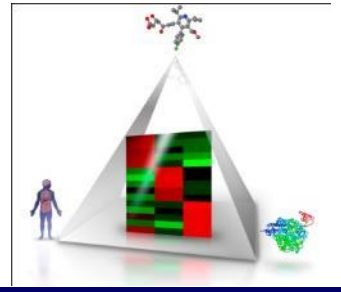
# Artificial Intelligence in Drug Discovery – Some Aspects of What (Somewhat) Works, Some Aspects Why it's Difficult

Andreas Bender, PhD
Natural Philosopher for Molecular Informatics
Department of Chemistry, University of Cambridge

Chief Scientific Officer, Terra Lumina, Berlin

Co-Founder of Healx, Ltd. and PharmEnable, Ltd.

UNIVERSITY OF CAMBRIDGE



Any statements made during this talk are
in my capacity as an academic

Further reading: Artificial Intelligence in Drug Discovery – What is Realistic, What are Illusions? (Parts 1 and 2)

Andreas Bender and Isidro Cortes-Ciriano

*Drug Discovery Today* 2021

# Contents

1. Current state of AI in drug discovery

2. Some examples of what (somewhat) works in early drug discovery

3. The Achilles heel of AI in drug discovery: *data*

4. Psychology, the hype cycle & the translational gap of methods

# 1. Current state: The 3rd wave of computers in drug discovery (80s, 2000, today) – time for realistic assessment has come

Fortune cover 1981

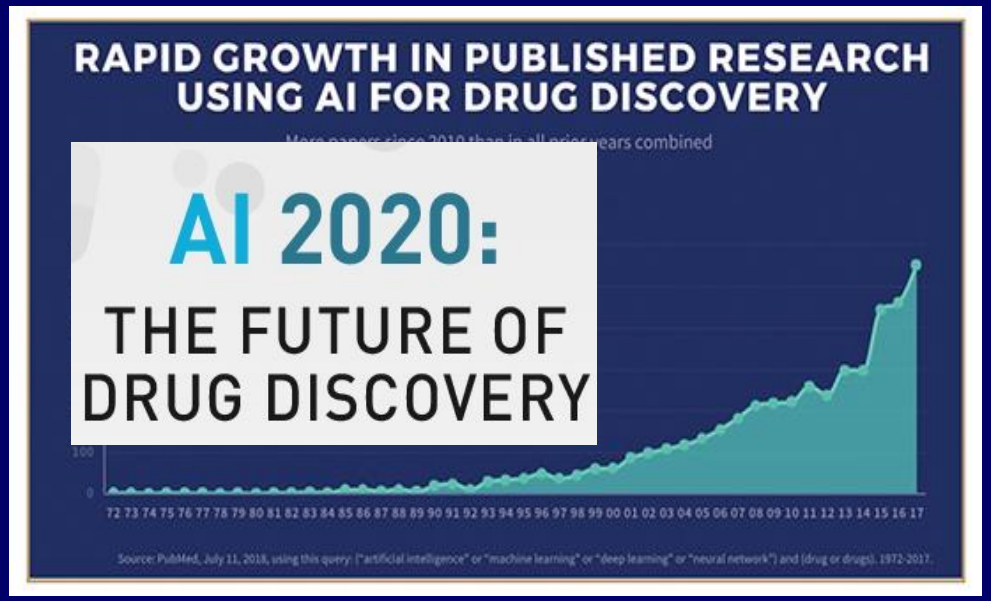

Recent headlines (2018-2020)

SPOTLIGHT · 30 MAY 2018

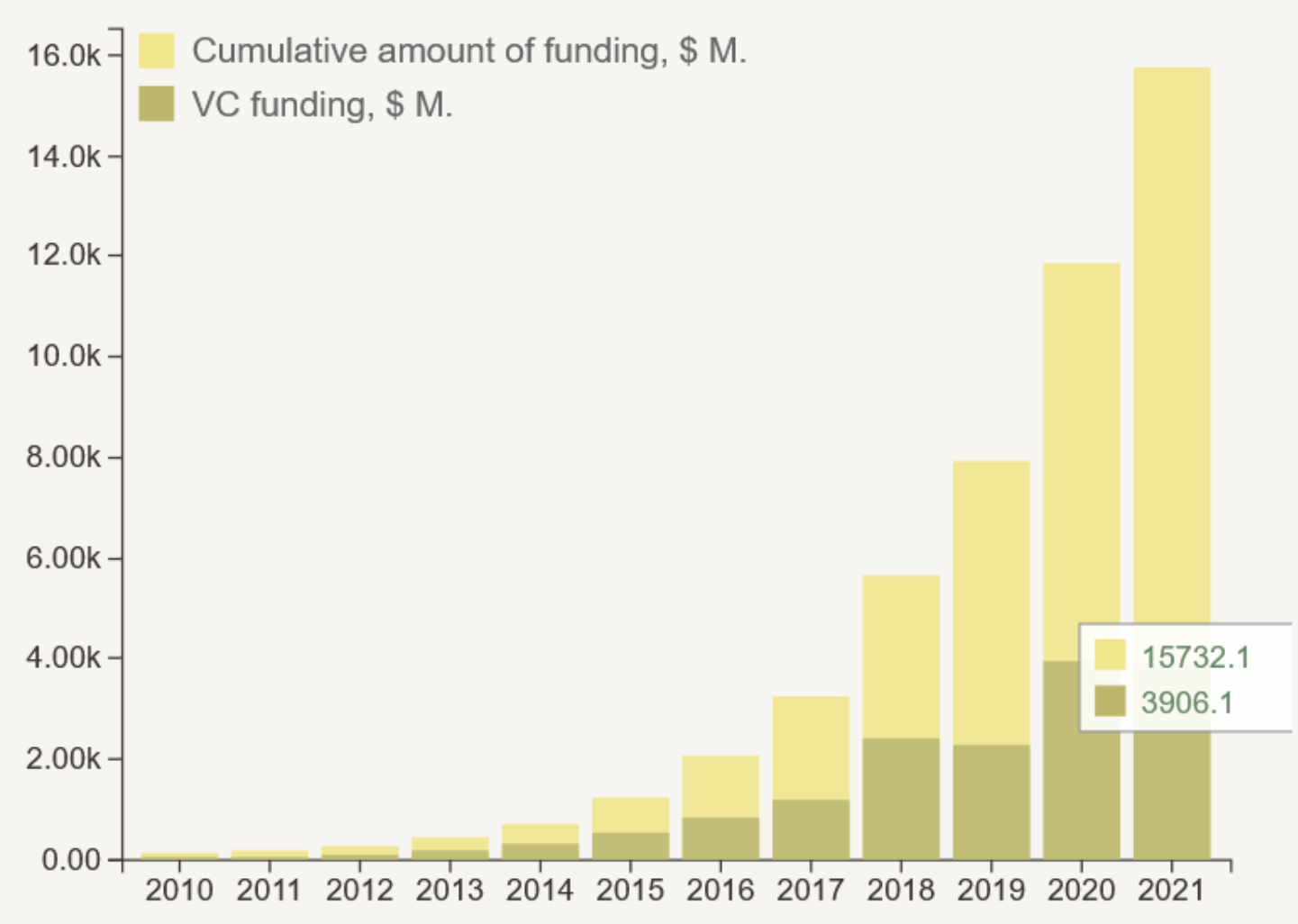
**How artificial intelligence is changing drug discovery**

