~100 Amino Acids	Max Protein Length	600+ Amino Acids	
Rapidly declines with size	Success Rate	High consistency across sizes	

<https://doi.org/10.1038/s41586-023-06415-8>

Foundational vs. General-Purpose AI

(A) Levels of Scale in Biological Intelligence

(B) Pragmatic Use

| What are Biological Foundation Models?

- Massive neural networks pre-trained on the "Raw Languages" of life.
- Trained on millions of protein sequences (Uniprot) or single-cells (CellXGene).
- May capture the underlying "Grammar" of biological state-spaces.
- Does not, necessarily, understand the true system underlying the observations it was trained on.

Biological foundation models are artificial intelligence systems trained on massive amounts of unannotated biological data using self-supervised learning. Much like Large Language Models (LLMs) like GPT-4 are trained on massive corpuses of human text to learn the rules of grammar and reasoning, biological foundation models learn the "languages of life"—the complex rules governing DNA, proteins, cells, and tissues.

Once pre-trained on these foundational rules, these models can be fine-tuned or prompted to perform a wide variety of downstream tasks, such as predicting the structure of a novel protein, identifying a disease-causing mutation, or designing entirely new drugs.

Because biology operates at multiple scales, these models are categorized by the specific data modalities they are trained on.

The Self-Supervised Learning Advantage

BIOLOGICAL FOUNDATION MODEL MASKING: SEQUENTIAL PREDICTION

STEP 1: FULL SEQUENCE

(Amino Acids or Genes)

G-L-V-Y-I-H-K-D-N-E

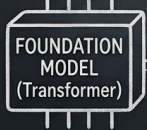
STEP 2: MASKING

A MODEL HIDES PART OF THE SEQUENCE

G-L-V-Y-?-H-K-D-N-E

STEP 3: PREDICTION

THE MODEL PREDICTS THE MISSING ELEMENT



G-L-V-Y-I-H-K-D-N-E

G-L-V-Y-I-H-K-D-N-E

BIOLOGICAL FOUNDATION MODELS: UNLABELED DATA INTAKE & PRE-TRAINING

STEP 1: RAW UNLABELED DATA (MASSIVE DATASET)

G-L-V-Y-I... G-L-V-Y...
M-S-T-P-A... M-S-T-P-A...
ATTCGG... G-L-V-Y-A...
T-S-A... ...

(No expensive clinical ground truth or laboratory labels required to start)

STEP 2: PRE-TRAINING (SELF-SUPERVISED LEARNING)

MODEL LEARNS
UNIVERSAL PATTERNS



(Transformer foundation
model like is pre-trained by
predicting data on itself)

STEP 3: GENERALIZED BIOLOGICAL UNDERSTANDING

PROTEIN STRUCTURE



FUNCTIONAL MOTIFS

...L-G-V-Y-M...

GENE INTERACTIONS

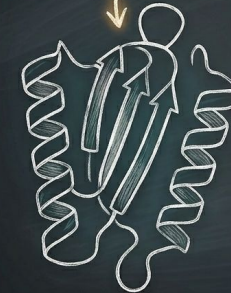


BIOLOGICAL EMERGENCE IN FOUNDATION MODELS

EMERGENT FEATURES (SPONTANEOUSLY IDENTIFIED)

EMERGENT PROTEIN FOLDING

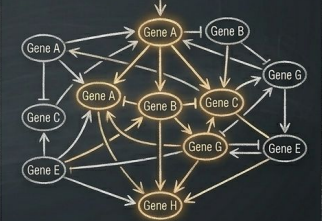
...G-L-V-Y-I-K-E...



Spontaneous recognition of complex
3D folding motifs from simple
sequence patterns.

EMERGENT PATHWAY MAPS

Promoter 1 → Gene X
Enhancer A → Gene Y
Suppressor B → Gene Z
Enhancer C → Gene T
Enhancer X → Gene Z



Regulatory and suppressor elements
spontaneously combine into
intricate pathway maps.

Masking

Model hides part of a sequence and predicts the missing amino acid/gene.

No Labels

Uses raw, unlabeled data. No need for expensive clinical ground-truth to start.

Emergence

Models spontaneously learn functional protein structures and cellular pathways.

The Potential of Unlabeled Data

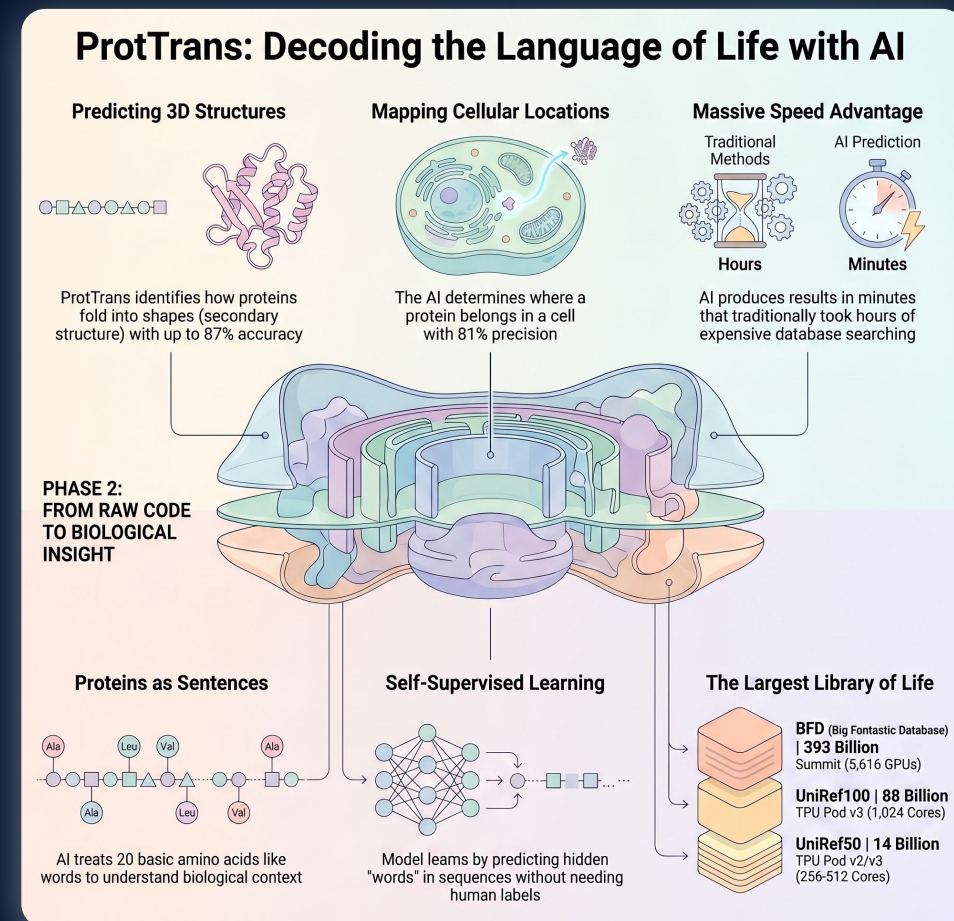
Foundation models thrive on the massive repositories of data we have generated over the last 30 years.

