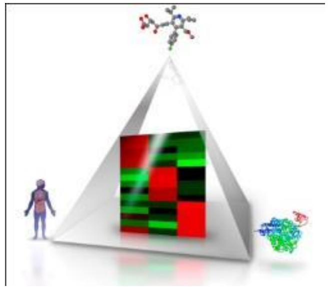
Artificial Intelligence in Drug Discovery: Where are we today? What else is needed to advance further?

Andreas Bender, PhD
Natural Philosopher for Molecular Informatics
Department of Chemistry, University of Cambridge
Fellow of King’s College, Cambridge
Director of Digital Life Sciences
Nuvisan, Berlin

UNIVERSITY OF CAMBRIDGE



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

The 3rd wave of computers in drug discovery (80s, 2000, today) – time for realistic assessment has come

Fortune cover 1981

The Blumenthal Revival at Burroughs
Bold Departures in Antitrust
Bunker Hunt's Savvy Sister
$2.50
October 5, 1981
FORTUNE
THE NEXT INDUSTRIAL REVOLUTION
Designing drugs by computer at Merck
USEDmagazines.com

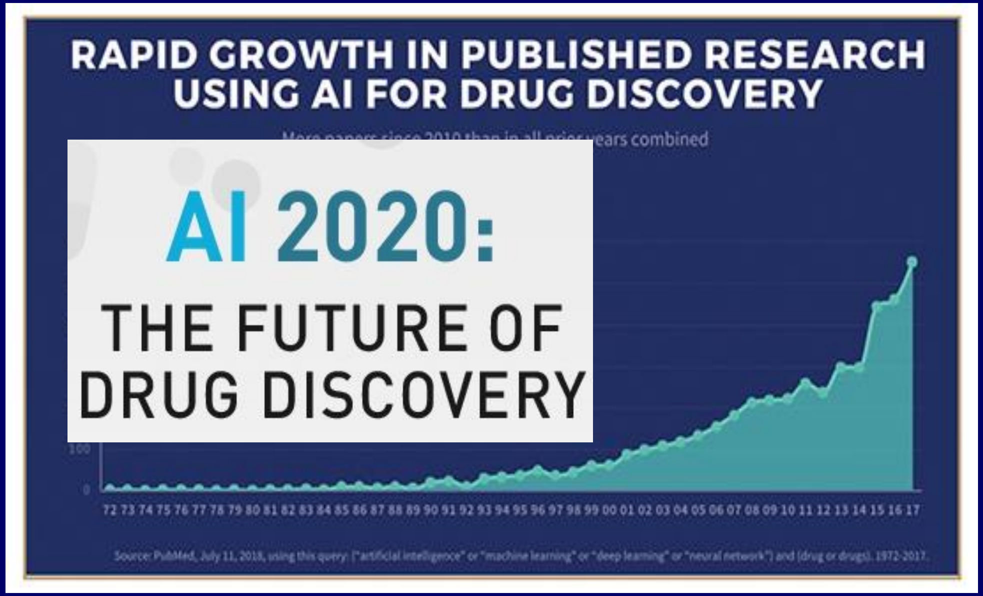
Recent headlines (2018-2020)

SPOTLIGHT · 30 MAY 2018

How artificial intelligence is changing drug discovery

World first breakthrough in AI drug discovery

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



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.

Old enough to remember 2000 biotech bubble, Human Genome Project, etc.

T. Reiss, Trends in Biotechnology, 2001:

“The number of drug targets will increase by at least one order of magnitude and target validation will become a high-throughput process.”

“More drug targets… 3,000–10,000 targets compared with 483”

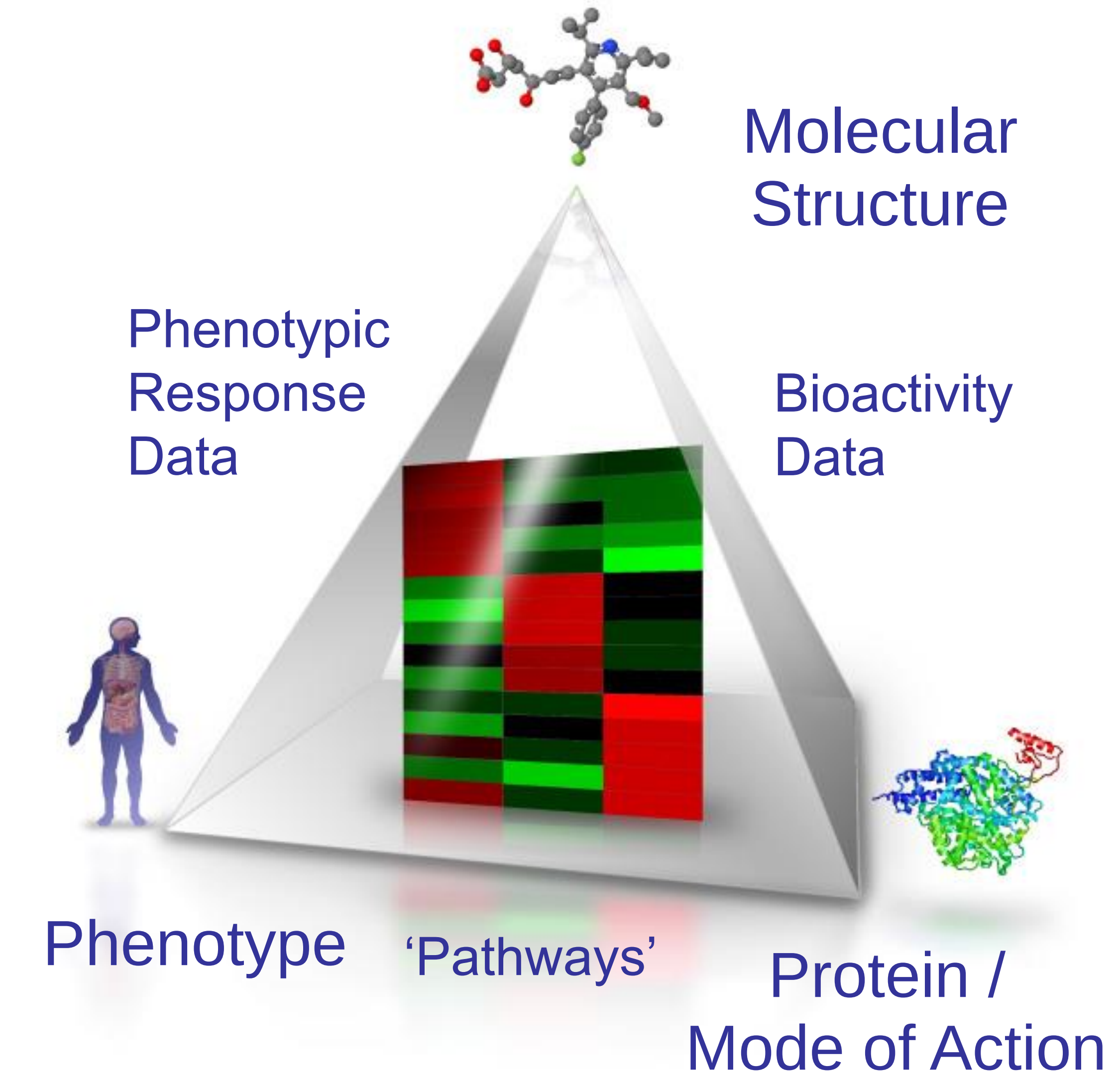
Recent (2017) estimates of drug targets put the number currently at around *667*

