XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery



AI in drug target discovery

Hopes, hype, and reality

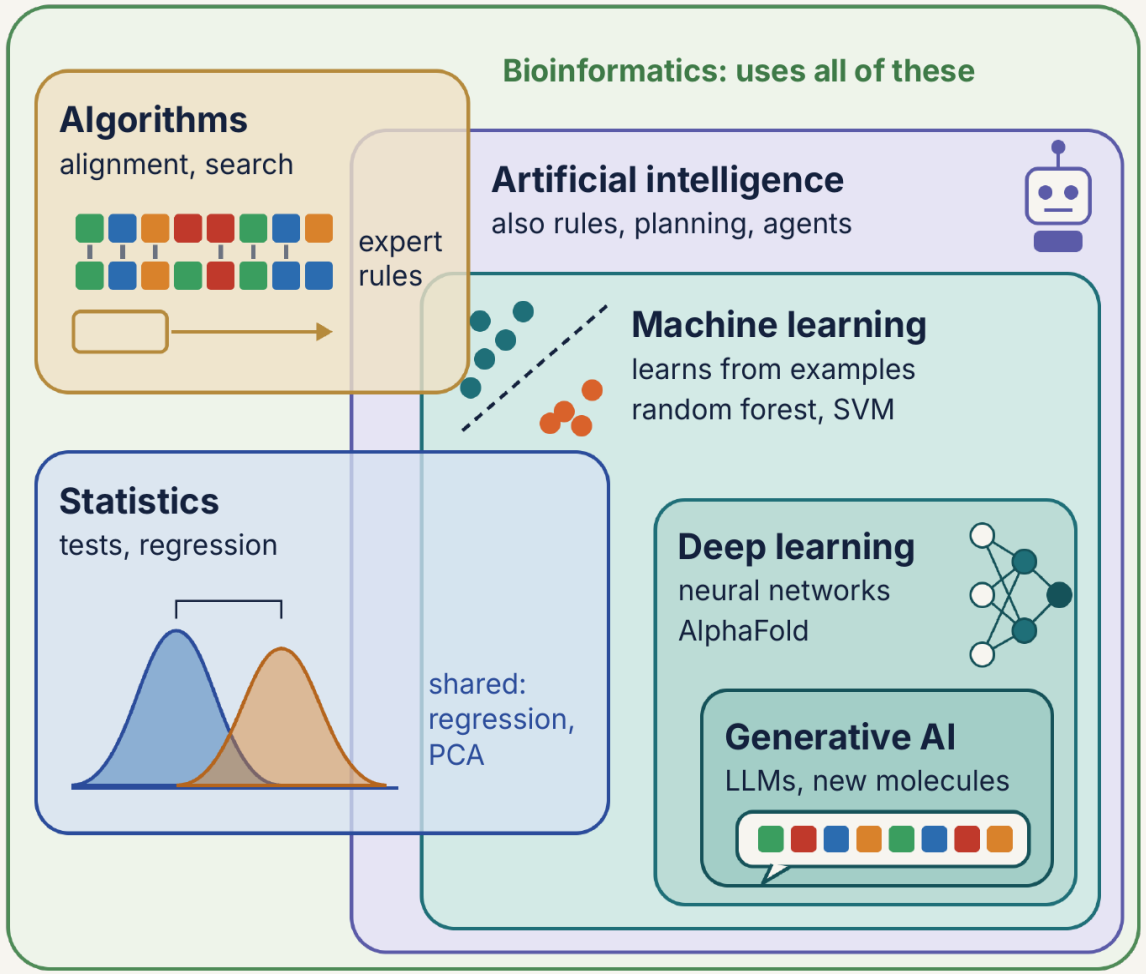
Alexander Kel[1] · Roman A. Zubarev[2] · Gleb Khegai[1] · Aleksandr Kalmykov[1] · Daria Stelmashenko[1]

[1] geneXplain GmbH, Wolfenbüttel, Germany

[2] Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden



What people call “AI”?



Algorithm = a lab protocol
Fixed steps, same result every time

Statistics = a controlled comparison
Is the difference between groups real?

Machine learning = a trainee
Learns from labelled samples, then judges new ones

Generative AI = an author
Writes new text, sequences or molecules

The pitch says → it really is

“AI biomarker” → logistic regression (statistics)

“AI-found subtypes” → clustering (ML)

“AI motif finder” → PWM scan (algorithm)

HOPES

AI-assisted target discoveries, 2023–2026

Target · disease	What the AI did	Best evidence	Honest note
TNIK · lung fibrosis	PandaOmics: omics, networks, text and graph ML ranking	Phase 2a (71 patients); Phase 3 started 2026	AI share not separated from classical analysis
CHK2 · Parkinson’s	XunZi: reasoning over literature and multi-omics data	Genetic and drug inhibition protect neurons in mice	Strong preclinical; no human data
ALOX5, OXTR · psoriasis	LLM “VirtualCRISPR” flags unexpected hits of a genome-wide CRISPR screen	Topical inhibitors reduce disease in mice	The screen found the hits; AI ranked novelty
OSM–OSMR · melanoma immunotherapy	MIDAS: multimodal graph neural network	Patient tumour explants: less T-cell dysfunction	No tumour regression shown
GPNMB · CAR-T	LLMs + single-cell atlases, human in the loop	CAR-T controls tumours in mice	Protein was already a drug target (ADC)

Ren 2024 Nat Biotechnol; Xu 2025 Nat Med; Huang 2026 Nat Biomed Eng; Zhao 2026 Nat Commun; Augustine 2026 Nat Mach Intell; Baker 2026 Cell

How a computer picked TNIK

1 · Data
patient lung tissue + millions of texts

→

2 · Scores
classic statistics + some machine learning

→

3 · Filter
only kinases, new, easy to drug

→

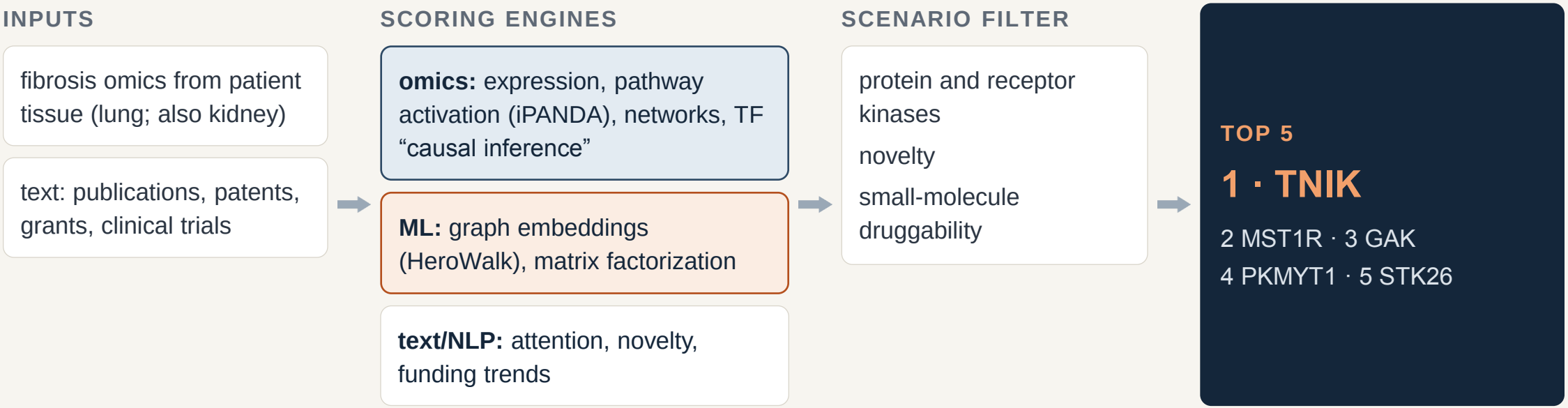
WINNER

TNIK

We see the winner, but **not which score made it win**.

Ren F et al., Nature Biotechnology 2024 · Insilico Medicine, PandaOmics platform

How PandaOmics chose TNIK in 2019



Many engines, one combined rank. Blue is classical bioinformatics, orange is machine learning.
How much each engine moved TNIK is not reported.

Ren F et al., Nat Biotechnol 2024 (doi:10.1038/s41587-024-02143-0); PandaOmics: J Chem Inf Model 2024; Insilico press release, March 2024

Seven years from ranking to Phase 3

2019	2021	Phase 1	Phase 2a	FVC	2026
TNIK ranked first	candidate; microdose study, Australia	78 healthy (NZ) + 48 (China): safety, PK	71 IPF patients, 12 weeks	+98.4 mL (60 mg) vs −20.3 mL placebo	Phase 3: 320 patients, 52 weeks

What it proves

- a computational pipeline can nominate a target that reaches patients
- target and molecule from one platform, fast
- a first efficacy signal in a small trial

What it does not prove

- that ML, not classical omics and networks, found TNIK
- TNIK's rank in an open, genome-wide search
- long-term efficacy, or that TNIK alone explains the effect

Ren F et al., Nat Biotechnol 2024; Xu Z et al., Nat Med 2025; Phase 3 per trial registry trackers (2026)

PAPER 2 · XUNZI · 1/3

A small AI with a big library

The reader

a language model that read **24 million** research papers

The map

a knowledge map: **620,000** biological objects, **6 million** known links

ANSWER

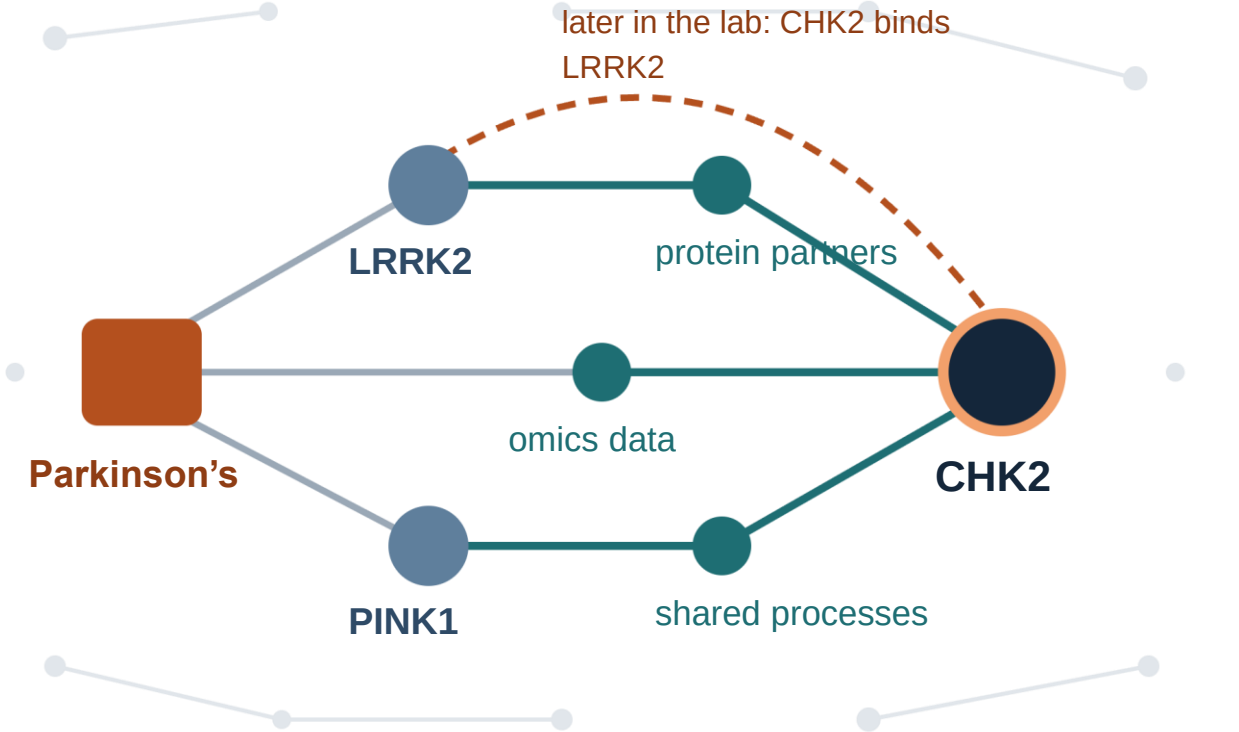
a ranked list of genes, and an explanation why

Not a giant chatbot: a small model made strong by **curated knowledge**.

Huang X et al., Nature Biomedical Engineering 2026 · details from a published summary, to verify

PAPER 2 · XUNZI · HOW THE MAP HELPS

The map finds a path that papers alone miss



reading papers only

0.70

reading + knowledge map

0.88

The map links a gene to the disease through **known interactions and processes**, even when few papers connect them directly.

Simplified illustration, not the paper's graph · scores: finding known Parkinson's genes (0.5 = guessing, 1 = perfect) · Huang X et al., Nat Biomed Eng 2026

PAPER 2 · XUNZI · 3/3

CHK2: a new Parkinson's target, tested in mice

1 · Suggest
the AI points to CHK2, an enzyme over-active in the sick brain area

→

2 · Turn down
less CHK2 gene activity: brain cells survive

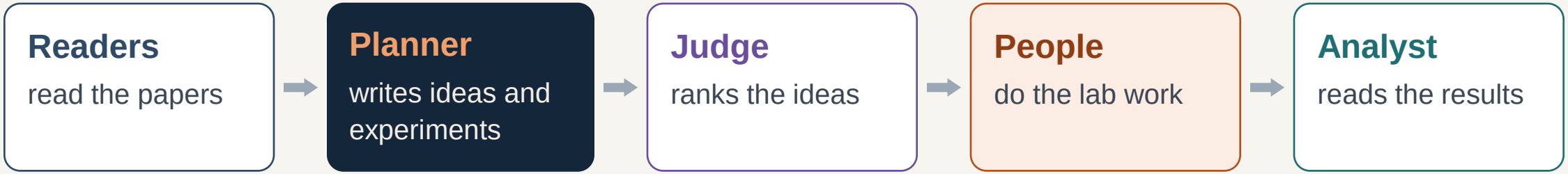
→

3 · Block
a CHK2-blocking drug: mice move better, in two disease models

Strong evidence, but **only in mice**, and people planned every experiment.

Huang X et al., Nature Biomedical Engineering 2026

A team of AI agents, like a small lab



results go back to the planner, who writes **new ideas**

Robin, Nature 2026 · FutureHouse (preprint arXiv:2505.13400)

PAPER 3 · ROBIN · 2/3

An eye disease, step by step

Step 1	reads 151 papers → idea: help the eye’s cleaning cells remove waste
Step 2	lab test of its top ideas → one known drug type works
Step 3	reads the new data → a new candidate gene, and a new drug idea: **ripasudil**, already used for glaucoma → it works

What is new: the AI **changed its idea** after reading its own experiment.

Robin, Nature 2026 · dry age-related macular degeneration, human retinal cell line

PAPER 3 · ROBIN · 3/3

People still matter, and must check the AI

The AI did

ideas, plans, data analysis, figures

People did

chose what to test, all lab work, wrote the instructions

A warning

7.5× vs 1.75×

same data: AI analysis vs human analysis

A real step forward, but tested only in cells, and the new gene was **not yet tested**.

Robin, Nature 2026 · preprint arXiv:2505.13400 (fold changes as reported there)

REALITY CHECK

What reached the clinic so far

117

AI-enabled assets entered clinical trials (to Dec 2025)

8

of them had completed Phase 2

0

AI-discovered drugs with full approval (mid-2026)

TNIK · rentosertib, Phase 2a

FVC +98.4 mL (60 mg) vs −20.3 mL placebo at 12 weeks; 71 patients, safety primary. Phase 3 (320 patients) started 2026.

The other side

BEN-2293 (target from a knowledge graph) failed Phase 2a in 2023. Phase 2 failure of AI drugs looks similar to other drugs.

ASCO 2026 abstract (JCO 44:16 suppl, 11072); Xu et al., Nat Med 2025; Nat Rev Drug Discov 2026 (s41573-026-01496-2). Counts depend on the tracker.

HYPE · THE VERDICT ON TNIK

A real success, but whose success?

Shown

- a computational pipeline chose an interesting target
- knockdown and models support an antifibrotic role
- the program reached patients

Not shown

- that deep learning was needed for the nomination
- that it beats a classical omics and network baseline
- a new executable disease mechanism

The missing experiment

Same data, same filters. Where does TNIK rank with:

1. expression + pathways only
2. + networks and TF upstream analysis
3. ML components only
4. everything except ML
5. everything (published)

A success of computationally assisted target discovery. Whether it is a success of deep learning is still open.

Independent link of TNIK to fibrosis: siRNA screen, liver fibrosis in mice (Buchl et al., Hepatol Commun 2022)

HYPE · BENCHMARKS

Do deep models predict perturbations better than simple ones?

2025 · NATURE METHODS

“No”

Several deep-learning models of genetic perturbation effects did not beat deliberately simple linear baselines.

Ahlmann-Eltze, Huber, Anders

2026 · NATURE BIOTECHNOLOGY

“Yes, with calibrated metrics”

Common metrics hide real signal; with well-calibrated metrics, deep models do beat uninformative baselines.

Shift Bioscience team

Both can be right. And both measure predicted **expression changes**, not whether an intervention **reverses a disease phenotype**.

Ahlmann-Eltze et al., Nat Methods 2025 (s41592-025-02772-6); Nat Biotechnol 2026 (s41587-026-03307-w; preprint bioRxiv 2025.10.20.683304)

TWO SPACES · THE TRADITIONAL VIEW

Huge chemistry, few targets

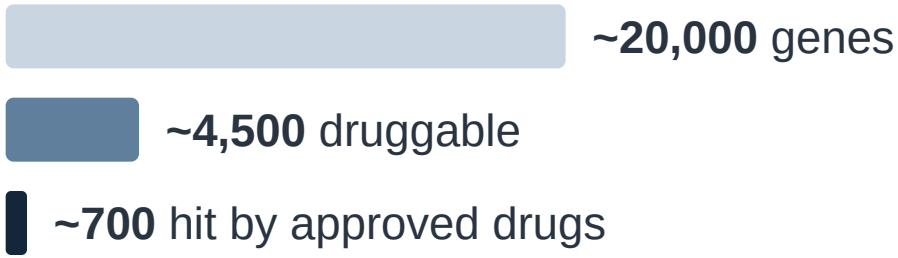
CHEMICAL SPACE

$\sim 10^{60}$

possible drug-like molecules

tens of billions already orderable; AI is good at searching here

TARGET SPACE



The usual picture: **infinite molecules, a short list of proteins**.

Bohacek et al., Med Res Rev 1996 (10^{60} estimate) · Finan et al., Sci Transl Med 2017 (druggable genome) · Santos et al., Nat Rev Drug Discov 2017 (approved-drug targets)

THE QUESTION

How many potential drug targets do we have?

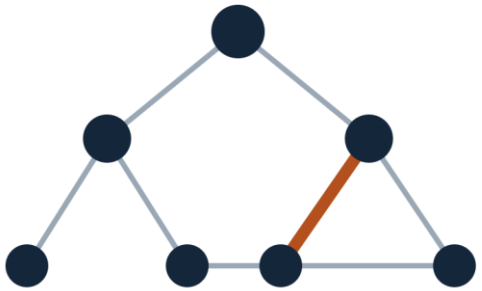
Thousands?

Millions?

Maybe astronomical.