200M+
0.28%

Available Protein
Sequences to
Learn From

Reviewed

<https://www.uniprot.org/uniprotkb>



<https://doi.org/10.1109/tpami.2021.3095381>

| Scaling Laws & 'Intelligence' $\neq$ Guarantee Success

100B

<https://doi.org/10.1101/2023.07.05.547496>

Parameters in Top
Protein Foundation
Models

Performance scales predictably with compute and data, making AI a reliable "R&D Infrastructure" investment, **for sequence based data.**

1B

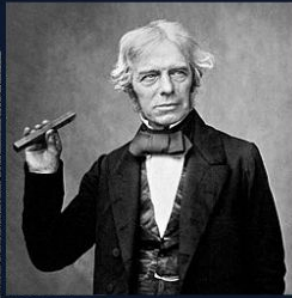
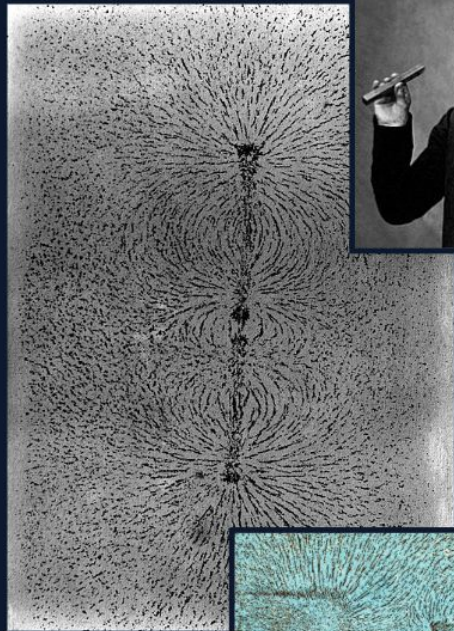
<https://doi.org/10.1038/s41467-026-69102-y>

Parameters in Top
Cellular Foundation
Models

Performance bottlenecks with compute and data, meaning **we can't just throw more compute at the problem.**

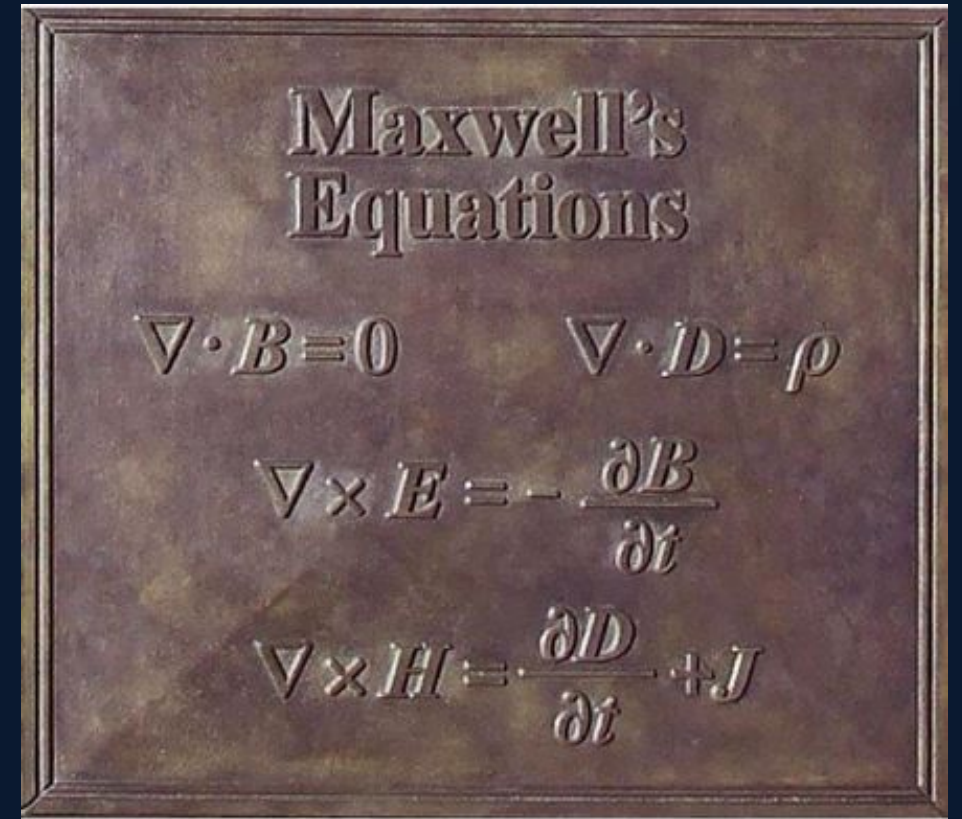
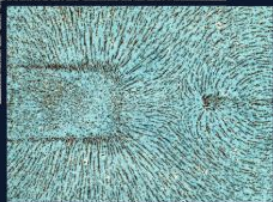
How we **record** & **operate** on data matters

Recall ↓



Michael Faraday (1851)

Iron filings showed the phenomenon but neither proved nor described the field.

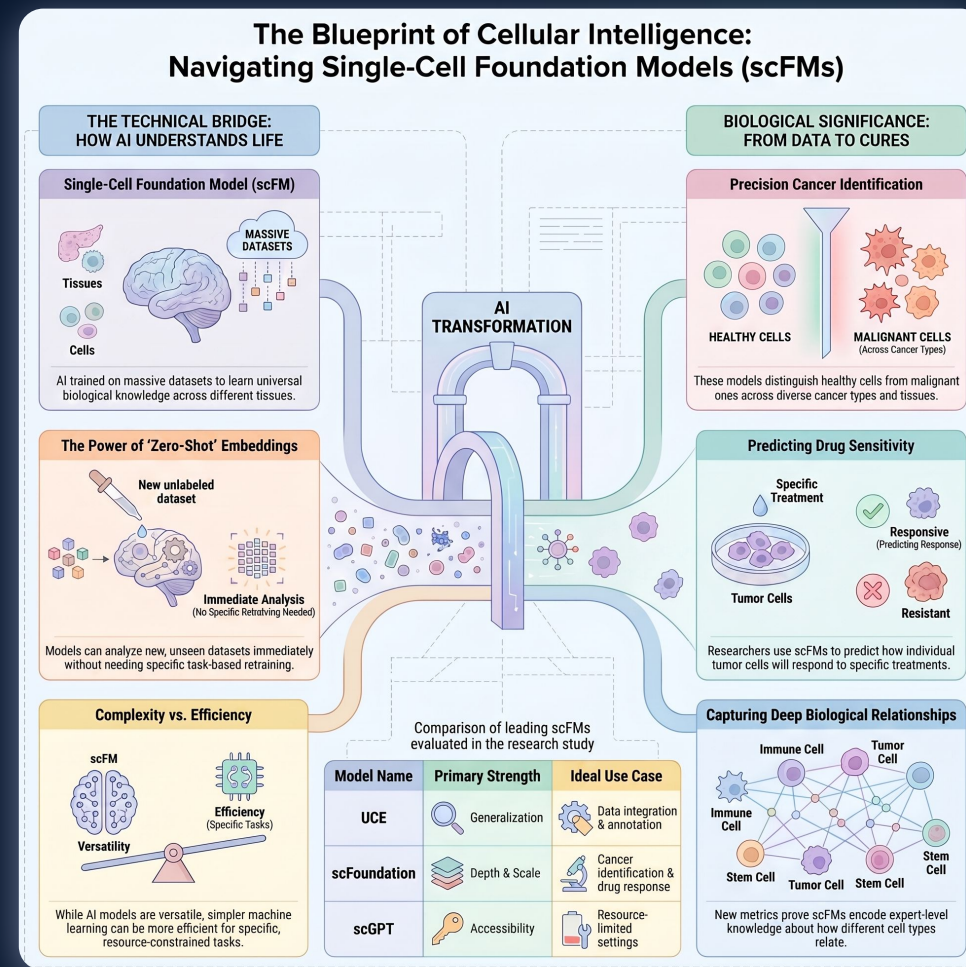


Issues in single-cell Foundational Models

"Biology-driven insights into the power of single-cell foundation models"

— Wu J., et al. (2025)