http://www.DrugDiscovery.NET/DataSignal

Outline: The data landscape, deep learning, biology… and humans

- Chemical and biological data: The flat-earth view
  - And where a flat earth is great!
- Chemical and biological data: The round-earth view
  - Drug discovery data and its complexity (... the elephant in the room…)
- Intermezzo: Deep Learning?
- Key learnings:
  1. The data we have is not the data we need
  2. … so what do we need, then?
  3. Model validation is poor….
  4. … and it is poor because of human biases, preferences

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"

(which are often insufficiently quantified, not directed, unconditional, don't have time/ concentration/biological setup dependence, *etc.*)

So what’s the point of it all?
We would like to answer questions

- “What is the reason upon treatment with A for phenotypic effect B?”

  -> *Mode of Action*

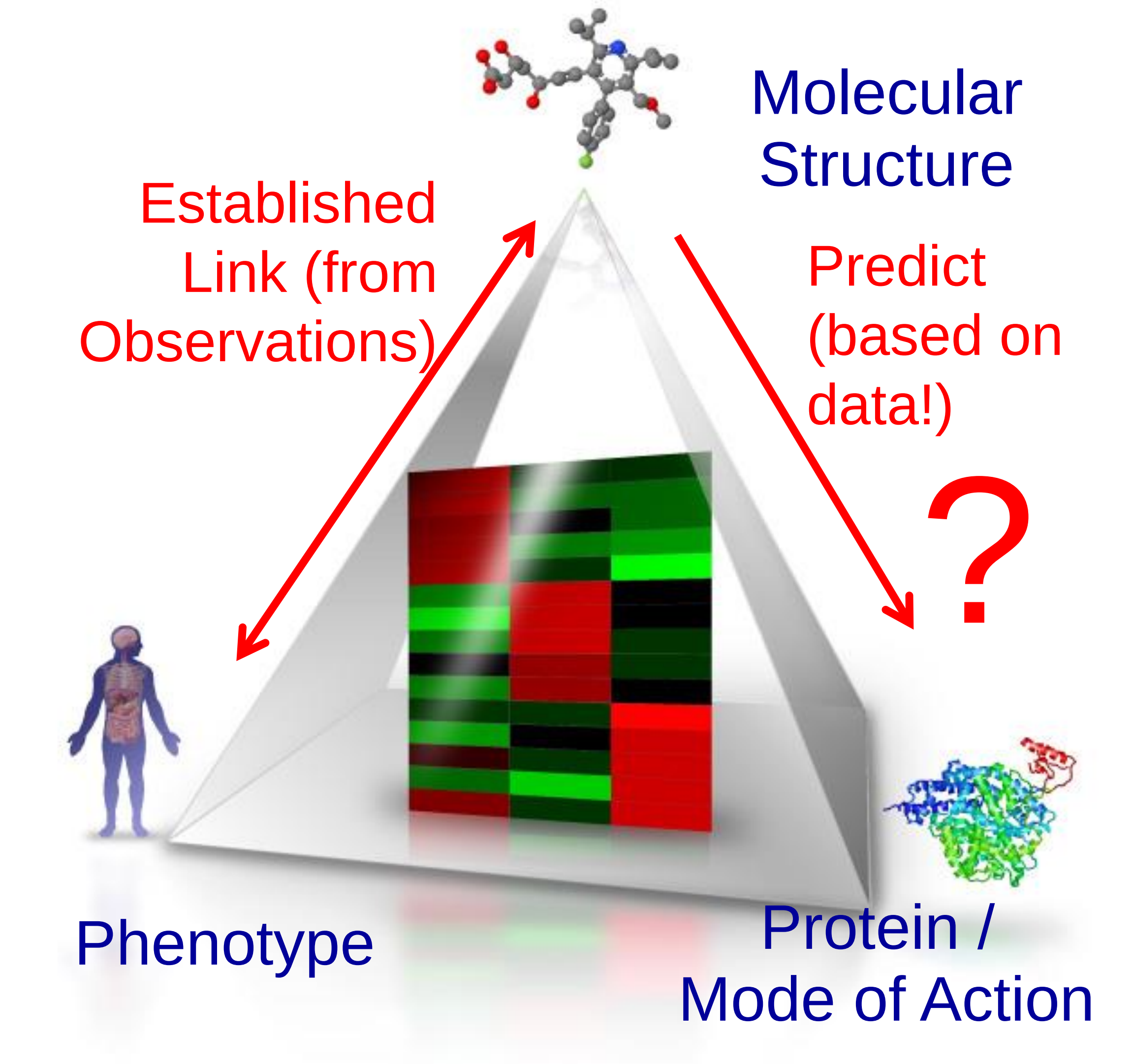
- “Which compound should I make to achieve effect C in a biological system?”

