

# Modelling Molecules? Great!

## What to Consider to Really Impact Drug Discovery

Andreas Bender, PhD

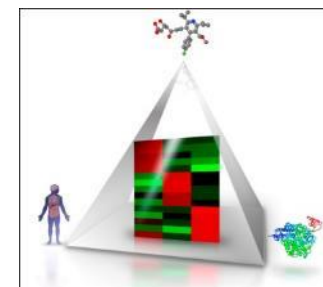
Natural Philosopher for Molecular Informatics  
Department of Chemistry, University of Cambridge

Chief Informatics & Technology Officer, Pangea Botanica, London/UK and Berlin/Germany

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

# My key scientific/society inflection points (so far)

- When I was doing my PhD I focused *on one thing* (deeply)...
  - ... afterwards I recognized how much more there is to *drug discovery*
- I 'grew up as a chemist, who also programmed since childhood' ...
  - ... e.g. in postdoc at Novartis I learned to appreciate how important it is to understand chemistry, biology/pharmacology, machine learning... *and beyond*
- When I started my first group leader position in Leiden/NL we set up the 'Pharma-IT Platform', between computer science and life sciences...
  - ... ever since then I am trying to bridge life sciences and computer sciences/ML, only together we are able to *really* make progress

# Key Learnings

1. Pick the right endpoint – either directly *in vivo* relevant, or you know how to translate to relevance
2. Anticipate future uses of the predictive model
3. Pick problem-relevant performance metrics (not generic ones!)
4. Care about the *process*, not only the *model* ('I have predicted' – fantastic, and now?)
5. Perform prospective validation, where possible

# Context: The 3<sup>rd</sup> 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

### RAPID GROWTH IN PUBLISHED RESEARCH USING AI FOR DRUG DISCOVERY

More papers since 2010 than in all prior years combined

## AI 2020: THE FUTURE OF DRUG DISCOVERY



Source: PubMed, July 11, 2018, using this query: ("artificial intelligence" or "machine learning" or "deep learning" or "neural network") and (drug or drugs), 1972-2017.

# Contents

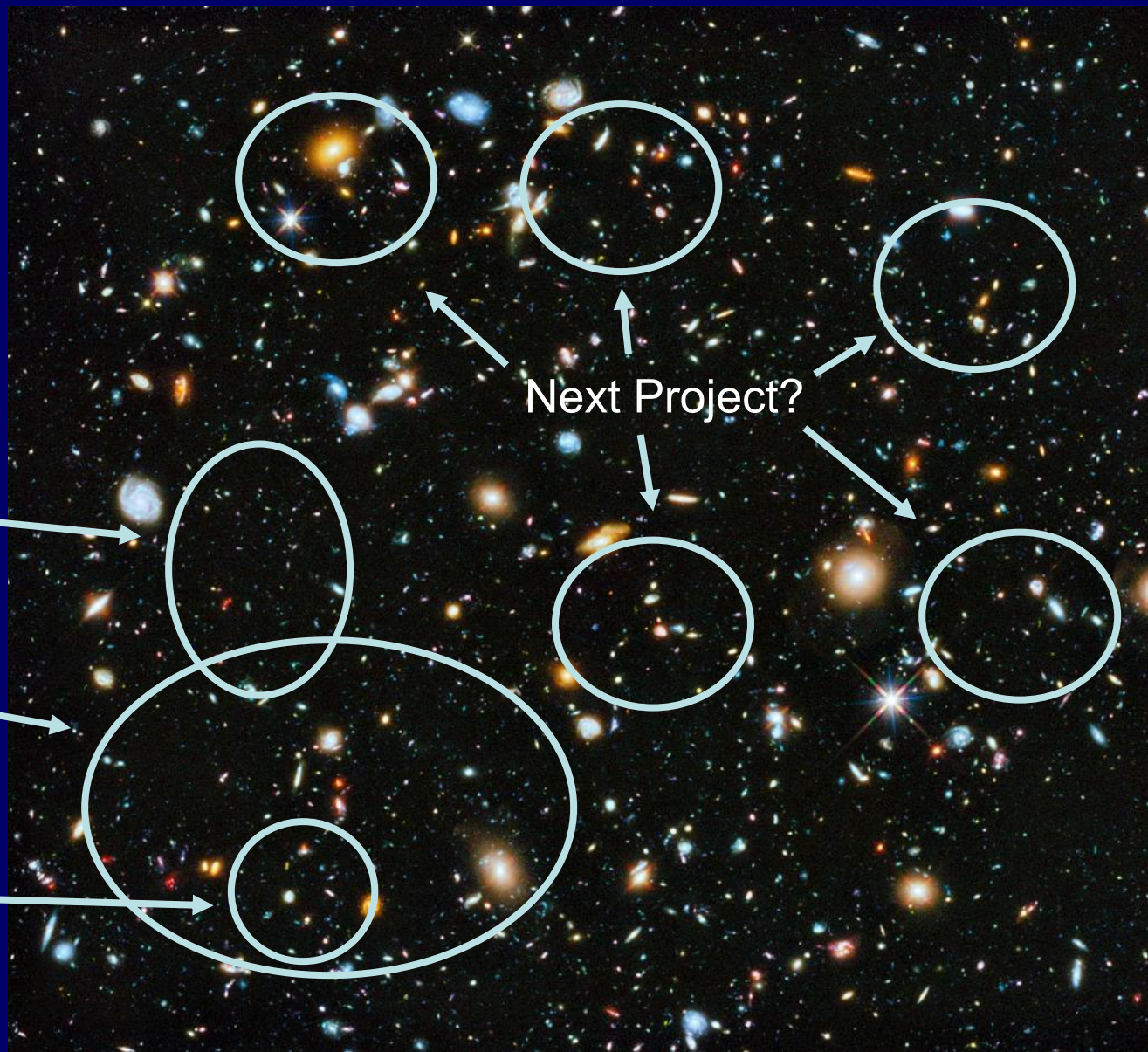
1. How do we know that a method *works*? What is 'validation'?
2. The Achilles heel of AI in drug discovery: data & proxy assays
3. Psychology, the hype cycle & the translational gap of methods