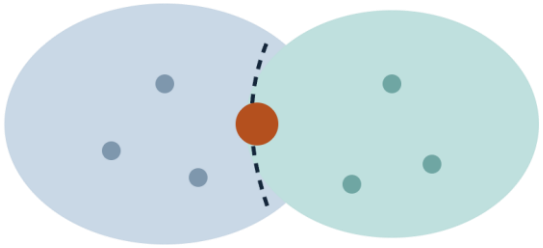
TARGET SPACE

The real targets are interactions, not molecules



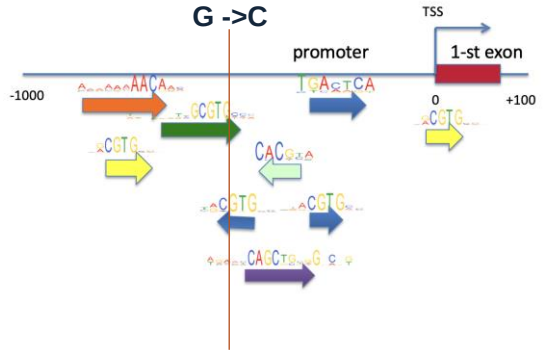
One connection

in a signaling network, in one cell state, can be the intervention point



One amino acid

at a protein–protein interface decides whether a signaling complex forms



One nucleotide

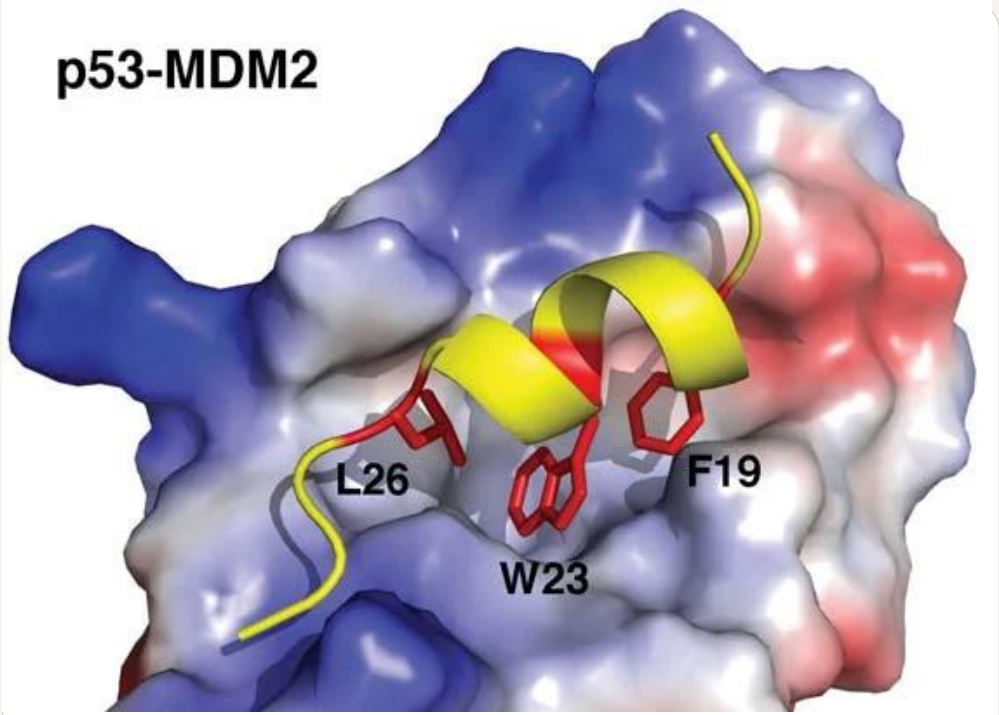
can create or destroy a transcription factor binding site and change a whole gene program

~20,000 genes × binding sites × interfaces × connections × cell states → a combinatorial space. **How do we find the few interactions that control the disease?**

Promoter face: geneXplain lecture illustration (TRANSFAC motifs). The size of the space is a conceptual argument, not a count.

INTERACTION TARGETS · PROTEIN–PROTEIN

p53–MDM2: three amino acids make a target



p53 helix: Phe19, Trp23, Leu26 **MDM2** hydrophobic pocket

The interaction

MDM2 binds the p53 activation helix, blocks it and sends p53 to degradation. Many tumours keep p53 wild type but silence it through MDM2.

The drug idea

Nutlins (2004) mimic the three p53 residues and fill the pocket: p53 is freed, not MDM2 removed.



Kussie PH et al., Science 1996 (p53–MDM2 structure); Vassilev LT et al., Science 2004 (nutlins); pipeline status: ApexOnco, April 2025

INTERACTION TARGETS · ALREADY IN THE CLINIC

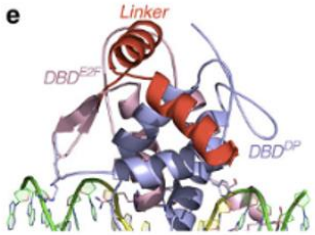
Approved drugs that act on an interaction

Interaction	Drug	Disease	US approval
BCL-2 – BH3 proteins (BIM, BAX)	venetoclax	CLL, AML	2016
menin – KMT2A (MLL)	revumenib	KMT2A-rearranged leukaemia	2024
PD-1 – PD-L1	pembrolizumab and others	many cancers	2014
α4β7 integrin – MAdCAM-1	vedolizumab	ulcerative colitis, Crohn’s	2014
LFA-1 – ICAM-1	lifitegrast	dry eye disease	2016
KRAS G12C: one amino acid, effector binding	sotorasib	lung cancer	2021
TTR tetramer interface (stabilised)	tafamidis	TTR amyloid cardiomyopathy	2019

Block, mimic, or stabilise: the drug changes **a contact between molecules**, often a few residues wide.

Transcription Factor (TF):

(3.5 Tryptophan cluster factors)



TF binding motif:

(TRANSFAC PWM: M00231)



Target gene: AMHR2

TF binding site:

(gain – score difference)

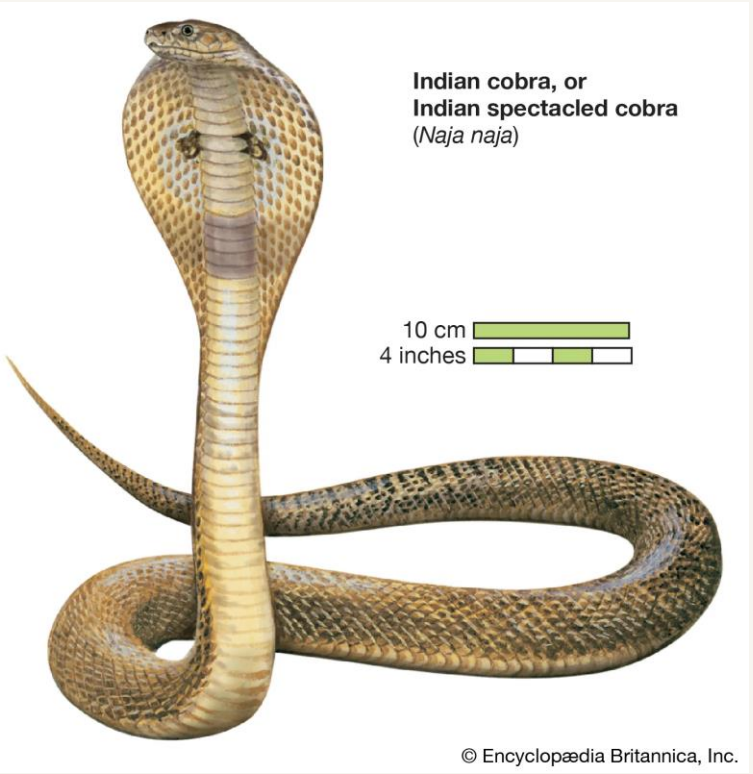
ctaGggaaaaca

(reference allele)

ctaCggaaaaca

(alternative allele)

Mutation: CR1910211:G>C



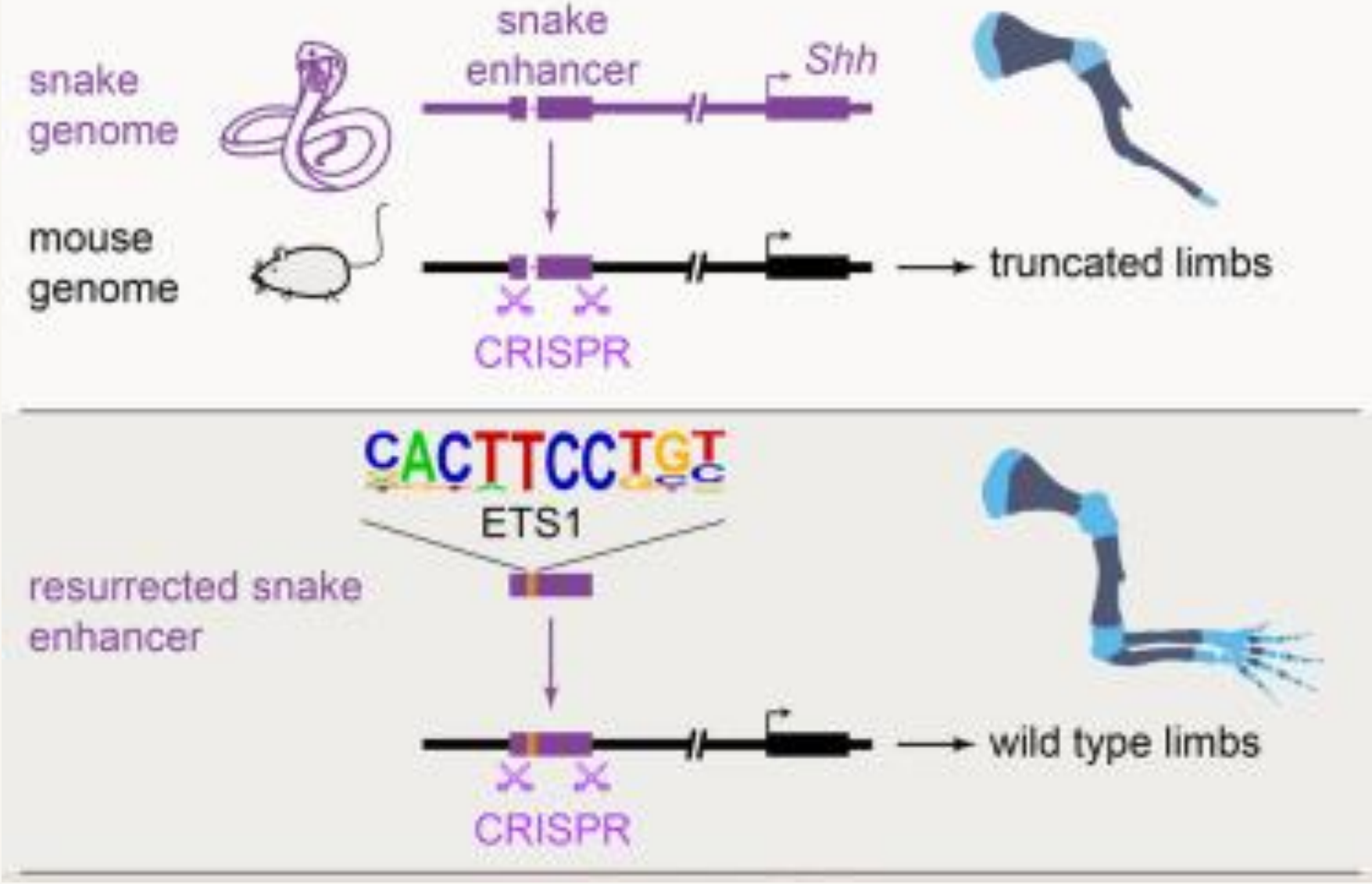
Cell

Volume 167, Issue 3, 20 October 2016, Pages 633-642.e11

Article

Progressive Loss of Function in a Limb Enhancer during Snake Evolution

Evgeny Z. Kvon [1], Olga K. Kamneva [2], Uirá S. Melo [1], Iros Barozzi [1], Marco Osterwalder [1], Brandon J. Mannion [1], Virginie Tissières [3], Catherine S. Pickle [1], Ingrid Plajzer-Frick [1], Elizabeth A. Lee [1], Momoe Kato [1], Tyler H. Garvin [1], Jennifer A. Akiyama [1], Veena Afzal [1], Javier Lopez-Rios [3], Edward M. Rubin [1 4], Diane E. Dickel [1], Len A. Pennacchio [1 4], Axel Visel [1 4 5 6]



1. SITE GAIN — *TERT* promoter

***TERT* promoter** (–124 bp, relative to ATG)

C228T

Wild-type (C)

5′ … C C C C C C T T C C G G … 3′

3′ … G G G G G A A G G G C C … 5′

No ETS motif (weak binding)

Mutant (T)

C228T –124 bp

5′ … C C C C T T C C G G … 3′

3′ … G G G G A A G G C C … 5′

GABP (ETS family)

ETS motif created (strong binding)

5′-CCCC TTCCGG-3′

ETS core

↑ *TERT* transcription
↑ telomerase activity

Cancer-associated driver
(e.g. glioma, melanoma, bladder, thyroid, other cancers)

Associated diseases:
glioma, melanoma, bladder cancer, thyroid cancer, others

Mutant sequence shown from published *TERT* promoter studies; C228T creates a de novo ETS binding site.

2. SITE LOSS — *GATA3–LMO1* enhancer

Enhancer in first intron of *LMO1* (rs2168101)

G risk allele (disease risk)

rs2168101

5′-G AATAA-3′

GATA3

5′ … A G A T A A … 3′

3′ … T C T A T T … 5′

GATA motif

Strong GATA3 binding

G allele
↓
GATA3 binding
↓
super-enhancer
↓
↑ *LMO1* expression
↓
Higher neuroblastoma susceptibility

T protective allele (protective)

rs2168101

5′-T TATAA-3′

GATA3 ✕

5′ … A T A T A A … 3′

3′ … T A T A T T … 5′

GATA motif

Loss of GATA3 binding

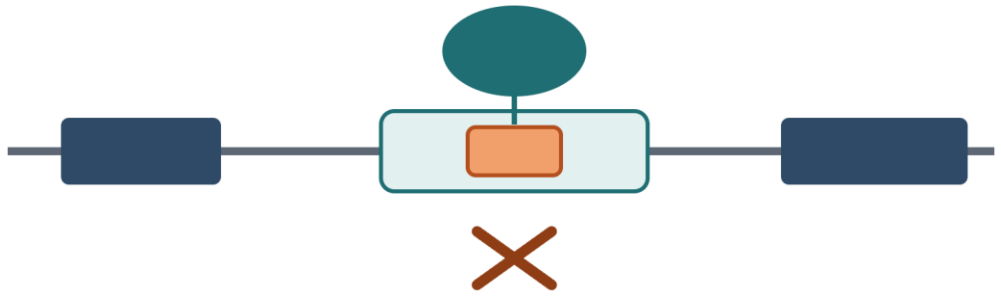
T allele
↓
loss of GATA3 binding
↓
↓ *LMO1*
↓
Lower neuroblastoma risk (protective)

Associated disease:
neuroblastoma
(GATA3 high expression = poor prognosis)

single-nucleotide change → **altered TF binding** → **altered cis-regulation** → **disease phenotype**

INTERACTION TARGETS · TF–DNA

The first CRISPR therapy targets one TF binding site



BCL11A exons **GATA1** site in erythroid enhancer **Cas9 cut**

Found by genetics

variants linked to fetal haemoglobin sit in an erythroid enhancer of BCL11A

The edit

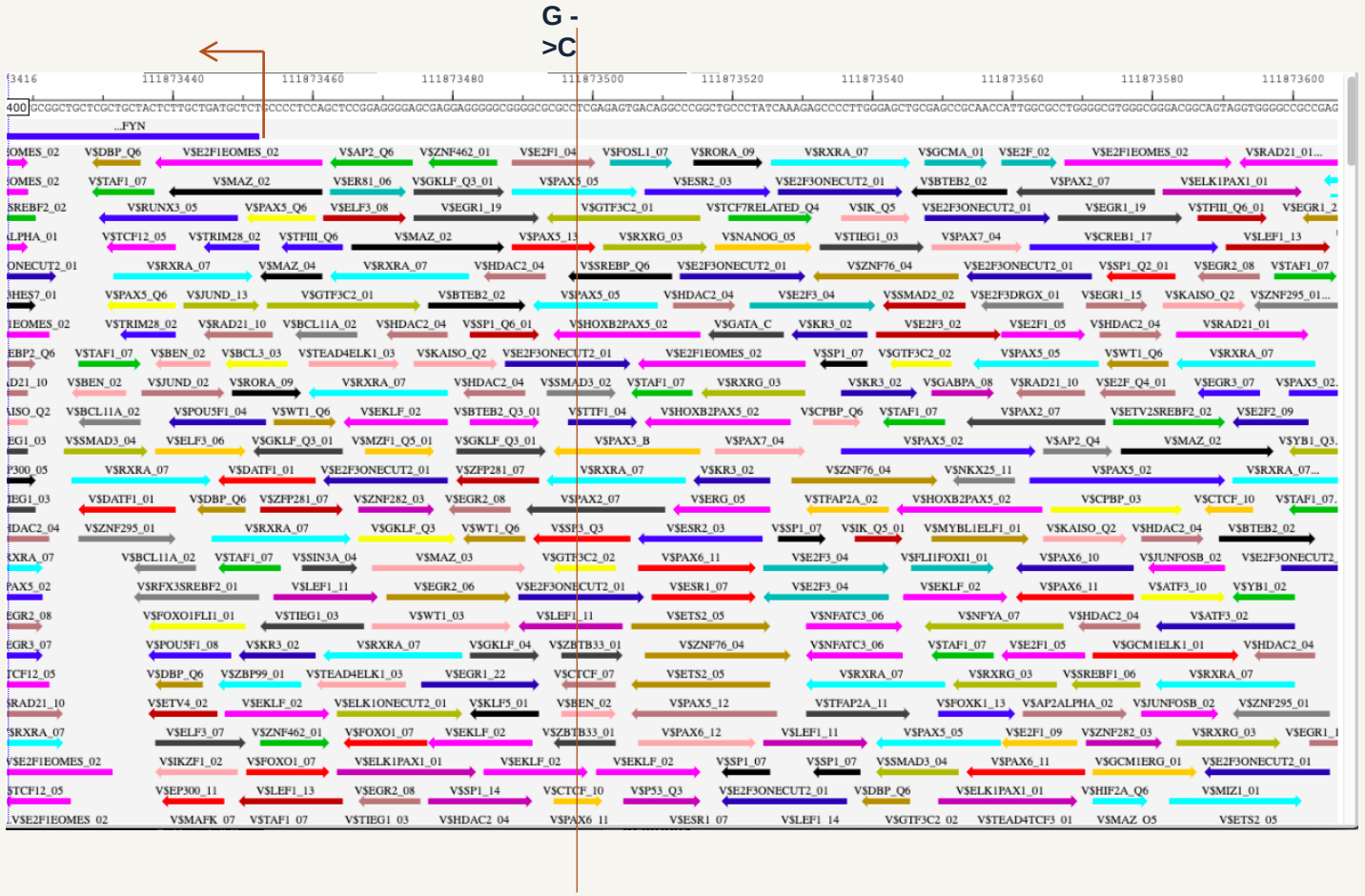
disrupt the GATA1 site → BCL11A off only in red-cell precursors → fetal haemoglobin back on

Approved

Casgevy (exa-cel): sickle cell disease, Dec 2023; β-thalassaemia, Jan 2024 (FDA)

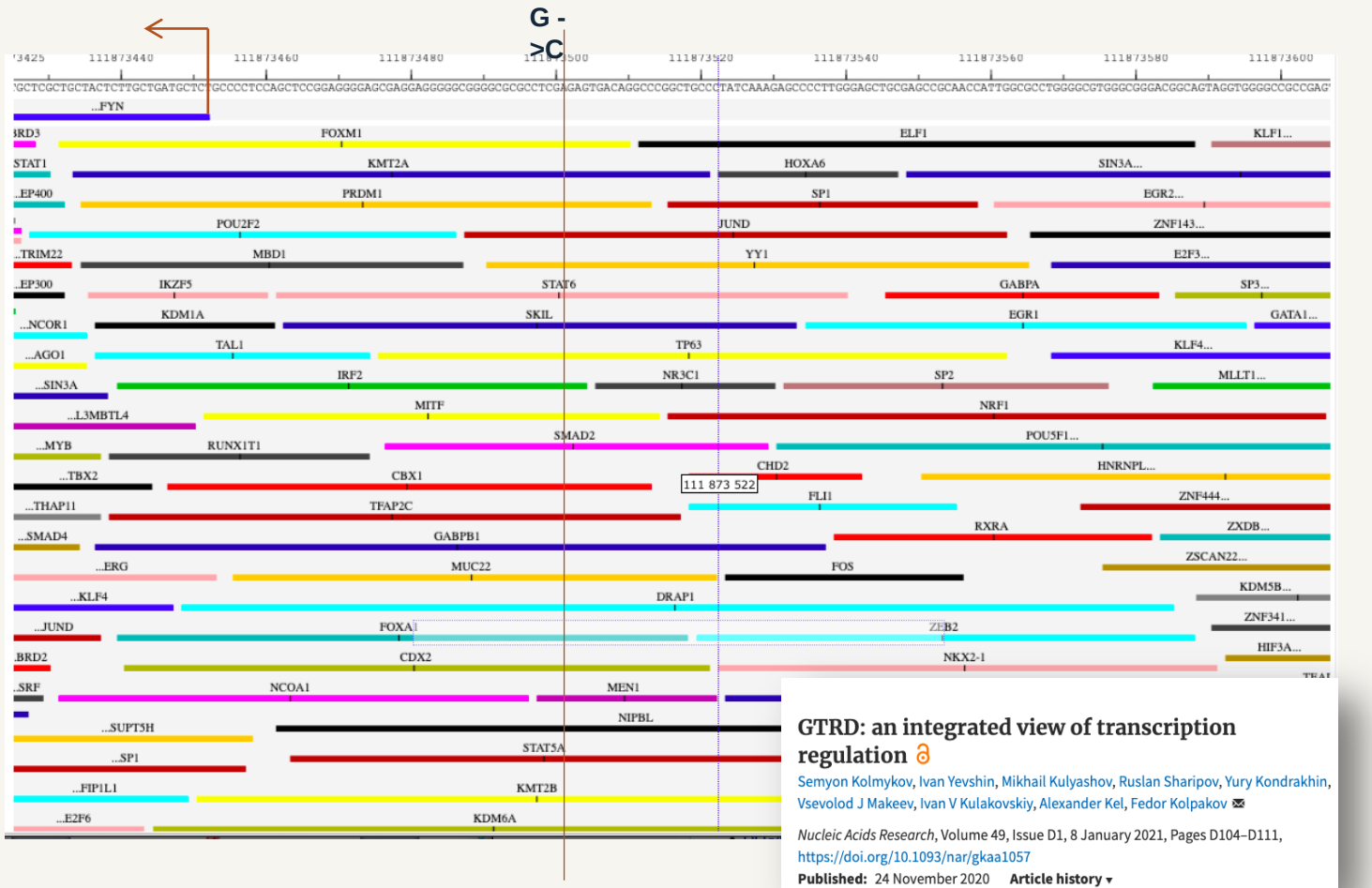
Not a gene knockout, not a protein inhibitor: the target is **one TF–DNA contact** in one cell type.