  -> *Chemistry*

- “Does patient D or patient E respond better to drug F?”

  -> *Phenotype / Phenotype Change*

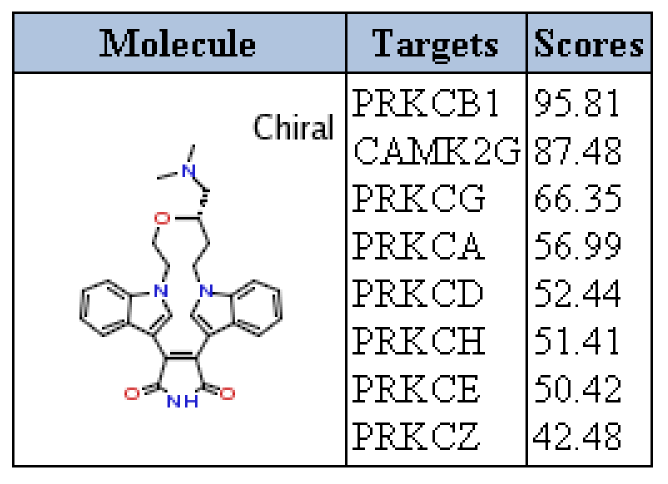
Starting from *in vivo* efficacy we can hypothesize the MoA, based on ligand chemistry



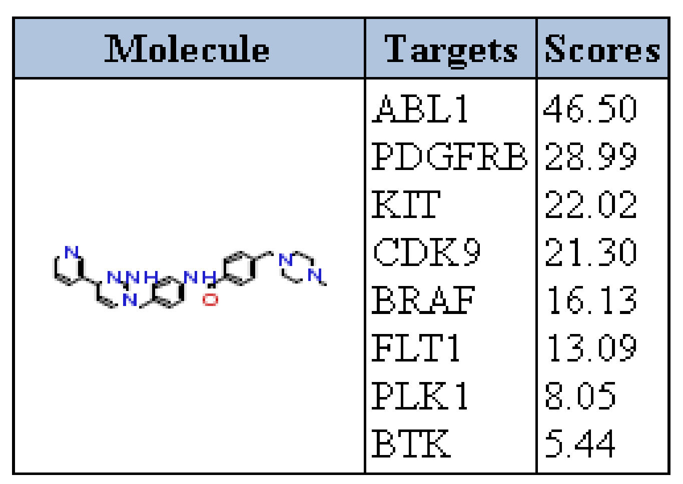
A. Koutsoukas *et* al., J Proteomics 2011 (74) 2554 – 2574.

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

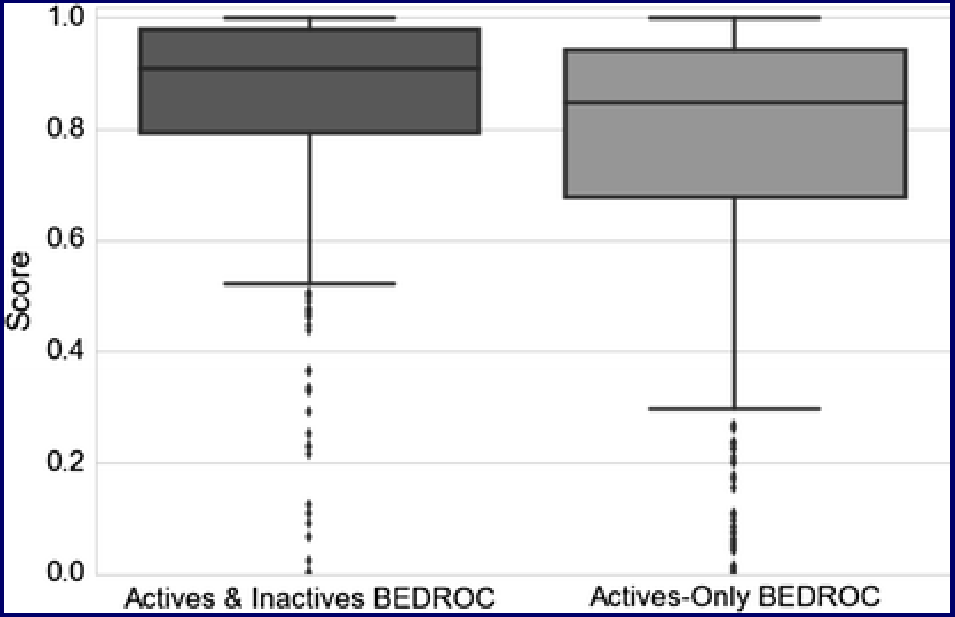
- E.g. work of Lewis Mervin, with AstraZeneca
- 2015, *J. Cheminformatics* (7) 51
- ChEMBL actives (~300k), PubChem inactives (~200m); 1,080 targets
- Can be retrained on in-house data
- https://github.com/lhm30/PIDGIN



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



Molecule	Targets	Scores
	ABL1	46.50
	PDGFRB	28.99
	KIT	22.02
	CDK9	21.30
	BRAF	16.13
	FLT1	13.09
	PLK1	8.05
	BTK	5.44



