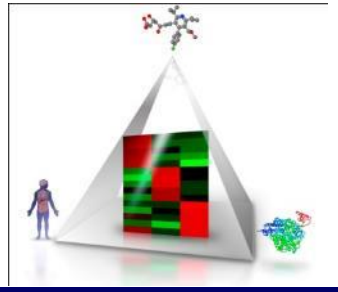
Artificial Intelligence in Chemistry: Discovery Processes and Tools Of the Future

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

Director of Digital Life Sciences
Nuvisan, 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

Outline

- On hypes and predicting the future
- An old process with AI is still… only an old process
- Problems along the way: Our data!
- What next?

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

Fortune cover 1981

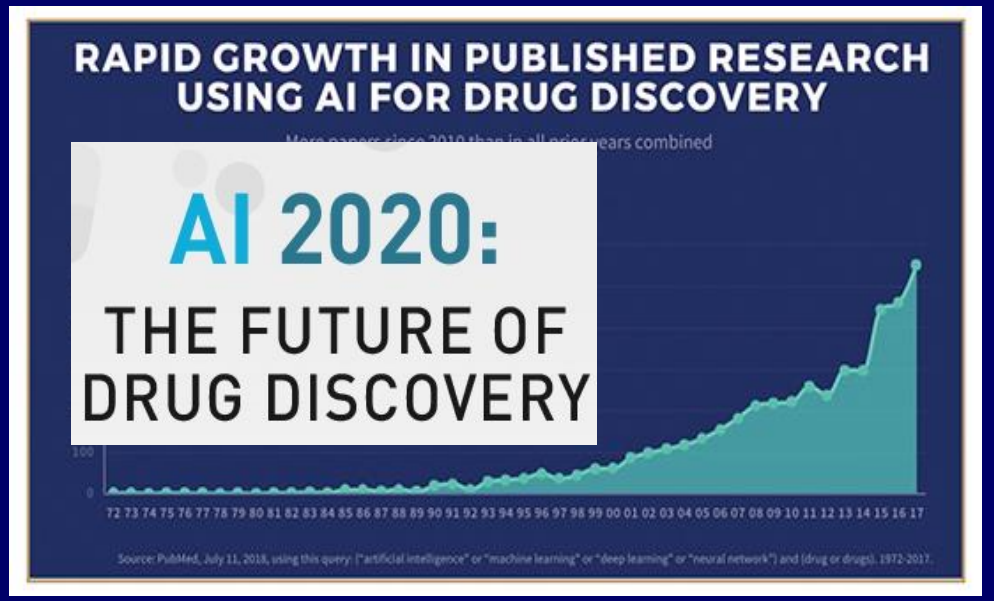
Recent headlines (2018-2020)

SPOTLIGHT · 30 MAY 2018

How artificial intelligence is changing drug discovery

World first breakthrough in AI drug discovery

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



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.”

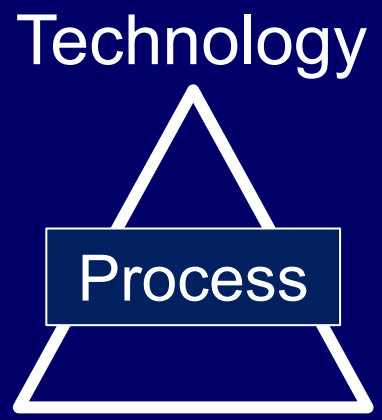
“More drug targets… 3,000–10,000 targets compared with 483”

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

http://www.DrugDiscovery.NET/DataSignal

Discovery Processes and Tools in Chemistry now

- Processes grown out of “technology (pull), cost/societal/legal environment (push), beliefs (into big numbers, analytical deconstruction, etc. ; push)”… plus history



Cost/societal/legal environment

Beliefs (into big numbers, analytical thinking, etc etc.)