# 1. How do we know that something *works*? What is 'validation'?

- Core question in science, core question for start-ups
- In *theory* we establish a method, use a benchmark, and know how well the method works
- In *practice* this doesn't really work in *drug discovery* –
  - Labels are either mostly only *in vitro*-relevant, or conditional ('depend' on dose, *etc*)
  - Validation is costly (phase II studies for efficacy; *plus controls*), so *little prospective data*
  - Difficult to sample distribution in chemistry/'project' space well (diversity, number), so performance *depends heavily on test set*
- Retrospective validation is equally futile (no prospective discovery, predictivity for future projects unknown, all behave differently)
- *Core reasons for problem: In chemical space proper sampling impossible, underlying distribution unknown; conditionality of in vivo data*



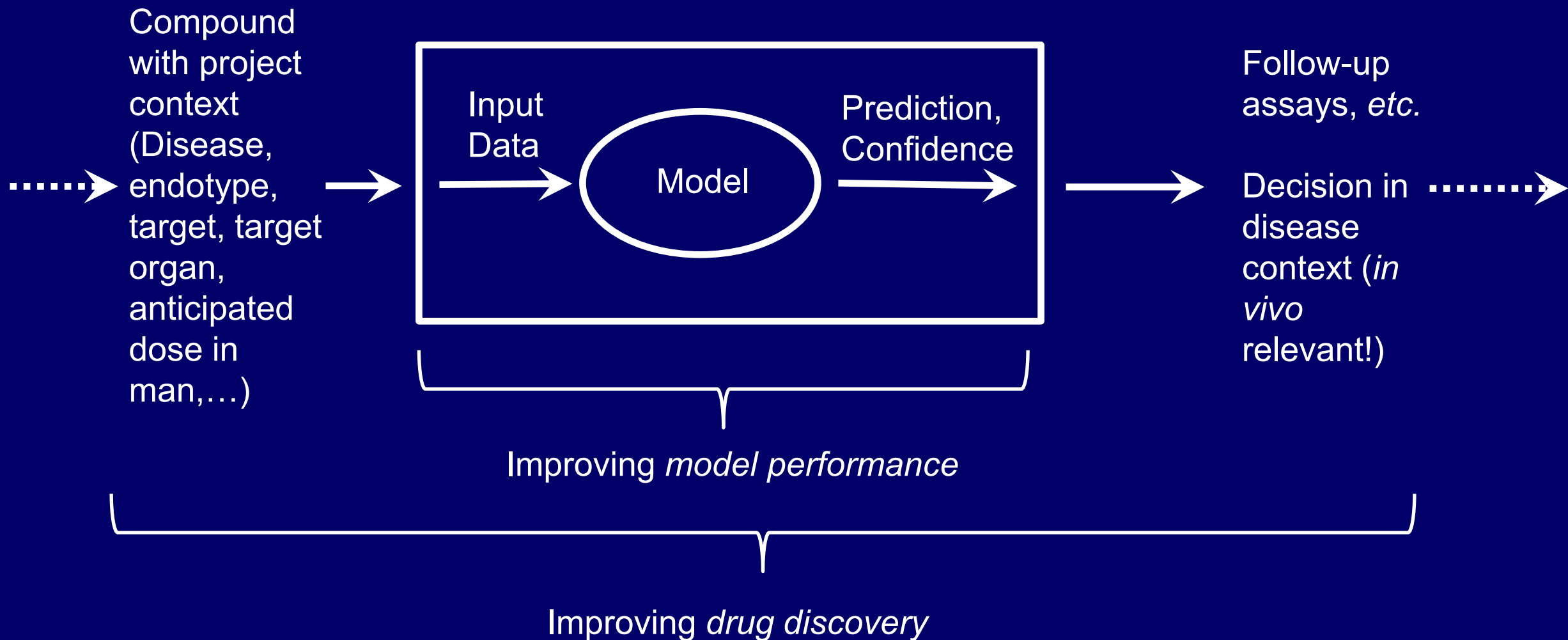
# Why 'validation' of a model has little value: You get the numbers you want (depending on the question you ask/data set you use)



- Chemical space is large; data sets are small
- Model is unable to generalize to unseen spaces
- ***Numerical distances mean something different in different areas of chemical space***
- ***'If you go 10m (e.g.  $T_c = 0.9$ ) from any one bridge (active compound), you... can be in lots of different places!'***
- "Every model is a local model"



# Model validation vs process validation (e.g. ligand structure-based property predictions)




# What to watch out for in validation – and why the model, *embedded into the process* matters

- ‘Proof by example’ abounds, without baseline
- Irrelevant endpoints abound (numerical improvements, endpoints that don’t directly translate into *in vivo*-relevant decision making)
- *Validation that matters* includes the *process and not only the model* in the validation (!)
- Success ascribed to the model (as part of a *process*), e.g. in virtual screening, where process variables have impact
- Small numbers
- Trivial successes (e.g. bioisosteric substitutions)
- No negative control
- ...

# Model validation – two resources

1. <http://www.drugdiscovery.net/HowToLie>
2. Nature Reviews Chemistry 2022 article

## Evaluation guidelines for machine learning tools in the chemical sciences

*Andreas Bender, Nadine Schneider, Marwin Segler, W. Patrick Walters, Ola Engkvist and Tiago Rodrigues* 

### ML model reporting guide

- Data set availability
- Code availability
- Comparison to baseline
- Appropriate metrics
- Appropriate comparisons
- Prospective evaluations
- Model interpretation

