12 February 2025 - AMLD 2025

Multi-megabase scale genome interpretation with genetic language models

Patrick Schwab, Senior Director and Head of Biomedical AI, GSK

Co-authors: Frederik Träuble, Lachlan Stuart, Andreas Georgiou, Pascal Notin (Harvard), Arash Mehrjou, Ron Schwessinger, Mathieu Chevalley, Kim Branson, Bernhard Schölkopf (Max Planck Institute), Cornelia van Duijn (Oxford), Debora Marks (Harvard)

GSK.ai Biomedical AI group

- AI for Health and Biology, Software, and Technology
- Based in Heidelberg/Germany, Zug/Switzerland and London/UK

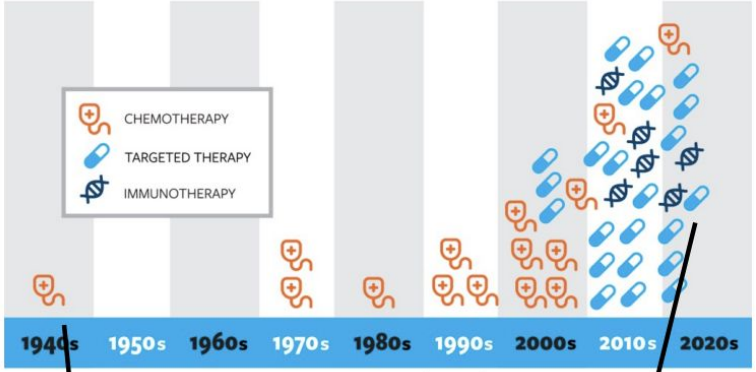
- We **help identify, monitor, and treat disease** with Clinical AI
- We **create a map of the immune system** using AI-guided experimentation
- We **advance the science of AI for Health** in partnership with leading European/Swiss institutions (e.g., ETH Zurich, Oxford, Cambridge, King's College)

the team

The good: drug development works for society

GSK

FDA approved therapies for lung cancer over time



Source: Lung Cancer Research Foundation

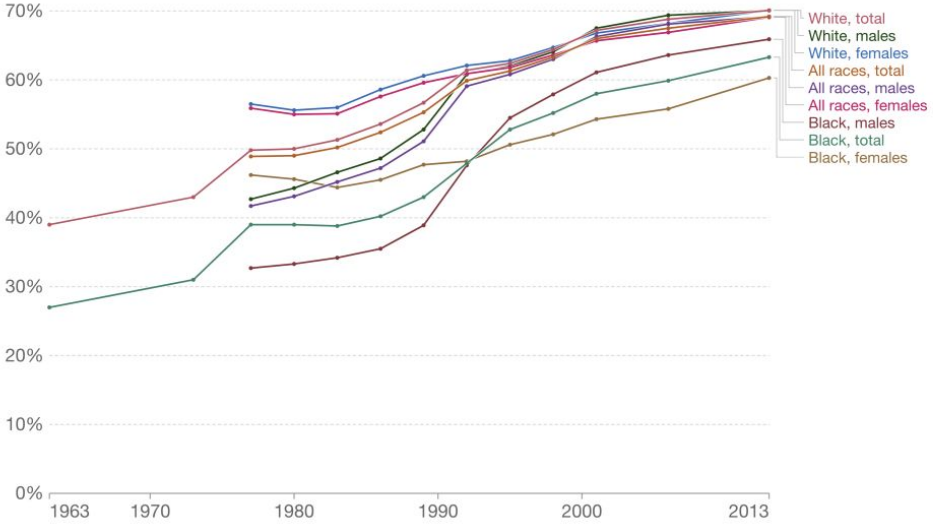
Mechlorethamine Hydrochloride
Candidate chemical warfare agent research

Mobocertinib
NSCLC with EGFR exon 20 insertion w/ progression after platinum therapy

Five-year cancer survival rates by sex and race, 1963 to 2013

Our World in Data

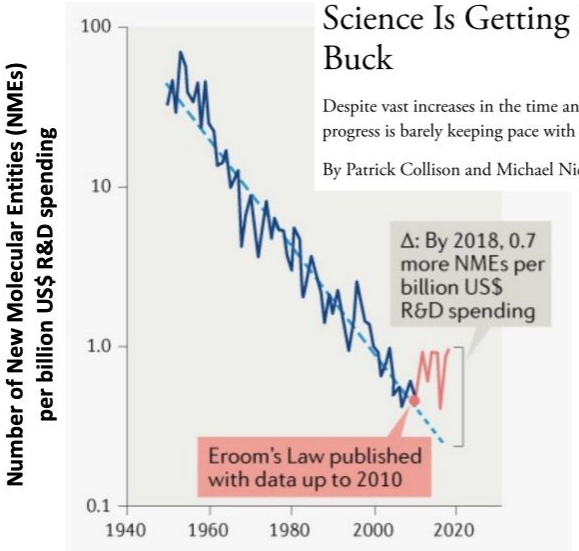
Percentage of cancer patients surviving at least five years following diagnosis of any cancer type. This is shown by sex and race in the United States.