World first breakthrough in AI drug discovery

By **Emma Morriss** - January 30, 2020



# Funding going into AI in drug discovery until 2021: ~$4bn VC funding, $16bn total (*very approx.)*



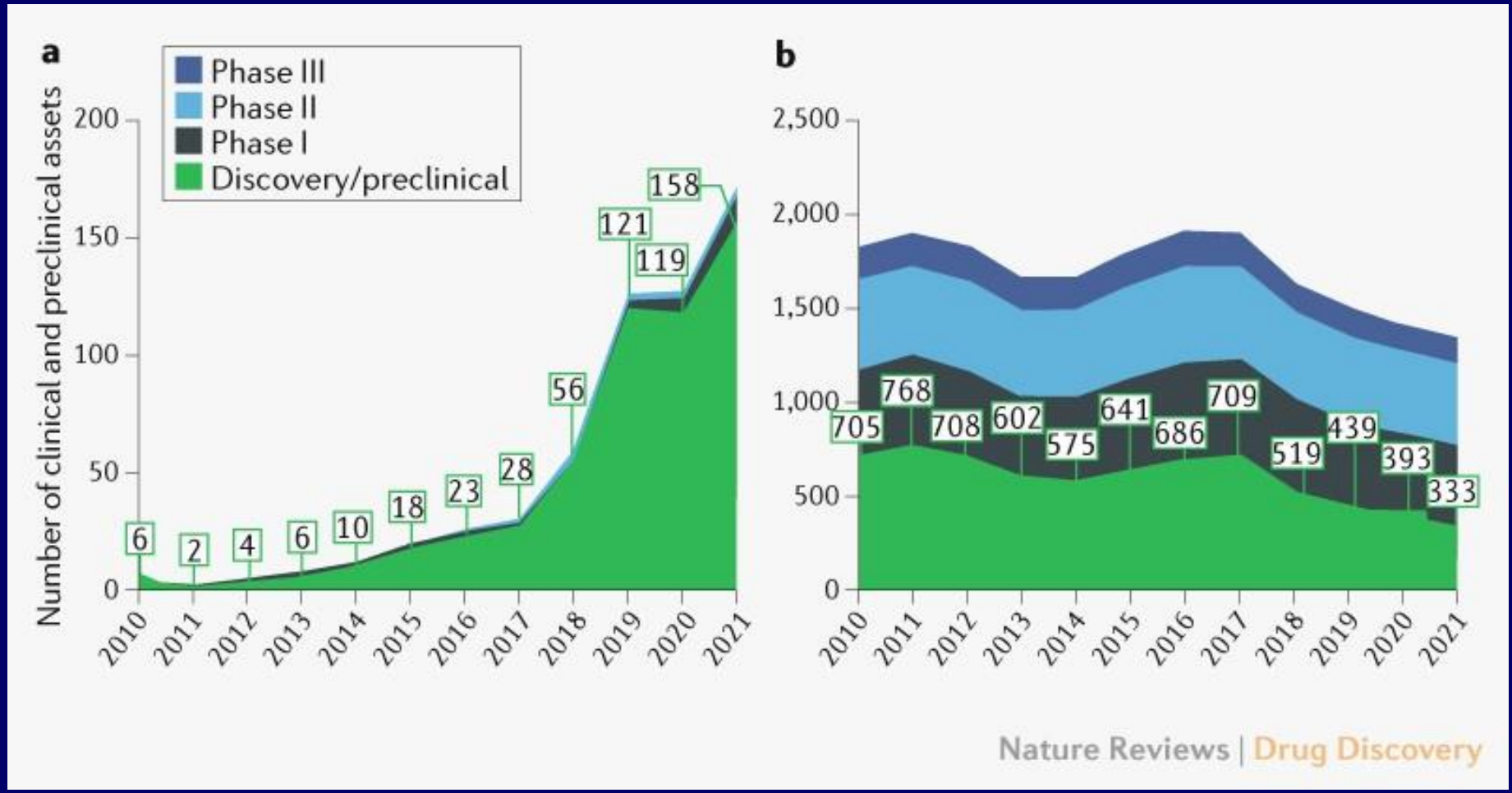

https://www.biopharmatrend.com/post/397-pharmaceutical-artificial-intelligence-in-2021-key-developments-so-far/

# Current discovery pipeline: AI-based start-ups vs big pharma

‘AI-native companies’ Top 20 pharma

Significant *number* of *discovery/ preclinical* programs of AI companies (~160 vs ~330)