# Key problem in chemical datasets: Biases!

## Influences all explainable AI approaches (!)

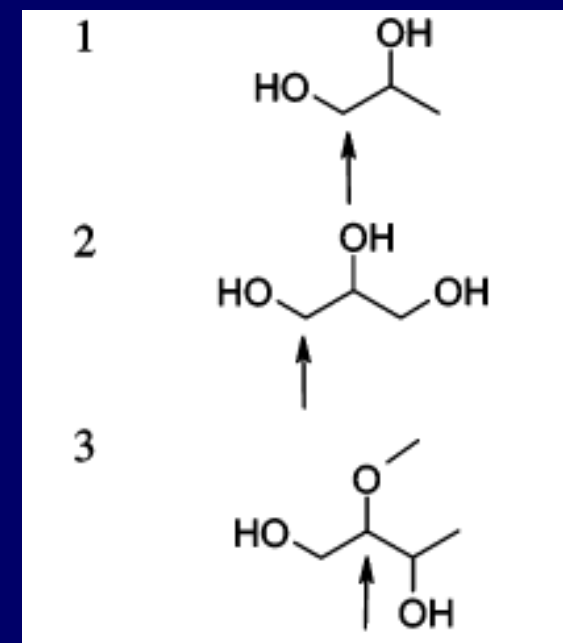
- Chemical space is  $10^{63}$  - however, our data (large is  $10^6$  compounds) clusters tremendously
  - Drugs? Fast followers, analogues
  - Published literature? Series (for SAR)
  - *Etc*

- Example (from own work): 649 bitter compounds vs 13k compounds from MDL Drug Data Repository

- Characteristic features for bitter compounds?

*Sugar rings! (due to glycosylation of natural products, which are often bitter; shown are fingerprint features which capture parts of those rings)*

Rodgers, *J. Chem. Inf. Model.* 2006, 46, 569.



# Competitions: Help or Hindrance?

Structure of most competitions:

- Use *pre-processed* dataset with *defined labels* based on a *proxy* endpoint and *validate method* on a *defined test set*, using *generic performance metrics*

Problems with this setup when related to real world:

- Pre-processed: ignores translation of experimental measurements into target values (uncertainty, choice of values, etc.)
- Defined labels: ignores conditionality of *in vivo* relevant data
- Proxy endpoint: ignores practical relevance of endpoint (*in vivo* translation)
- Defined test set: Tries to approximate real-world discovery projects, but by definition at the same time doesn't
- Generic performance metrics: Doesn't tailor how model performance is measured to real-world problem that needs to be solved



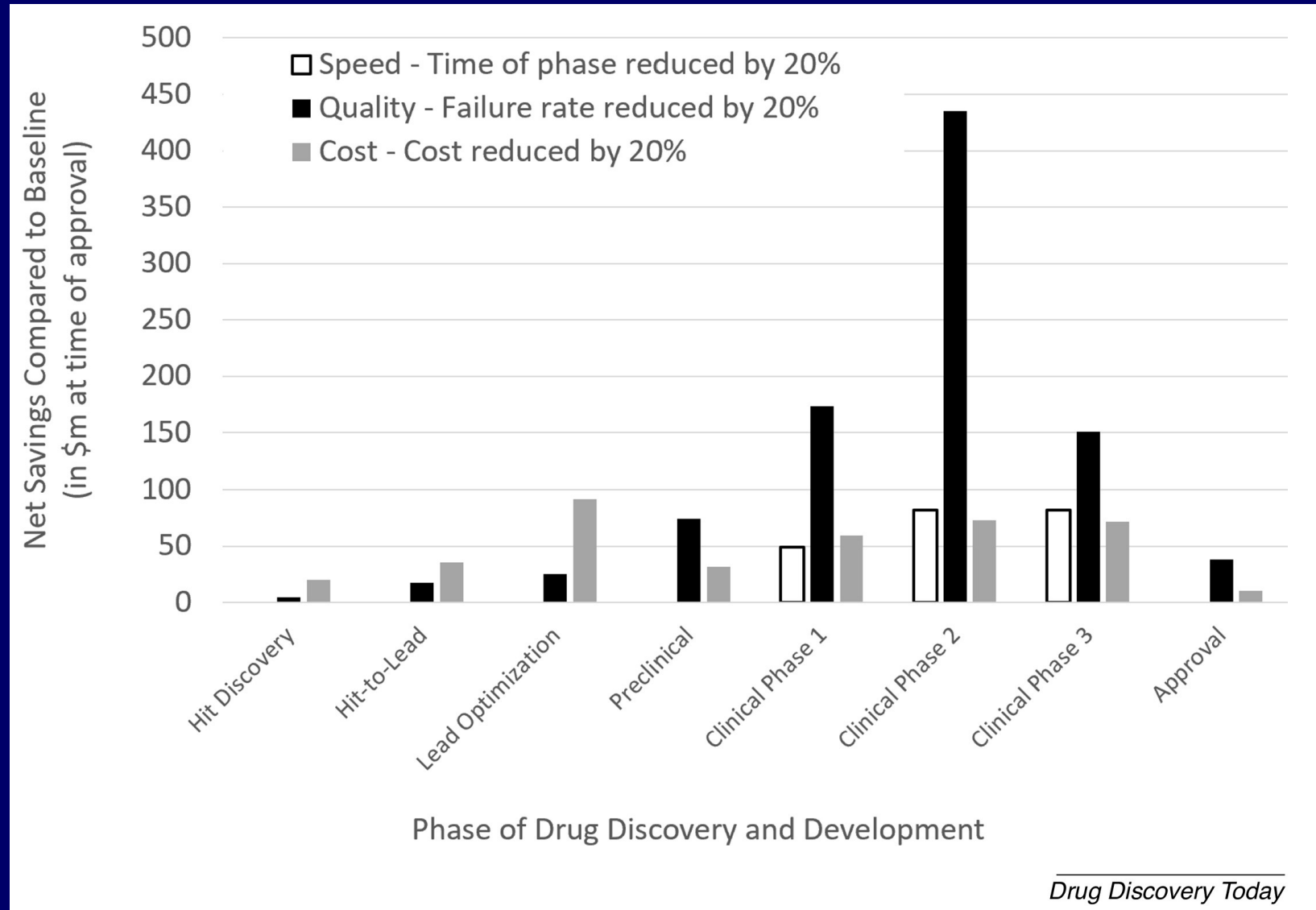
# Problem-relevant performance metrics

- Virtual screening, finding hits: “Many possible solutions in – for practical purposes – infinite search space”; you can screen say  $10^6$  compounds; probably sufficient recall in e.g. top 1% (assuming  $10^8$  search space) is what matters (*but also diversity, etc.*)
- Target prediction: Elucidating modes of action of a compound; you can test handful of predictions, also recall in top 0.1% (5/5,000 or so) matters
- Safety endpoints: Very different! Do you want a ‘warning flag’ generator? (Or avoid the ‘worrying machine’?) Depends on follow up assays!
- Note: Generic (global) performance measures, like AUROC, AUPRC, class-averaged accuracy *etc.* virtually never matter in practice!