Also data publicly available

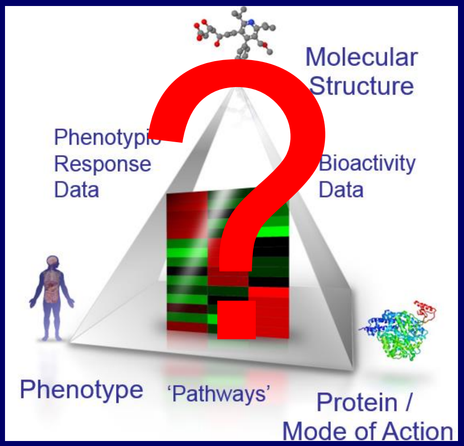
So: Using bioactivity data for ligand-protein activity modelling ‘*is relatively possible*’

- We make use of existing data (millions of data points!)
- 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: 'T*echnology 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?**

Object

Representation

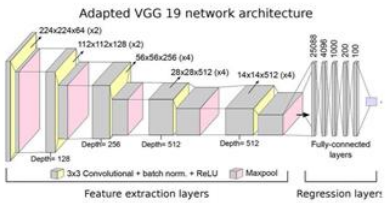
Model

Object Label

Image Domain

→

→



ResNet? AlexNet? CapsuleNet?

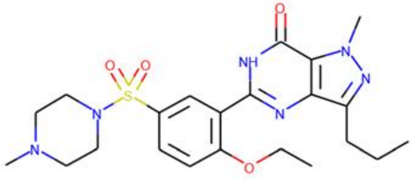
→

Cat

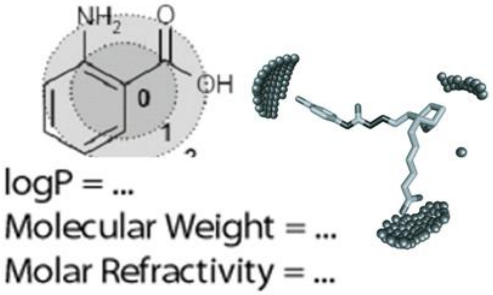
Representation and model are *intrinsically linked* (ie, model uses native object representation by pixels)

Largely *Unconditional* labels

Drug Discovery: Chemical Domain



?
→



logP = ...
Molecular Weight = ...
Molar Refractivity = ...

?
→

Artificial Neural Network/DNN? Support Vector Machine? Random Forest? Bayesian Classifier?...

→

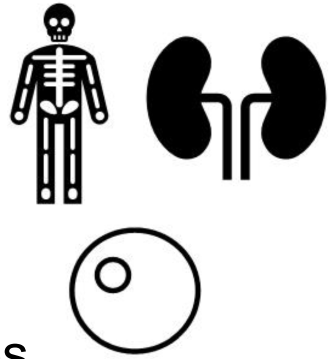
Property A

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

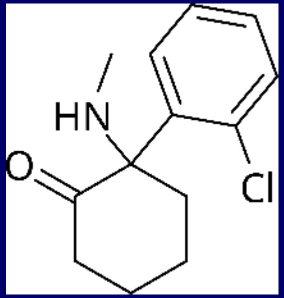
Heavily conditional labels (eg dependent on genotype, dose, endotype, sex, age, comedications, lifestyle, ...)

?

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

Are our understanding and data good enough? The many facets of ketamine



- Ketamine both used as (rather safe) anaesthetic (iv 2mg/kg), approved since 1970, as well as a street drug
- In 2000 effect as antidepressant, when dosed significantly lower, also bronchodilator (acute asthma); iv 0.5mg/kg
- Ketamine long been thought to act via blocking the NMDA receptor - *but* other NMDA blockers such as memantine and lanicemine have not been successful in clinical trials
- Also the opioid system implicated in action of ketamine (naltrexone/opioid antagonist influences its effects)
- Furthermore, a metabolite of ketamine has recently been found to be active in animal models of depression
- … etc etc. (disease endotype, co-medication, accumulation, …)

Das, J. Repurposing of Drugs–The Ketamine Story. *J. Med. Chem.* 2020 (ASAP Article)

Example of labelling problems: 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)
  - With known targets 1...n, but also unknown targets (n+1…z)
  - 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?**

Data/'AI' in early discovery vs efficacy/safety

Early discovery/proxy space (usually *in vitro*)

- Often 'simple' readouts (eg protein activity), hence…
- Large number of data points for training models

- *Models have clear labels* (within limits of model system, eg 'ligand is active against protein at IC50<10uM', or solubilities, logP, or the like)
- Good for model generation: *Many*, *clearly categorized* data points

Efficacy/safety (usually *in vivo*)

- Quantitative data (dose, exposure, …)
- More complex models (to generate data), *fuzzy labels* (classes 'depend', on exposure, multiple eg histopathological endpoints) – hence…

- *Less, and less clearly labelled data*: Difficult from machine learning angle
- Data: *Recording* vs data *suitable for mining* – eg animal data tricky, even within single company

Problem setting in early discovery vs safety

Early discovery/proxy space

- Discovery setting – ‘find me suitable 100s or 1000s out of a million’ (eg screening)

- Anything fulfilling (limited) set of criteria will do ‘for now’, predicting *presence of something*