Very little Phase 1, less Phase 2, no Phase 3



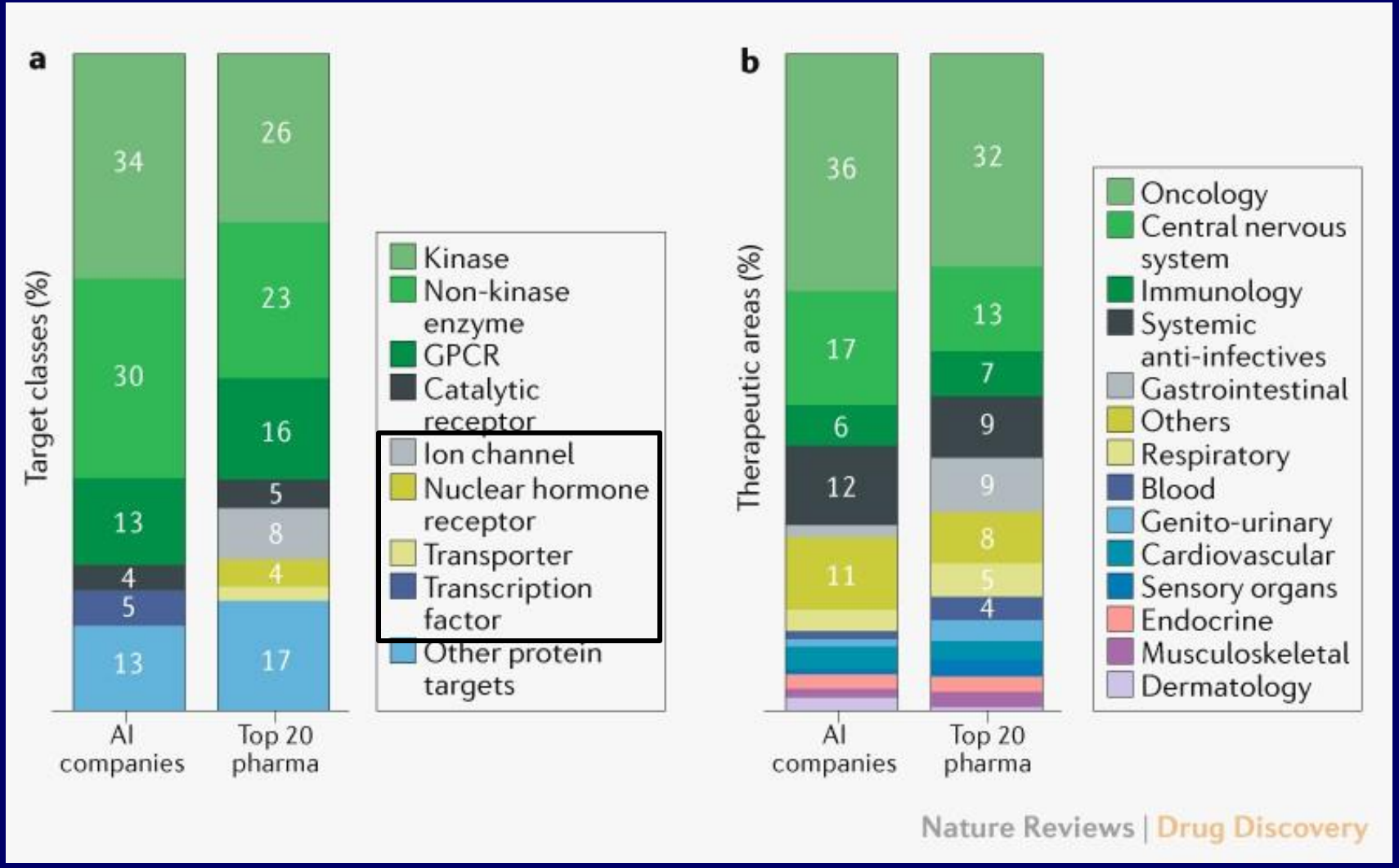

-> Little *in vivo* safety (Phase 1) data *yet*; virtually no *in vivo* efficacy (Phase 2/3) data *yet*

Jayatunga et al., AI in small-molecule drug discovery: a coming wave? *Nature Reviews Drug Discovery* 7 Feb 2022

# Distribution of target profile similar, but focus on areas of more data, less complex target pharmacology

More kinases and enzymes in AI-driven companies:
(a) Quite *data-rich*
(b) Less complex pharmacology than other target classes

+ Transcription factors
- No ion channels, NHRs and transporters




Jayatunga et al., AI in small-molecule drug discovery: a coming wave? *Nature Reviews Drug Discovery* 7 Feb 2022

# Conclusion about the world as it is

- Lots of activity in early stage pipeline of AI-first companies, but often already explored targets, close analogues
- Appropriate question to ask: Where is the novelty?
- Data is often limiting factor – in both chemical and target space (leads to work on well-explored targets, with more data, less complex pharmacology)

- Is *input* (e.g. funding) success, or *output*?
- The first ‘AI-designed drug’ will be celebrated by the media, but…

… tens of billions went into funding AI in drug discovery, so even the null model would lead to an expected tens of approved drugs

# 2. Examples of What (Somewhat) Works

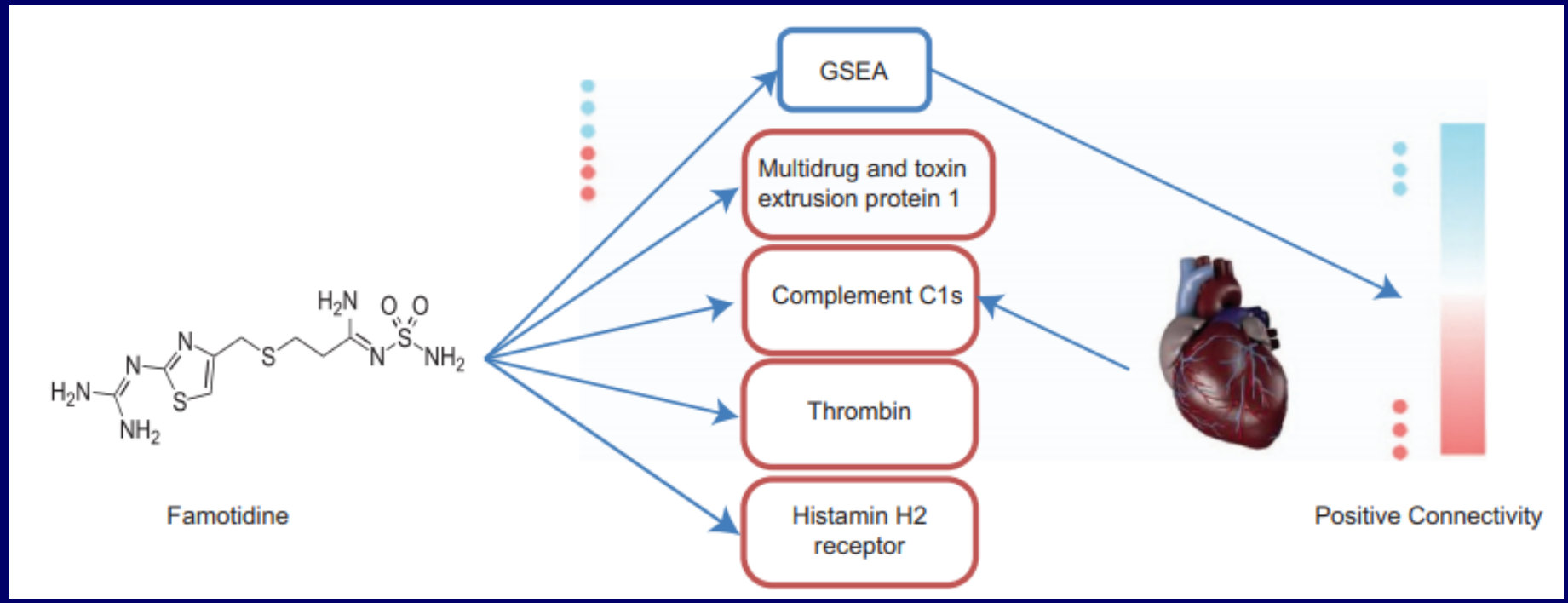
A) Gene expression-based compound selection (Repurposing/personalized medicine)