Bauer DE et al., Science 2013 (BCL11A erythroid enhancer); Frangoul H et al., NEJM 2021; Casgevy prescribing information 2024



5260 TFBS motifs in the 1000bp FYN promoter

GTRD database – comprehensive ChIP-seq database



GTRD: an integrated view of transcription regulation

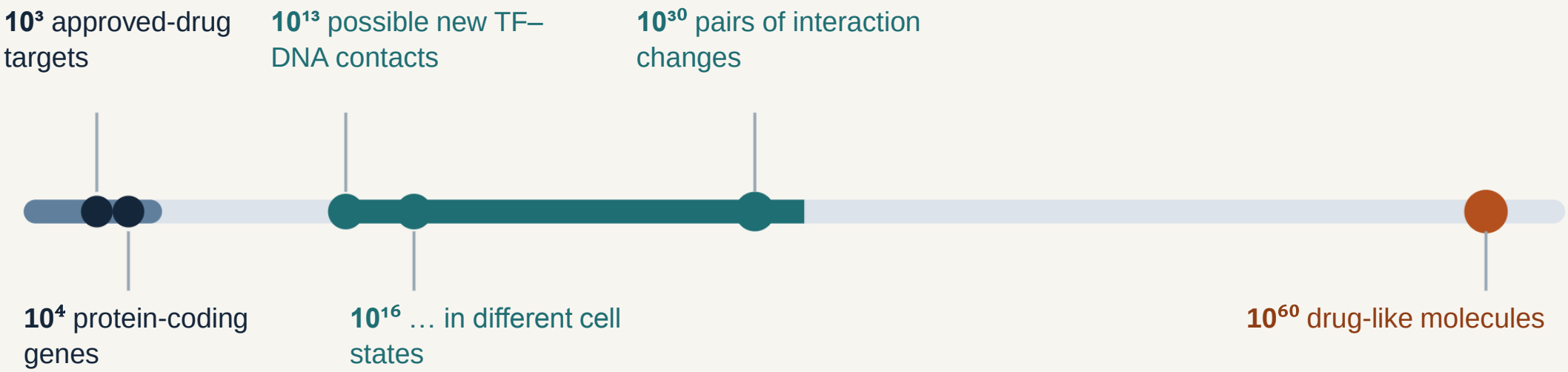
Semyon Kolmykov, Ivan Yevshin, Mikhail Kulyashov, Ruslan Sharipov, Yury Kondrakhin, Vsevolod J Makeev, Ivan V Kulakovskiy, Alexander Kel, Fedor Kolpakov

Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D104–D111, https://doi.org/10.1093/nar/gkaa1057

Published: 24 November 2020 **Article history**

TWO SPACES · SIDE BY SIDE

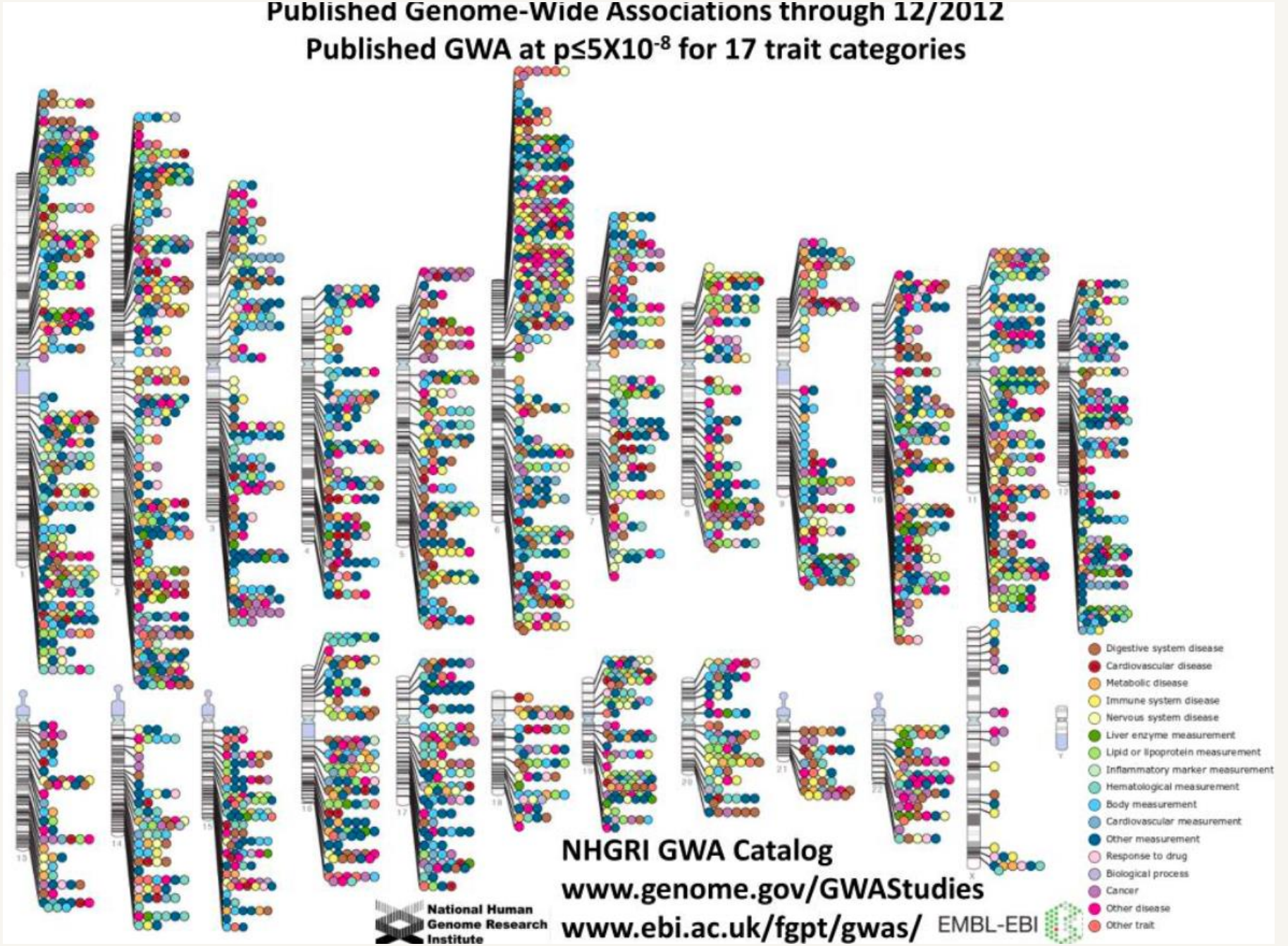
Target space is astronomical too



~10^{10} possible single-letter DNA changes × ~1,600 transcription factors ≈ 10^{13} · × hundreds of cell states ≈ 10^{16} · two changes combined ≈ 10^{30}

Not 700 proteins: **a space of interactions**, almost as large as chemistry. We need Big Knowledge to search it.

Order-of-magnitude illustration, not a count. Genome 3×10^{9} bp × 3 alternative letters; ~1,600 human TFs (Lambert et al., Cell 2018).


Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10-8 for 17 trait categories
Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
National Human Genome Research Institute
EMBL-EBI

2020



95% of top disease associated mutations are in **non-coding** regions.

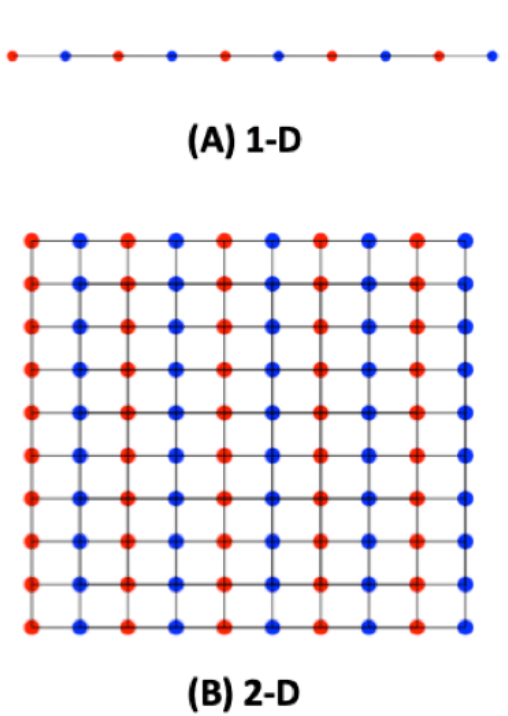


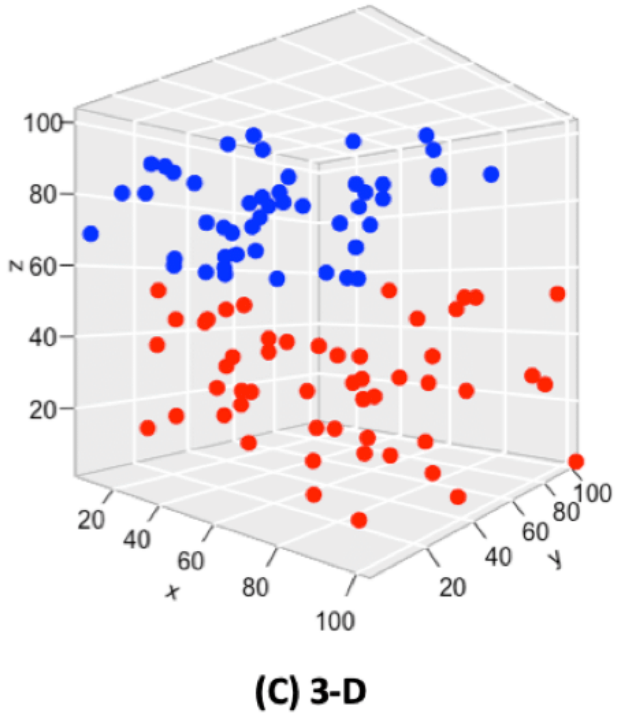
Dark matter of genome!

Based on NHGRI GWA Catalog, www.genome.gov/GWAStudies, www.ebi.ac.uk/fgpt/gwas/

THE CURSE OF DIMENSIONALITY

Add enough dimensions, and even random points separate





1-D AND 2-D

Red and blue are mixed: **no line** separates them

3-D

One more axis, and a **plane** separates them

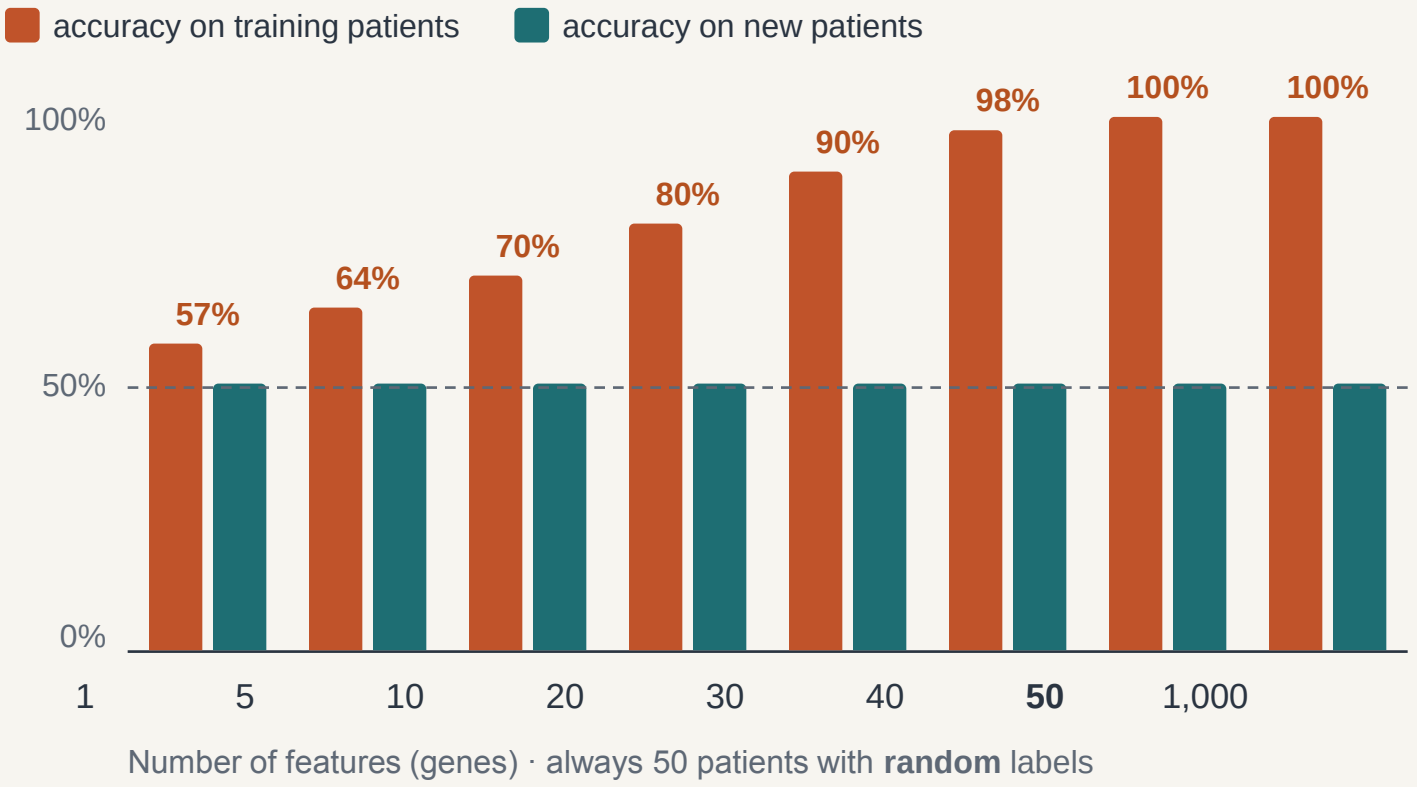
THE REASON

Each feature adds a new direction, and the points move apart. With **n samples** and **n − 1 features**, **any** labels can be separated, also random ones.

Bellman R, 1957 (“curse of dimensionality”). Cover TM. IEEE Trans Electron Comput 1965;14:326–334 (separating capacity)

CONSEQUENCE FOR AI TRAINING · SAMPLES ≪ FEATURES

A perfect model of pure noise



OVERFITTING

Training looks perfect, new patients get a coin flip

MANY "ANSWERS"

Thousands of gene sets fit equally well; most random signatures also predict outcome

NO REPRODUCIBILITY

Biomarkers and targets fail in the next cohort

WAY OUT

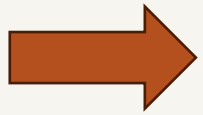
Big Knowledge cuts dimensions: 20,000 genes → TF sites → a few master regulators. Then test on independent cohorts.

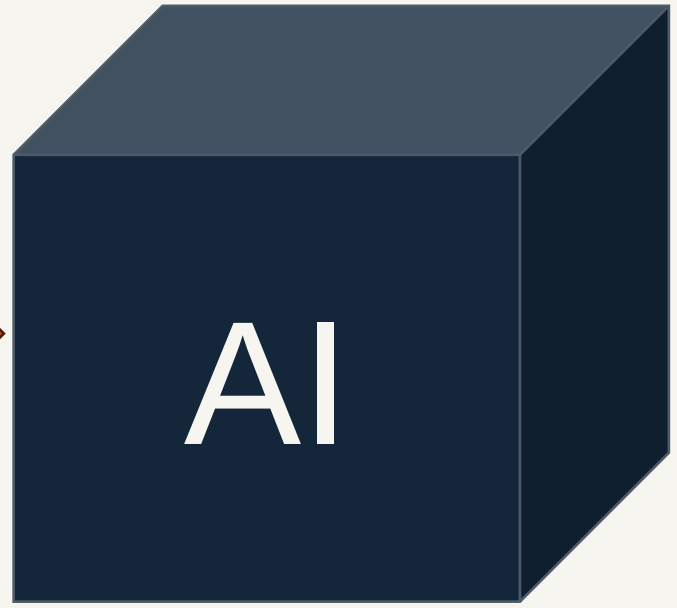
Simulation: linear classifier, random data, 200 runs per point. Venet D et al. PLoS Comput Biol 2011;7:e1002240 (random signatures)

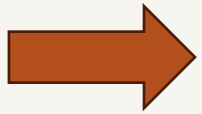




Big data



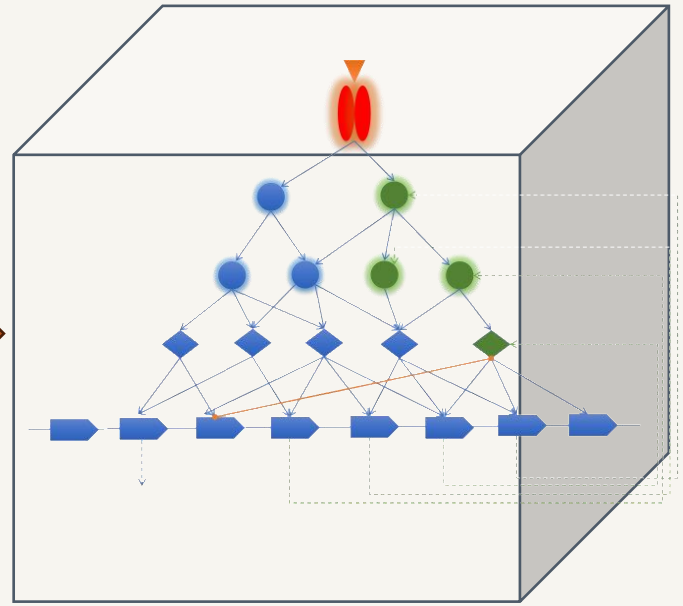




- Late detection
- New drugs
- Metastases
- Drug resistance

Explainable **AI**

Big Data



Big Knowledge

- **Late detection**
- **New drugs**
- **Metastases**
- **Drug resistance**

REALITY · BIG KNOWLEDGE

Every success above stands on decades of curated biology

XunZi

reasons over 24.4 million publications plus multi-omics data

MIDAS

learns on gene interaction networks and perturbation phenotypes

VirtualCRISPR

uses the literature to tell what is already known

Baricitinib

for COVID-19, proposed from a knowledge graph in 2020, later approved

Without this knowledge, powerful models find associations but **cannot explain the mechanism**.