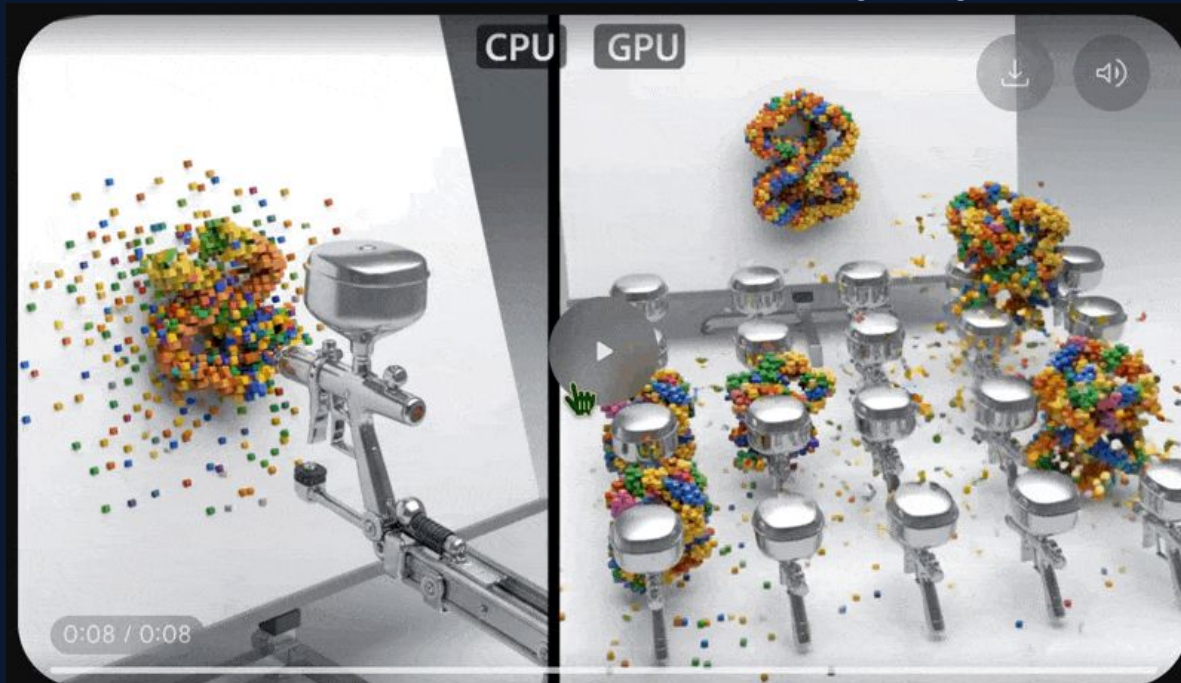
A study of six single-cell foundation models. scFMs are robust and versatile tools for diverse applications. **But simpler machine learning models are more efficient at adapting to specific datasets, particularly under resource constraints.**



<https://doi.org/10.1186/s13059-025-03781-6>

Serial vs Parallel Computation

VEO 3.1 MP4 $\Rightarrow$ 30 min ; 4 iterations ; Inability to adjust



Foundation models provide near-instant results at extreme infrastructure costs. Their flexibility varies within their domains but other, more traditional, methods are required to maximise flexibility at reduced cost until they (A) produce enough consistent data to train foundation models and (B) scales-of-efficiency justify the investment.

Gemini Code Simulation $\Rightarrow$ 10 min ; 1 iteration ; 4 adjustments



| General-Purpose LLMs (Gemini, Claude)

-  Experts in Human Knowledge and Code.
-  Excellent logical reasoning and planning.
-  Fluent in the scientific literature.

Important Role: The Assistant

They act as the "Cognitive Orchestrator" that writes the code to analyze your biological data.

LLMs as Cognitive Orchestrators



Synthesis

Summarizing thousands of PubMed papers in seconds.



Coding

Writing robust Python/R scripts for complex omics analysis.



Planning

Designing multi-step R&D workflows and experiments.

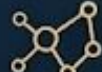
The Languages of Life

DNA, RNA, and Proteins as Tokens

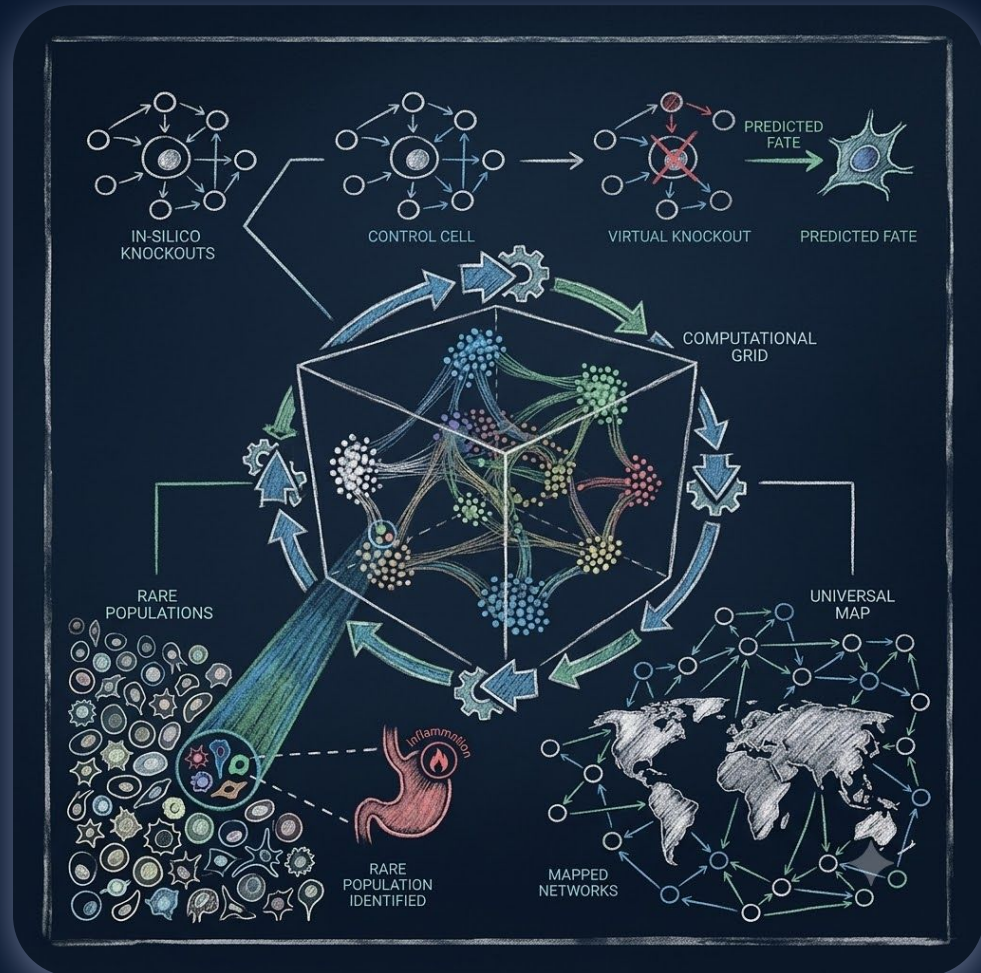
| Tiers of Biological Foundation Models ⇒ Architectures

Modality	Data Type / Representation	Application
Genomics	Nucleotide sequences (A, T, C, G)	Identifying disease mutations
Proteomics	Amino acid strings / 3D coordinates	Structural prediction & binder design
Transcriptomics	Gene expression matrices	Predicting drug effects on cells
Cheminformatics	SMILES strings / Molecular graphs	<i>De novo</i> small molecule drug discovery
Pathology	Histology slides / Gigapixel images	Oncology diagnosis & prognosis