B) Generative models – ‘the computer suggests compounds’

C) Using cell morphology data for PROTAC safety assessment

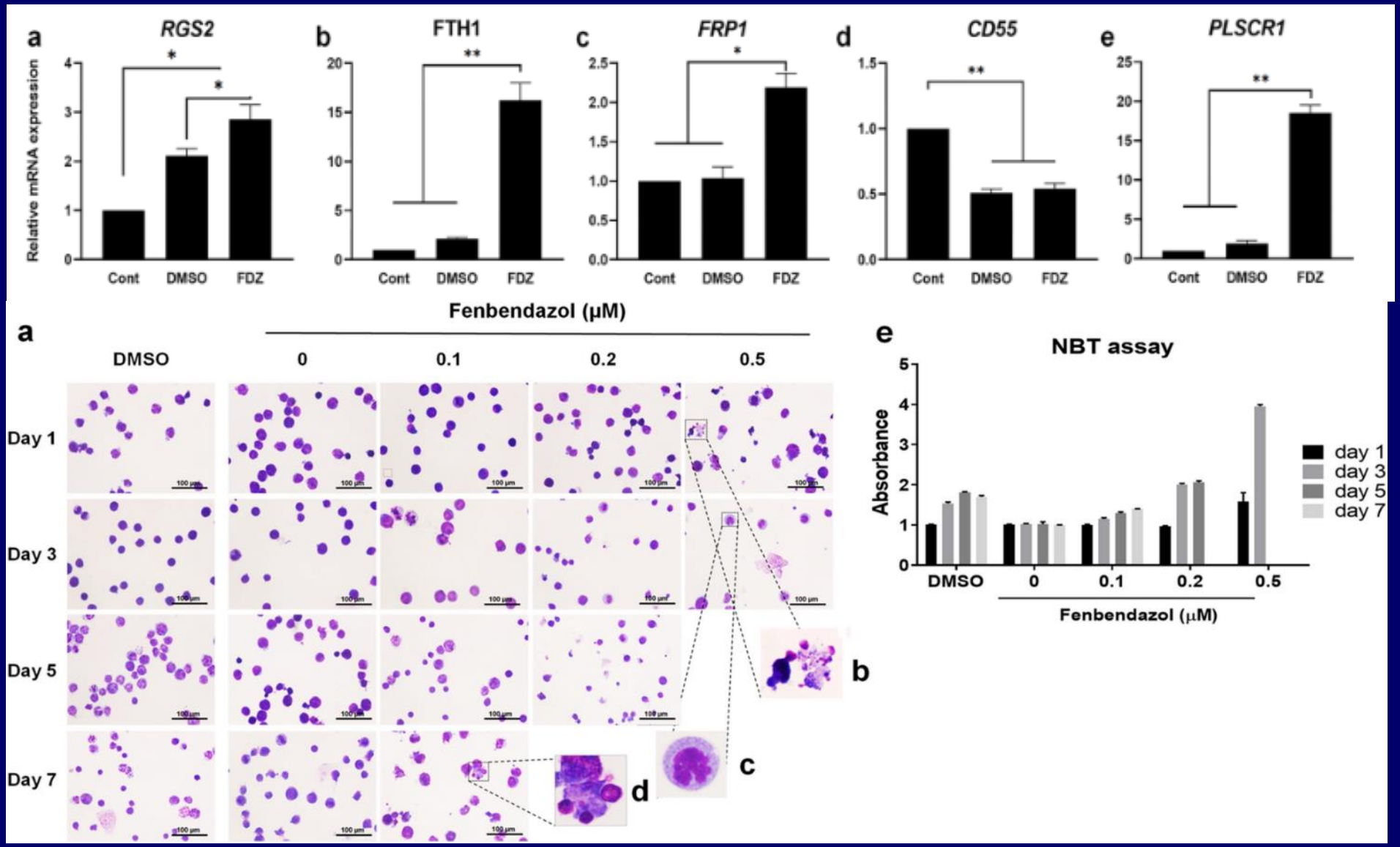
# A) Transcriptional compound selection for differentiation therapy in AML

- In acute myeloid leukemia (AML) cells, differentiation is blocked in cellular maturation stage
- Assumption: We can select small molecules to overcome differentiation block
- Method: Expression-data based matching of compounds to disease/patient