Source: National Cancer Institute

Innovative medicines play a crucial role in enabling humans to live longer, healthier lives.

The bad: we are doing less, with (exponentially) more



SCIENCE

Science Is Getting Less Bang for Its Buck

Despite vast increases in the time and money spent on research, progress is barely keeping pace with the past. What went wrong?

By Patrick Collison and Michael Nielsen

"Eroom's law":
Exponential drop in R&D productivity.

- **Failure is the norm**: Probability of success for a new medicine is **~5.5%**
- **Median cost** of a new drug is **$1.1 billion**
- Trend stable for decades with recent reversal (potentially) due to **increasing personalization** and **molecular/genetic evidence**.

Ringel, Michael S., et al. "Breaking Eroom's Law." Nature reviews. Drug discovery (2020).
Wouters et al. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018". JAMA (2020)
Ringel et al. "Breaking Eroom's law." Nature reviews Drug Discovery (2020)

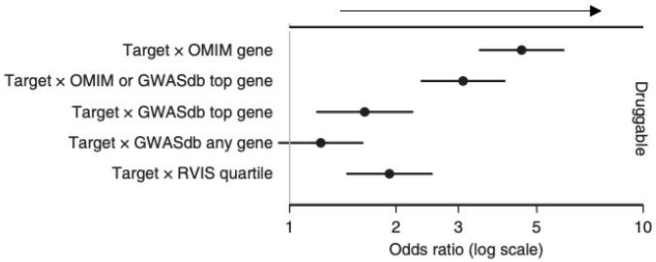
What do we know works?

1

Population-scale Biobanks – Variation in Genetic Background (>100'000 people)

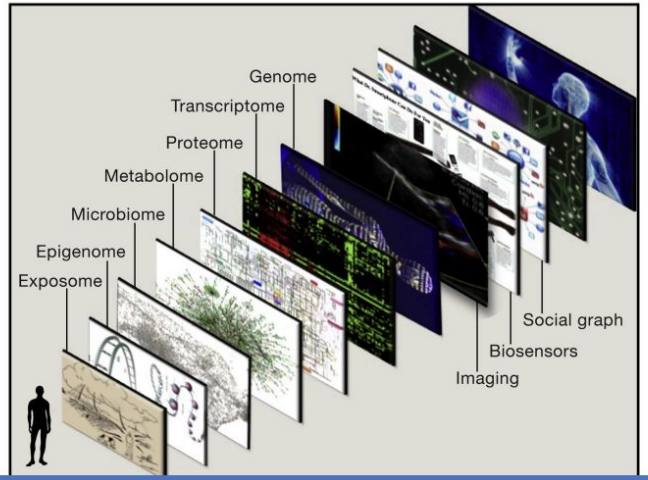
FINNGEN biobank uk All of Us RESEARCH PROGRAM

Higher probability of approval (EU/US)



2

Targeted Biomarkers supporting Therapy
2x (immune) to 8x (oncology) higher probability of trial success



A causal & targeted link between disease of interest and a molecular mechanism substantially improves our chances of successful translation.