Commonalities & Disparities by Data Type

SHARED COMMONALITIES  PRE-TRAINING  MASSIVE UNANNOTATED BIOLOGICAL DATA  ATTENTION MECHANISM (TRANSFORMER BACKBONE)  DOWNSTREAM FINE-TUNING FOR SPECIFIC TASKS 					
	GENOMICS (DNA)	PROTEOMICS (PROTEINS)	TRANSCRIPTOMICS (SINGLE-CELL)	CHEMINFORMATICS (SMALL MOLECULES)	PATHOLOGY (TISSUE IMAGES)
DATA TYPE & KEY IDEA	 DNA Sequences (sequence-to-sequence)	 Folded Protein Chain and 3D folding Structure	 Cell state: (Permutation-Invariant Set)	 Small Molecules (Molecule graphs and charts)	 Tissue Images (Histology slide patches)
STRUCTURE & GEOMETRY	Captures long-range dependencies.	Amino Acid Chains & 3D Folding Structure	Gene Expression Profile (Permutation-Invariant Set)	Non-linear Molecular Graph (Nodes & Edges)	2D Grid of Pixels and Spatial Relations
TOKENIZATION METHOD	1D Linear Nucleotide Sequence	1D Sequence-to-3D Folding Map	Bag-of-Genes, Non-sequential Set	Non-linear Molecular Graph (Nodes & Edges)	2D Grid of Pixels and Spatial Relations
KEY DISPARITY (non-sequential?)	K-mer-based Processors	Amino Acid Tokenization	Gene Tokenization (Set-of-Genes Processors)	Nodes and Edges Tokenization	Spatial Patch Tokens
ATTENTION MECHANISM DIFFERENCES	Retains 1D Sequential Order. Long-range sequence attention.	Implicit 3D Folding Context. Biased 1D sequence tokens	Set-of-Genes Bag Attention. Permutation-invariant attention.	Graph & Geometric Attention. Graph and geometric relationships.	2D Spatial Patch Attention (Local vs Global)
MODALITY-SPECIFIC ARCHITECTURES	 DNA-Specific Transformers (1D Sequence Model)	 Protein Language Protein Language Models (Sequence-to-Structure Prediction)	 Single-Cell Set-Based Models	 Graph Neural Networks (GNNs)	 Vision Transformers (ViTs)

Learning Single-Cell State Spaces



- 🧪 In-silico knockouts: Predicting cell fate without the pipetting.
- 🧪 Discovery of rare cell populations driving autoimmune flare-ups.
- 🧪 A "Universal Map" of gene regulatory networks.

2024: scGPT & Omics Integration

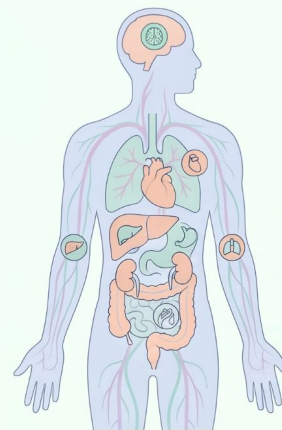
“ *scGPT: toward a foundation model for single-cell biology.*

— Cui, H., et al. (2024)

Standardizes batch correction and cell-type annotation across diverse datasets.

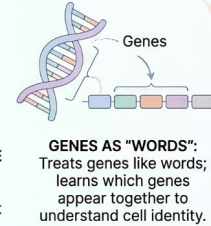
scGPT: The AI Foundation Model Decoding the Language of Cells

How it Works: The “GPT” for Genetics

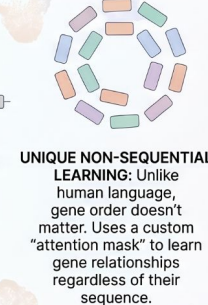


Transformer

MASSIVE KNOWLEDGE BASE: Pretrained on over 33 million human cells across 51 different organs to capture total cellular diversity.



GENES AS “WORDS”: Treats genes like words; learns which genes appear together to understand cell identity.



UNIQUE NON-SEQUENTIAL LEARNING: Unlike human language, gene order doesn't matter. Uses a custom “attention mask” to learn gene relationships regardless of their sequence.

Biological Significance & Performance



PREDICTING THE “UNSEEN”: Predicts how a cell will react to novel genetic changes, accelerating drug discovery.

SUPERIOR ACCURACY: In identifying cell types (e.g., healthy vs. Multiple Sclerosis cells), scGPT consistently hits **~85% accuracy**, outperforming specialized models.



SCALING EFFECT: Research shows the more data scGPT sees, the smarter it gets at uncovering hidden gene networks and pathways.

Model Performance (Cell Type Annotation Accuracy)

scGPT (Generalist; 33M cells)	~0.85+
scBERT (Transformer-based)	~0.75 - 0.80
TOSICA (Task-specific)	~0.70 - 0.78

NotebookLM

<https://doi.org/10.1038/s41592-024-02201-0>

Batch Correction & Multimodal Maps

- ≡ Integrates transcriptomics with ATAC-seq and protein data.
- ≡ Solves the batch-effect problem that has plagued R&D for decades.
- ≡ Enables cross-species mapping from mouse to human with high fidelity.

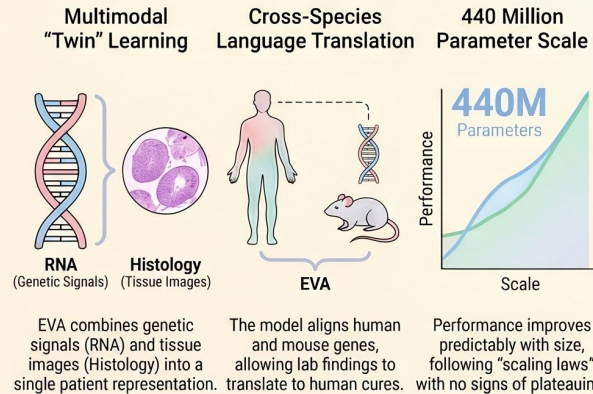
EVA: The Universal Immune Model

A specialized foundation model specifically for systems immunology and inflammation. Harmonizes clinical pathology with single-cell dynamics to predict organ-level immune response.

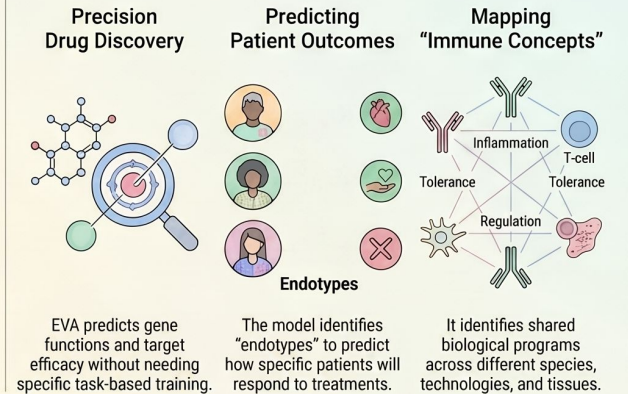
EVA: A Universal AI Brain for the Immune System



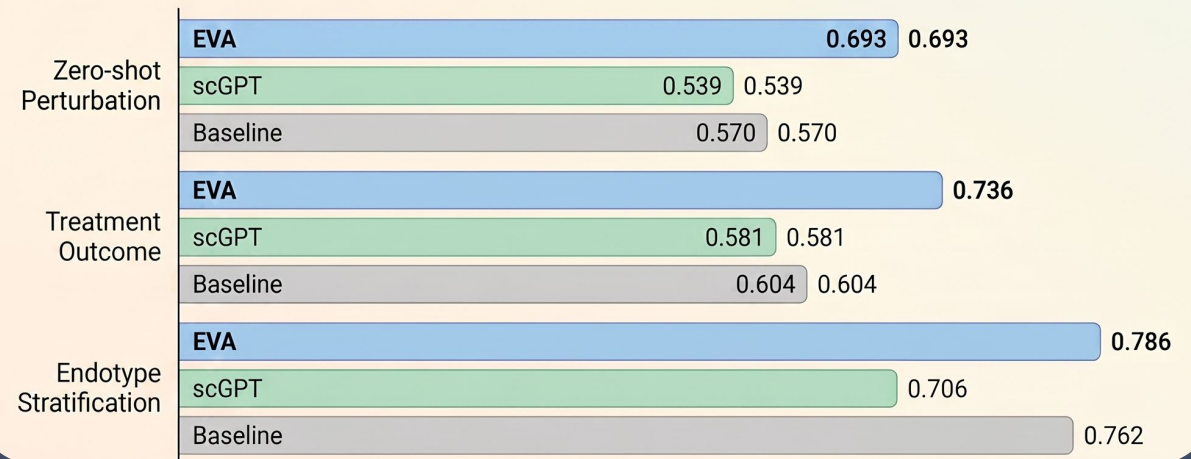
The Technical Foundation: How EVA Learns