Complex processes are often difficult to impossible to change

- Big companies (e.g. pharma) difficult to change processes
  - Change one part, and the whole system breaks down/needs to readjust over a *long* time period
  - Human resistance to change

- Disruption on the way (‘simulations/data first’ companies)
  - Can build processes from scratch
  - Can attract staff that is on board with processes from day 1
  - Advantages *newly built* environments

If you use Artificial Intelligence to support a crappy process you get…a crappy process ‘enhanced’ by AI

Change of tools available requires re-think of the process

NOT

Old process + new tool -> Old process supported by new tool

BUT

Old process + new tool -> New process supported by new tool

Example from drug discovery

1. ‘Anecdotal’ piece of knowledge as starting point (e.g. paper describing gene involved in disease) and
2. Brute force (HTS) to generate molecular starting point, then
3. Optimization by expert knowledge

Can (and *should, in suitable cases, where data > knowledge*) move to

1. Data-driven starting point (say, knowledge graphs, etc etc)
2. Optimization based on data (and expert knowledge)

Caveat: Limitations of **data** in (a) quantity (can be overcome), and (b) predictivity (requires sufficient understanding of system)

Data… the sometimes ugly child

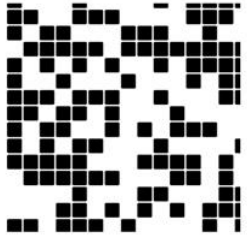
- Core difference between e.g. materials and drug discovery:
  - In materials we (often) known what matters – and we can (often) measure what we need to measure
  - In drug discovery we (rarely) know what matters – and most of our data comes from proxy assays

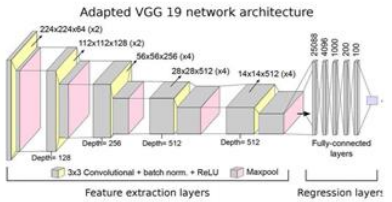
- *Huge* difference for AI in different domains when it comes to generating data, and building models!

Y. Lazebnik, “Can a biologist fix a radio?” Cancer Cell 2002

Object | Representation | Model | Object Label

Image Domain





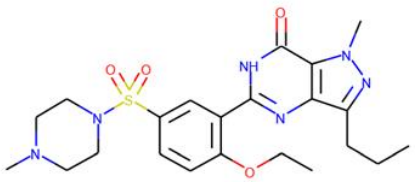
ResNet?
AlexNet?
CapsuleNet?

→ Cat

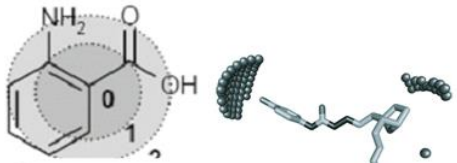
Largely *Unconditional* labels

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

Drug Discovery: Chemical Domain



?



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

?

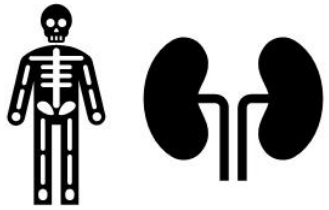
Artificial Neural Network/DNN? Support Vector Machine? Random Forest? Bayesian Classifier?...

→ Property A

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

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)
  - 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?**