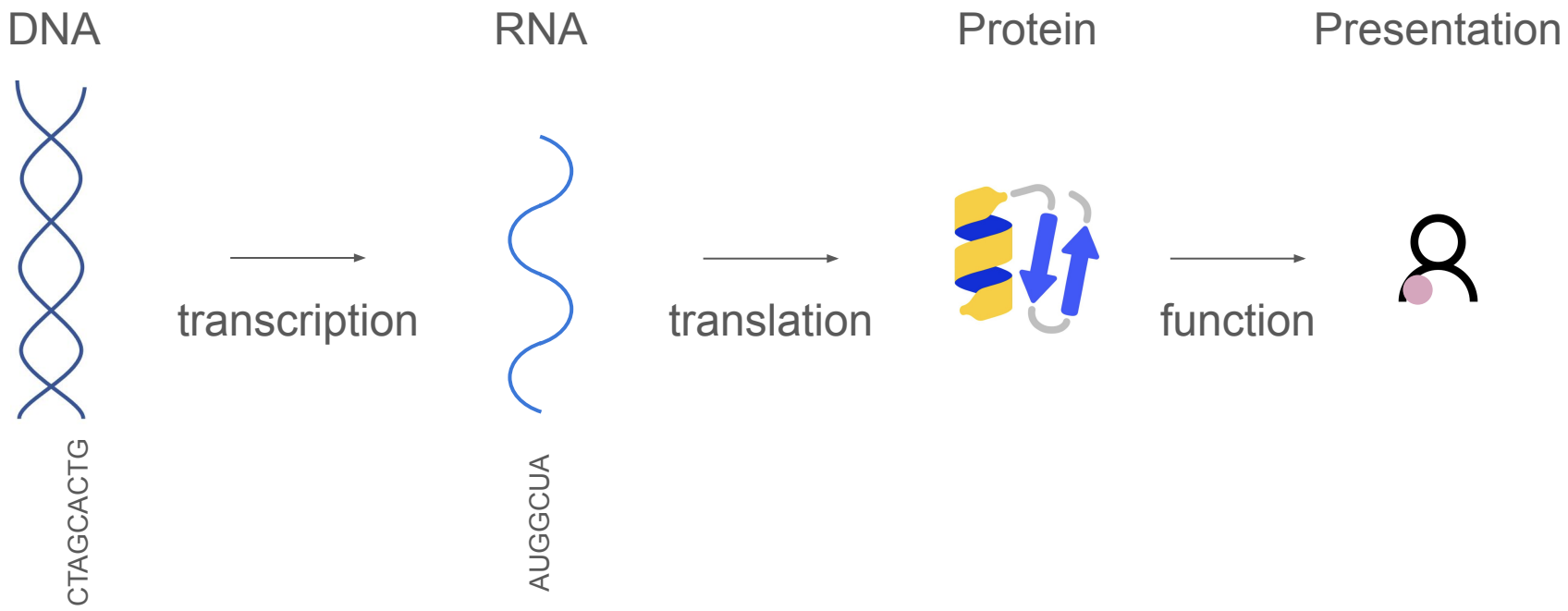
GSK

Wong et al. "Estimation of clinical trial success rates and related parameters " Biostatistics (2018)
Topol, Eric J. "Individualized medicine from prewomb to tomb." Cell 157.1 (2014): 241-253.
Nelson, Matthew R., et al. "The support of human genetic evidence for approved drug indications." Nature genetics 47.8 (2015): 856-860.
Ringel, Michael S., et al. "Breaking Eroom's Law." Nature reviews. Drug discovery (2020).

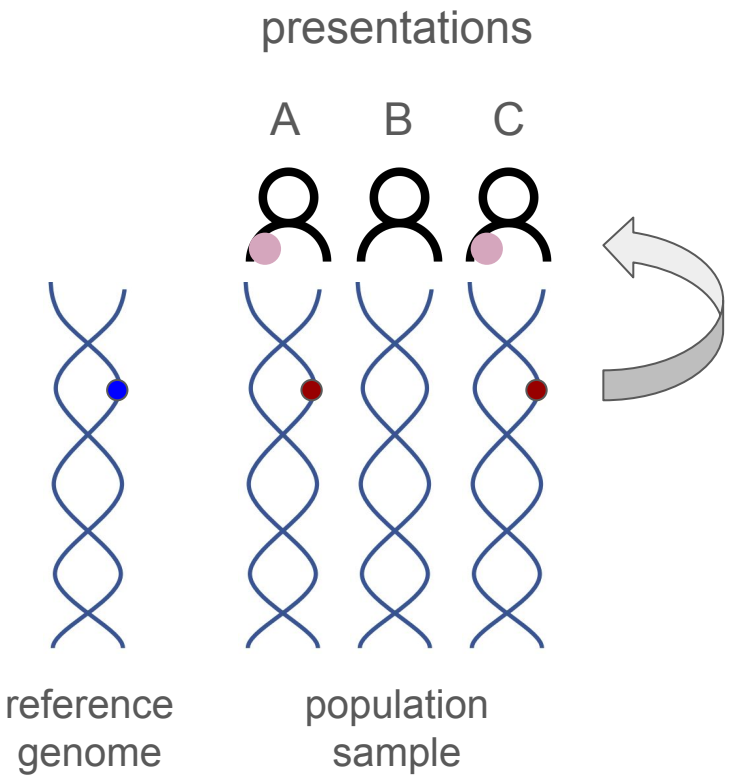
What if we could predict whole-organism mechanisms based on individual biological backgrounds?