- Computationally *generative* models often fine

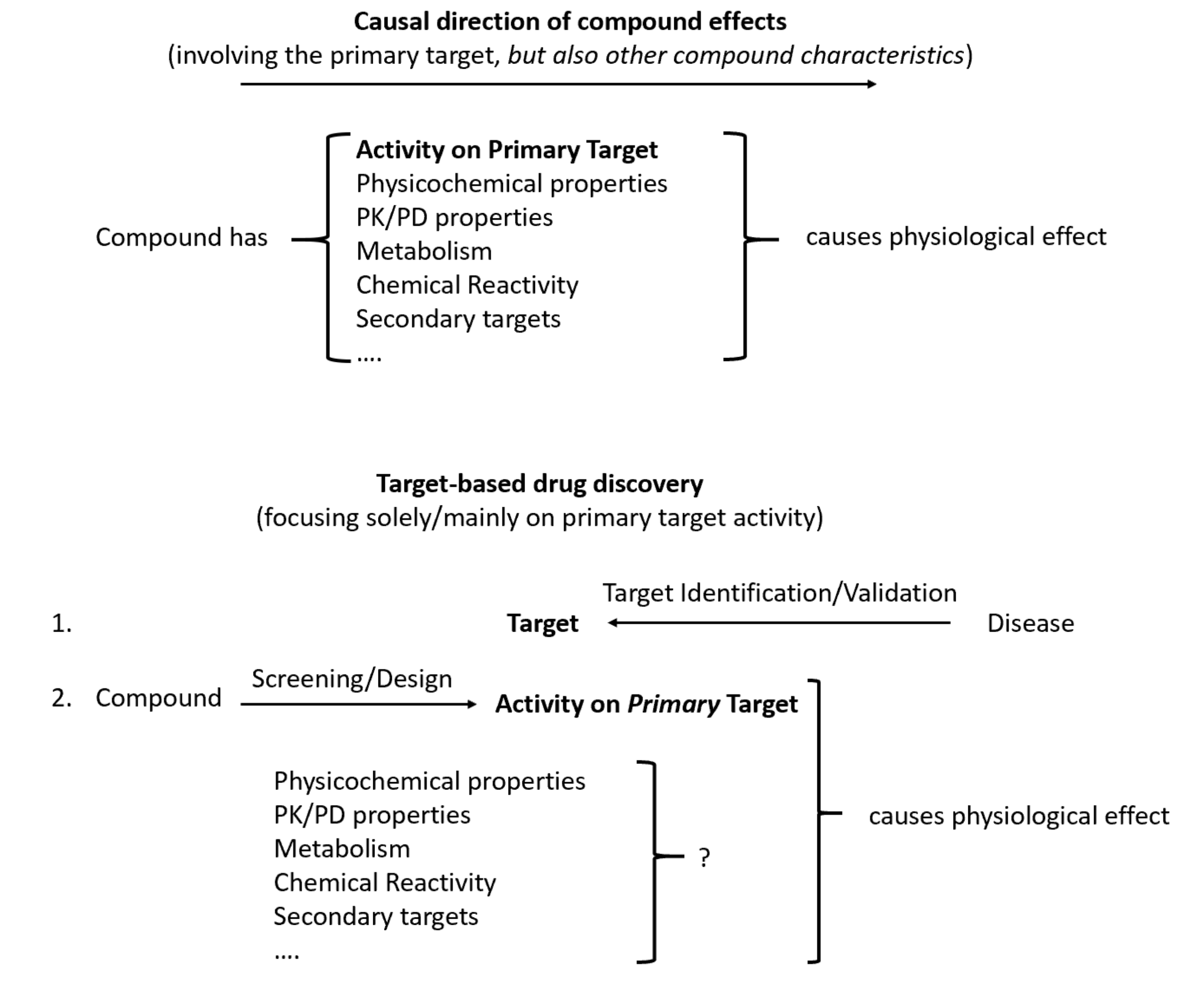
Efficacy/safety

- Need to predict for *this particular data point, quantitatively!*

- *Long list of criteria to rule out*, *based on limited data*… predicting *absence of ‘everything’* (eg different modes of toxicity)

- *Predictive* models (more tricky than generative!)

AI in drug discovery: Data availability drives the field of ‘AI in drug discovery’ … but a ligand is not a drug!



Causal direction of compound effects

(involving the primary target, *but also other compound characteristics*)

Compound has

Activity on Primary Target
Physicochemical properties
PK/PD properties
Metabolism
Chemical Reactivity
Secondary targets
....

causes physiological effect

Target-based drug discovery

(focusing solely/mainly on primary target activity)

1. **Target** ← Target Identification/Validation — Disease

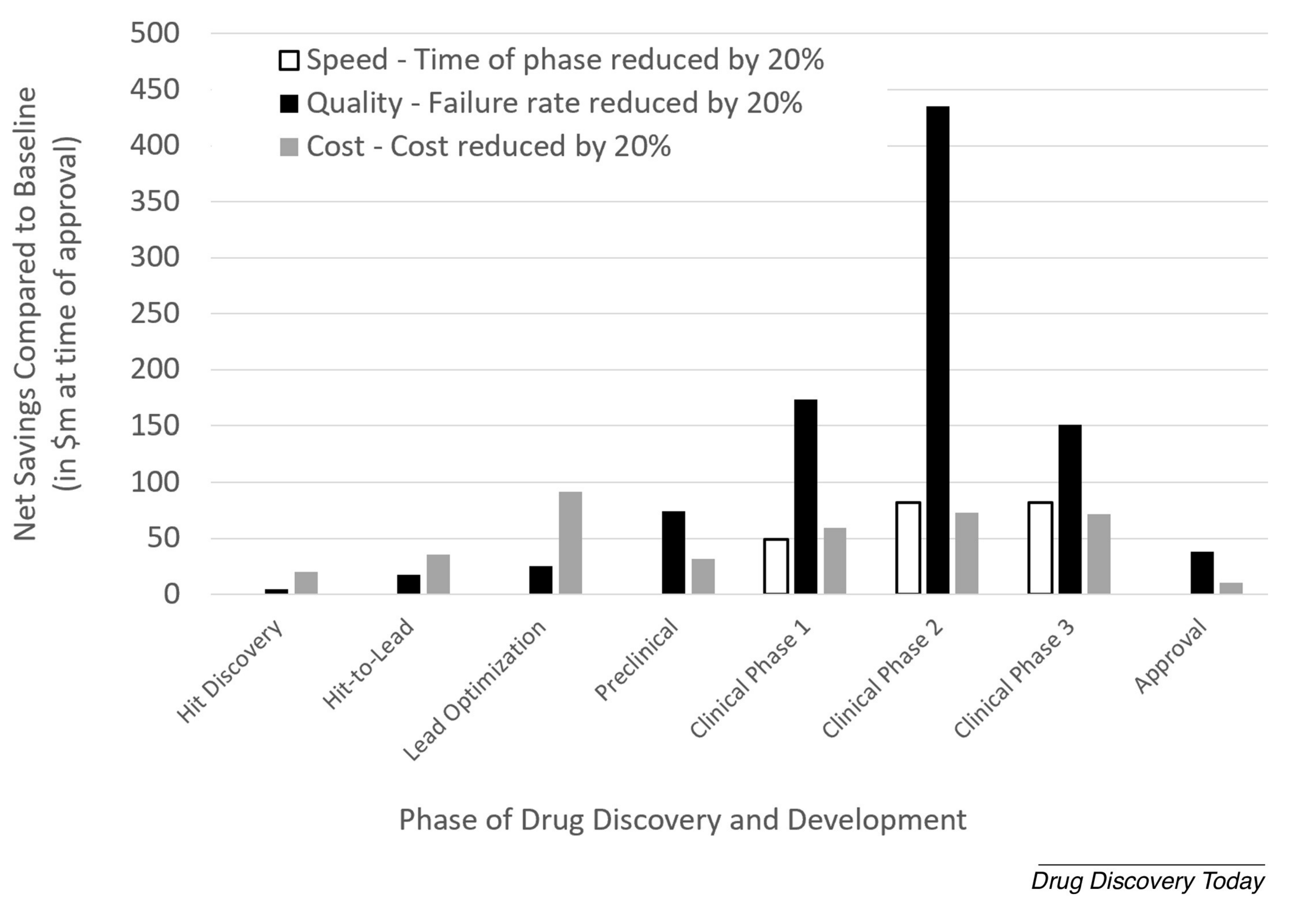
2. Compound → Screening/Design → **Activity on *Primary* Target**