Big Data → explainable AI ← Big Knowledge: geneXplain lecture illustration. Baricitinib: Richardson et al., Lancet 2020.

GENOME ENHANCER · UPSTREAM ANALYSIS

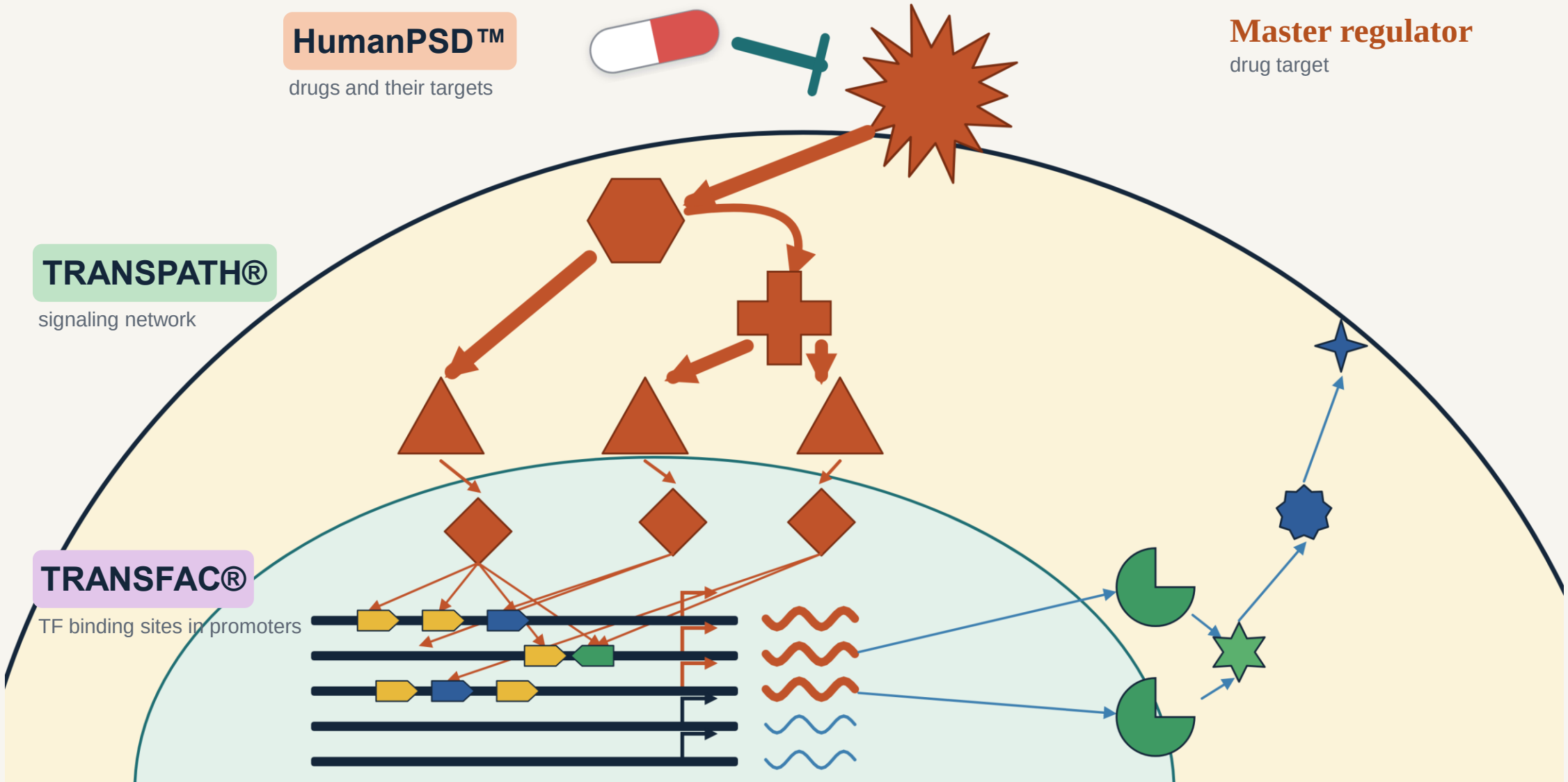
From changed genes back to the master regulator

HumanPSD™
drugs and their targets

Master regulator
drug target

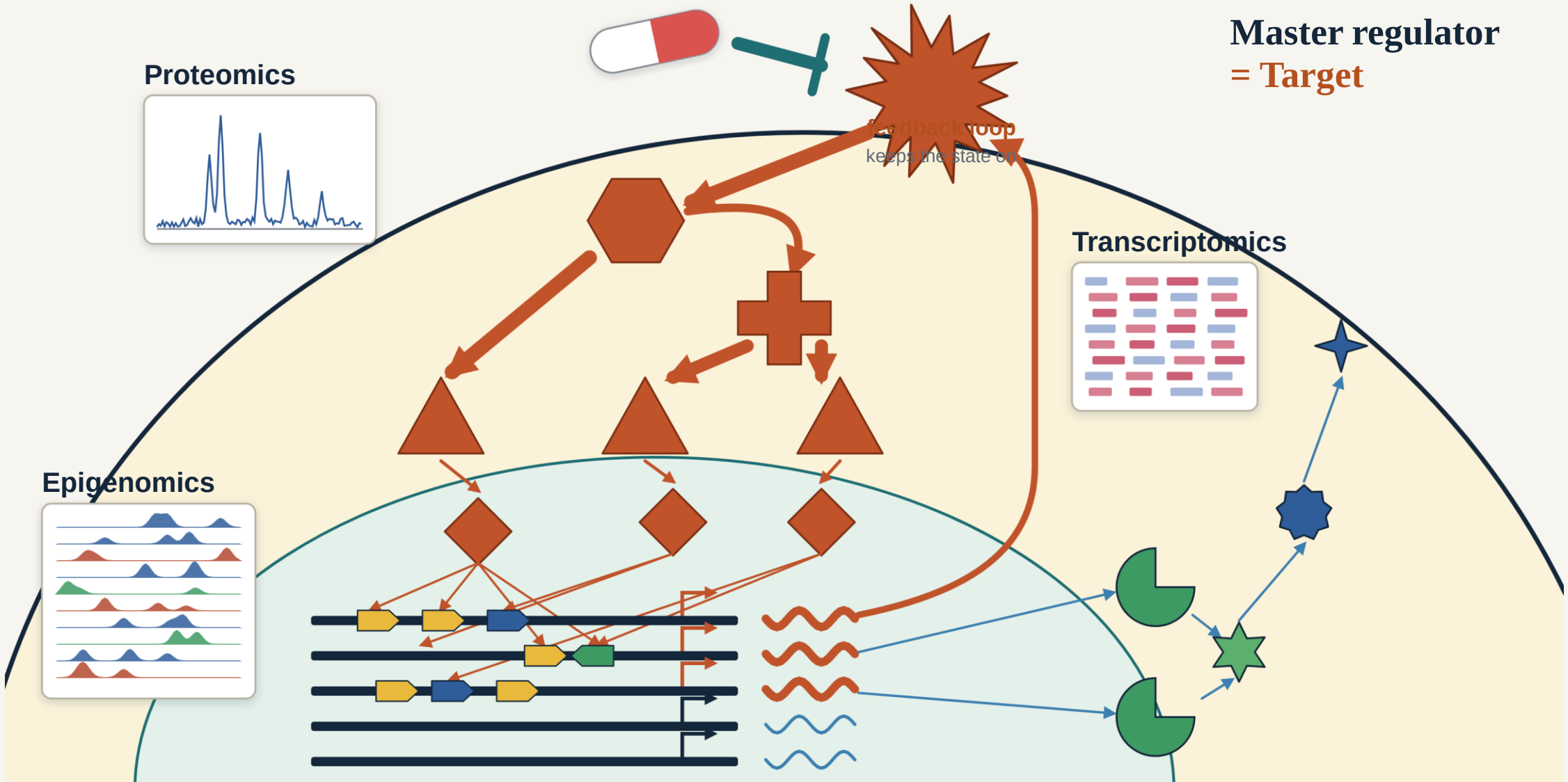
TRANSPATH®
signaling network

TRANSFAC®
TF binding sites in promoters



GENOME ENHANCER · MULTI-OMICS

Any omics data in, one explainable target out



WHY DISEASE IS HARD TO PREDICT

A healthy cell turns like a wheel. A cancer cell crashes.



1 · THE CYCLE

Like the cell cycle: G1 → S → G2 → M, turn after turn. **Easy to model.**



2 · HEALTHY SYSTEM

Wheels, steering and brakes work together: **behaviour is predictable.**



3 · DISEASE

The wheels still turn, but the car tumbles: **unexpected and unpredictable.**

Cancer cells still use the cycle but the control around it is broken, so they divide in an unpredictable way. A model learned from the normal “wheel” cannot predict the “crash”.

ONE PAPER · ZIKRY ET AL., PNAS 2024

When the usual route is blocked, some cells take another one

WHAT THEY DID

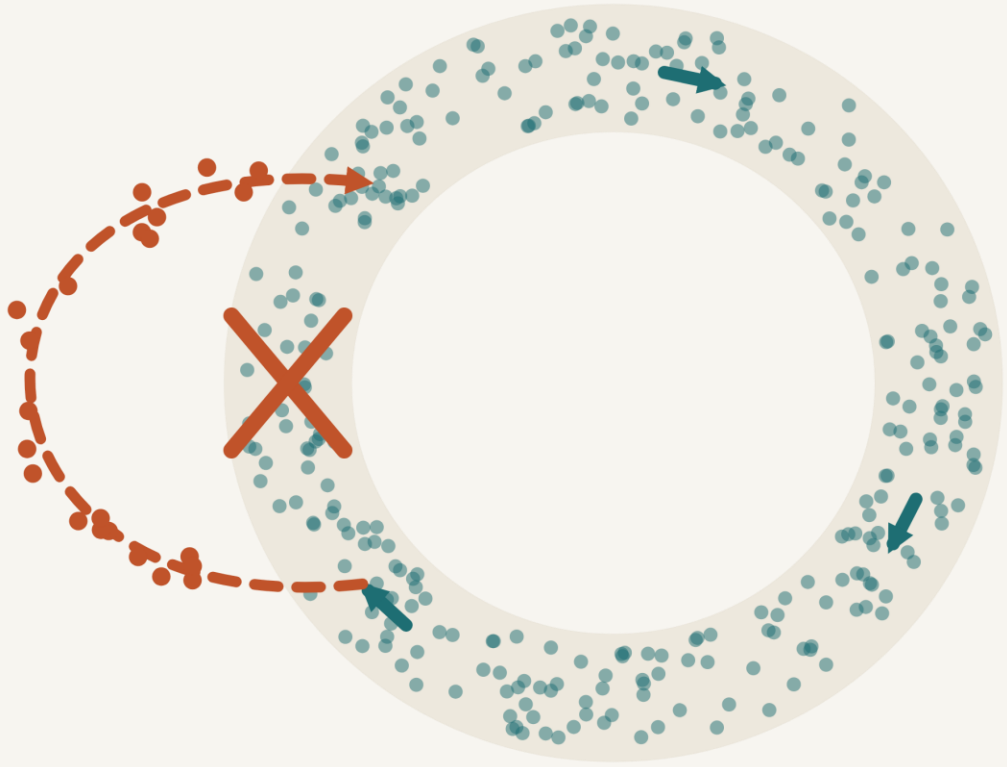
Single-cell imaging of cell cycle regulators in ER+/HER2− breast tumour cells and in **patient tumours**, treated with the CDK4/6 inhibitor **palbociclib**

WHAT THEY FOUND

Most cells stop. A **fraction keeps dividing**: they build up E2F1, Rb and CDK2 early and reach S phase without the usual CDK4/6 step

CELL CYCLE PLASTICITY

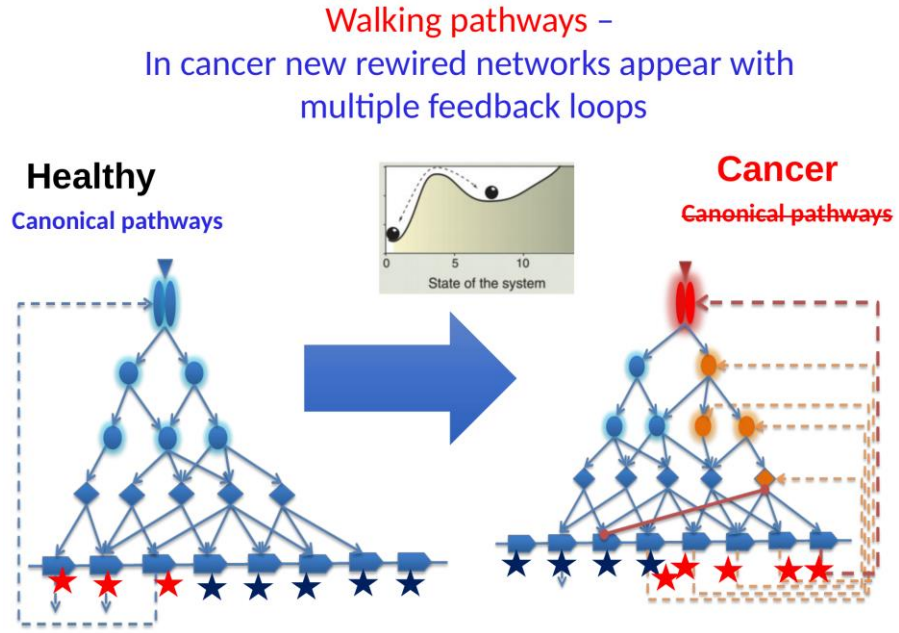
Variability between the regulators opens **alternative paths** through the cycle. These cells are then sensitive to CDK2 inhibition.



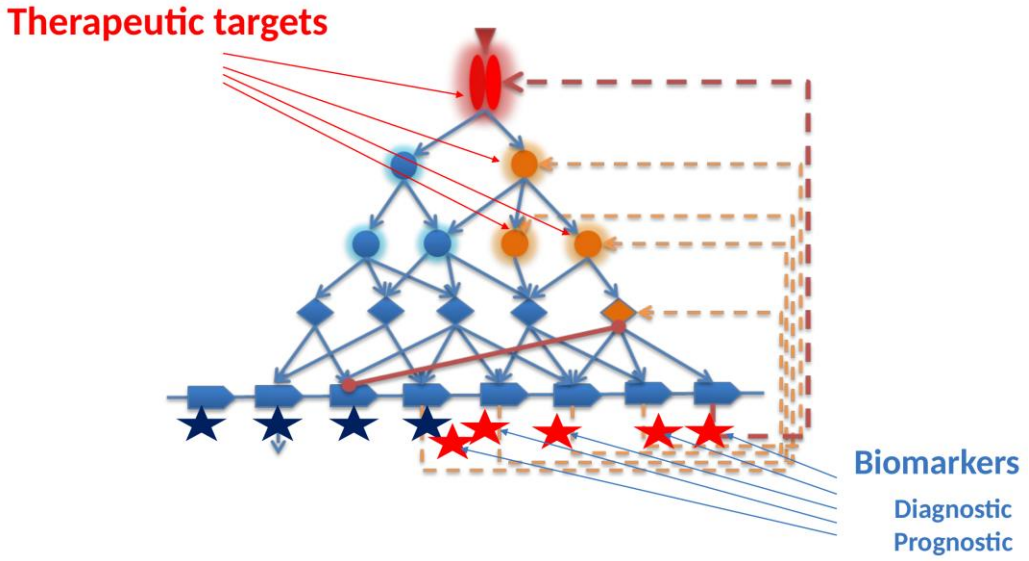
most cells: canonical path, arrested resistant fraction: alternative path

Scheme drawn after the concept in Zikry TM, Wolff SC, …, Spanheimer PM, Purvis JE. PNAS 2024;121:e2309261121. Not a reproduction of their figure.

Disease rewires the network, and the loops become targets



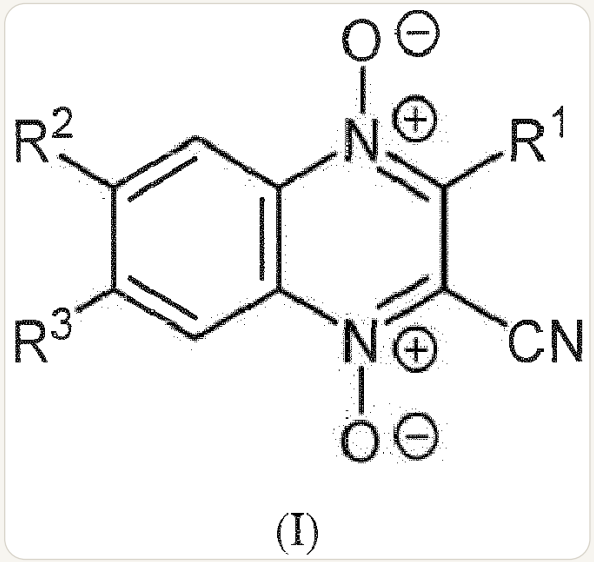
New master regulators and **positive feedback loops** lock the cell in a second stable state



Targets: nodes and edges inside the loops. **Biomarkers**: the genes they switch on.