Towards a generative model of biomedical mechanisms

The central dogma - (simplified) information flow in biology



Natural experiments offer unique view on mechanisms



Single Nucleotide Polymorphism (SNP)
Natural variation in a single location connects presentation to a potential source

GSK

… CTAGCACTGCTAGCACTGCTAGCACTG  CTAGCACTGCTAGCACTGCTAGCACTG …

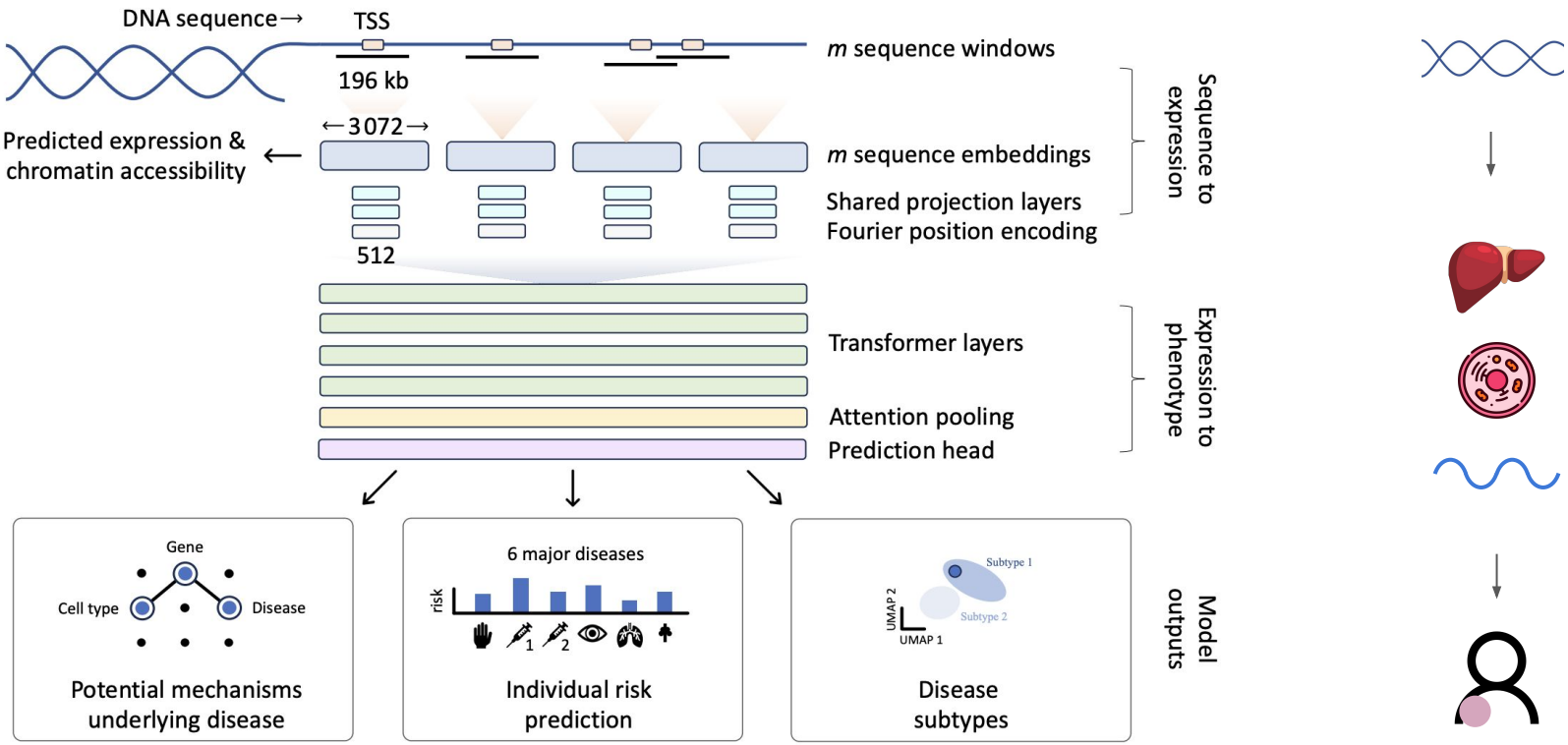
… however, the human genome spans <u>3 billion</u> base pairs …