Biological Significance: Impact on Medicine



EVA Outperforms Others in Key Drug Development Tasks



<https://doi.org/10.48550/arXiv.2602.10168>

Remember Attention? It's a Problem

Context is indeed everything

Where it all began

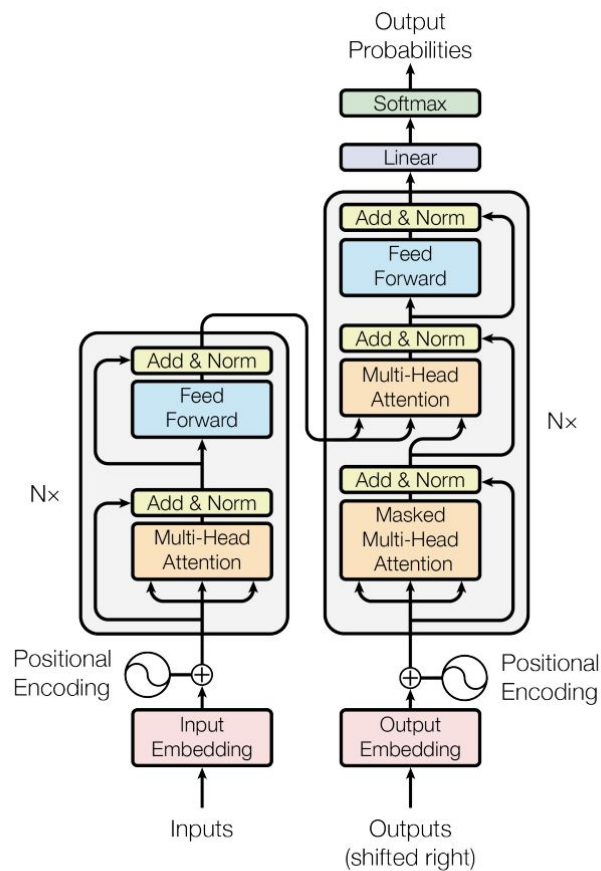


Figure 1: The Transformer - model architecture.

<https://doi.org/10.48550/arXiv.1706.03762>

“ Attention is all you need ”

— Vaswani, A., et al. (2017)

Replaced slow sequential processing with a parallelizable self-attention mechanism that enabled models to scale efficiently across massive datasets, creating the architectural blueprint for all modern foundational models.

Attention, Context, and Search Space... Oh My!

Designed how the attention mechanism in AI models forces into localized, suboptimal search spaces based on context.

Bright peaks represent highly accurate clusters of **objective facts and logic**

User Prompt +
Context Window

$O(N^2)$
attention
matrix

➤ **The Blindspot:** Because the attention mechanism over-indexes on the provided context, the model becomes mathematically incapable of 'seeing' the correct answers outside the semantic gravity well. **The search space has collapsed.**

Next Token
Generation

Pre-trained Latent Space (The Universe of Possibilities)

Before a prompt is entered, the model possesses an open, unconstrained probability distribution across its entire training dataset.

The Attention Mechanism / Softmax Filtering

Context forces the model to calculate the relevance of every token against every other token. This heavily warps the probability landscape, creating a localized gravity well.

The Constrained Search Space (Path Dependency)

Agentic AI: The Connecting Bridge

Autonomous Reasoners in the Lab

aka. The answer to “Should I use traditional ML, Foundational, or General Purpose AI”?

| Solving the Hallucination Bottleneck

- 🛡️ Problem: LLMs can "invent" non-existent protein-protein interactions.
- 🛡️ Solution: Agents use "Grounded Search" to verify facts.
- 🛡️ Agentic Workflow: Plan -> Execute -> Verify -> Correct.

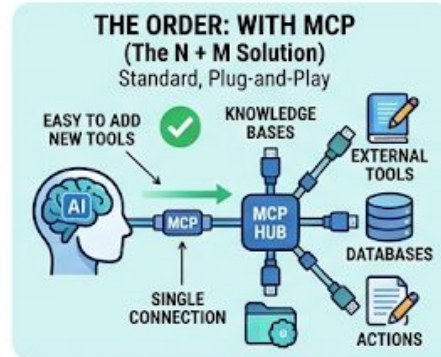
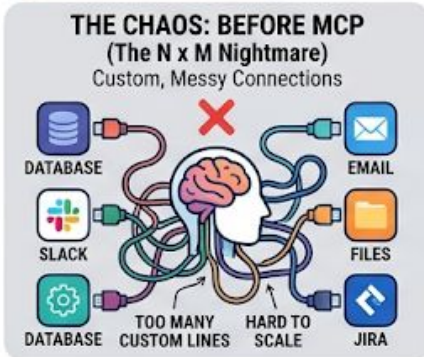
The Agentic Tool-Use Architecture

MCP: THE UNIVERSAL STANDARD FOR CONNECTED AI

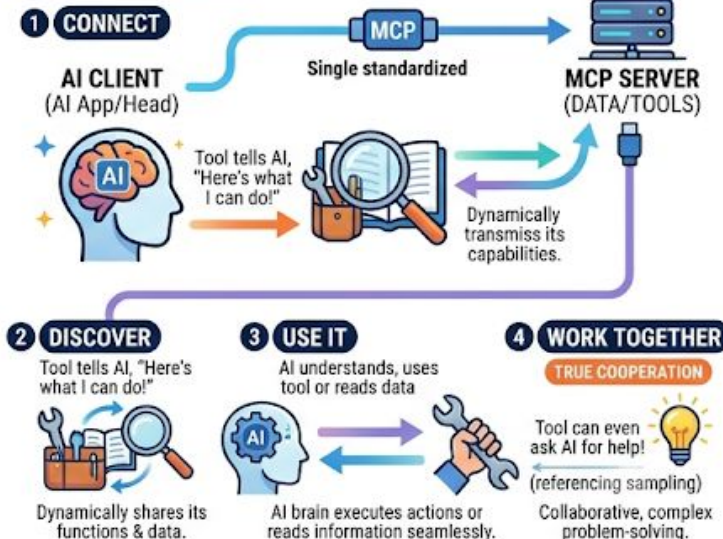
1 Think of MCP as the USB-C Port for Your AI!



2 Before vs. After



3 How it Works (Simple Terms)



WHAT: Agents connect to validated tools: BLAST, Ensembl, UniProt, Python libraries, and Foundational Models.

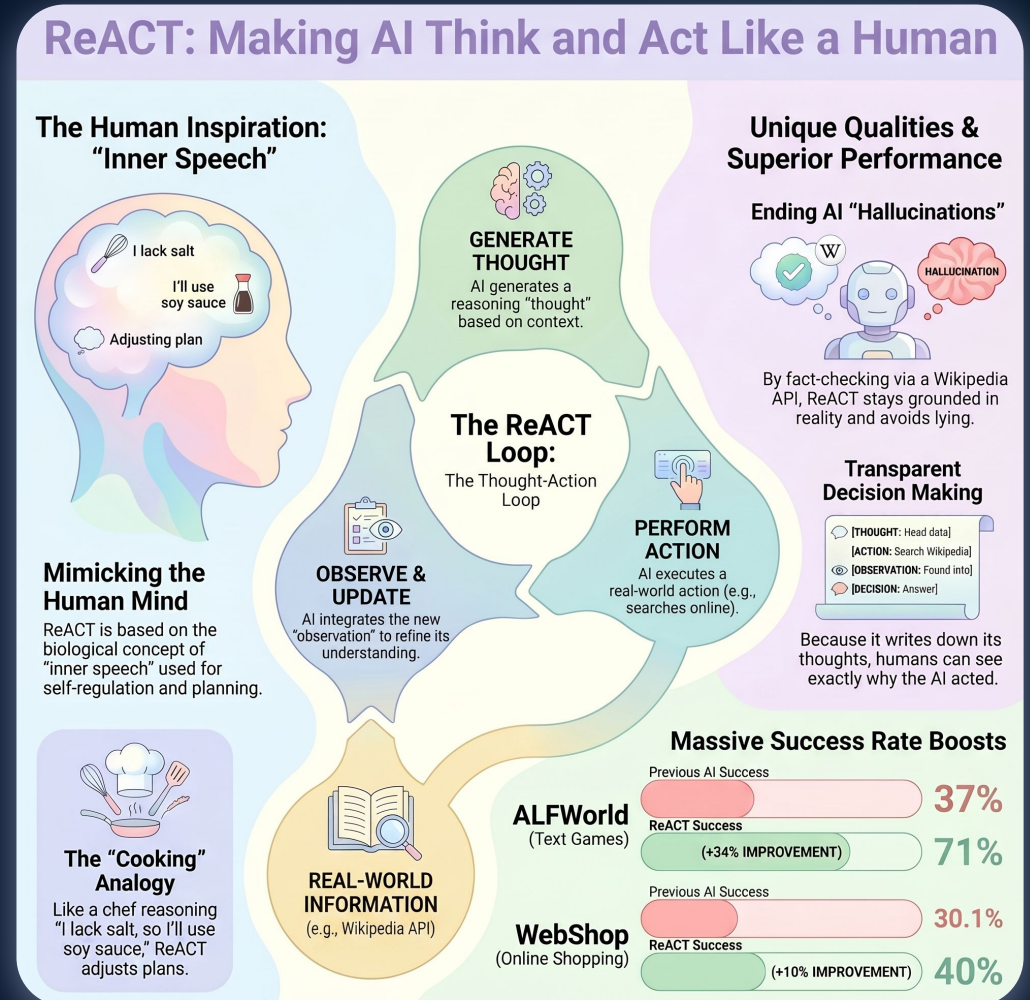
ENABLES: They act as the "Agent-in-the-Loop" for high-throughput screening and clinical stratification.