LCTA-2940 · OUR NEW COMPOUND

A selective thioredoxin reductase inhibitor



Quinoxaline 1,4-dioxide scaffold
Patent EP4201929B1

TrxR1 thioredoxin reductase

× **blocked by LCTA-2940**

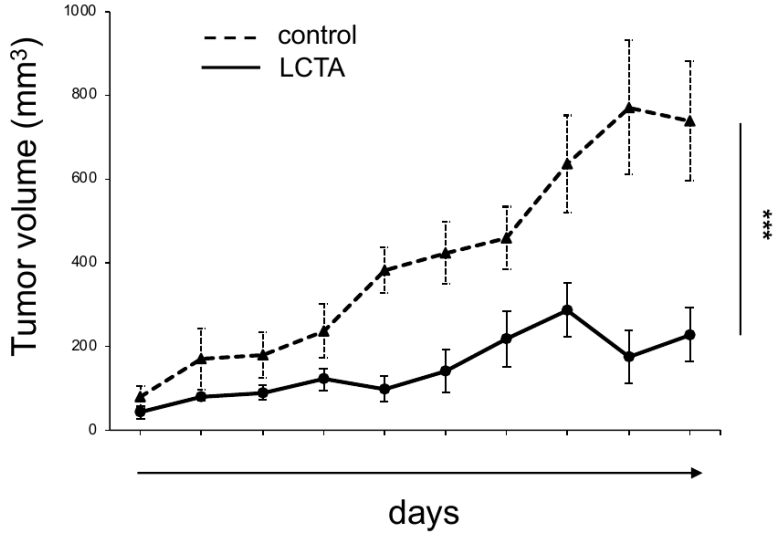
Thioredoxin is not recycled

↓

ROS go up
cells cannot remove peroxides

↓

Oxidative stress kills cancer cells



Colorectal cancer xenografts: tumours grow much less

≤ 10 µM
LC50 in 5 cancer cell lines

1 of 5
lines adapted after 16 weeks

Kel A et al. Quinoxaline thioredoxin reductase inhibitors, EP4201929B1 (2024). Lyu H, …, Kel A, …, Zubarev RA. OVERDRIVE, manuscript 2026

LCTA-2940 · HYPOXIA

Strongest in hypoxic cells, which resist radiation



Hypoxic cells already have **high ROS** and depend on their antioxidant systems

LCTA-2940

Blocks TrxR, adds ROS on top: **acts very well in hypoxic cells**

Suppresses tumour invasion

Sensitizes to radiation

Hypothesis: targeting the hypoxic fraction of paediatric glioma will increase tumour response to radiotherapy

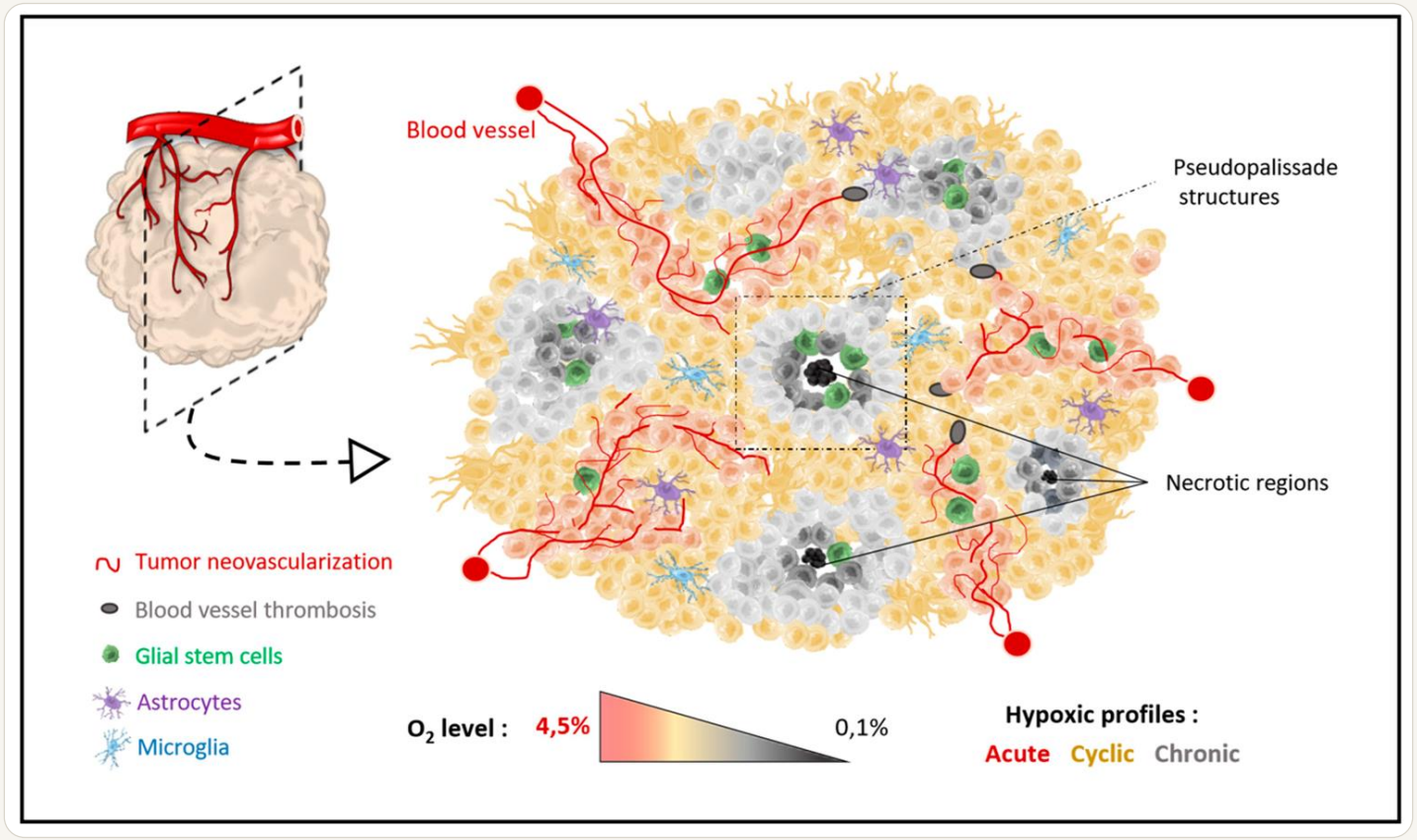


Figure: Fuchs et al. Cancers 2020;12(4):979 (hypoxia in glioma)

OVERDRIVE · THE APPROACH

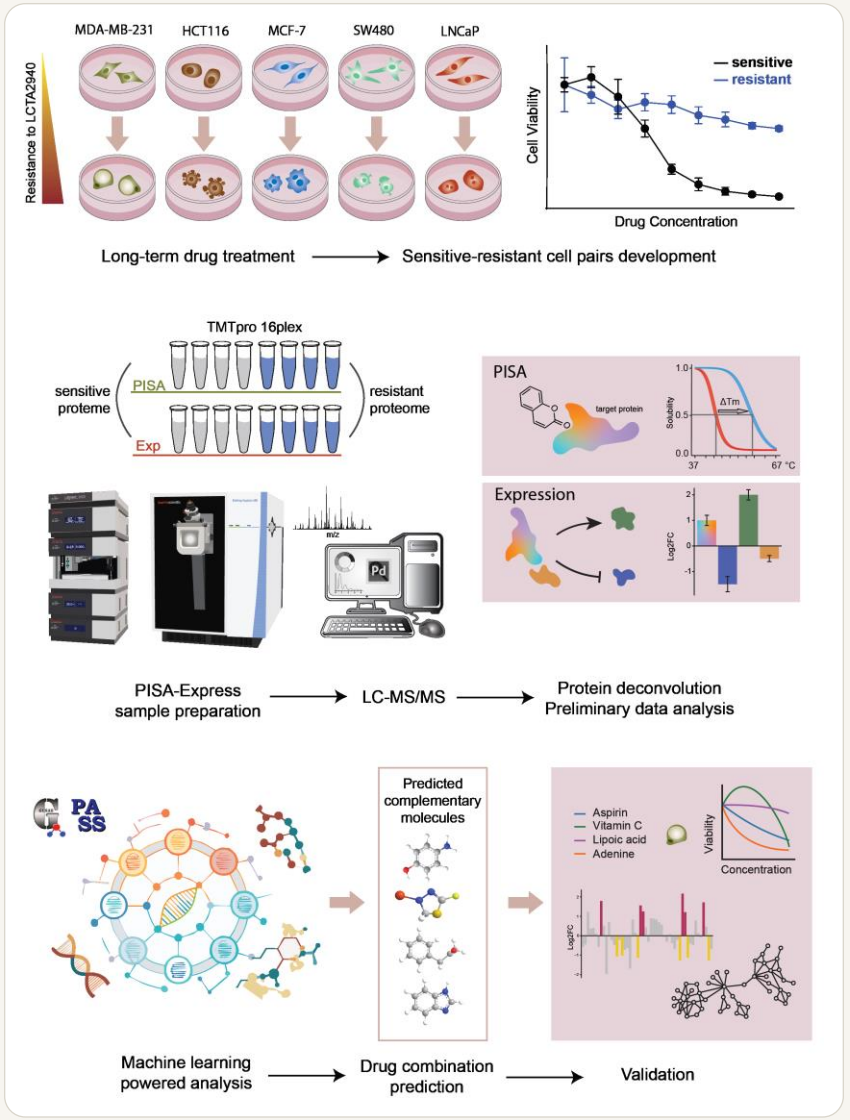
Let cells become resistant, then let the machine find how to break it

1 **Expose** cancer cells to rising drug doses for 8–16 weeks

2 **Measure** ~8,000 proteins: amount and solubility (PISA-Express)

3 **Genome Enhancer**: TF sites, genetic algorithm, master regulators = **targets of resistance**

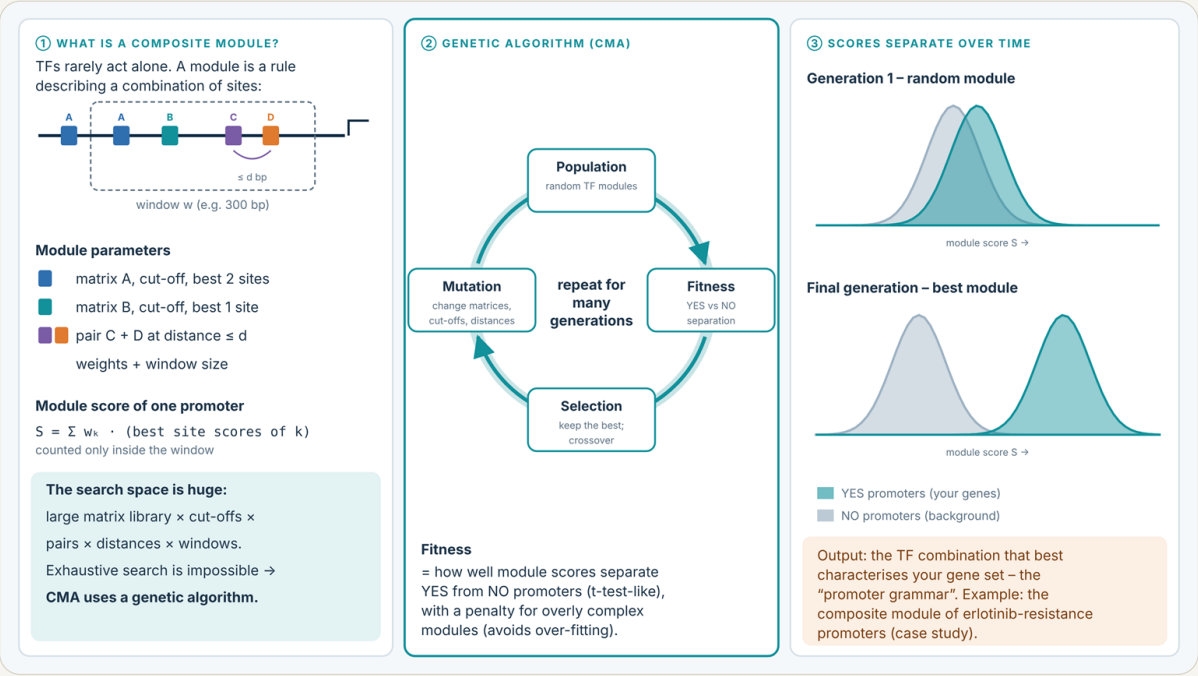
4 **Predict** non-toxic molecules that hit those targets, and **test** them with the drug



Lyu H, …, Kel A, Stelmashenko D, Selivanova G, Zubarev RA. OVERDRIVE. Manuscript, June 2026 (Karolinska Institutet, geneXplain)

INSIDE GENOME ENHANCER · MACHINE LEARNING

A genetic algorithm finds the regulatory code of resistance



Composite Module Analyst (CMA): candidate TF-site combinations mutate, recombine and are selected by how well they separate the two promoter sets.

1 · TF SITES

Changed proteins → genes → promoters scanned with **TRANSFAC**

2 · GENETIC ALGORITHM

Evolves the **combination** of TF sites that explains resistance (5-FU: ESR1, TAL1, STAT3)

3 · MASTER REGULATORS

Walk upstream in **TRANSPATH** to the hubs = **targets of resistance**

4 · MOLECULES

PASS + HumanPSD rank molecules for these targets; toxic ones are filtered out

Kel A et al. Genome Enhancer / upstream analysis (geneXplain). TRANSFAC®, TRANSPATH®, HumanPSD™, PASS

PROOF OF PRINCIPLE · 5-FU IN COLORECTAL CANCER CELLS

The machine was not told the drug, and still found uracil

TARGETS OF RESISTANCE

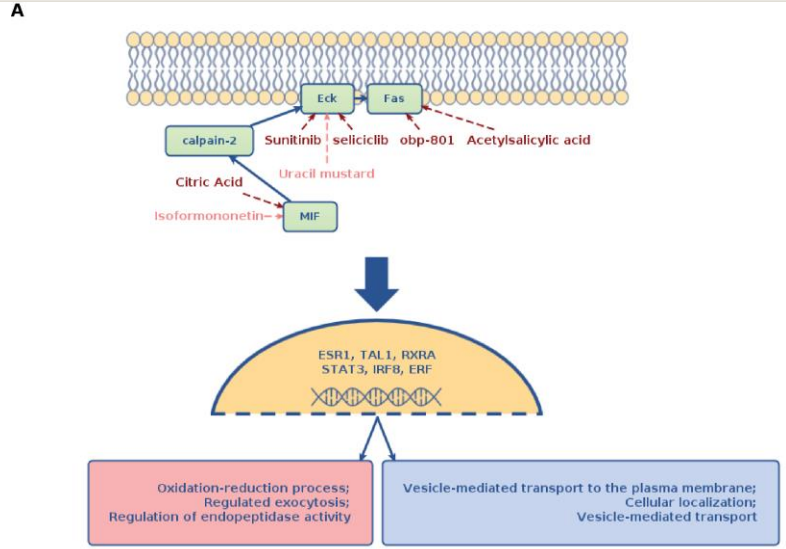
Master regulators **CAPN2, FAS, Eck**, plus **MIF** as a key node

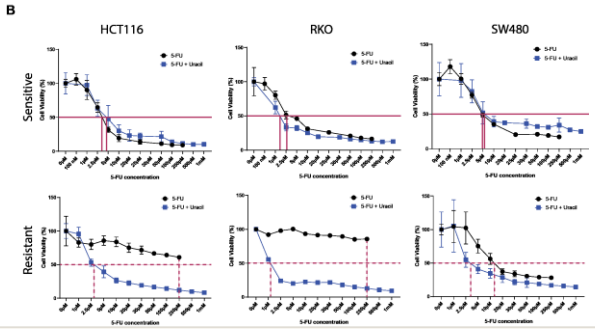
URACIL FAMILY, 5 TIMES IN TOP 150

of ~15,000 molecules, including **5-FU itself** and **eniluracil**, a drug made to boost 5-FU

VALIDATED

5-FU + uracil (1:4): resistant LC50 from **15 – >250 μM** back to **2.5 μM** (1.5 μM in RKO)





Lyu H, …, Kel A, …, Zubarev RA. OVERDRIVE. Manuscript, June 2026. Fig. 4. HCT116, RKO, SW480 cells

NEW DRUG · LCTA2940, A TRXR1 INHIBITOR

The network points straight at the drug’s own pathway

LCTA2940 is **not in the database**. Yet the top of the network is **auranofin** and **PX-12**, two inhibitors of the same thioredoxin system.

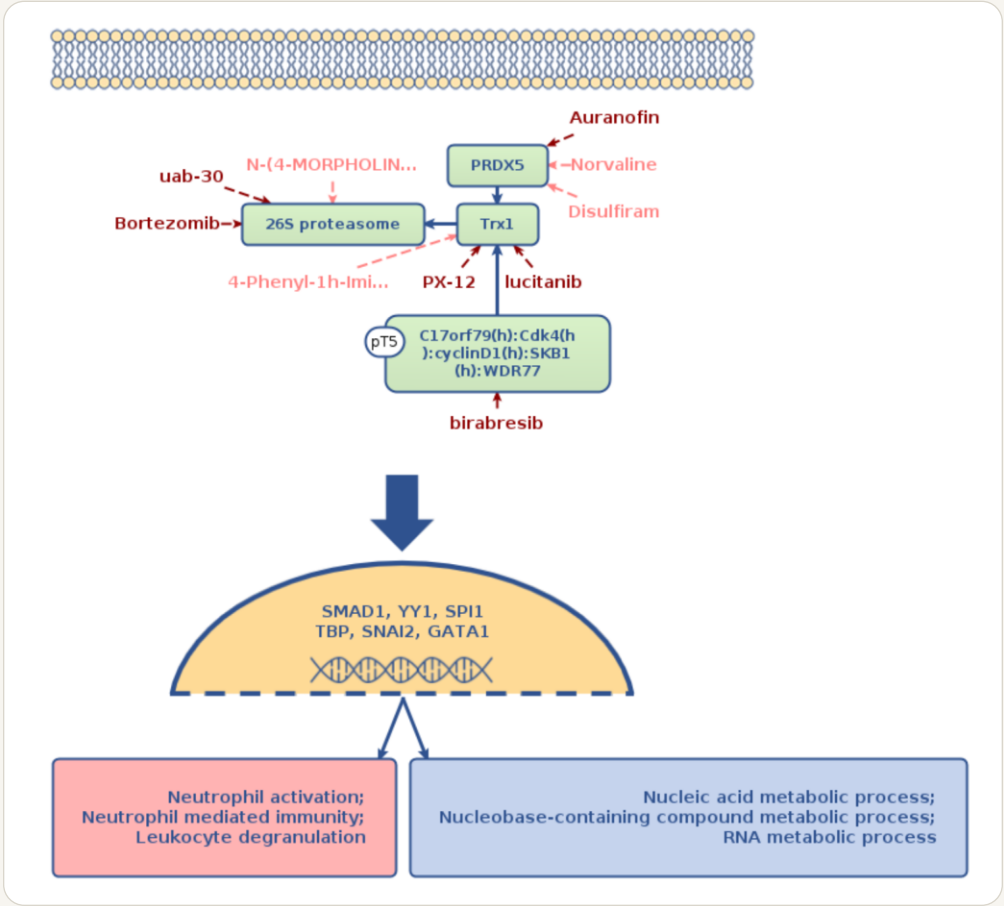
Protein level in adapted cells, sensitive cells = 100%

−32%	360%	209%	204%
TrxR1, drug target	NQO1	HMOX1	GCLM
$p < 10^{-4}$	$p < 10^{-7}$	$p < 10^{-3}$	$p < 10^{-6}$

NQO1, HMOX1 and GCLM are all **Nrf2 targets**

Resistance = rewired ROS defence + **low nucleotide metabolism** (a slow, hypometabolic state)

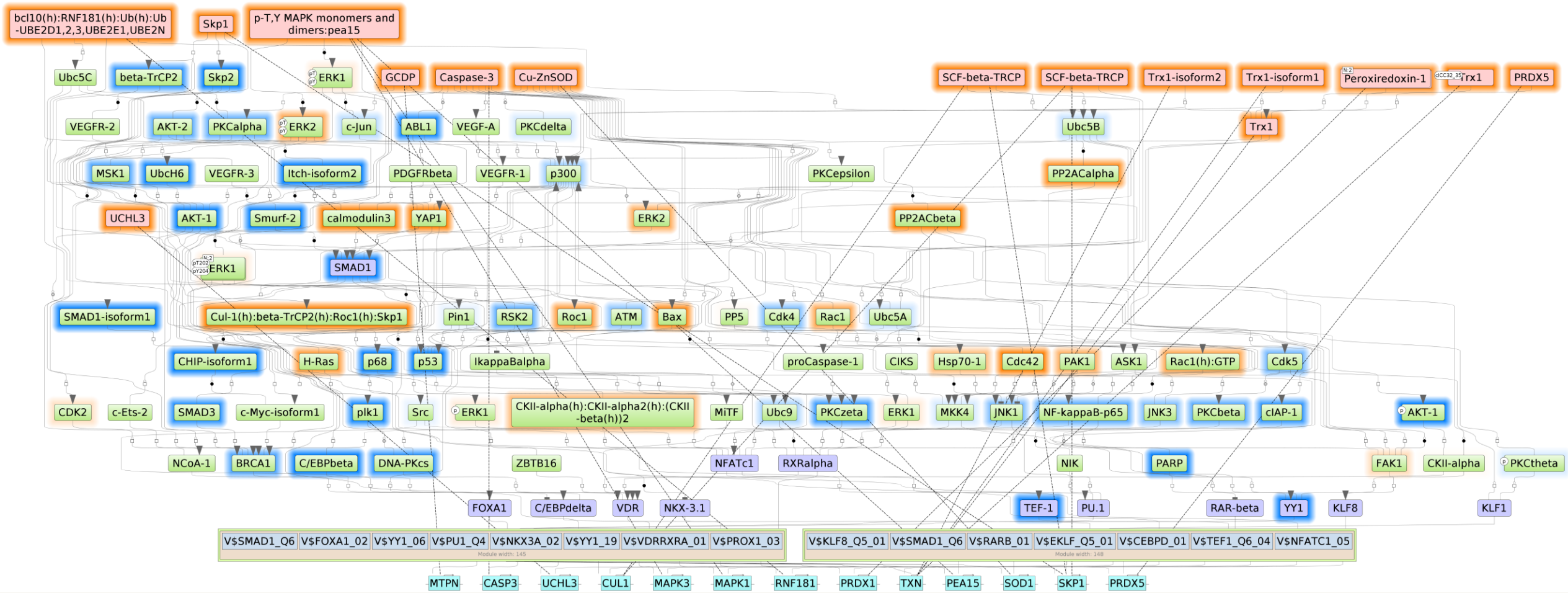
MDA-MB-231 cells after 16 weeks: adapted, partially resistant. Other 4 cell lines did not become resistant.