If your genome were a library, it would encompass

📚📚 **6’000 books with 250 pages each.**

a single letter misses the whole (*your!*) story

Phenformer - a multi-scale genetic language model to read whole genomes

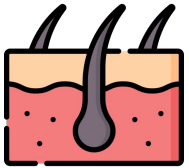


Träuble et al. “Multi-megabase scale genome interpretation with genetic language models.” arXiv preprint (2025). URL https://arxiv.org/abs/2501.07737

Phenformer reads genome to explain <u>how</u> disease comes to be

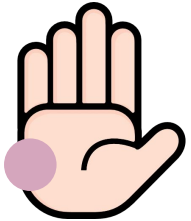
Phenformer

=>



+

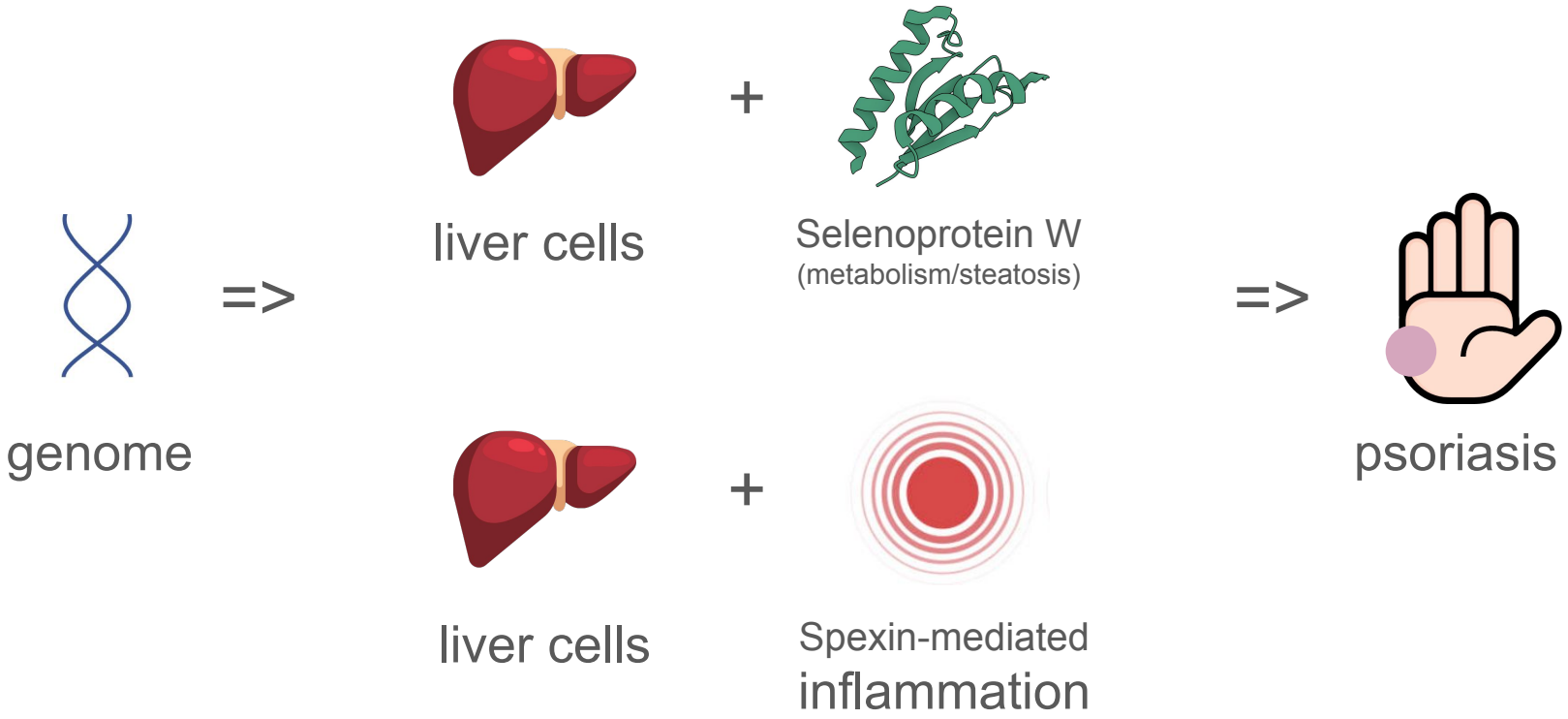
=>



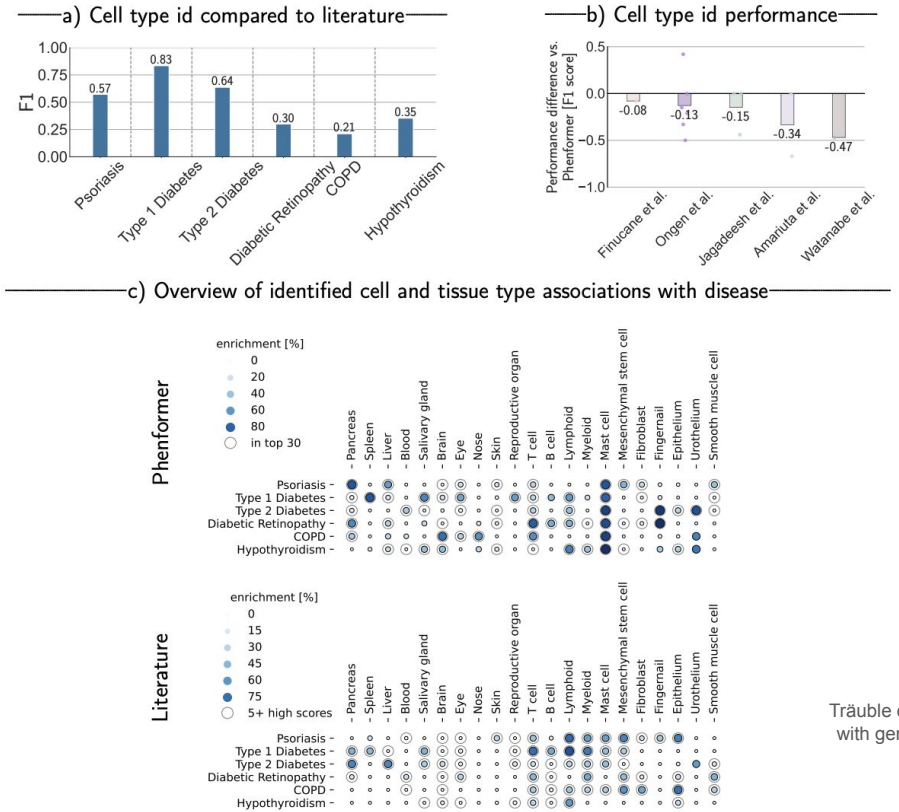
genome skin cells inflammation psoriasis

Incredibly, Phenformer explains presentations we do not fully understand

Clinically, psoriasis patients face 3x risk of fatty liver disease. Not known why.

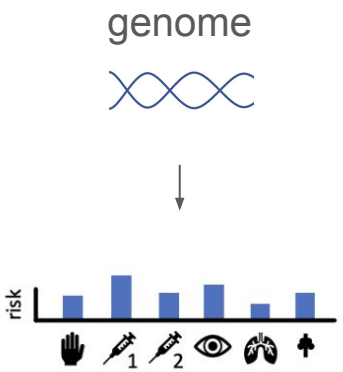


Benchmarking cell & tissue ID demonstrates leading performance compared to SOTA methods vs. literature-reported associations