## 2. The Achilles heel of AI in DD: Data and proxy assays

*“...it's the data, stupid!”*

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



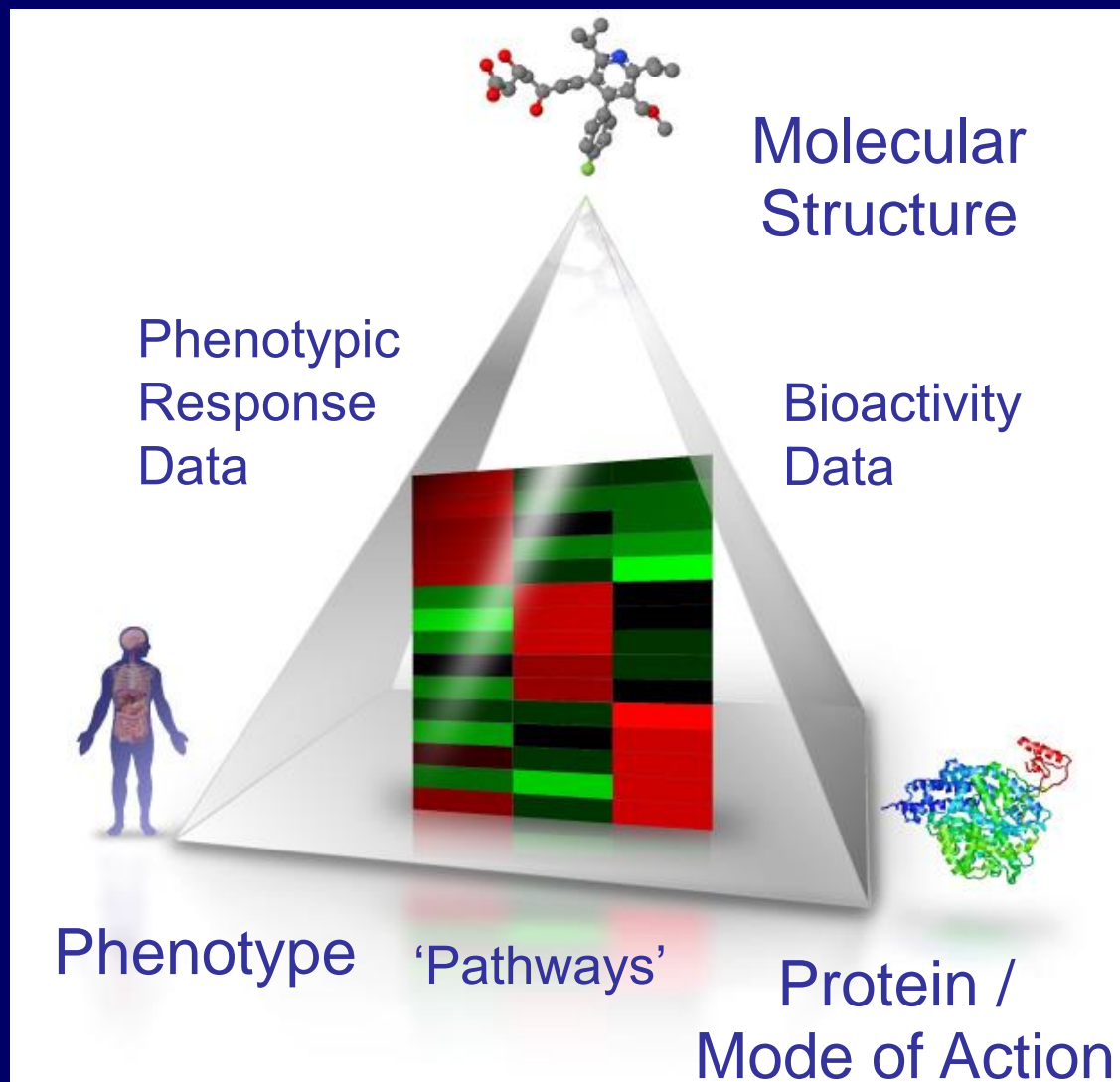
Bender and  
Cortes, Drug  
Discovery  
Today 2021

***In vivo*-relevant decisions matter most!**

**But... is this where our *data* for models is?**

- Chemical and biological data: The flat-earth (~'in vitro') view
  - And where a flat earth is great!
- Chemical and biological data: The round-earth (~'in vivo') view
  - Drug discovery data and its complexity (... the elephant in the room...)
- Why algorithms from image and speech recognition don't really translate to *drug* discovery

# A simple view on the world: Linking Chemistry, Phenotype, Targets / Mode of Action (myself, until *ca.* 2010)



a.k.a. “The world is flat”

= “We believe our labels”

“Compound A is toxic”,  
“Compound B binds target X”,  
“Compound C treats disease Y”, ...