Lyu H, …, Kel A, …, Zubarev RA. OVERDRIVE. Manuscript, June 2026. Figs. 6–7

NEW DRUG · LCTA2940, A TRXR1 INHIBITOR

Upstream analysis revealed master regulators – potential drug targets



FROM PREDICTION TO VALIDATION

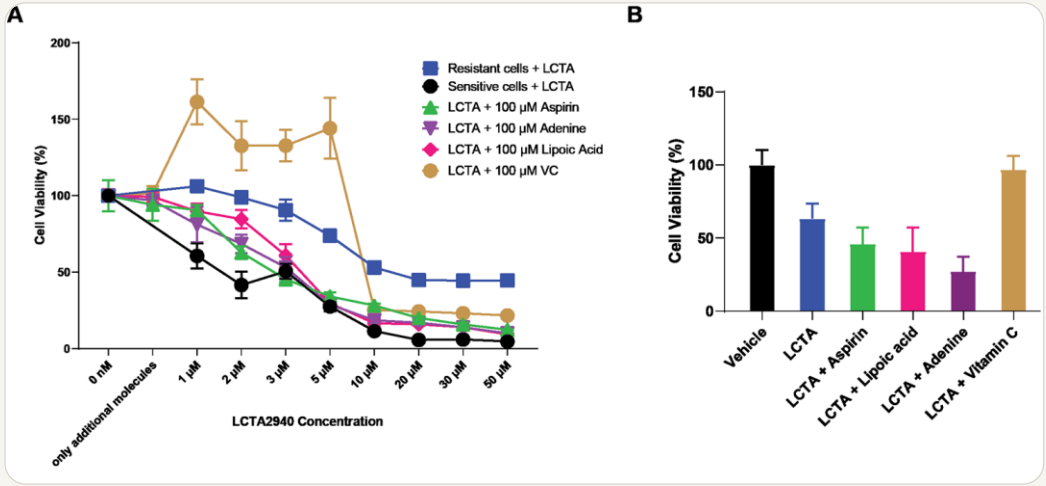
******: predicted, tested, and specific

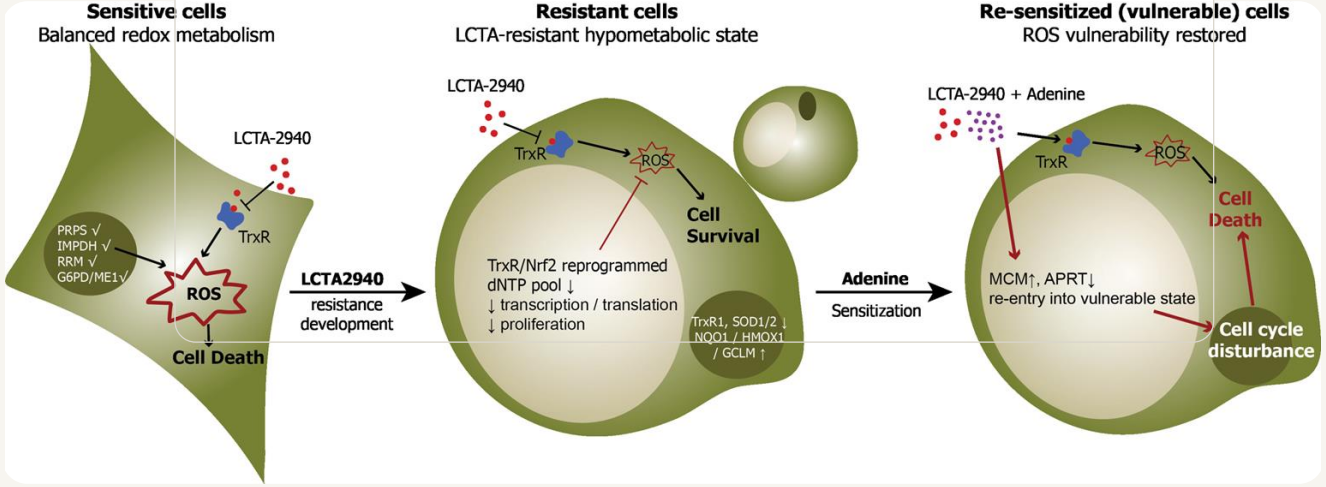
>360 candidates → **PASS** non-toxic, ≥2 targets → **3** to test

Three found compounds lowered LC50. ***** was best**: ~2× potency at 3 μM, not toxic alone.

CONTROLS: NO EFFECT WITH

glucose, glutamine (growth boosters) · xanthine (other purine) · 5-FU-resistant cells · sensitive cells





In vitro, one cell line. Toxicity in normal cells and in vivo effect are not yet tested.

Lyu H, …, Kel A, …, Zubarev RA. OVERDRIVE. Manuscript, June 2026. Figs. 8–9, Suppl. Figs. S6–S8

SUMMARY

Hopes, hype, and reality

Hopes

- integrate data and knowledge at scale
- single-cell views of escape states
- agents that plan and choose experiments

Hype

- a ranking or a good story called “discovery”
- retrospective agreement called accuracy
- success credited to deep learning without ablation

Reality

- successes stand on decades of curated knowledge
- a target is an interaction whose change reverses the phenotype
- experiments remain the judge

AGENTIC AI

Four levels of scientific autonomy

LEVEL 1

Run the analysis

choose tools, write code, process omics

Biomni · QSP-Copilot

LEVEL 2

Generate hypotheses

collect evidence, debate, rank

Co-Scientist · OriGene · XunZi

LEVEL 3

Build mechanistic models

signed networks, simulate perturbations

FLASH-P · VCR-Agent

LEVEL 4

Close the loop with the lab

propose experiment, read new data, revise

Robin · BioPathfinder

Most systems today: useful, growing fast

Few and early: this is where targets need to be found

Big Data + Big Knowledge + AI reasoning + experiments

The breakthrough will come when AI tells us not only *which molecules are involved*, but which interactions we need to change to reverse the disease.

Thank you

LCTA2940 project partners [names] · geneXplain team [names] · contact: [e-mail]

Thank you!



EVIDENCE FROM SINGLE CELLS

Cancer cells divide well with only part of the cycle machinery

IN REAL TUMOURS

Non-canonical combinations of cell cycle regulators in single tumour cells: the parts are **decoupled** from the normal pattern

Low “coherence” goes with more aggressive tumours

Gaglia et al., Nat Cell Biol 2022

ESCAPE FROM A DRUG

Under CDK4/6 inhibitor, a fraction of breast tumour cells keeps dividing through **alternative cycle paths**

“Cell cycle plasticity”, single-cell imaging of cells and patient tumours

Zikry et al., PNAS 2024

RE-ROUTING IN HOURS

Block CDK2 and cells **re-route through CDK4/6**, restoring the cycle within a few hours

Different tumours depend on different CDKs: RB-loss and CCNE1 models ignore CDK4/6

Arora et al., Cell 2023 · Kumarasamy et al., Nat Commun 2025

Even genetically: cells lacking CDK2, 3, 4 and 6 still divide

Only CDK1 is essential (Santamaría et al., Nature 2007). The wheel turns with most of its usual parts missing, so the **target is the broken control, not the cycle itself**.

Gaglia G et al. Nat Cell Biol 2022;24:316–326 · Zikry T et al. PNAS 2024;121:e2309261121 · Arora M et al. Cell 2023 · Kumarasamy V et al. Nat Commun 2025;16:1382 · Santamaría D et al. Nature 2007;448:811–815

CASE STUDY · LCTA2940

One target inhibited. How do the cells escape?

COMPOUND

LCTA2940, a selective thioredoxin reductase (TXNRD1) inhibitor

CELLS

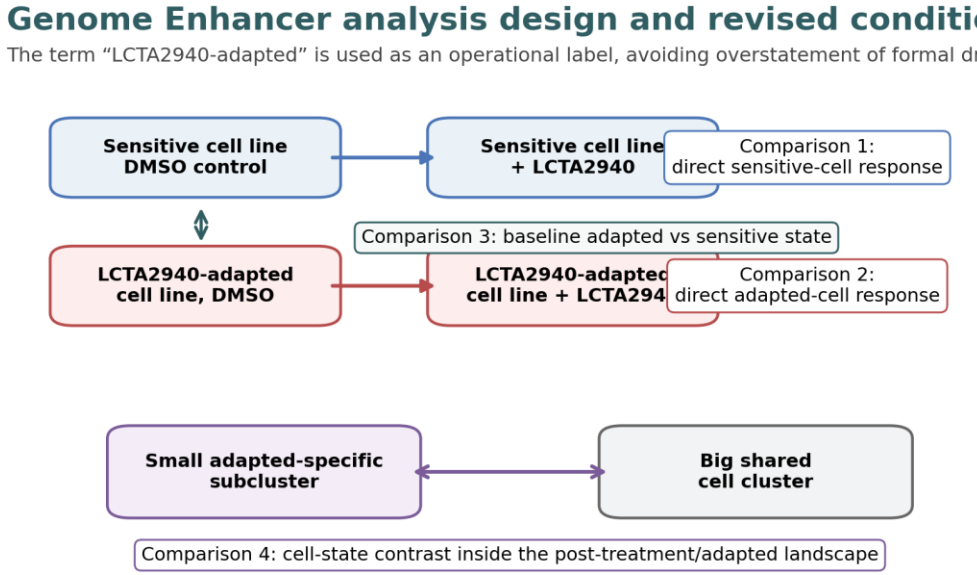
MDA-MB-231 breast cancer: sensitive line vs a line **adapted** by low-dose selection

DATA

scRNA-seq, 4 conditions (± LCTA2940), 22,291 cells after QC

ANALYSIS

pseudobulks of 500 cells → Genome Enhancer: DEGs, TF sites, master regulators



Four comparisons. “Adapted” = partially resistant; formal resistance not established.

CASE STUDY · TWO RESPONSES

Same inhibitor, two different regulatory programs

Sensitive cells: broad stress

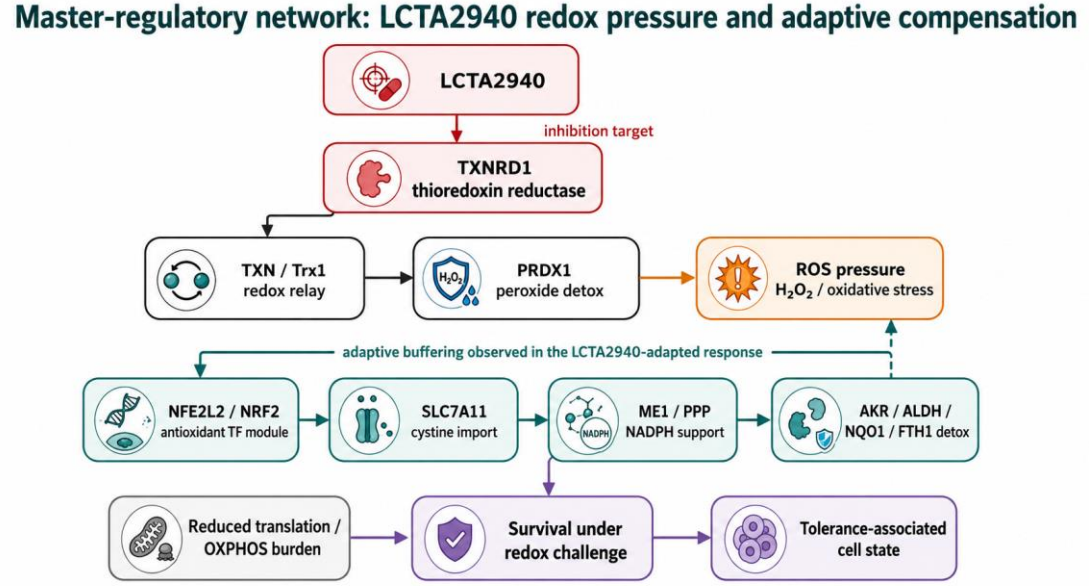
756 up · 346 down

mitochondrial, translation and stress programs; TP53, JUN, YBX1 up; STAT1/3, FOS family down; VEGFA, TNF/TNFR1 among regulators

Adapted cells: focused defence

573 up · 436 down

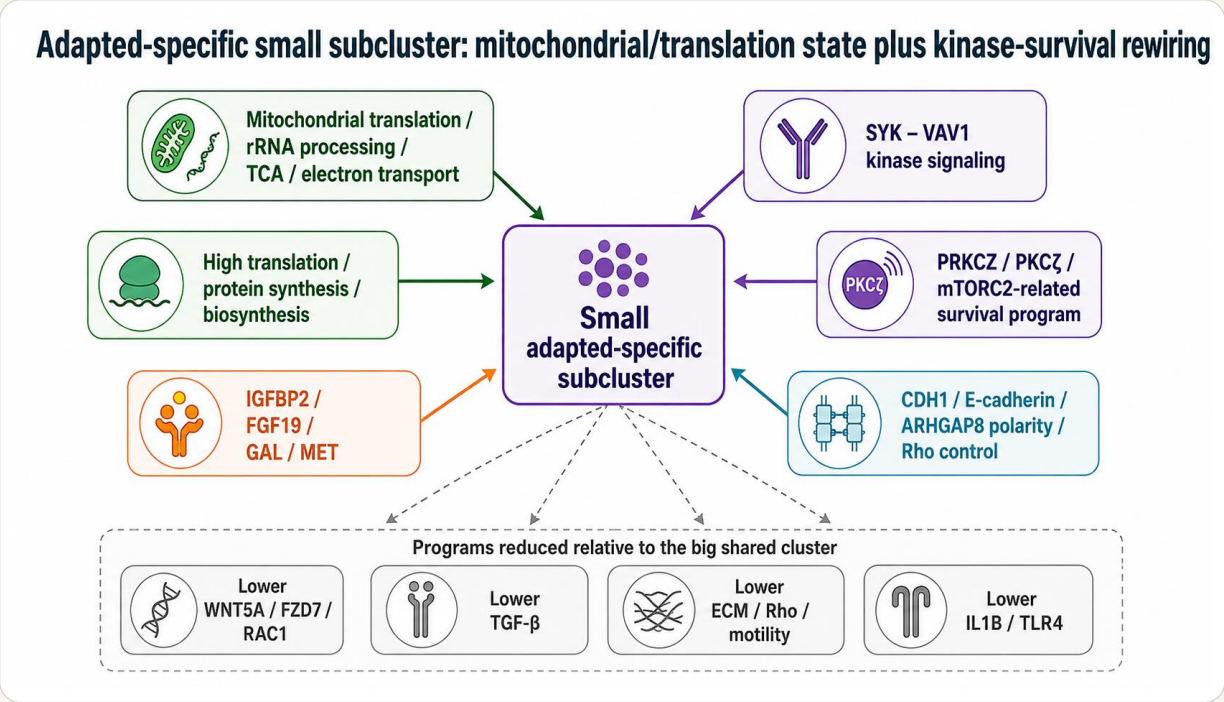
NRF2 (NFE2L2) with MAFK/MAFG; TXNRD1, TXN, PRDX1, SLC7A11, NQO1, AKR1C; regulators CSF2, IKBKG



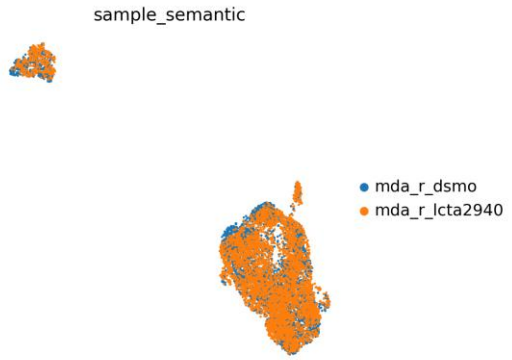
Working redox model from upstream analysis

CASE STUDY · SINGLE-CELL RESOLUTION

A small adapted population with its own signaling



A survival role of this cluster and its dependence on SYK or PRKCZ are hypotheses, not yet tested.

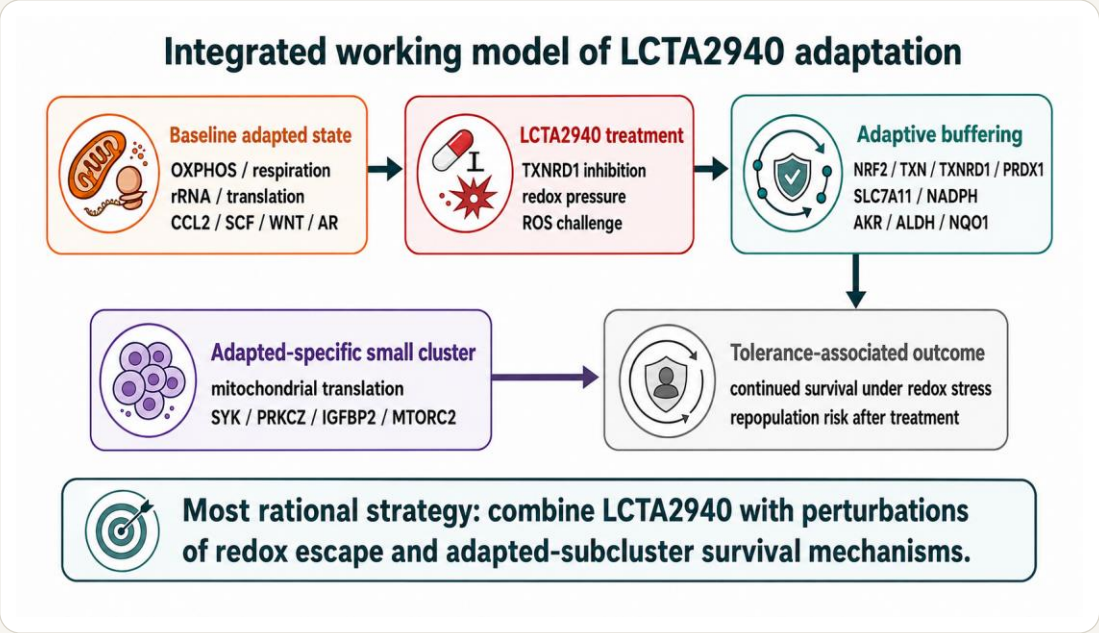


Adapted cells, UMAP: the small cluster holds DMSO and treated cells

- 3,946 up · 3,756 down vs the large cluster
- up: mitochondrial translation, biosynthesis
- **SYK–VAV1, PRKCZ–mTORC2, IGFBP2**, CDH1
- lower TGF-β, WNT5A–RAC1, ECM: not simple EMT

CASE STUDY · FROM MECHANISM TO TARGETS

The model points to interactions we can test



Redox escape route

perturb PRDX1, SLC7A11 / glutathione, thioredoxin axis

Small-cluster survival

perturb SYK, PRKCZ, mTORC2, alone and with LCTA2940

Success looks like

stronger effect in combination and loss of the small cluster

One culture per condition: pseudobulks are not biological replicates

"Adapted" is not proven resistance

Network edges are knowledge-based hypotheses

No AI agent was used in this analysis

OPTIONAL BLOCK A

State of the art, paper by paper

AI-assisted targets (6 slides) · agentic systems (9 slides) · one scorecard. Each slide: what they did, what they found, how far the evidence goes.

STATE OF THE ART · AI-ASSISTED TARGETS

TNIK in the clinic: rentosertib in IPF

Nature Medicine 2025 | peer-reviewed | evidence: Phase 2a, Phase 3 running

What they did

- GENESIS-IPF: 71 patients, 22 sites in China
- placebo vs several doses, 12 weeks
- primary endpoint: safety and tolerability

What they found

- FVC **+98.4 mL** at 60 mg vs **−20.3 mL** placebo
- mainly mild diarrhoea and liver-test changes
- Phase 3: 320 patients, 47 centres, 52 weeks (2026)

Our reading

- first target from an AI pipeline tested in patients
- small, short, safety-primary trial
- clinical success ≠ proof that AI chose the target

Xu Z et al., Nat Med 2025 (doi:10.1038/s41591-025-03743-2); Phase 3 NCT07687459 (per trackers); Ren F et al., Nat Biotechnol 2024

STATE OF THE ART · AI-ASSISTED TARGETS

XunZi: CHK2 in Parkinson’s disease

Nature Biomedical Engineering 2026 | peer-reviewed | evidence: rescue in mice

What they did

- “AI biologist” reasoning over 24.4 M publications
- plus ~614 TB of multi-source data: 21,008 genes, 5,850 diseases
- hypotheses for Parkinson’s and lung cancer

What they found

- aberrant CHK2 and IRAK4 activation across PD models
- genetic and drug inhibition of CHK2: less neuron loss, better motor function in mice

Our reading

- strong preclinical validation, genetic and chemical
- IRAK4 not tested; no human data
- AI proposes, humans test: no closed loop, no baseline

Huang X, Qin J, Tang F et al., Nat Biomed Eng 2026 (doi:10.1038/s41551-026-01769-6)

STATE OF THE ART · AI-ASSISTED TARGETS

CRISPR + AI: ALOX5 and OXTR in psoriasis

Nature Communications 2026 | peer-reviewed | evidence: rescue in mice

What they did

- genome-wide knockout screen: ~77,000 guides, ~19,000 genes, human keratinocytes
- readout: IL17RA on the cell surface
- “VirtualCRISPR” LLM asked: would the literature predict this hit?