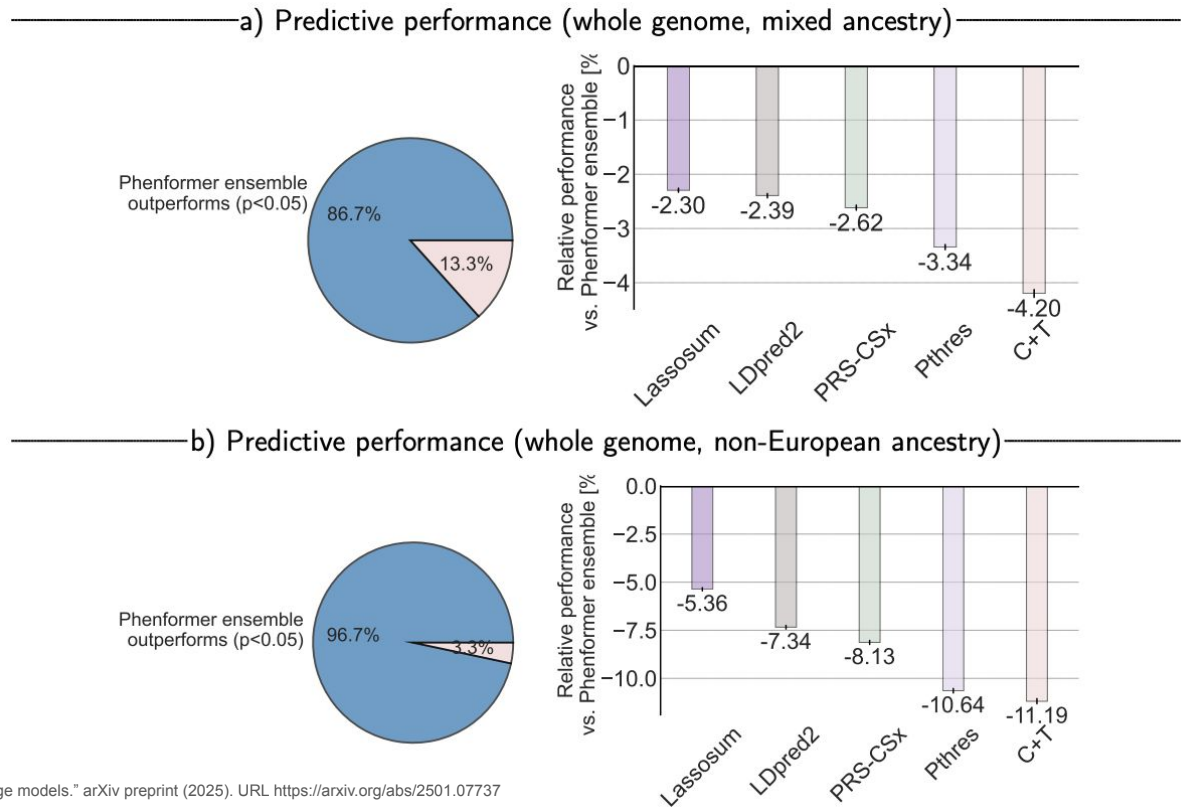


Träuble et al. “Multi-megabase scale genome interpretation with genetic language models.” arXiv preprint (2025). URL https://arxiv.org/abs/2501.07737

Using Phenformer as a personalised, whole-genome risk predictor considerably enhances predictive accuracy & generalisability