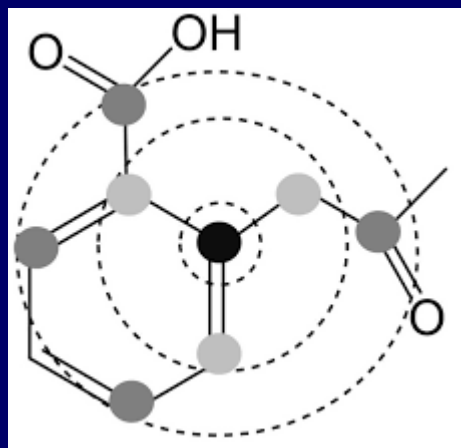
Works in cases where data is large-scale, and homogenous, and we have meaningful labels

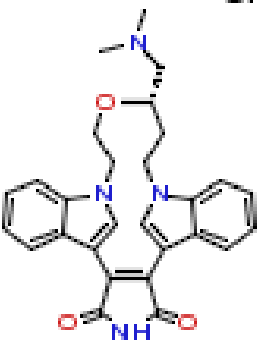
Does not consider data conditionality, e.g. dose, PK, translatability from model system to *in vivo* setup, endotype, genotype, *etc. etc.*

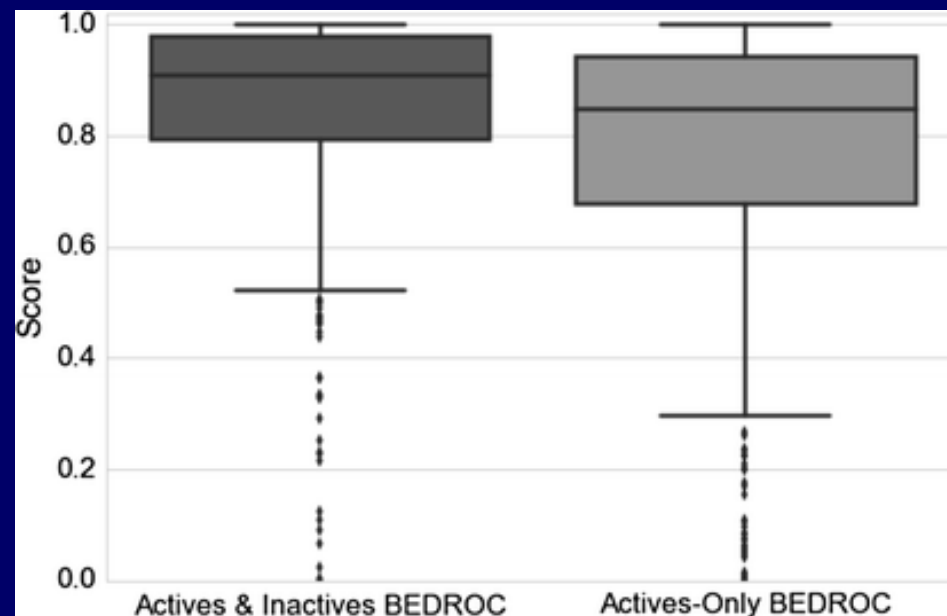


# The 'flat earth' view can *still* help! Eg Public target prediction model, based on ~200 mio data points

- E.g. work of Lewis Mervin, with AstraZeneca, *J. Cheminformatics* (7) 51
- ChEMBL actives (~300k), PubChem inactives (~200m); 1,080 targets
- Many classes (>1,000); more inactives than actives (100:1-1,000:1); very imbalanced classes (20-10,000 compounds/class); analogue bias
- <https://github.com/lhm30/PIDGIN>



Molecule	Targets	Scores
 Chiral	PRKCB1	95.81
	CAMK2G	87.48
	PRKCG	66.35
	PRKCA	56.99
	PRKCD	52.44
	PRKCH	51.41
	PRKCE	50.42
	PRKCZ	42.48



## So: Using bioactivity data for ligand-protein activity modelling '*is relatively possible*'

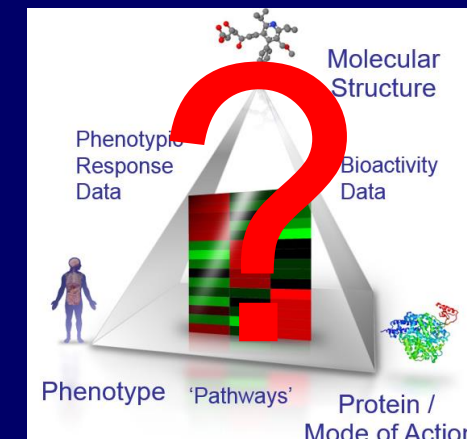
- On-target bioactivities (links between chemical structure and protein targets) are *relatively large-scale*, and *relatively homogenous*
- Hence, generating models for on-target bioactivities is 'possible'
- Can also be used for design (eg multi-target ligands)

### ***BUT:***

- Only covers known chemical space /suffers from various data biases (analogues, data set sizes, etc.)
- Labels are *still* heterogenous
- *In vivo* relevance of predictions needs to be established (!!!; PK, target engagement *in vivo*, competing ligand/knock-out, etc.)

# BUT...The world is not flat. What now?

- 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.)
- 'Pathways' from very heterogenous underlying information; dynamic elements not captured etc.
- Effects are state-dependent (variation between individuals, age, sex, co-medication...) – PK is often rather neglected in AI approaches
- 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: 'Technology push' vs 'science pull' (!)
- **Are our labels – 'drug treats disease X', 'ligand is active against target Y', ... - meaningful?**
- **Conditionality: Causality, confidence, quantification, ....?**
- **Computer science is tremendously powerful... but is our data?**