Y KalantarMotamedi, *et al.* Sci Rep (2021) 11, 12537

# (a) Mechanistic analysis matched gene expression hypothesis; (b) differentiation to granulocytes observed

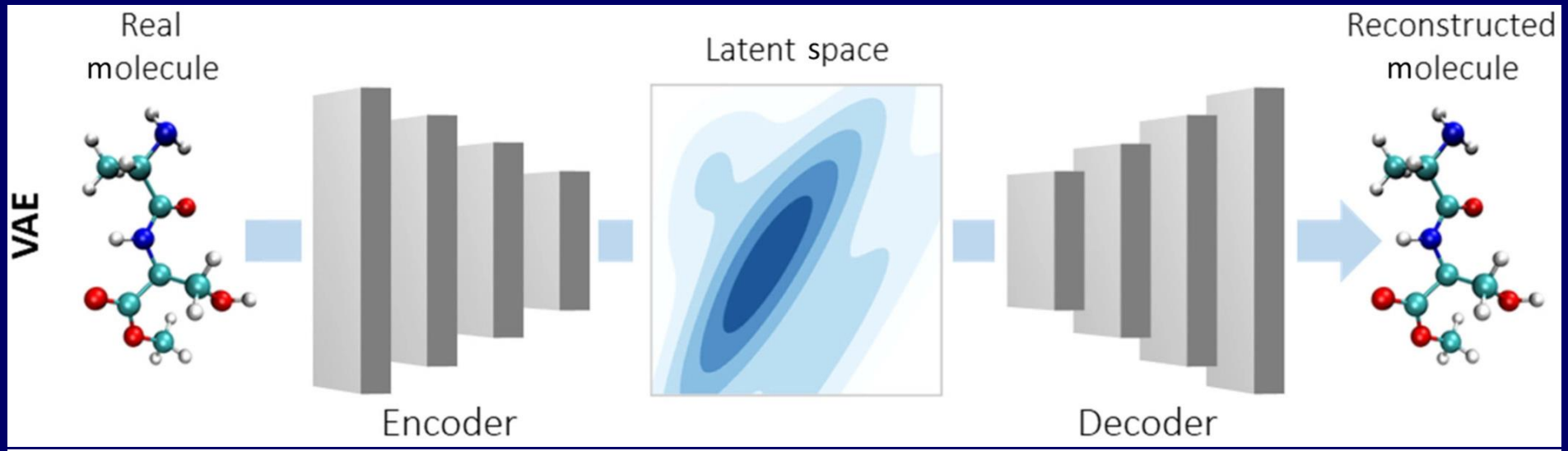


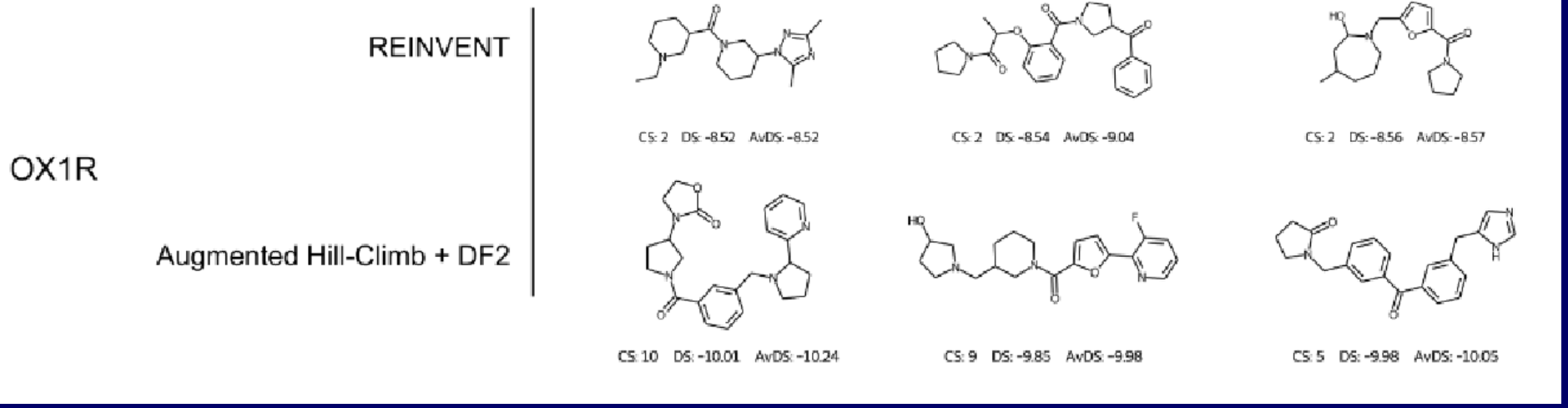

healx

- Based on general idea (plus NLP, further developments) Healx.io was established in 2014
- Drug repurposing for rare diseases
- $50m Series B funding 2019; ~120 staff

# B) Generative Models for *De Novo* Design

Work by Morgan Thomas, with SoseiHeptares




Based on known structures and data distributions, the algorithm (e.g. neural network) devises new structures

Morgan Thomas *et al.* Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. https://doi.org/10.26434/chemrxiv-2022-prz2r

# Brief comment on language models (here GPT-3) and whether they are sentinent

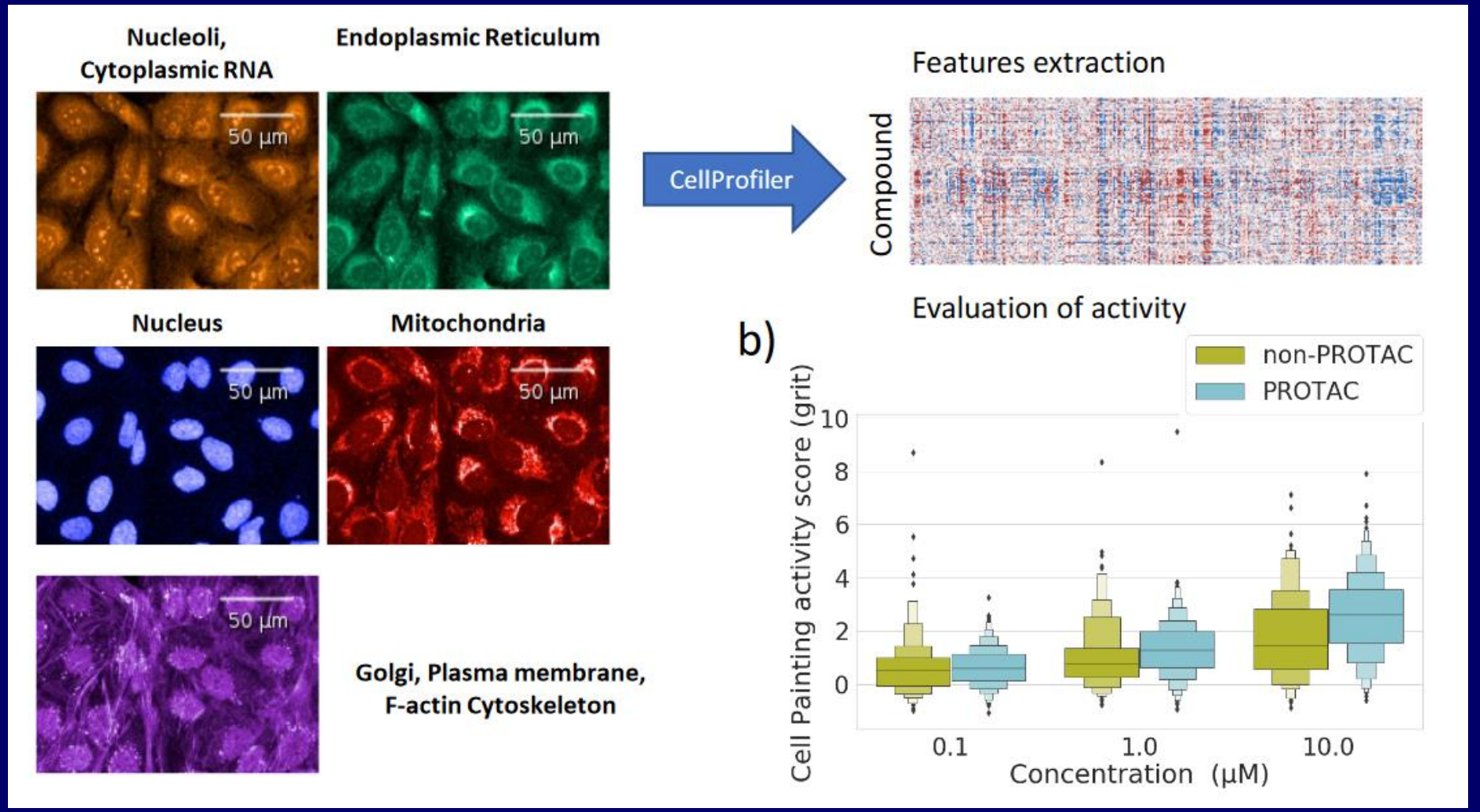
- Q: What's the world record for walking across the English Channel?
- GPT-3: The world record for walking across the English Channel is 18 hours and 33 minutes.