What they found

- strong but unexpected hits: **ALOX5** (zileuton) and **OXTR** (cligosiban)
- topical gels reduced disease in a mouse model, close to anti-IL17RA after a week

Our reading

- the experiment discovered; the AI measured surprise
- a smart use of knowledge: filter what is already known
- preclinical only

Zhao C et al., Nat Commun 2026 (doi:10.1038/s41467-026-75249-5), Biohub

STATE OF THE ART · AI-ASSISTED TARGETS

MIDAS: OSM–OSMR in melanoma immunotherapy

Nature Machine Intelligence 2026 | peer-reviewed | evidence: human tumour explants

What they did

- multimodal graph neural network
- inputs: gene interaction networks, patient multi-omics, immune biology, perturbation phenotypes
- ranks approved targets above OpenTargets

What they found

- **OSM–OSMR** signaling axis
- blocking it in TRACERx melanoma explants: fewer dysfunctional $CD8^+$ T cells, less CCL4

Our reading

- a knowledge-heavy model: networks are the prior
- human tissue is a strong intermediate step
- no tumour regression or survival yet

Augustine M et al., Nat Mach Intell 2026;8:670–689 (doi:10.1038/s42256-026-01201-3), Crick / UCL

STATE OF THE ART · AI-ASSISTED TARGETS

LLMs + single-cell atlases: GPNMB for CAR-T

Cell 2026 | peer-reviewed | evidence: tumour control in mice

What they did

- >10,000 candidate antigens from tumour and normal single-cell data
- filters: tumour specificity, normal tissue, feasibility
- 1,000 independent LLM runs → consensus; humans review

What they found

- **GPNMB** on top, expressed across many tumour types
- GPNMB CAR-T eliminated tumours in mouse models of melanoma, monoblastic leukaemia and colorectal cancer

Our reading

- new modality, not a new target: GPNMB had an ADC (glembatumumab vedotin)
- repeated LLM runs = a guard against hallucination
- normal-tissue expression: safety open

Baker et al., Cell 2026 (doi:10.1016/j.cell.2026.06.002), University of Pennsylvania

STATE OF THE ART · AI-ASSISTED TARGETS

Geneformer: in-silico treatment of cardiomyopathy

Nature 2023 | peer-reviewed | evidence: iPSC cell model

What they did

- transformer pretrained on ~30 million single-cell transcriptomes
- fine-tuned with little patient data
- in-silico deletion: which genes move diseased cells to healthy?

What they found

- candidate therapeutic targets for cardiomyopathy
- knockout of predicted targets improved contraction in an iPSC cardiomyocyte disease model

Our reading

- first “foundation model” to propose and test targets
- cell model only; in-vivo relevance open
- later zero-shot benchmarks found clear limits

Theodoris CV et al., Nature 2023 (doi:10.1038/s41586-023-06139-9); Kedzierska KZ et al., Genome Biol 2025

STATE OF THE ART · AGENTIC AI

Co-Scientist: epigenetic targets in liver fibrosis

Nature 2026 · Advanced Science 2025 | peer-reviewed | evidence: human organoids

What they did

- agents generate, debate, rank and evolve hypotheses
- scientists asked: epigenetic targets for liver fibrosis?
- test: TGF-β-stimulated human hepatic organoids

What they found

- HDAC and BRD4 inhibitors were anti-fibrotic; vorinostat strongest
- DNMT1 inhibitor toxic; two expert-chosen targets failed
- also: re-derived an unpublished phage mechanism

Our reading

- the target class was given by humans
- repurposing of known mechanisms, honest controls
- a strong hypothesis partner, not yet a discoverer of new targets

Co-Scientist: Nature 2026 (doi:10.1038/s41586-026-10644-y); liver fibrosis: Adv Sci 2025 (doi:10.1002/advs.202508751)

STATE OF THE ART · AGENTIC AI

Robin: a hypothesis loop for dry AMD

Nature 2026 | peer-reviewed | evidence: cells in vitro

What they did

- literature agents propose a strategy: boost RPE phagocytosis
- agent proposes drugs and experiments; humans run them
- data agent analyses the new RNA-seq and updates the hypothesis

What they found

- **ripasudil** (ROCK inhibitor) and KL001 increased phagocytosis
- RNA-seq: **ABCA1** lipid efflux up, proposed as a new target

Our reading

- rare example: the agent changed its mind after new data
- authors call it “semi-autonomous”
- ABCA1 not perturbed; no in-vivo data

Robin, Nature 2026 (doi:10.1038/s41586-026-10652-y); preprint arXiv:2505.13400 (FutureHouse)

STATE OF THE ART · AGENTIC AI

OriGene: a “virtual disease biologist”

bioRxiv 2025, revised 2026 | preprint | evidence: organoids, humanized mice

What they did

- 5 agents: coordinator, planner, reasoner, critic, reporter
- >600 tools and 18 curated databases
- learns from human and lab feedback

What they found

- **GPR160**, liver cancer (from 125 candidates): organoids, tumour fragments, humanized mice, more T cells
- **ARG2**, colorectal cancer (from 86): inhibitor reduced viability of 4 patient organoids

Our reading

- broad experimental follow-up
- compound selectivity and novelty versus literature not controlled
- not yet peer-reviewed

OriGene, bioRxiv (doi:10.1101/2025.06.03.657658); Lingang Lab, Shanghai Jiao Tong, Fudan and others

STATE OF THE ART · AGENTIC AI

BioPathfinder: NKG2A in exhausted CAR-T cells

bioRxiv 2026 | preprint | evidence: in vivo models

What they did

- evidence-tracked base: CAR-T papers, patient scRNA-seq, clinical data
- agents generate, review and rank mechanisms
- virtual perturbation of candidate genes

What they found

- an NK-like program in exhausted CAR-T cells
- KLRC1, KLRD1, KLRG1 ranked; experts chose **KLRC1 (NKG2A)**
- blockade improved anti-tumour activity and persistence in vivo

Our reading

- provenance of every piece of evidence: the right design
- NKG2A is a known checkpoint; new context
- humans made the final choice

BioPathfinder, bioRxiv 2026 (doi:10.64898/2026.07.15.738646)

STATE OF THE ART · AGENTIC AI

FLASH-P: agents build causal networks

bioRxiv June 2026 | preprint | evidence: retrospective

What they did

- agents read the literature for a trait and species
- build signed, directed causal networks in under an hour
- simulate: signal propagation, Hill ODEs, random walk

What they found

- 12 networks, 7 species
- **90%** correct direction for 1,088 published perturbations
- recovered known pleiotropic effects

Our reading

- closest to an executable mechanistic model
- the test perturbations come from the same literature
- mostly plants; no prospective test yet

Mitsanis C, Fortuna N, Beveridge C, Kainer D. FLASH-P, bioRxiv 2026 (doi:10.64898/2026.06.13.731799)

STATE OF THE ART · AGENTIC AI

MechAInistic: agents on executable models

arXiv 2026 | preprint | evidence: in silico only

What they did

- Architect and Reviewer agents turn a question into a workflow
- run on genome-scale models of immune cells, disease vs healthy

What they found

- RA B cells: mitochondrial rewiring, OGDH (devimistat)
- MS Th17 cells: NADP-isocitrate dehydrogenase; ivosidenib as repurposing idea

Our reading

- the right idea: agent reasons, the model computes
- models were given by humans; no wet lab
- ivosidenib targets mutant IDH1: the drug mapping is doubtful

Loecker J … Helikar T. MechAInistic, arXiv:2607.18249 (2026)

STATE OF THE ART · AGENTIC AI

VCR-Agent: explanations that are checked

arXiv 2026 · Valence Labs | preprint | evidence: computational

What they did

- explain a compound's effect as an action graph of 20 typed steps
- retrieval from knowledge graphs and PubMed
- verifiers: predicted binding and measured expression

What they found

- 18,950 compound–cell pairs from Tahoe-100M
- verifiers removed **28.2%** of wrong binding claims, corrected 87.3% of expression steps

Our reading

- shows how often LLM mechanisms are wrong
- checking against data and knowledge is the way
- explanations, not new targets

VCR-Agent, arXiv:2604.11661 (2026); code: github.com/valence-labs/VCR-Agent

STATE OF THE ART · AGENTIC AI

BioDiscoveryAgent: choosing the next experiment

ICLR 2025 | peer-reviewed conference | evidence: simulated screens

What they did

- LLM agent picks which genes to perturb, 5 rounds of 128
- uses literature search, code and a critic agent
- outcomes looked up in existing CRISPR screens

What they found

- **~21%** more hits than Bayesian optimization, 6 datasets
- ~46% more on non-essential genes
- gene pairs: 2× better than random

Our reading

- the gain comes from prior biological knowledge in the LLM
- experiments were replayed, not run
- a key skill for closed-loop discovery

Roohani Y et al., BioDiscoveryAgent, ICLR 2025 (arXiv:2405.17631)

STATE OF THE ART · AGENTIC AI

Biomni: a general biomedical agent

bioRxiv 2025 · Stanford | preprint | level 1: runs analyses

What they did

- >150 tools, >100 software packages, 59 databases
- plans, writes and runs code from a plain request

What they found

- >15,000 users, ~100,000 workflows in 9 months
- strong: databases, sequences, cloning
- weaker: clinical judgement, deep biological synthesis

Our reading

- real productivity for every lab
- weakest exactly where mechanisms are needed
- a good host for curated knowledge tools

Huang K et al., Biomni, bioRxiv 2025.05.30.656746; usage figures: Stanford HAI, 2026

STATE OF THE ART · SCORECARD

One table, nine targets

Target · system	Perturbation	Best evidence	Non-AI baseline	Loop by agent
TNIK · PandaOmics	knockdown + drug	patients, Phase 2a	not reported	no
CHK2 · XunZi	genetic + drug	mouse rescue	not reported	no
ALOX5, OXTR · VirtualCRISPR	CRISPR + drugs	mouse rescue	the screen itself	no
OSM–OSMR · MIDAS	pathway blockade	human explants	ranking only	no
GPNMB · LLM consensus	CAR-T	mouse tumours	not reported	humans in loop
GPR160 · OriGene	compound	humanized mice	not reported	partly
KLRC1 · BioPathfinder	antibody	mouse in vivo	not reported	humans chose
HDAC, BRD4 · Co-Scientist	compounds	human organoids	expert controls	partly
ABCA1 · Robin	not yet	cells in vitro	not reported	yes, semi-auto

Speaker’s reading of the published papers, October 2026. “Not reported” = no head-to-head comparison with a non-AI pipeline found.

OPTIONAL BLOCK D

Three key papers in depth

TNIK

AI target ranking that reached patients

XunZi

reasoning model + knowledge graph; CHK2 validated in mice

Robin

an agent loop that revised its hypothesis after new lab data

OPTIONAL BLOCK B

Why Big Knowledge is needed

The grammar analogy · what goes wrong when models learn from omics or genomes alone · where prior knowledge helps

BIG KNOWLEDGE · THE GRAMMAR OF BIOLOGY

Data give the vocabulary. Knowledge gives the grammar.

> *"Omics datasets alone give a model the vocabulary of biology without the grammar."*
>
> *Such models "can speak fluently and still be wrong."*
>
> genexplain.com/knowledge-graph

Language	Gene regulation
letters	A, C, G, T
words	TF binding sites, motifs
grammar	TF combinations: order, spacing, cooperation
sentences	signaling: receptor → kinase → TF → gene
meaning	the phenotype, the disease mechanism

A model trained on expression learns which words occur together. Grammar tells **what a change means**, and **what to change**.

LIMITS OF DATA ALONE · 1

In omics, correlations are cheap

PLOS COMPUT BIOL 2011

Random signatures predict outcome

Most randomly chosen gene sets were significantly associated with breast-cancer outcome, and so were signatures of unrelated biology, such as social defeat in mice.

Venet D, Dumont JE, Detours V

PNAS 2006

Gene lists are unstable

Thousands of samples are needed for a robust prognostic gene list; lists from different cohorts for the same question hardly overlap.

Ein-Dor L, Zuk O, Domany E

With 20,000 parameters and 100 samples, **association with outcome is not evidence of mechanism**. A bigger model does not fix this.

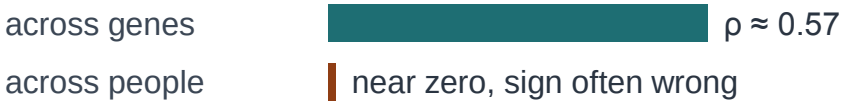
LIMITS OF DATA ALONE · 2

Models trained on genomes miss what matters for a person

NATURE GENETICS 2023

Personal genomes, personal expression

Enformer, Basenji2, ExPecto, Xpresso on 421 individuals (Geuvadis):



Huang C … Ioannidis NM. Enformer row shown.

GENOME BIOLOGY 2025

DNA language models

Pre-trained genomic language models gave no substantial advantage over a well-tuned CNN trained on one-hot DNA for regulatory-genomics tasks.

Tang Z, Somia N, Yu Y, Koo PK

A target lives in the effect of a variant or an interaction in **this** patient. Learning sequence statistics is not learning regulatory grammar.

LIMITS OF DATA ALONE · 3

Single-cell foundation models: scale is not structure

GENOME BIOLOGY 2025

Zero-shot scGPT and Geneformer

For cell-type clustering and batch integration without fine-tuning, the models were often unreliable, and simpler methods frequently did better.

Kedzierska KZ, Crawford L, Amini AP, Lu AX

NATURE METHODS 2025

Perturbation prediction

Deep models of genetic perturbation effects did not beat deliberately simple linear baselines; well-calibrated metrics (2026) narrow but do not close the gap to biology.

Ahlmann-Eltze C, Huber W, Anders S

Tens of millions of cells teach co-expression. They do not, by themselves, teach **direction, cause and intervention**.

WHERE KNOWLEDGE HELPS

Add prior knowledge, and models improve where data are thin

NAT BIOTECHNOL 2024

GEARS

A Gene Ontology graph links genes never perturbed to known ones: predicts unseen perturbations; precision for genetic interactions up >50%.

Roohani Y, Huang K, Leskovec J

NATURE 2021

P-NET

A neural network wired by pathways: interpretable, held on two external prostate cohorts, and pointed to MDM4, confirmed in cells.

Elmarakeby HA … Van Allen EM

ICLR 2025

BioDiscoveryAgent

An agent with prior biological knowledge found ~21% more screen hits than a knowledge-free Bayesian optimizer.

Roohani Y et al.

Knowledge works twice: as an **inductive bias** that generalizes, and as an **explanation** we can test.

BIG KNOWLEDGE · THE GRAMMAR BOOK

Grammar is not scraped. It is read, checked and linked to experiments.

38+ years of expert curation	**12,469** TF weight matrices (TRANSFAC)	**1.27 M** signaling reactions (TRANSPATH)
1,000+ person-years of curation effort	**231,413** enhancers and silencers	**2,796** disease models (HumanPSD)

Release 2026.1 statistics as published on genexplain.com/knowledge-graph (October 2026)

OPTIONAL BLOCK C

Explainability and robustness

Why black boxes fail on new cohorts · why in target discovery the explanation is the result · Walking Pathways: mechanism-based biomarkers on an independent cohort

EXPLAINABILITY · WHY IT MATTERS

Black boxes learn shortcuts, and fail on the next cohort

NAT COMMUN 2019

“Clever Hans”

An image classifier recognised horses by a copyright tag in the corner, not by the horse.

Lapuschkin S et al.

PLOS MED 2018

Pneumonia from X-rays

A deep model also learned which hospital the image came from; performance dropped in other hospitals.

Zech JR et al.

NAT MACH INTELL 2021

COVID-19 radiographs

Models chose shortcuts such as text markers and patient position over disease signal, and failed on external data.

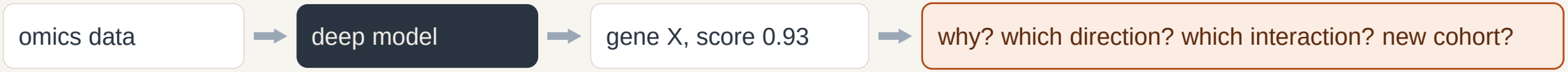
DeGrave AJ, Janizek JD, Lee S-I

In omics the shortcuts are batch, centre, platform and proliferation. For high-stakes decisions, prefer **models that are interpretable by design** (Rudin, Nat Mach Intell 2019).

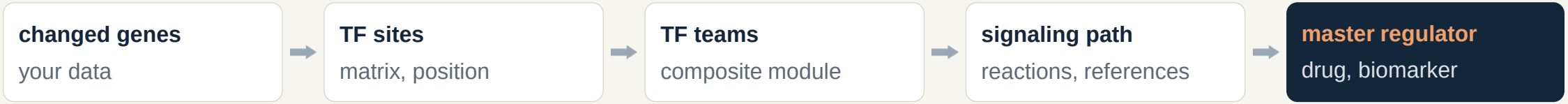
EXPLAINABILITY · IN TARGET DISCOVERY

In target discovery, the explanation is the result

BLACK BOX



UPSTREAM ANALYSIS



Every step can be **checked** against a publication, and every step can be **perturbed** in the lab.

EXPLAINABILITY · ROBUSTNESS

Why mechanism-based markers should travel to new cohorts

Correlation-based markers

- chosen from ~20,000 features in one cohort
- absorb cohort noise: batch, centre, platform
- many equally good lists; little overlap between cohorts

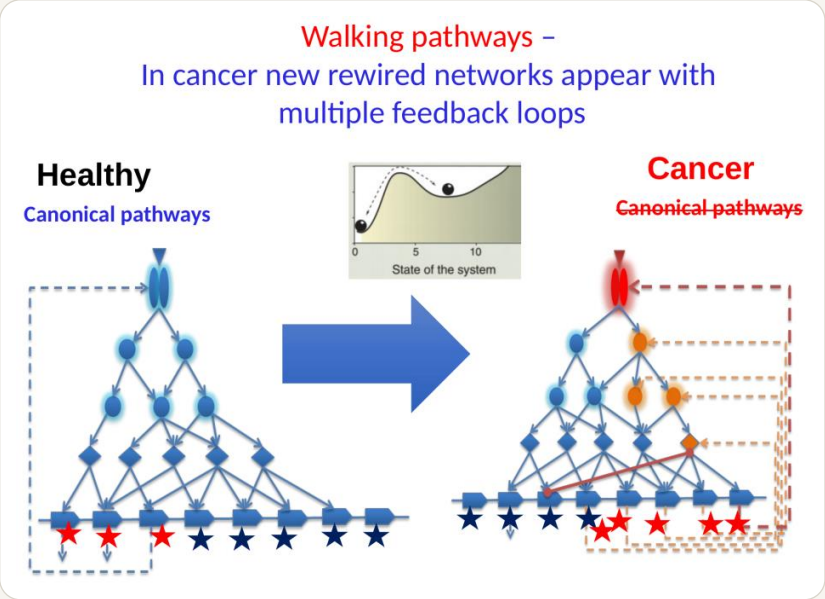
Mechanism-based markers

- constrained by known regulation: sites, loops, master regulators
- many genes must agree with one mechanism: fewer free parameters
- patients share the mechanism, not the noise

This is a theoretical expectation. It must be **tested on independent cohorts**, as in the next example.