Physicochemical properties
PK/PD properties
Metabolism
Chemical Reactivity
Secondary targets
....

?

causes physiological effect

The *quality* of *in vivo-relevant* decisions matters more than *early speed*!



Drug Discovery Today

Deep Learning?

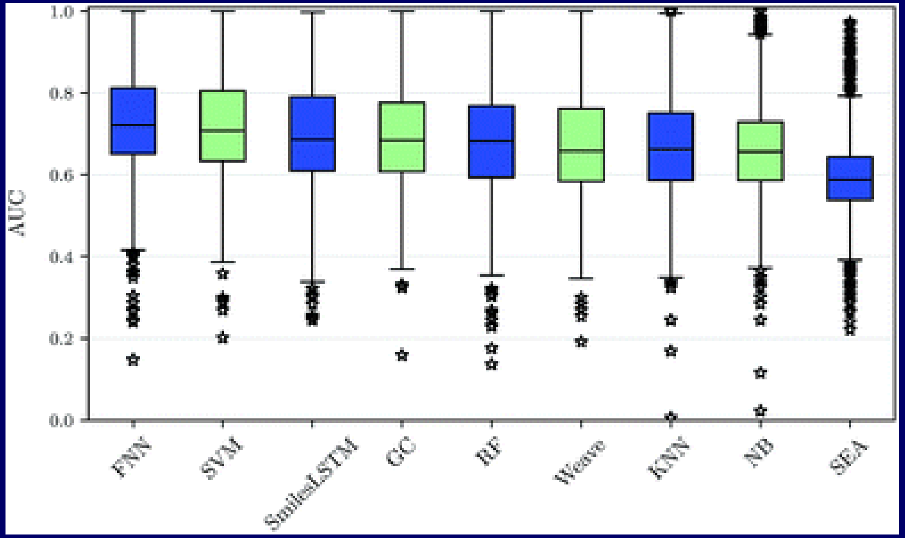
- Can work well

- Sometimes works well *numerically*, but it doesn’t really address the underlying question

- Is sometimes pushed in a biased ways in publications

There *are* areas in drug discovery where deep learning can work well

Andi Mayr et al. “**Large-scale comparison of machine Learning methods for drug target prediction on ChEMBL**”



But trade-off – taking computational time, parameter optimization into account eg for model updates, is it worth it?

- *Statistical significance* is one thing … but does it translate into *practical relevance*?
- "Is your machine learning telling you anything you didn’t already know?“ Anthony Nicholls' slides from 'AI in Chemistry' conference in Cambridge September 2019; put online with Ant's permission: http://drugdiscovery.net/data/cambridge_ai.pdf