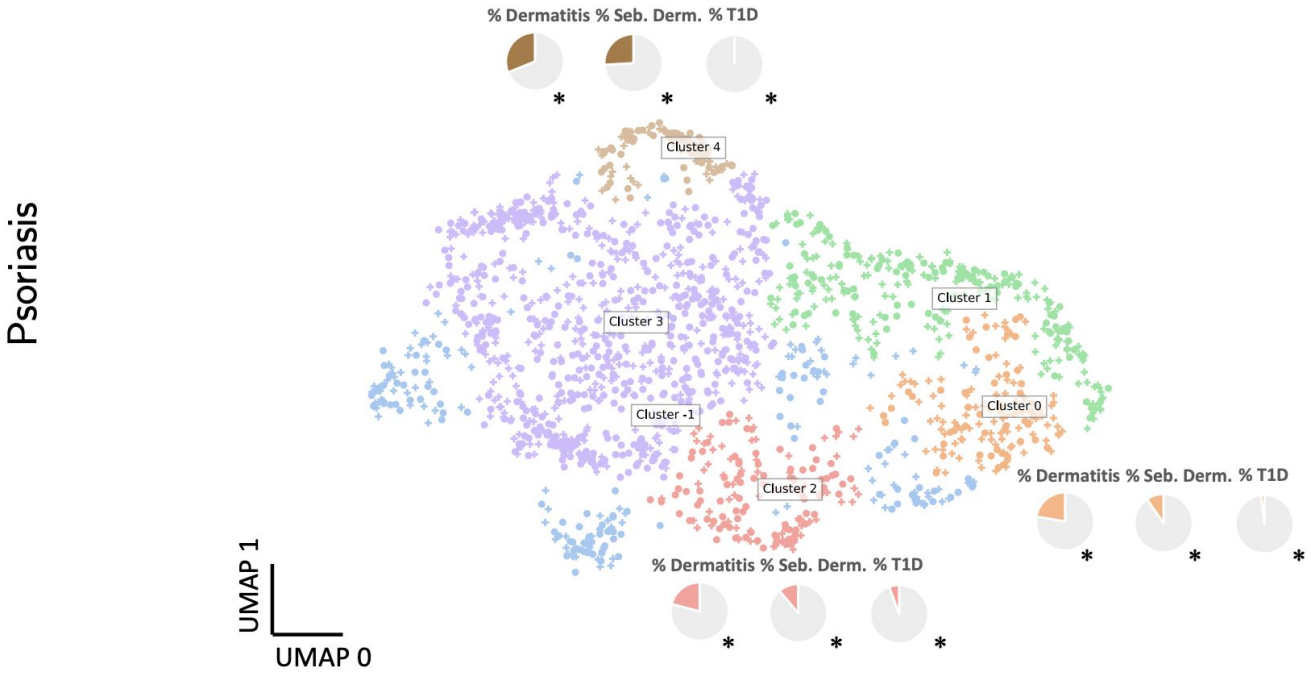


Individual risk prediction



Träuble et al. “Multi-megabase scale genome interpretation with genetic language models.” arXiv preprint (2025). URL https://arxiv.org/abs/2501.07737

When embedding whole genomes using Phenformer, we obtain a map of mechanistic subtypes with putatively* different underlying mechanisms



* as evidenced by significant ($p<0.05$) differences in comorbidity rates

Träuble et al. "Multi-megabase scale genome interpretation with genetic language models." arXiv preprint (2025). URL https://arxiv.org/abs/2501.07737

Summary

- Emerging class of *generative models of mechanisms* hold potential to enable understanding of *how* diseases come to be
- Such models are capable of interpreting the vast biological background of single individuals
- Using large compute (18 GPU months), Phenformer processes 3 billion base pairs and predicts for 6’000+ tissues and cell types
- Individualised understanding of *what* and *how* biological change happens could enable *hyper-targeted therapy* in the future

The coming era of predictive medicine

Frederik Träuble, Lachlan Stuart, Andreas Georgiou, Pascal Notin (Harvard), Arash Mehrjou, Ron Schwessinger, Mathieu Chevalley, Kim Branson, Bernhard Schölkopf (Max Planck Institute), Cornelia van Duijn (Oxford), Debora Marks (Harvard)

"Multi-megabase scale genome interpretation with genetic language models."
available at: **https://arxiv.org/abs/2501.07737**

Patrick Schwab
patrick.x.schwab@gsk.com
@schwabpa

We are hiring: software engineers, AI/ML scientists, internships, …

the team

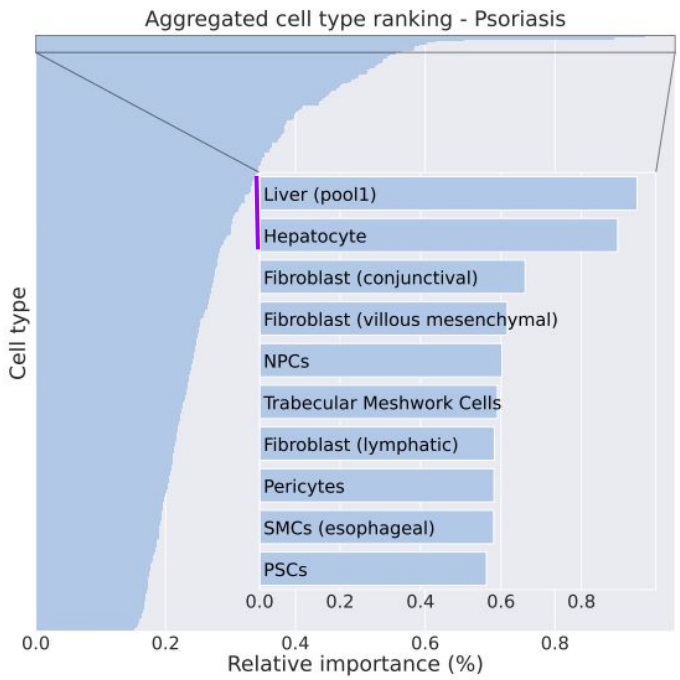
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

HARVARD MEDICAL SCHOOL

UNIVERSITY OF OXFORD

GSK

Following the <sequence-cell-expression-phenotype> paths enables generation of potential disease-associated mechanisms



- We find many **expected results** - e.g. fibroblasts/psoriasis
- ... but, strikingly, Phenformer also **substantiates many yet-unexplained clinical findings**
  - e.g. **liver-involvement in psoriasis** which to-date lacks mechanistic understanding (psoriasis pts are 2x higher risk for liver disease)
  - ... & many more surprising findings (optic nerve in COPD, appendix/nails in T1D)!

On same test data, Phenformer significantly outperforms existing PRS methods for risk prediction, while maintaining transportability