Our…

Understanding of predictive endpoints

Influences our

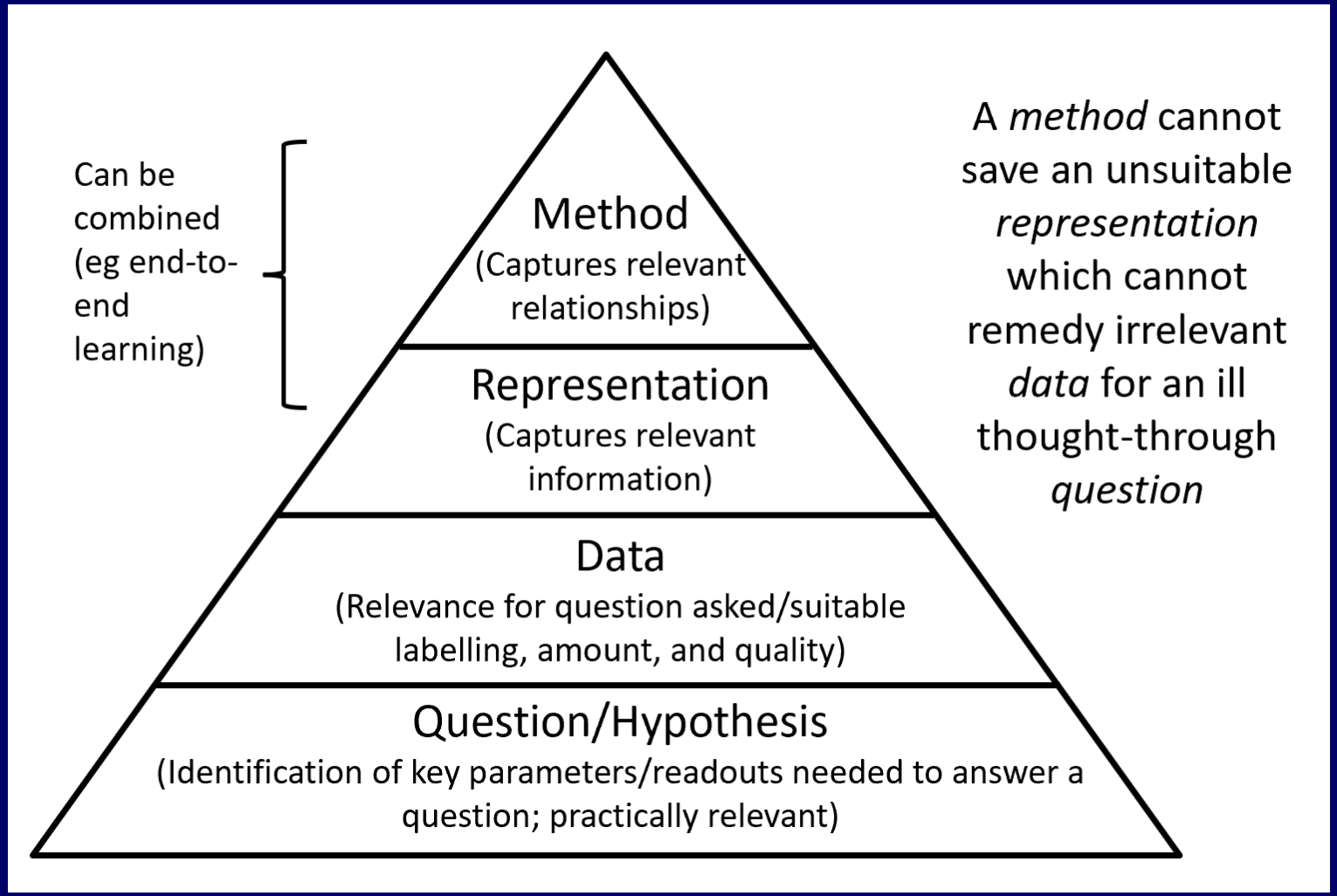
Ability to generate predictive data

Influences our

Ability to train predictive models

The *question* needs to come first… and then the data, then the representation, and then the method

Bender & Cortes Drug Discovery Today 2021



Lots of attention currently here…

But we need to care more about this

What next?

- We need suitable data to support our processes
- To achieve this, we need to understand which data we need (endpoints!), and which data points to generate
- Differences between disciplines (e.g. materials and drug discovery!)

- Requires re-engineering processes *in areas which we sufficiently understand:* Starting with data, and a problem that can be answered (approximated) with the data available

What next (examples from drug discovery)?

- We need *relevant* data (predictive for the *in vivo* situation), which is *possible to generate large-scale*

- ‘omics data: *Yes, but* experimental conditions (e.g. cell line)/dose/time point often don’t extrapolate to relevant situations
- Cellular morphology data: *Yes, but* we need to understand better what the applicability domain is/which interventions are visible in the readout
- Organ-on-a-chip: *Yes (!), but* still under heavy development, details to be seen

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

- Processes grow out of available tools, and taking existing circumstances *at a time* into account
- When tools change, processes need to change – this has only (very) partially been achieved today
- We should not support existing processes with AI, but rather re-think processes from scratch

- Areas differ when it comes to data (amount, labelling, predictivity) – those areas with higher predictivity, better ability to label, and most data will benefit most and earliest from AI
- See also: Bender and Cortes, “Artificial Intelligence in Drug Discovery: What is Realistic, What are Illusions?” Parts 1 and 2, Drug Discovery Today 2021

Thank you for listening!
Any questions?

Email: mail@andreasbender.de
Web: http://www.DrugDiscovery.NET
Twitter: @AndreasBenderUK