# Example of conditional labels: adverse reactions

- **“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?**

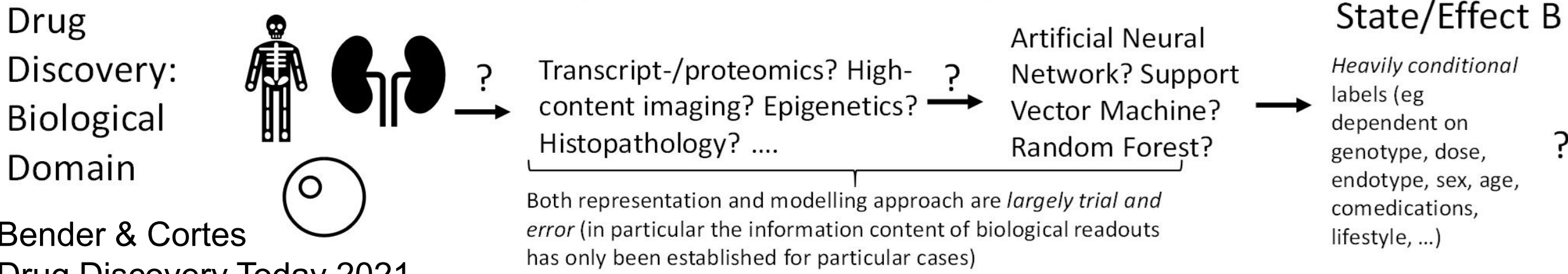
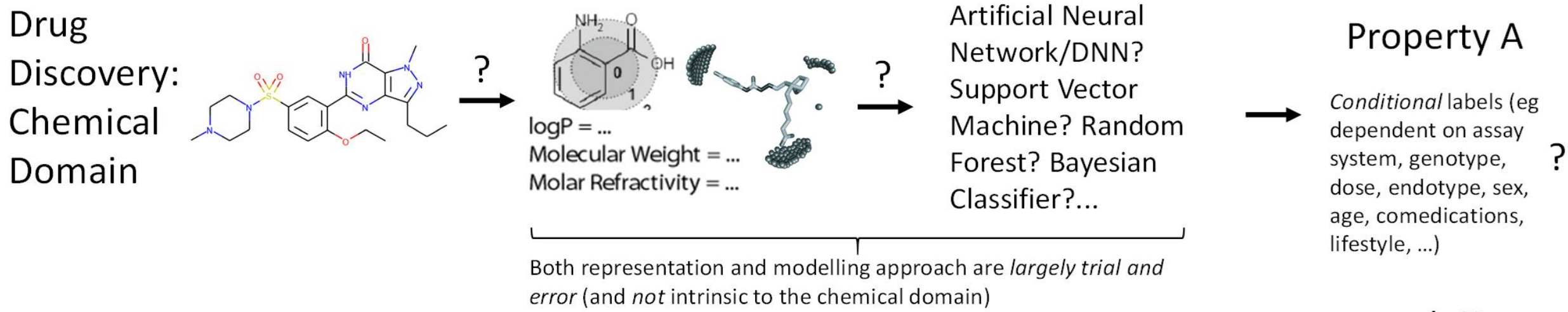
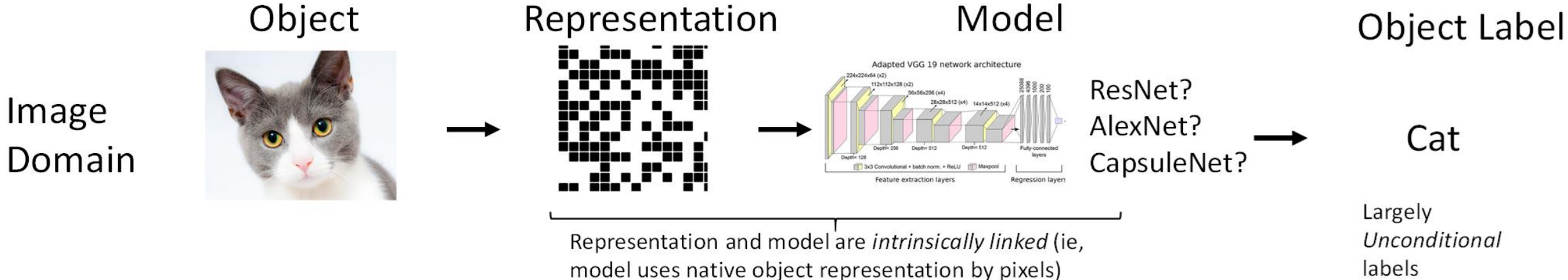
# Biological data has no inherent representation, underlying distributions are unknown, sampling is biased, data is conditional

TABLE 2

Comparison of data and representations in the image, speech, and chemistry/biology domains<sup>a,b</sup>

Domain	Representation relevant for objective	Representation comprehensive	Underlying distribution known	Sampling of underlying distribution	Conditionality of data	Quantitative dependence of label on external context
Images	Pixels describe object (but dependent on orientation)	Yes within domain (images contain all information about visual object)	No	Biased but good (large data sets available)	Partial	None (labels can be assigned in binary fashion)
Speech	Yes (waveform captures all aspects of speech)	Yes	No	Biased and good (large data sets available)	Partial (context); local and global structure	None (words can be assigned entirely based on waveform)
Chess/GO	Yes (locations and functions of pieces are fully defined)	Yes (positions of pieces entirely describe state of system)	Can be calculated in principle, because there is a large but finite set of movements	Can be exhaustively sampled (in principle)	No	N/A
Drug discovery: chemistry	Depends on context: which features/representation of compounds is relevant is often unknown	Partially (conformations, protonation states, etc. are frequently unknown)	No (chemical space not known in its entirety; can only be calculated as approximation)	Biased and small (100 s; up to $10^6$ – $10^9$ out of $10^{63}$ [49])	Partially (e.g., lipophilicity depends on protonation states, etc.)	Depends on context
Drug discovery: biology	Which aspect of biology contains information for which endpoint is frequently unknown	No (level of biological type of data generated, temporal, and spatial domain not explored)	Very partial (e.g., amino acid distributions in evolution)	Biased (depends heavily on experimental set-up)	Yes (e.g., gene expression depends on treatment, cell type, etc.)	Very large (biological system is heavily influenced by system, experimental set-up)