ROBUSTNESS IN PRACTICE · COLORECTAL CANCER

Walking pathways: methylation markers that held in blood



Markers chosen from positive feedback loops around master regulators, not by correlation alone

DISCOVERY · TUMOUR TISSUE

~300 patients (SysCol): DNA methylation and RNA-seq → upstream analysis → feedback loops

INDEPENDENT COHORT · BLOOD

cell-free DNA, 102 CRC vs 100 controls (Moscow, Novosibirsk), pyrosequencing

RESULT · 6-CPG PANEL

ENO1, MYC, CALCA, PDX1, TCF7, ZNF43: mean accuracy **92.3%**; sensitivity and specificity above 92%

Small validation cohort; a prospective screening study is still needed.

Kel A, Boyarskikh U, Stegmaier P, … Esteller M, … Wingender E. BMC Bioinformatics 2019;20(Suppl 4):119 (doi:10.1186/s12859-019-2687-7)

BACKUP · REFERENCES 1/3

Targets, trials and benchmarks

- Ren F et al. TNIK inhibitor targets fibrosis. Nat Biotechnol 2024. doi:10.1038/s41587-024-02143-0
- Xu Z et al. TNIK inhibitor for IPF, randomized phase 2a trial. Nat Med 2025. doi:10.1038/s41591-025-03743-2
- PandaOmics methods. J Chem Inf Model 2024. doi:10.1021/acs.jcim.3c01619 · iPANDA: Ozerov IV et al. Nat Commun 2016
- Buchl SC et al. TNIK, procollagen-I trafficking and hepatic fibrogenesis. Hepatol Commun 2022
- Huang X et al. XunZi, an AI biologist. Nat Biomed Eng 2026. doi:10.1038/s41551-026-01769-6
- Zhao C et al. AI-guided CRISPR screening in psoriasis. Nat Commun 2026. doi:10.1038/s41467-026-75249-5
- Augustine M et al. Immunotherapy target identification with ML (MIDAS). Nat Mach Intell 2026;8:670–689
- Baker et al. AI-guided CAR-T target discovery (GPNMB). Cell 2026. doi:10.1016/j.cell.2026.06.002
- Theodoris CV et al. Transfer learning in network biology (Geneformer). Nature 2023
- Ahlmann-Eltze C, Huber W, Anders S. Perturbation models vs simple baselines. Nat Methods 2025
- Calibrated metrics for perturbation models. Nat Biotechnol 2026. doi:10.1038/s41587-026-03307-w
- AI clinical evidence review. Nat Rev Drug Discov 2026. doi:10.1038/s41573-026-01496-2 · ASCO 2026, abstr. 11072
- Richardson P et al. Baricitinib for 2019-nCoV acute respiratory disease. Lancet 2020;395:e30–e31
- Lloyd K et al. Clarithromycin repurposing in IBD. Dis Model Mech 2020;13:dmm044040

BACKUP · REFERENCES 2/3

Agents and upstream-analysis methods

- Co-Scientist, a multi-agent system for hypothesis generation. Nature 2026. doi:10.1038/s41586-026-10644-y
- AI-assisted drug re-purposing for human liver fibrosis. Adv Sci 2025. doi:10.1002/advs.202508751
- Robin: a multi-agent system for automating scientific discovery. Nature 2026. doi:10.1038/s41586-026-10652-y
- OriGene: a self-evolving virtual disease biologist. bioRxiv preprint. doi:10.1101/2025.06.03.657658
- Mitsanis C, Fortuna N, Beveridge C, Kainer D. FLASH-P. bioRxiv preprint 2026. doi:10.64898/2026.06.13.731799
- VCR-Agent: mechanistic action graphs for virtual-cell perturbations. arXiv:2604.11661
- BioPathfinder: agentic analysis of CAR-T dysfunction. bioRxiv preprint 2026. doi:10.64898/2026.07.15.738646
- Biomni: a general-purpose biomedical AI agent. bioRxiv 2025.05.30.656746
- BioDiscoveryAgent: adaptive perturbation experiment design. ICLR 2025
- QSP-Copilot: agents for quantitative systems pharmacology models. CPT Pharmacometrics Syst Pharmacol 2025. doi:10.1002/psp4.70127
- Kel AE et al. MATCH: a tool for searching transcription factor binding sites. Nucleic Acids Res 2003;31:3576–3579
- Waleev T et al. Composite Module Analyst: TF binding site combinations by genetic algorithm. Nucleic Acids Res 2006

Preprints are not peer-reviewed; status checked October 2026.

BACKUP · REFERENCES 3/3

Knowledge, explainability and robustness

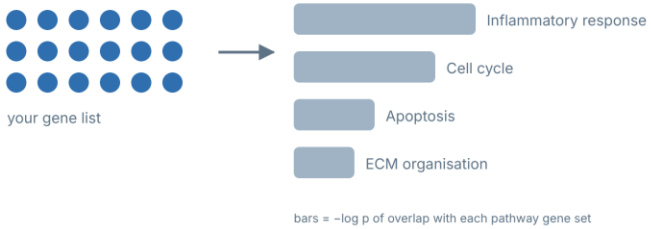
- Venet D, Dumont JE, Detours V. Most random gene expression signatures are associated with breast cancer outcome. PLoS Comput Biol 2011
- Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list. PNAS 2006
- Huang C et al. Personal transcriptome variation is poorly explained by genomic deep learning models. Nat Genet 2023;55:2056
- Tang Z, Somia N, Yu Y, Koo PK. Representational power of pre-trained DNA language models. Genome Biol 2025
- Kedzierska KZ et al. Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biol 2025;26:101
- Roohani Y, Huang K, Leskovec J. GEARS: transcriptional outcomes of multigene perturbations. Nat Biotechnol 2024
- Elmarakeby HA et al. Biologically informed deep neural network for prostate cancer (P-NET). Nature 2021;598:348
- Lapuschkin S et al. Unmasking Clever Hans predictors. Nat Commun 2019
- Zech JR et al. Generalization of a deep model for pneumonia across hospital systems. PLoS Med 2018
- DeGrave AJ, Janizek JD, Lee S-I. AI for COVID-19 radiographs selects shortcuts over signal. Nat Mach Intell 2021
- Rudin C. Stop explaining black box models for high-stakes decisions. Nat Mach Intell 2019
- Kel A et al. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of CRC. BMC Bioinformatics 2019;20(Suppl 4):119
- Loecker J … Helikar T. MechAInistic. arXiv:2607.18249 · Roohani Y et al. BioDiscoveryAgent. ICLR 2025

FROM CORRELATION TO MECHANISM

Enrichment shows what changed. Upstream analysis asks what drives it.

GO / PATHWAY OVER-REPRESENTATION

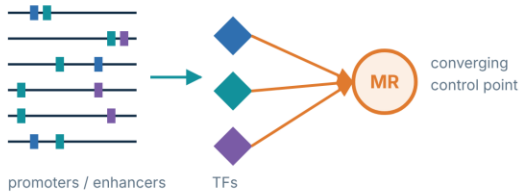
"Which processes are affected?"



Output
A list of affected processes – essential context.
Based on gene-set membership; promoter sequences and signalling direction are not part of the test.

UPSTREAM ANALYSIS

"What is driving it?"

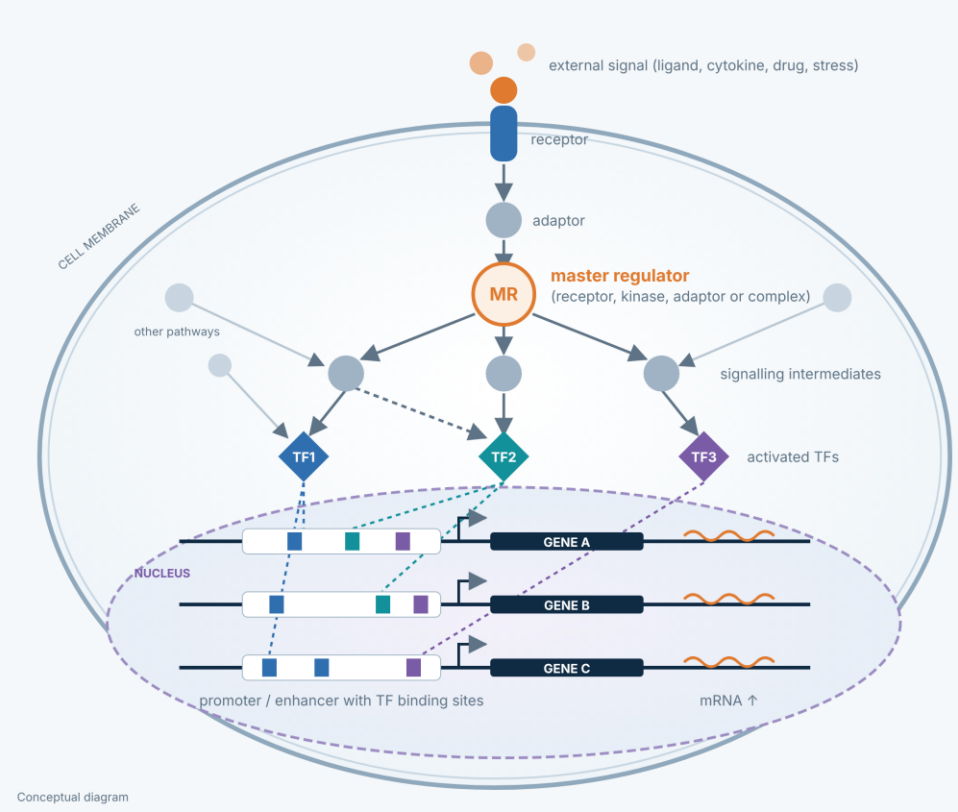


Output
A ranked list of molecules that could control the programme, with the network linking them to your genes – mechanistic hypotheses to test. The two are complementary.

	GO / pathway over-representation	Upstream analysis
Evidence used	gene-set membership	**TF sites + TF combinations + directed reactions**
Regulatory DNA (promoters, enhancers)	not used	**TRANSFAC matrices, MATCH, CMA**
Typical result	"inflammation is involved"	**"molecule X may drive the inflammatory genes"**
Best used for	describing the response	**choosing what to perturb next**

UPSTREAM ANALYSIS · THE IDEA

Work backwards, from changed genes to the controlling switch



RESULT

master regulators and interactions to perturb

STEP 2 · SIGNALING NETWORK

which upstream molecules reach many of these TFs?

STEP 1 · REGULATORY DNA

TF combinations in promoters and enhancers

START · YOUR DATA

changed genes: bulk, single-cell, proteomics

STEP 1 · REGULATORY DNA

Which transcription factors control the changed genes?

① MOTIF → SITES (MATCH)

Each TF is described by a weight matrix built from curated binding sites



TRANSFAC weight matrix (AP-1-like)



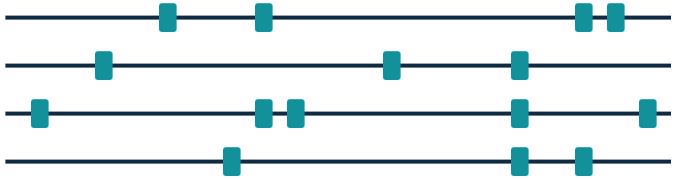
score 1.00 ✔ site score 0.44 ✘

Every promoter/enhancer window is scored; matrix-specific cut-offs keep false positives under control.

② ENRICHMENT VS BACKGROUND

Are these sites more frequent in your genes than in unchanged genes?

YOUR genes (YES set)



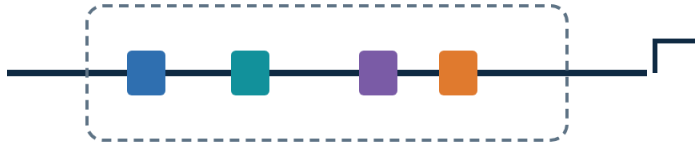
UNCHANGED genes (NO set)



YES/NO ratio and p-value for every matrix → ranked list of enriched TF motifs → the TFs that bind them

③ TF COMBINATIONS (CMA)

TFs work in teams: search for the combination that best separates YES/NO



composite module in a ~200-300 bp window

A genetic algorithm evolves combinations of matrices, cut-offs and distances; the best module describes the “promoter grammar” of your gene set.

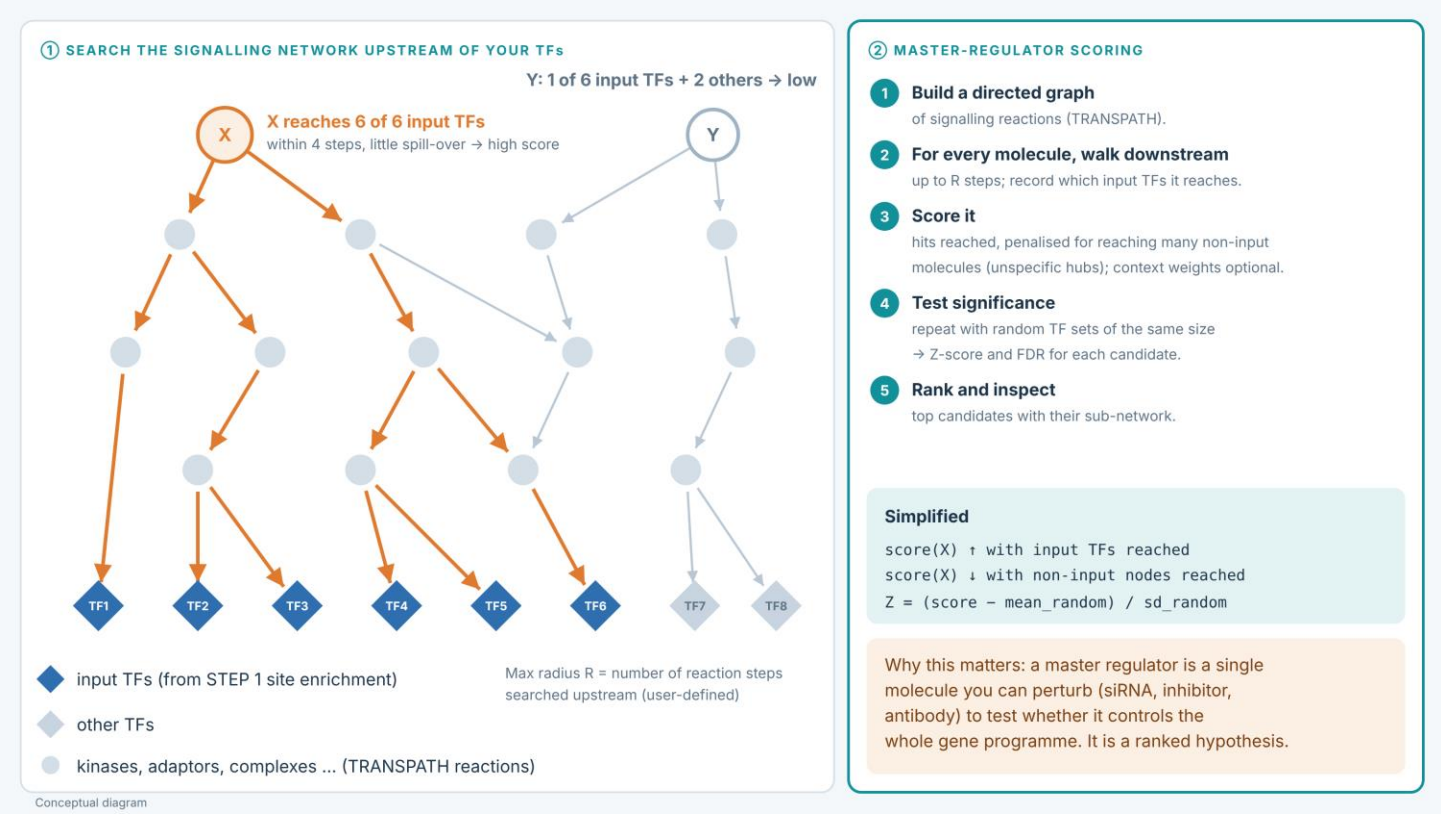
Output: candidate TFs + their binding sites in your genes → input for STEP 2

Conceptual diagram. Methods: MATCH (Kel et al., Nucleic Acids Res 2003); CMA (Waleev et al., Nucleic Acids Res 2006, Web Server issue). Detailed algorithms → appendix.

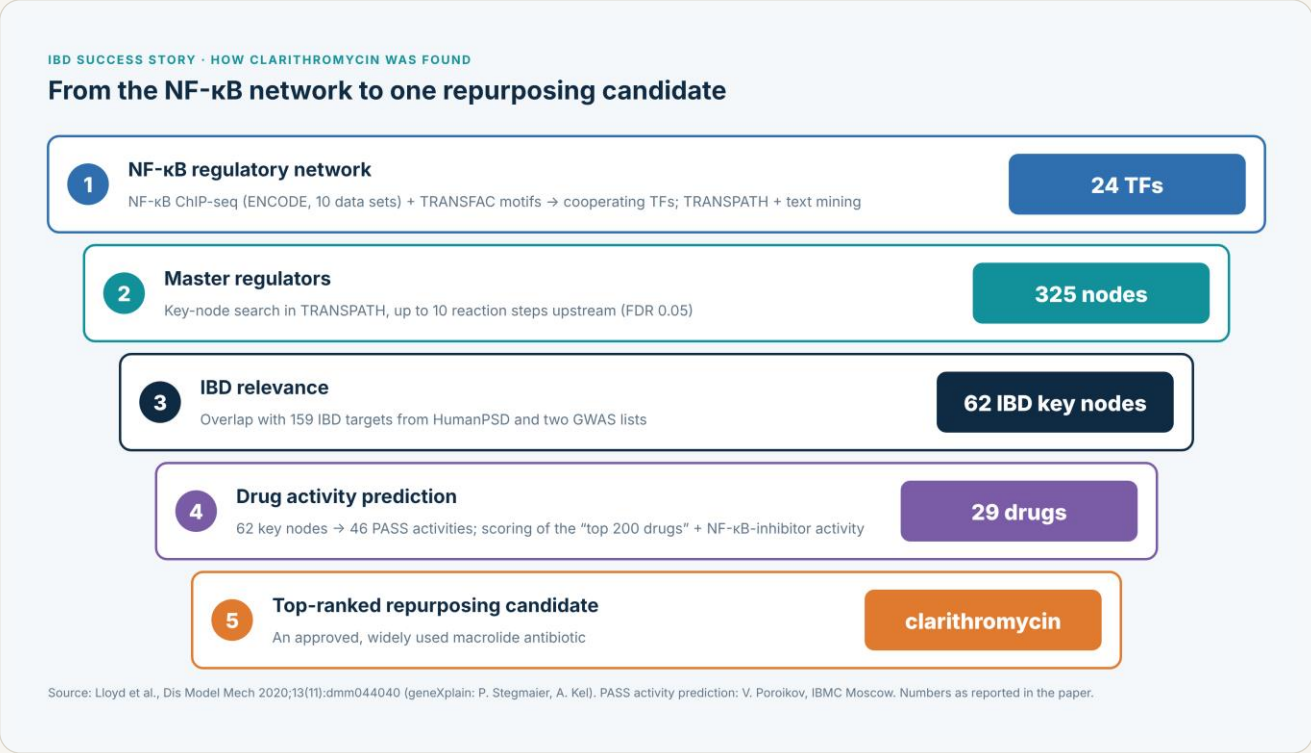
A predicted site is a sequence match, not measured TF binding; confidence grows when sites, combinations, expression and network evidence agree.

UPSTREAM ANALYSIS · STEP 2

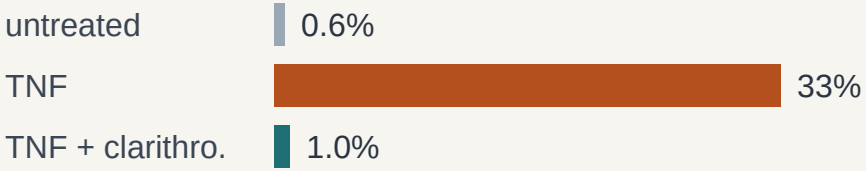
Which molecule upstream controls these transcription factors?



IBD: from the NF-κB network to a tested candidate



Human intestinal organoids, 6 donors: cells with active NF-κB



- less NF-κB activity in mouse macrophages and LPS-challenged mice
- colitis model: less weight loss, healthier gut lining

No deep learning. Preclinical only; antibacterial effect is a possible confounder.

Lloyd et al., Dis Model Mech 2020;13:dmm044040 (EU SysmedIBD). Drug activity prediction: PASS (V. Poroikov, IBMC Moscow).