2022: The ReAct Framework

"ReAct: Synergizing reasoning and acting in language models."

— Yao, S., et al. (2022)

Established the pattern of using an LLM to generate "thoughts" that drive external "actions."



<https://doi.org/10.48550/arXiv.2210.03629>

| Self-Correction in Data Analysis

- ↻ Agents can catch their own coding errors and fix them in real-time.
- ↻ They evaluate the "Biological Plausibility" of their results.
- ↻ Reduces the burden on bioinformatics staff by 80%.

Autonomous R&D Teams



Team of Agents

Assigning specialized roles:
Literature Agent, Coding Agent,
Clinician Agent.



Consensus

Agents cross-verify each other to
eliminate individual hallucinations.



Acceleration

Parallelized discovery pipelines
that run 24/7 without human
fatigue.

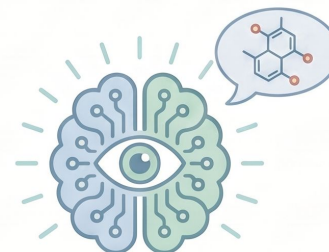
2023: Coscientists & Harnesses

"Autonomous chemical research
with large language models.

— Boiko, D. A., et al. (2023)

Showed that AI agents can autonomously design, plan,
and execute chemical research workflows. Called it a
“Coscientist” presenting a ‘harness’ to control and
enable several instances of different agents directed
toward specific tasks over long running periods.

Coscientist: The AI That Masters the Laboratory



The Planner Hub

A GPT-4 “brain” receives simple text prompts like “perform a Suzuki reaction.”



Autonomous Toolkit

The AI searches the internet for procedures and reads hardware manuals to learn.



Physical Execution

Coscientist writes Python code to control robotic liquid handlers and cloud laboratories.

Unique Qualities and Global Impact



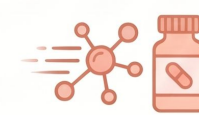
Advanced Chemical Reasoning

It justifies reagent choices based on concepts like reactivity, selectivity, and electronic properties.



Self-Correcting Mastery

When code fails, the AI consults documentation and automatically fixes its own errors.



Accelerated Discovery

Speeds up the development of life-saving medicines by optimizing chemical yields automatically.

Performance Comparison: Synthesis of Common Pharmaceutical Compounds

Compound	Standard GPT-4 Performance	Coscientist (Search-Enabled)
Aspirin	High (400-500)	Perfect (400-500)
Ibuprofen	Failed (400-500)	Passing (400-500)
Nitroaniline	Moderate (400-500)	Perfect (400-500)

<https://doi.org/10.1038/s41586-023-06792-0>

The Pipeline

The application of AI to Drug Development

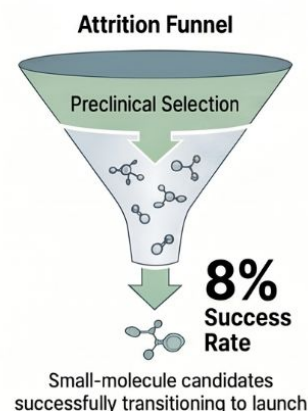
Pharma Economics vs. Hype-onomics

The Pharma R&D Evolution: From Financial Barriers to the AI-Enabled Pipeline

The 2010 Economic Reality (Historical Baseline)

The \$1.8 Billion Financial Barrier

Fully capitalized cost for one NME to market



13.5-Year Development Cycle

Average duration from initial discovery to market entry

The Pharmaceutical Value Equation:

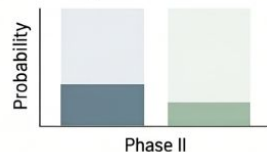
$$P = \frac{\alpha(WIP \times p(TS) \times V)}{CT \times C}$$

WIP = Work-In-Process
p(TS) = Success Probability
V = Cycle Time
C = Cost

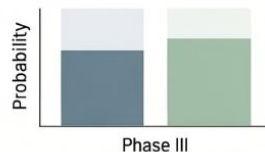
CT = Cycle Time
C = Cost

Phase II & III Attrition Benchmarks

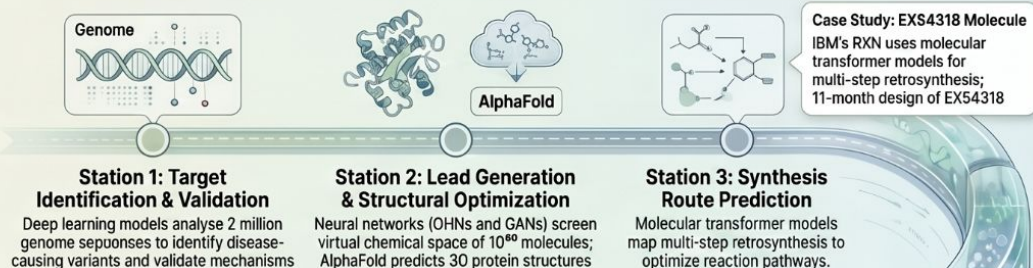
Phase II:
34% Success Rate
2.5 years



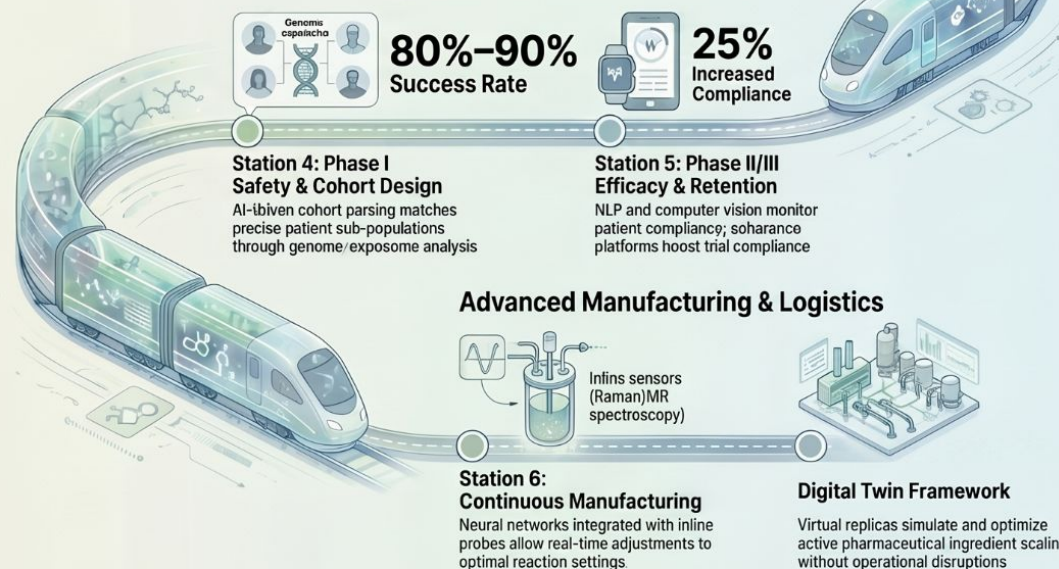
Phase III:
70% Success Rate
2.5 years



Preclinical Discovery (The AI R&D Sweet Spot)



Clinical Development (The Acceleration Zone)



The Historical Integration Gap:
Translating Retrospective Financial Reality into Predictive AI Architecture

<https://doi.org/10.1038/nrd3078>

<https://doi.org/10.1016/j.drudis.2020.10.010>

<https://doi.org/10.3390/pharmaceutics16101328>

| Competitive Intel & Landscaping



Autonomous monitoring of clinical trial registries and preprints.