Modelling synergy of anti cancer compounds using deep learning

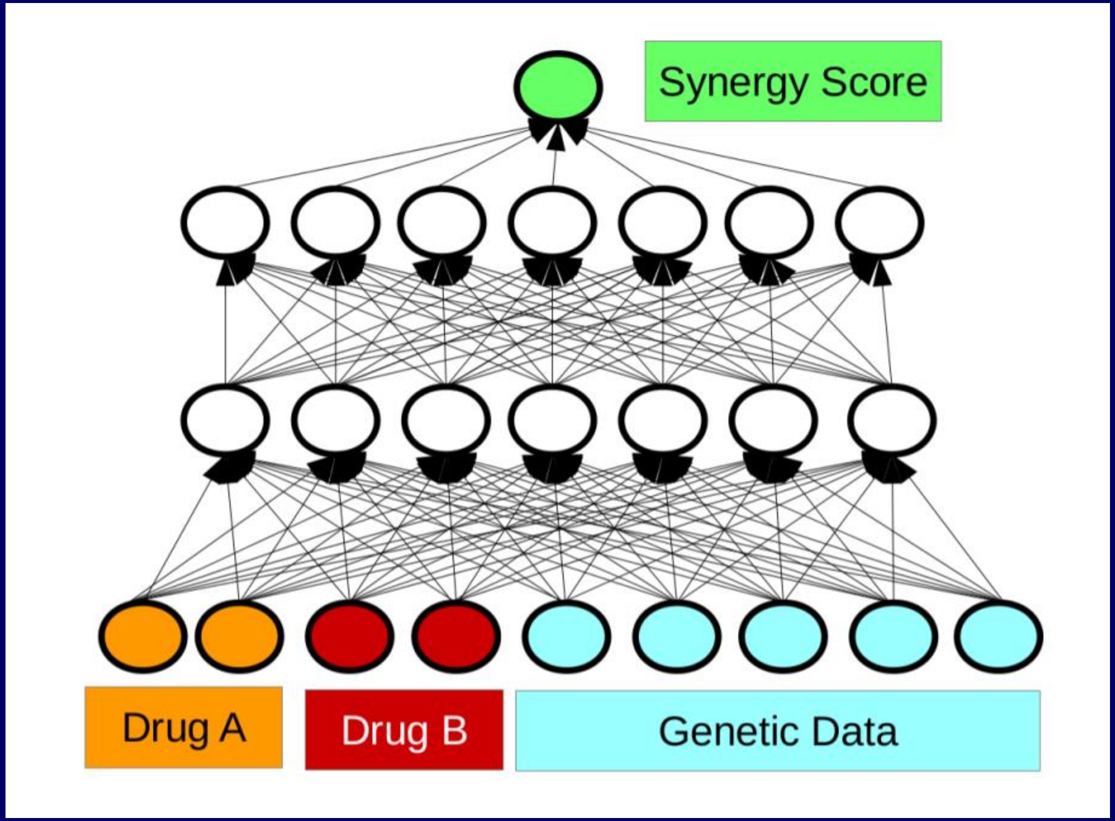
- Sometimes synergy between drugs is desired (in cancer, infectious diseases, …) to ideally improve efficacy/decrease side effects of treatment

- Merck, AZ, NCI ALMANAC, … recently published combination datasets which were can use to model combination effects

- Self-critical evaluation of our work: So does this matter in drug discovery, in practice – in the *real world*?

- Preuer *et al.*, Bioinformatics 2018

Models Used: Deep Neural Networks (‘DeepSynergy’)



Compared to: median polish, Elastic nets, Random Forest, SVM, Gradient Boosting Regression

DeepSynergy model results: Classification and quantitative model

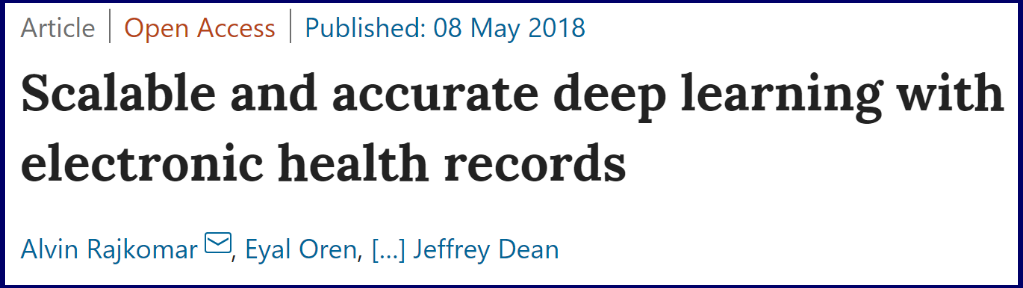
- Synergy score of 30 as threshold: True Positive Rate 0.55, True Negative Rate 0.95

- *‘1 out of 2 positive synergistic predictions is correct, on average, while 19 out of 20 non-synergistic predictions are also correct, and can be rightly discarded, when looking for synergistic compound combinations’*

- But: Practical relevance? Synergy is dose dependent; and ***does it translate to in vivo situation***….? (Greater question: Do simple endpoints, which we possibly need to generate data for ‘AI’, really help??)

- **Sometimes we maybe only play a ‘My numbers are higher than yours’ game in the end…**

“You see what you want to see” – biased reporting



Article | Open Access | Published: 08 May 2018

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar ✉, Eyal Oren, [...] Jeffrey Dean

Abstract: “Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient’s final discharge diagnoses (frequency-weighted AUROC 0.90).”

Logistic regression baseline (last page in SI): “For the full feature enhanced baselines, for predicting inpatient mortality at 24 hours after admission, the AUROC was 0.93 (95%CI 0.92-0.95) for Hospital A and 0.91 (95%CI 0.89-0.92) for Hospital B. For predicting unexpected readmissions within 30-days the AUROCs at discharge were 0.75 (95%CI 0.73-0.76) for Hospital A and 0.75 (95%CI 0.74-0.76) for Hospital B. For long length-of-stay at 24 hours after admission, the AUROC was 0.85 (95%CI 0.84-0.85) for Hospital A and 0.83 (95%CI 0.83-0.84) for Hospital B.”

Discussion

1. The data we have is not the data we need
2. … so what data do we need, then?

3. Model validation is poor….
4. … and it is poor because of human bias

(for details see Bender and Cortes, Drug Discovery Today 2021)