- Q: When was the Golden Gate Bridge transported for the second time across Egypt?
- GPT-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

- Q: Why does President Obama not have a prime number of friends?
- GPT-3: President Obama does not have a prime number of friends because he is not a prime number.

Hofstadter, The Economist, 9 June 2022

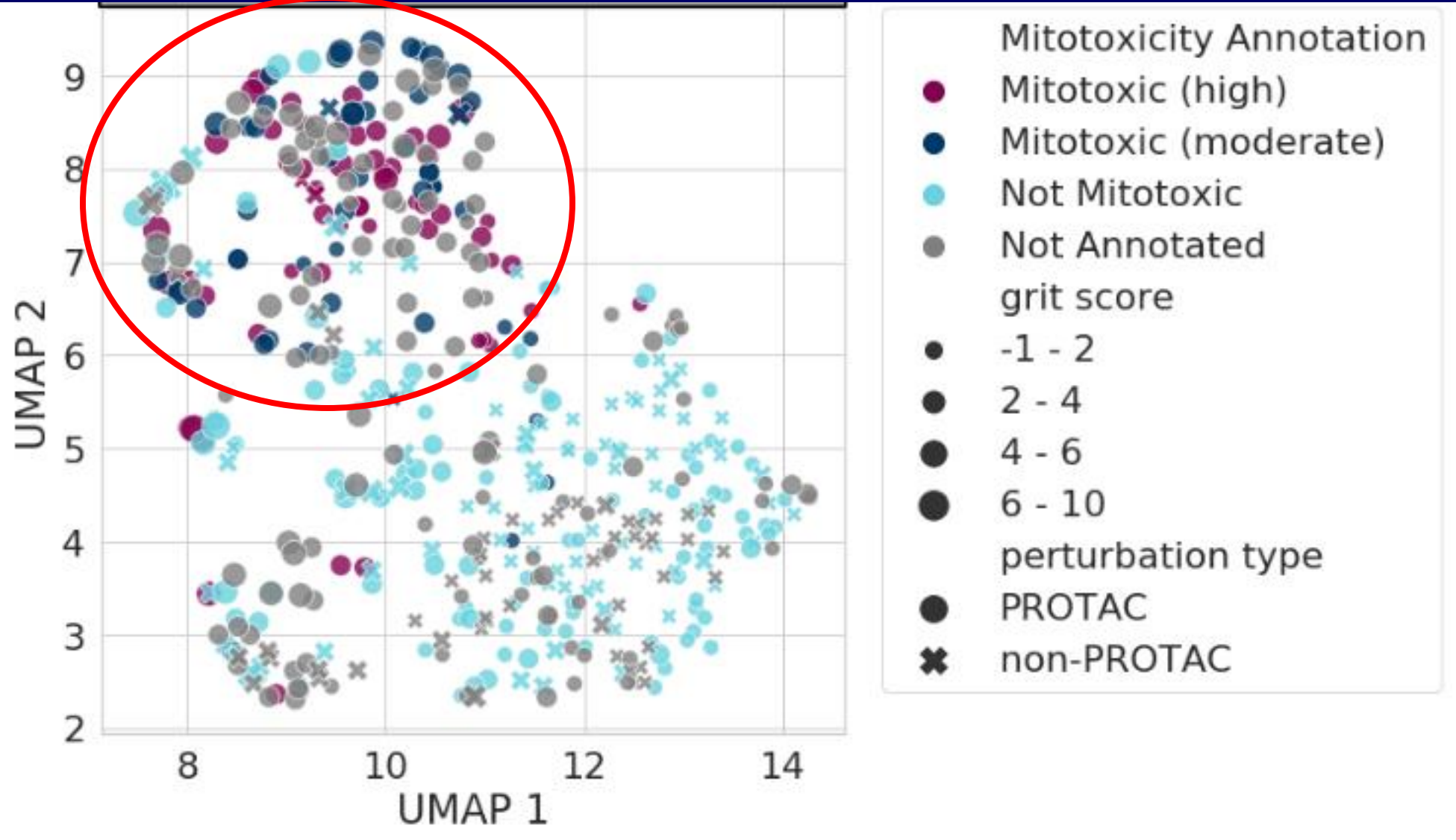
# C) PROTAC Mitochondrial Safety Assessment Using Cell Profiling (Morphology) Data




Work of Marianna Trapotsi, Srijit Seal, with AstraZeneca

Cell morphological profiling enables high-throughput screening for PROteolysis TArgeting Chimera (PROTAC) phenotypic signature, https://doi.org/10.1101/2022.01.17.476610