Automated synthesis of the drug landscape for any target cytokine.



Real-time alerts for IP white-space discovery.

Good AI Practice , GxP

Without ethics and governance complement the scientific method

Consider The Potential of AI to Damage Healthcare

The Journey of Your Health Data: Mapping the NHS Landscape



Step 1: The Origin (Care Providers)

Data is recorded at 6,544 GP practices and 216 hospital trusts across England.

Step 2: The Extractors (Intermediaries)



Organizations like NHS Digital collect and package this data for secondary uses.

Step 3: The Consumers (End Users)

Over 460 organizations, including 216 universities and 106 pharmaceutical companies, use this data.



Universities & Research: 216 entities
Observational research and clinical audits



Pharma & Life Sciences: 106 entities
Scientific research and vaccine evaluation



Non-Profit & Clinical: 44 entities
Policy research and health evaluation

Challenges & The Path Forward



The Transparency Gap

Most data usage occurs 2-3 steps away from the patient, making it hard to track.



35% Secure Access Rate

Only about a third of data consumers currently use highly secure "data environments."



Recommendation: Secure Data Environments (SDEs)

Shifting to "data-access" instead of "data-sharing" to keep information inside secure NHS platforms.

[https://doi.org/10.1016/S2589-7500\(23\)00157-7](https://doi.org/10.1016/S2589-7500(23)00157-7)

AI without control

Accelerates the goals of the individual or group increasing duplicative efforts and bias.

- Deepens Transparency and Public Trust Crisis
- Escalates Privacy and Breach Risks
- Wastes Funding and Dilutes Value
- Perpetuates Data Bias and Masks Deficit
- Fragments Collaboration and Reproducibility

The Future of AI

Your guess is as good as mine, but ...

THE HYPE CYCLE COMPOUNDED

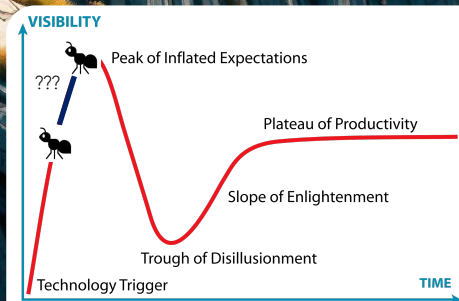
Don't focus on technology, focus on valid needs and adopt AI as necessary.
Watch, wait, learn - together.

A hype cycle depicts phases of enthusiasm for a new tech: initial excitement, over-expectation, disillusionment, and productive use.

- AI's development is marked by overlapping hype cycles, each depicting phases from excitement to productive use.
- From rule-based systems to deep learning, each era has its own cycle, with daily AI discoveries adding to the complexity.
- AI advancement is accelerating due to computational growth, data availability, and intersections with other technologies.
- This compounded and quickening trajectory makes predicting AI's future increasingly challenging.

We are in no position to accurately estimate the end-game for AI or decide on the most impactful use of AI in healthcare.

We can only assume the impact will be extreme & **we must learn**.

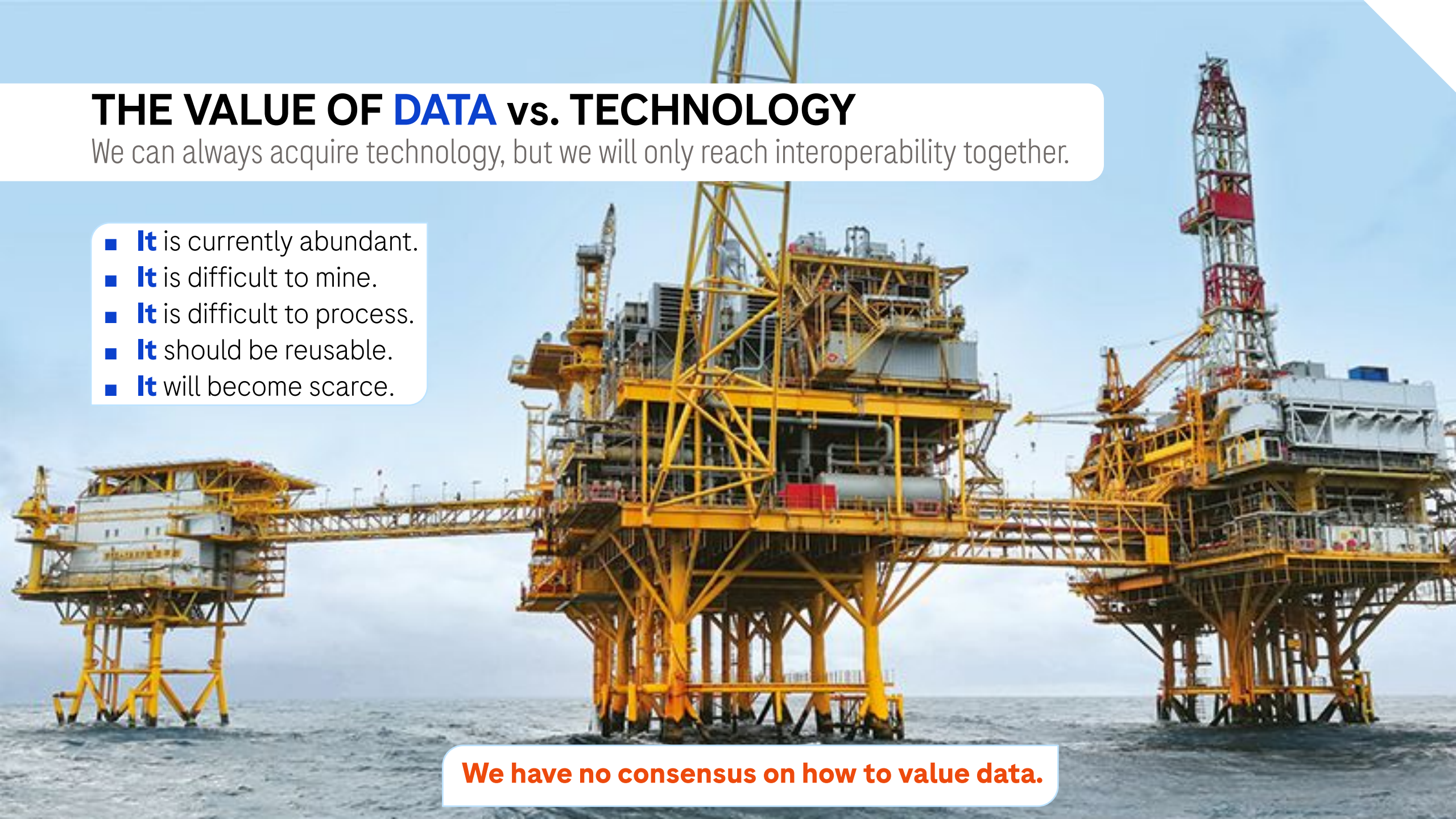


THE VALUE OF **DATA** vs. TECHNOLOGY

We can always acquire technology, but we will only reach interoperability together.