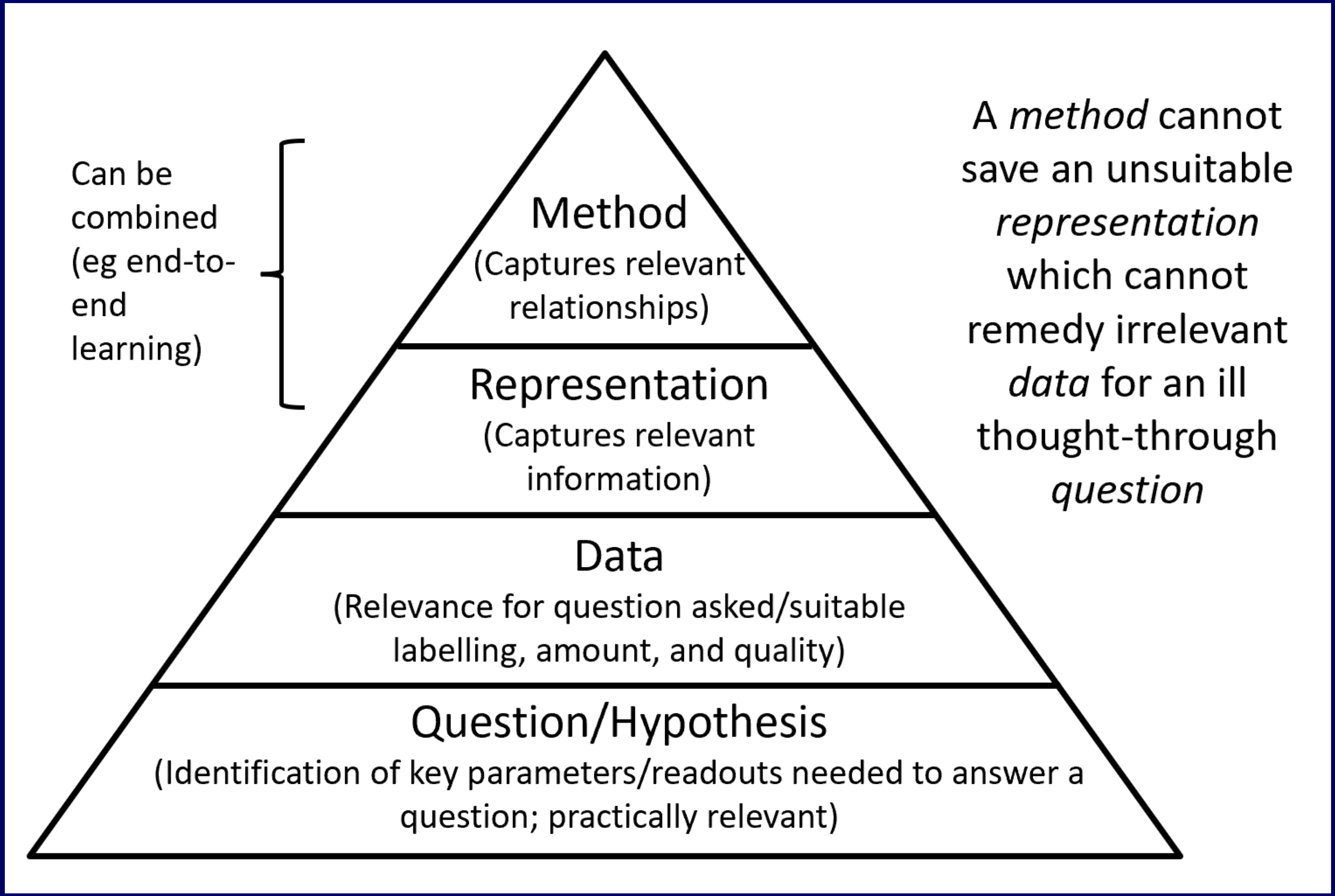
Much of the data we generate is generated for the wrong reasons (or in wrong ways)

- Often proxy measures (to reduce cost); historical data gets repurposed now ‘for AI’
- Not always relevant system/dose/time point/endpoint etc.

- “Models of models” – “the *in silico* model of the Glu/Gal mitotoxicity model” … is then meant to predict the *in vivo* situation
- We need to care more about modelling the actual endpoint of interest (say, organ risk), not the proxy (say, assay) endpoint!

- Often hypothesis-free (‘here we have our pile of data … anyone wants to have a go at it?’) instead of hypothesis-driven
- Often ‘technology push’, instead of ‘science pull’

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



Can be combined (eg end-to-end learning)

Method
(Captures relevant relationships)

Representation
(Captures relevant information)

Data
(Relevance for question asked/suitable labelling, amount, and quality)

Question/Hypothesis
(Identification of key parameters/readouts needed to answer a question; practically relevant)

A *method* cannot save an unsuitable *representation* which cannot remedy irrelevant *data* for an ill thought-through *question*

Lots of attention currently here…

But we need to care more about this

What do we really *validate* if we talk about ‘AI in *drug discovery*’?

- Discovering *ligands* or *drugs*?
- Often no meaningful baseline comparison
- Prospective validation often small, and/or (manually) biased; ‘proof by example’ style abounds

- Ascribing success of *validation* to computational *model* (!)
- BUT: “*Model* validation is *process* validation”!

- ***“How to Lie With Computational Predictive Models in Drug Discovery”***
- ***http://www.DrugDiscovery.NET/HowToLie***

Is it the method… or is it by chance?

- If a drugs results from the ‘drug discovery pipeline’ it is *the result of a long series of choices*

- Claim: “AI discovers a drug against X!”
- What is responsible?

Impossible to say!

- Viewpoint A: ‘We don’t have a baseline for control!’
- Viewpoint B: ‘But it worked – look at the compound!’
  - Both true at the same time!

- Problem: Biased reporting; *focus on trivial wins.. and we have an illusion of progress!*

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

- Hype bring 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?’)
- ‘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

Summary

- We need to analyse our data (as we did for many years before), absolutely!
- ‘AI’/deep learning 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)
- Currently a lot of computer science-driven approaches, some of which are more applicable in drug discovery than others (real translation is necessary, *but also better experimental design!*)
- Consortia on even larger scale are needed (for targeted data generation, not just sharing what is there already)

Resources

Artificial Intelligence in Drug Discovery – What is Realistic, What are Illusions?

Part 1: Ways to make an impact, and why we are not there yet

Part 2: a discussion of chemical and biological data

Andreas Bender and Isidro Cortes, *Drug Discovery Today* 2021 (in press)

http://www.DrugDiscovery.NET/AIReview

“How to Lie With Computational Predictive Models in Drug Discovery”

http://www.DrugDiscovery.NET/HowToLie

Thank you for listening!

Any questions?

Contact: ab454@cam.ac.uk

Personal email: mail@andreasbender.de

Web: http://www.DrugDiscovery.NET

Twitter: @AndreasBenderUK