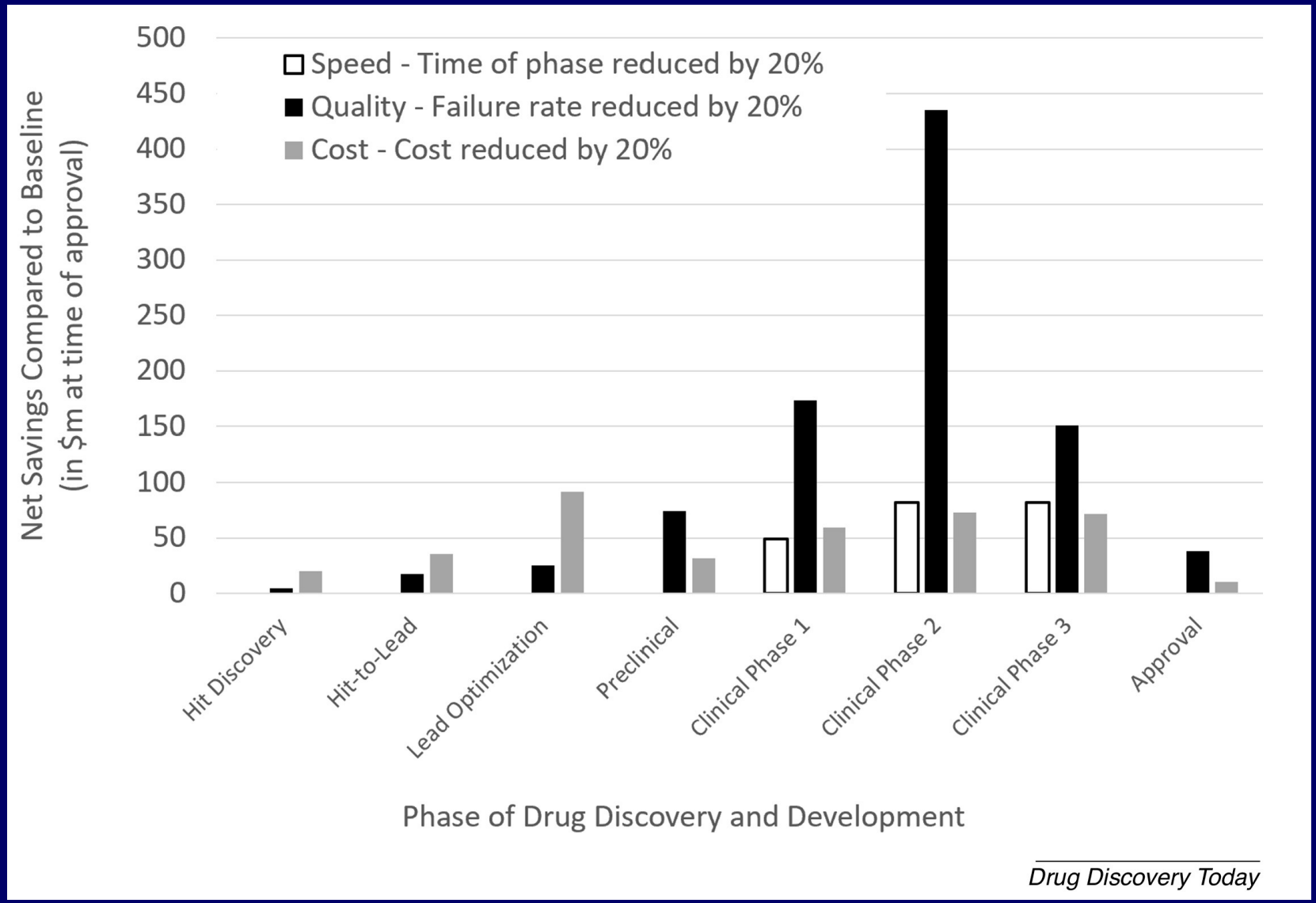
# Mitotoxic and non-mitotoxic PROTACs separated; useful as early ‘red flag’ in discovery



# 3. The Achilles heels of Artificial Intelligence in Drug Discovery: *conditional* and *proxy* data

*"…it's the data, stupid!"*

# The *quality* of *in vivo-relevant* decisions matters more than *speed* and *cost*!




1. We need to increase success rates *in the clinic* to have maximum impact with AI (left)

2. The data we have is (largely) unsuitable for this purpose proxy (next slides)

Bender and Cortes, Drug Discovery Today 2021

# Problem 1: Much of the data we have has been generated with proxy assays, is only marginally predictive for *in vivo* situation

- There is *what we are really interested in* - say, mitochondrial safety, Drug-Induced Liver Injury (DILI), …
- And there is what we *measure as an assay endpoint* – say, cytotoxicity in a Glu/Gal (differential cytotoxicity) assay to *approximate* mitochondrial safety; Bile Salt Export Pump (BSEP) inhibition to *approximate* DILI, …

- Take-away: 'Proxy' assays measure only part of reality, in a particular assay, with particular conditions
- Not to be confused with property itself!!!
- Problem: Proxy endpoint (a) taken as 'ground truth' in AI in drug discovery, (b) embedding into project context neglected

# Problem 2: *In vivo* data is (a) conditional. We have (b) too *little* of it and we (c) cannot label it properly

- ***Example: "Does* drug Y cause adverse reaction Z? Yes, or no?"**
- Pharmacovigilance Department: Yes, *if* we have…
  - A patient with this *genotype* (which is generally unknown)
  - Who has this *disease endotype* (which is often insufficiently defined)
  - Who takes *dose X* of *drug Y* (but sometimes also forgets to take it)
  - Then we see *adverse reaction (effect) Z* …
  - But only in x*% of all cases* and
  - With *different severity* and
  - *Mostly if co-administered with a drug from class C*, and then
  - More frequently in *males* and
  - Only *long-term*
  - (Etc.)

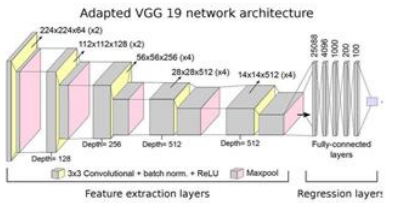
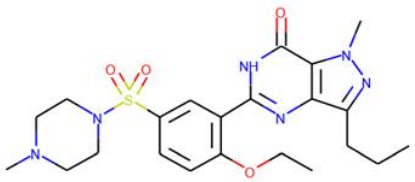
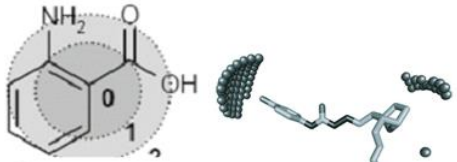
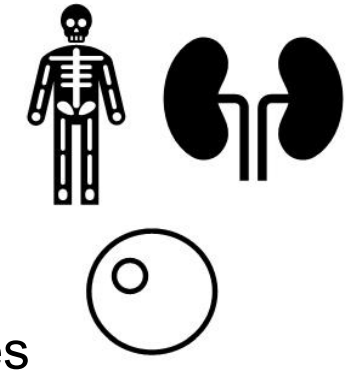
- **So – does drug Y cause adverse event Z?**

# Problem 2 goes across *all* areas of data we have – pharmacology, PK, phenotyping disease, … all are conditional!

- Links between drugs/targets/diseases are quantitative, incompletely characterized
- Subtle differences in eg compound effects (partial vs full agonists, off-targets, residence times, biased signalling, etc.)
- Effects are conditional (variation between individuals, age, sex, co-medication…)
- Phenotyping is sparse, subjective (deep phenotyping?)
- We don't understand biology ('the system'), we don't know what we *should* label, and measure, hence …
- We label what we *can* measure: 'T*echnology push*' vs '*science pull*' (!)