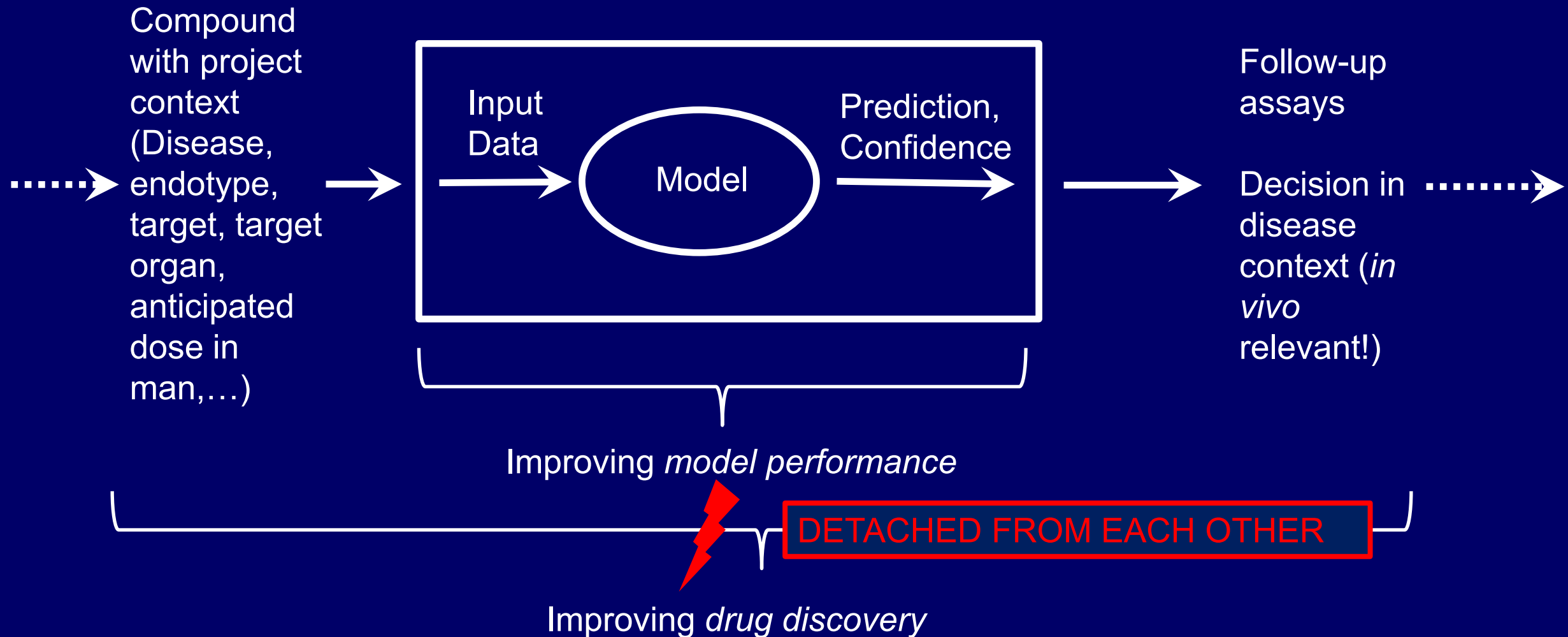




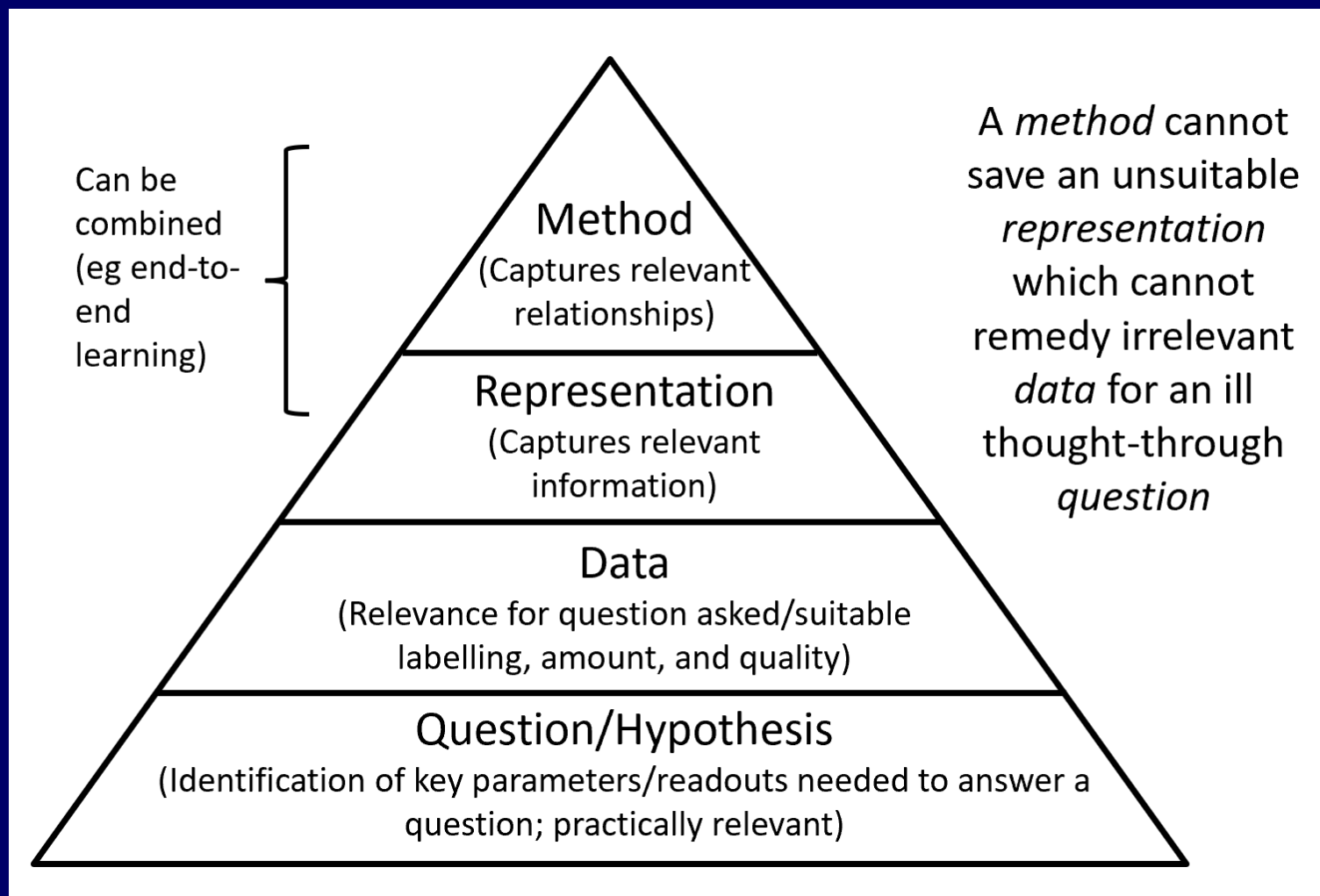
# Much of the data we have has been generated with proxy assays. Why is this a problem for AI in drug discovery?