- **It** is currently abundant.
- **It** is difficult to mine.
- **It** is difficult to process.
- **It** should be reusable.
- **It** will become scarce.

We have no consensus on how to value data.

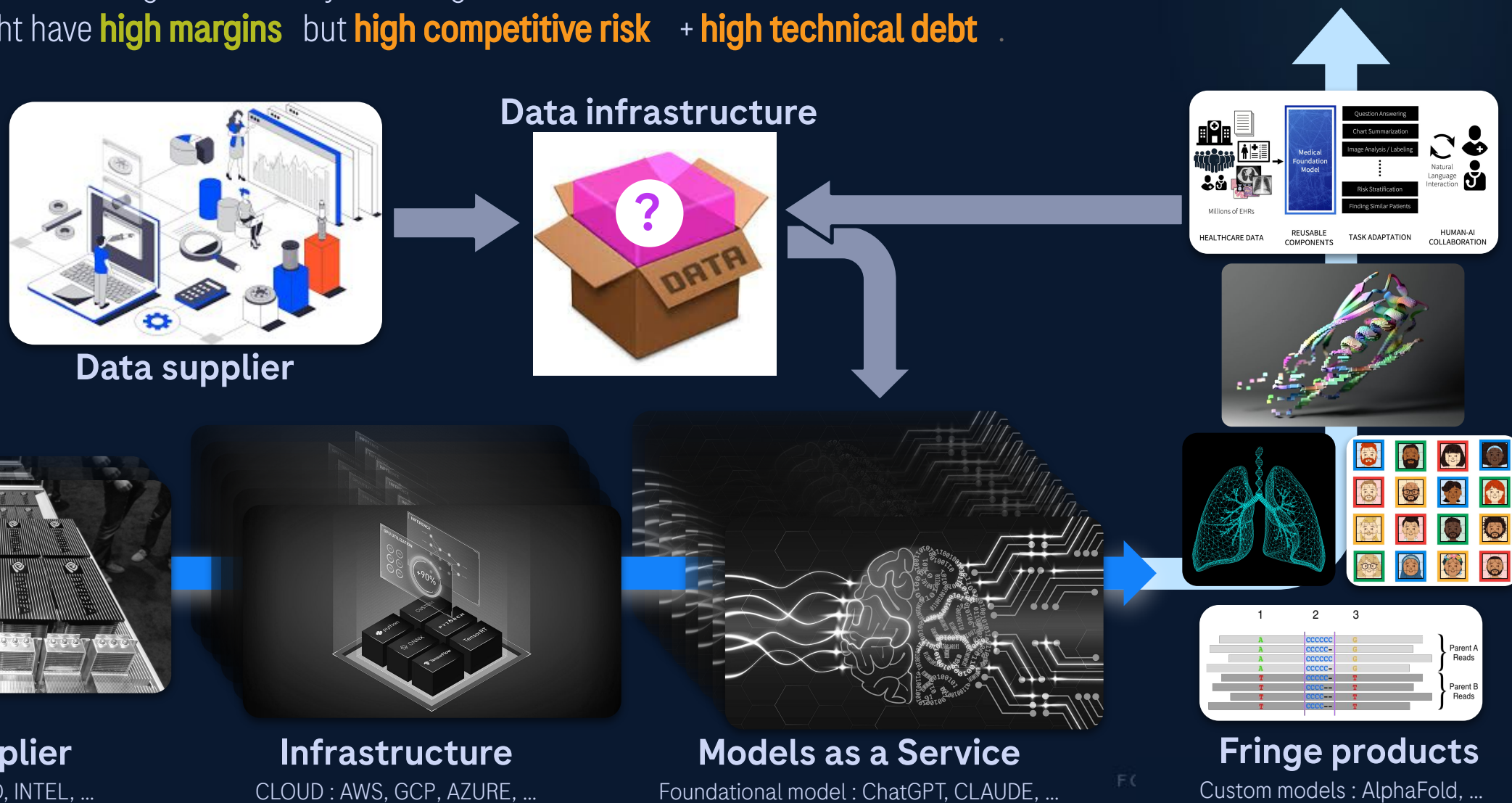


THE AI SUPPLY CHAIN - FRINGE BENEFITS

The safe business of artificial intelligence is in infrastructure & data.

Infrastructure has a high bar to entry, the fringe is difficult to ID, and data has no consensus.

All else might have **high margins** but **high competitive risk** + **high technical debt**.



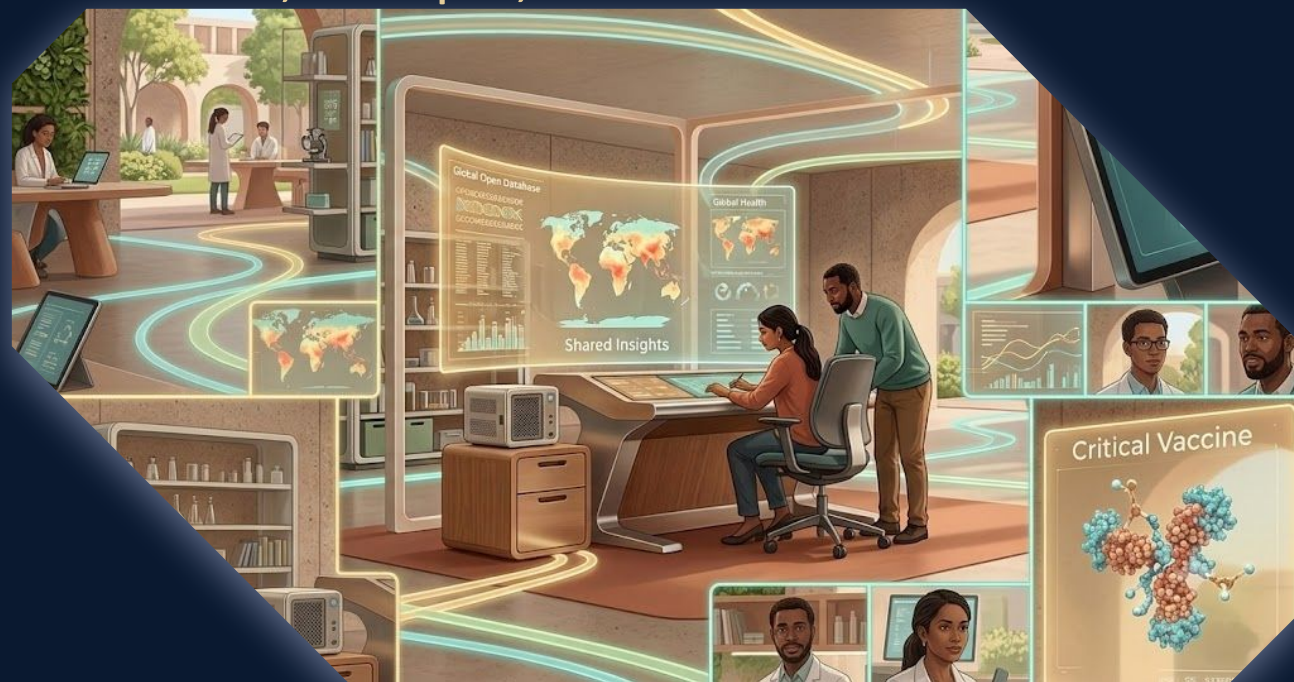
| Data Sovereignty & Secure R&D

100%

Local & Private Execution in Enterprise AI

New frameworks allow running these agents locally, ensuring proprietary clinical trial data never leaves the pharma's firewall.

⇒ Smaller, cheaper, more accessible



Summary

-  Historical foundation: We have the data and algorithms.
-  The Change: From classifiers (Predictive ML) to creators (GenAI).
-  Foundation Models: Pre-trained "Biological Engines."
-  The Bridge: Agentic AI orchestrates the entire discovery loop.
-  GxP: With great efficiency comes the responsibility to work together
-  The Future: More data and more cost effective models

Questions?

Next Up: Hands-on Workshop in Google Colab

Thank you for your *attention*.