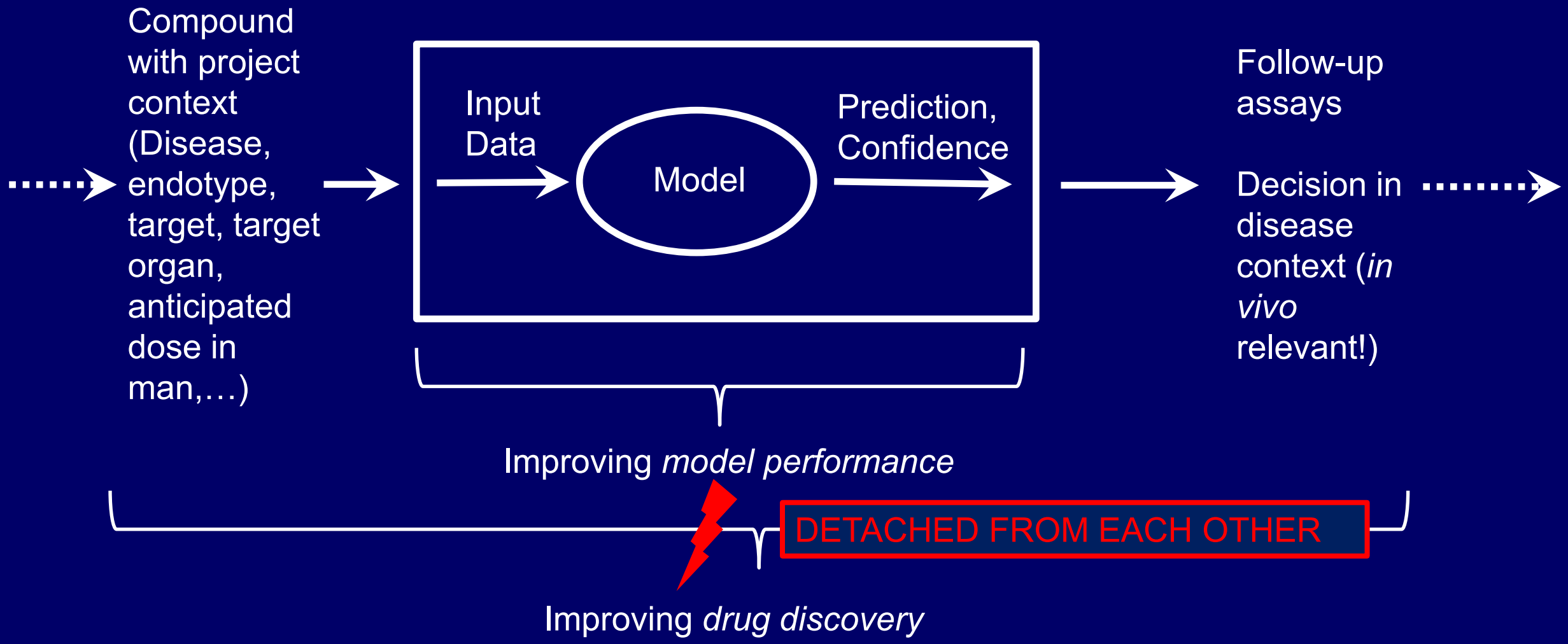
- **Are our labels – 'drug treats disease X', 'ligand is active against target Y', … - meaningful?**

- **Computer science is tremendously powerful… but is our data?**

| | Object | Representation | Model | Object Label |
|---|---|---|---|---|
| Image Domain |  |  |   ResNet? AlexNet? CapsuleNet? | Cat<br>Largely *Unconditional* labels |
| | | Representation and model are *intrinsically linked* (ie, model uses native object representation by pixels) | | |
| Drug Discovery: Chemical Domain |  ? |  logP = ... Molecular Weight = ... Molar Refractivity = ... ? | Artificial Neural Network/DNN? Support Vector Machine? Random Forest? Bayesian Classifier?... | Property A<br>*Conditional* labels (eg dependent on assay system, genotype, dose, endotype, sex, age, comedications, lifestyle, ...) ? |
| | | Both representation and modelling approach are *largely trial and error* (and *not* intrinsic to the chemical domain) | | |
| Drug Discovery: Biological Domain |  ? | Transcript-/proteomics? High-content imaging? Epigenetics? Histopathology? .... ? | Artificial Neural Network? Support Vector Machine? Random Forest? | State/Effect B<br>*Heavily conditional* labels (eg dependent on genotype, dose, endotype, sex, age, comedications, lifestyle, ...) ? |
| | | Both representation and modelling approach are *largely trial and error* (in particular the information content of biological readouts has only been established for particular cases) | | |

Bender & Cortes
Drug Discovery Today 2021

# Why AI in drug discovery (often) falls short of expectations: Disconnect of (a) proxy data vs real-world endpoint; (b) model vs process



# 4. Psychology, the hype cycle and a methods translational gap

# The bigger picture: ‘AI’ is where it is due in no small part due to human psychology

- Hype brings you money and fame – realism is boring
- FOMO (‘the others also do it!’) and ‘beliefs’ often drive decisions (‘maybe they *really* have the secret sauce?’)
- Beware of the ‘hot air strategy’ of start-ups.. (*hot air + FOMO -> perception of ‘secret sauce’*)!
- ‘Everyone needs a winner’ (‘*after investing X million we need to show success to the CEO/VP/our investors/…’)*
- Selective reporting of successes leads to everyone declaring victory (but in reality no one knows what’s actually going on)
- Difficult to really ‘advance a field’ (or admit defeat!) with little real comparison of methods… *we cannot even properly measure progress!*

Bender and Cortes, Drug Discovery Today 2021

# Summary

- We need to analyse our data (as we did for many years before), absolutely!
- ‘AI’ *is a valuable tool* in the toolbox
- The *real* game changer for translation to patients will come only once we understand biology/biological data better (and generate it, and encode it, and analyse it)
- From the data side, consortia on even larger scale are needed (for targeted data *generation*, not just sharing what is there already)
- Methods need to *translate into reality*, we need to go *from model validation to process validation*

Thank you for listening!
Any questions?

Contact: ab454@cam.ac.uk
Personal email: mail@andreasbender.de
Web: http://www.DrugDiscovery.NET
Twitter: @AndreasBenderUK