- 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

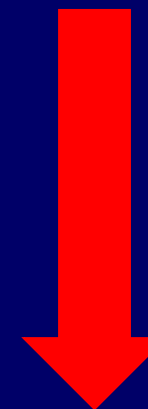
# Why meeting the proxy endpoint (and any derived models) is neither sufficient (nor necessary!) for success in a drug discovery project



The *question* needs to come first... and then the data, then the representation, and then the method  
<http://www.DrugDiscovery.NET/HowToLie>



Lots of attention currently here...



But we need to care more about this

## **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 (1)

- 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?')
- 'Ideal' Start-up Strategy Equation:  
$$\text{'Hot air (from start-up) + FOMO (from big pharma) = Perception of Secret Sauce'}$$
- NB: Multiple levels, individual psychology, as well as organizational psychology matter



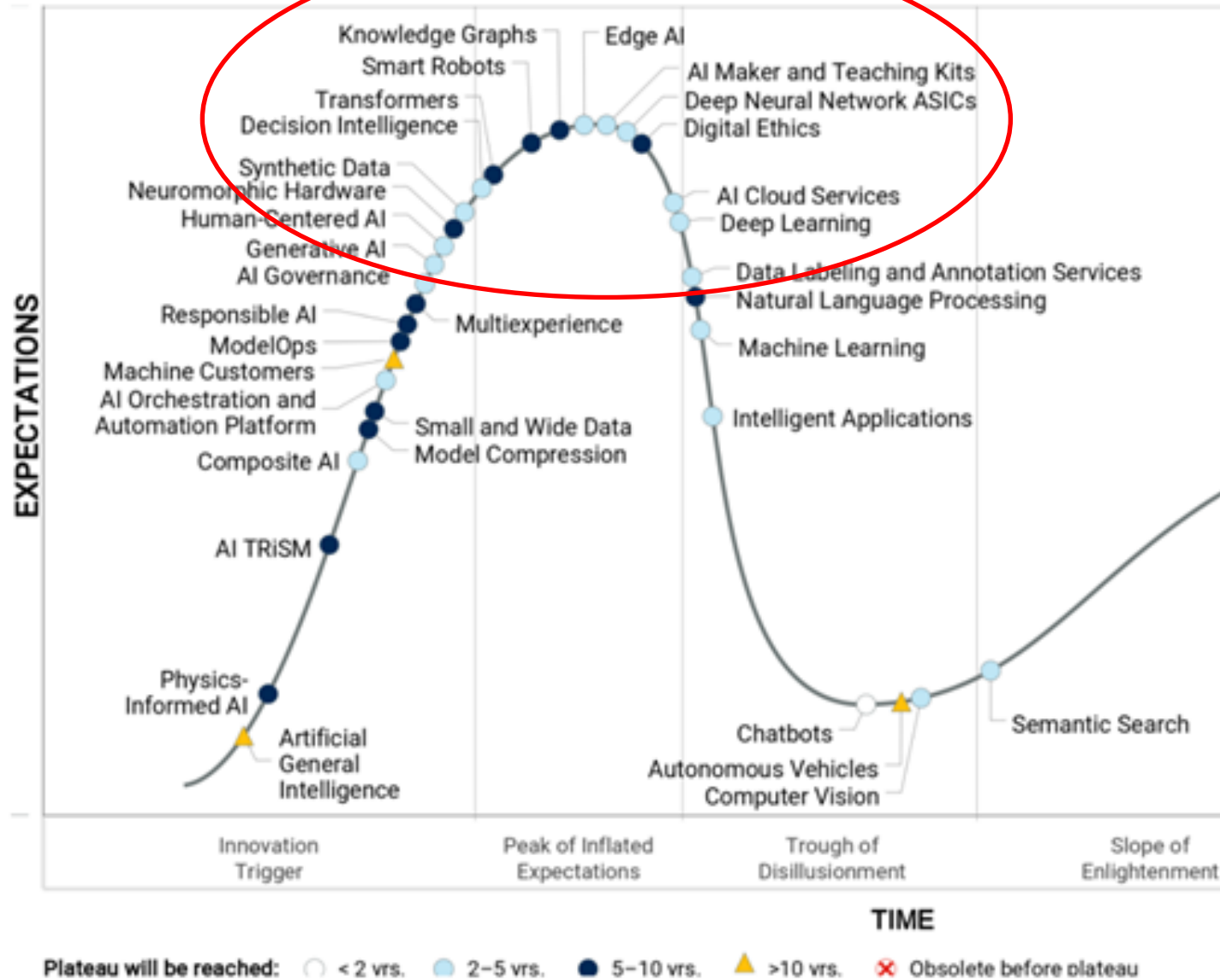
# The bigger picture: 'AI' is where it is due in no small part due to human psychology (2)

- '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' with little real comparison of methods

# AI on the Hype cycle (Gartner, 2021)

Notes:

- Y axis are expectations, not 'results'
- *Does not exist in this form, only in perception, with huge spread in the details*
- Agree with general place; *but aspects clearly working (DL for images, ML for target prediction, cloud services useful in practice, etc etc.)*
- Near future will further explore applicability of given method in a given context



# 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*

# Key Learnings

1. Pick the right endpoint – either directly *in vivo* relevant, or you know how to translate to relevance
2. Anticipate future uses of the predictive model
3. Pick problem-relevant performance metrics (not generic ones!)
4. Care about the *process*, not only the *model* ('My model has predicted!' – fantastic, and now?)
5. Perform prospective validation, where possible

Thank you for listening!

Any questions?

Contact: [ab454@cam.ac.uk](mailto:ab454@cam.ac.uk)

Personal email: [mail@andreasbender.de](mailto:mail@andreasbender.de)

Web: <http://www.DrugDiscovery.NET>

Twitter: [@AndreasBenderUK](https://twitter.com/AndreasBenderUK)