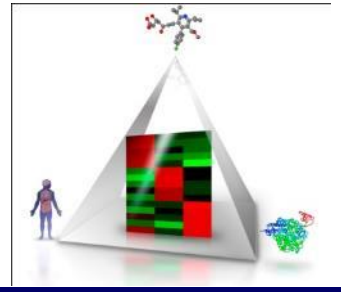
Computational Toxicology – Opportunities, Pitfalls, and What to Watch out for in Model Predictions

Andreas Bender, PhD

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

Director of Digital Life Sciences
Nuvisan, Berlin

UNIVERSITY OF CAMBRIDGE



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

On data, endpoints, models, and predictions

- Preamble: Will ‘AI’ and computational models eat the world? Some historical context

- Data and Endpoints
  - Coverage, conditionality, error, and predictivity
- Models and Validation
  - Descriptors, machine learning, validation
- Predictions and Applications
  - “Questions to ask your friend, the modeller”

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

Fortune cover 1981

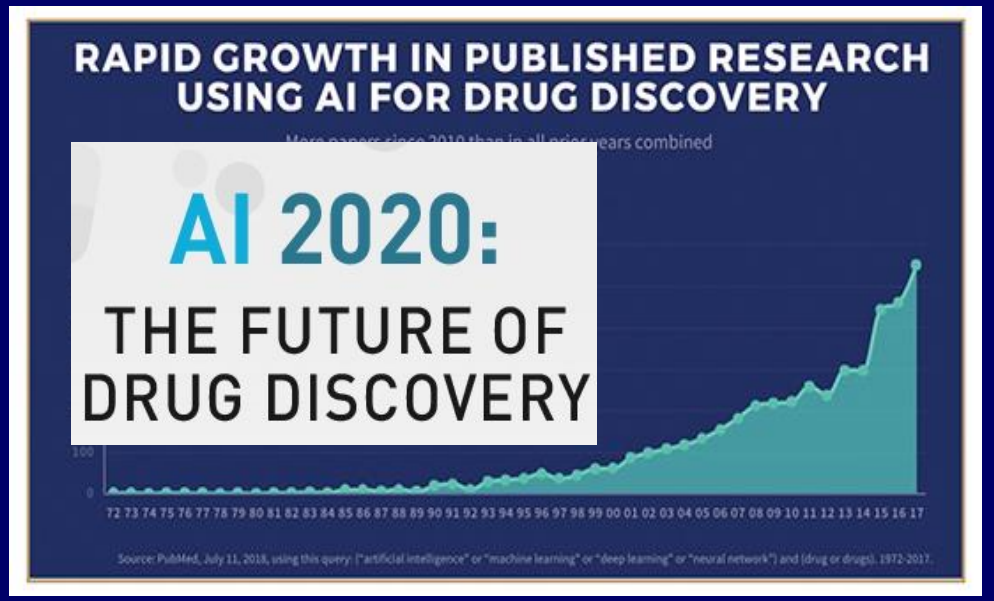
Recent headlines (2018-2020)

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

-> How to go from *technology and potential* to *applications/better decisions*?

-> What are the limitations of what we do, that we need to keep in mind?

Key aspects of computational models (will be revisited later)

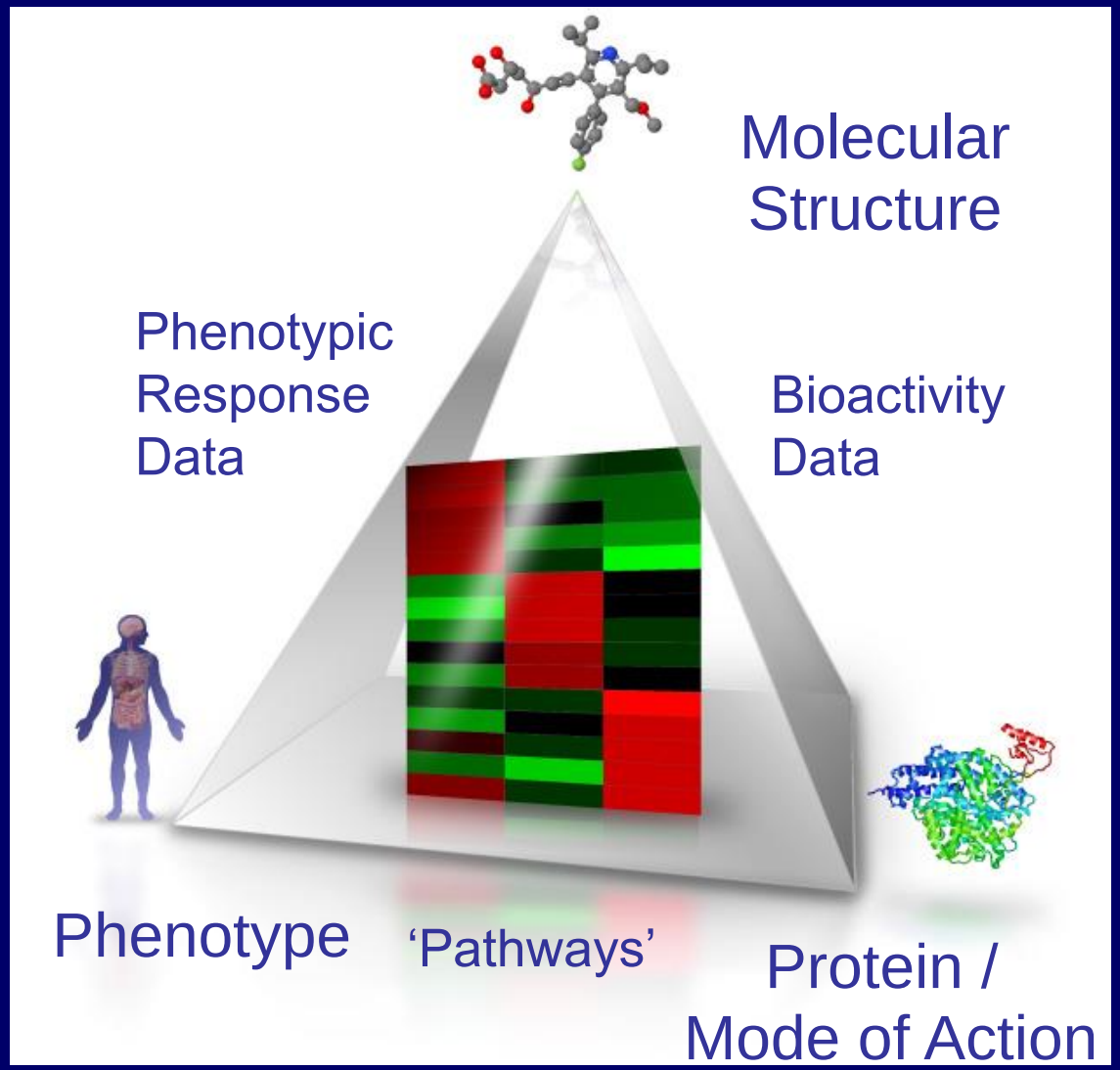
- Models only know what *explicitly* gets described to the model:

- *Input (data) space*
  - E.g. chemical structures, and chemical space coverage, which are in a dataset is all that is available to a model (models don't reason *beyond* that data);
- *Descriptor space*
  - The *representation* of information is the space in which an algorithm reasons (and if molecules are represented in one way, the model will not be able to reason beyond that);
- And *endpoint/output variable space*
  - The algorithm *needs labels*, and 'believes' the numbers we give to it – including assay errors or inconsistent data, *confounding factors*, irrelevant endpoints, …

On data, endpoints, models, and predictions

- Data and Endpoints
  - Coverage, conditionality, error, and predictivity
- Models and Validation
  - Descriptors, machine learning, validation
- Predictions and Applications
  - “Questions to ask your friend, the modeller”

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



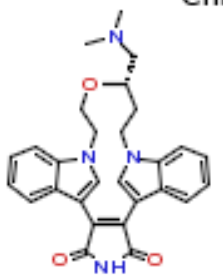
a.k.a.
"The world is flat"

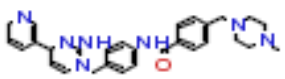
= "We believe our labels"

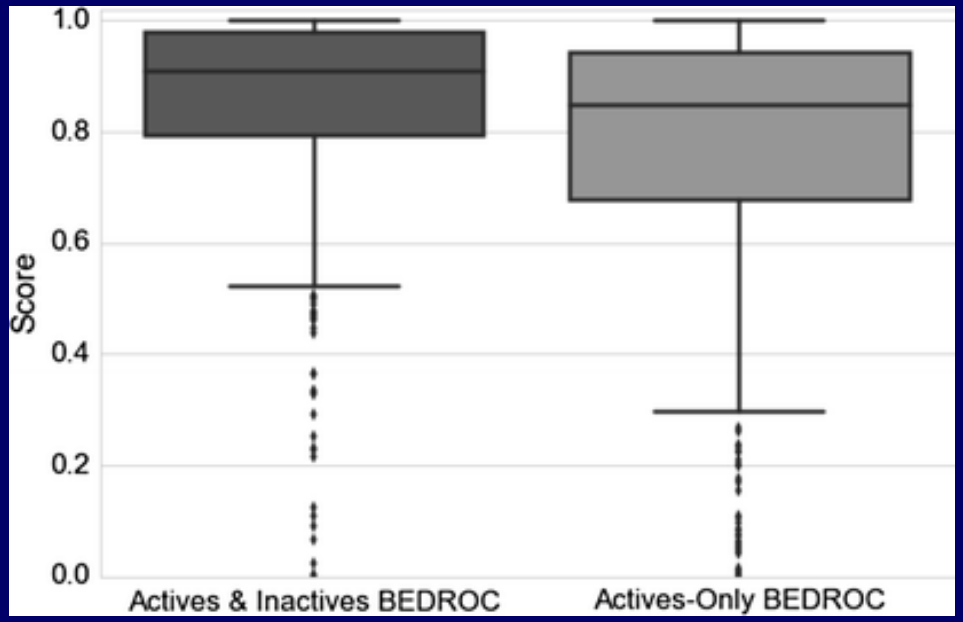
(which are often insufficiently quantified, not directed, unconditional, don't have time/ concentration/biological setup dependence, *etc.*)

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

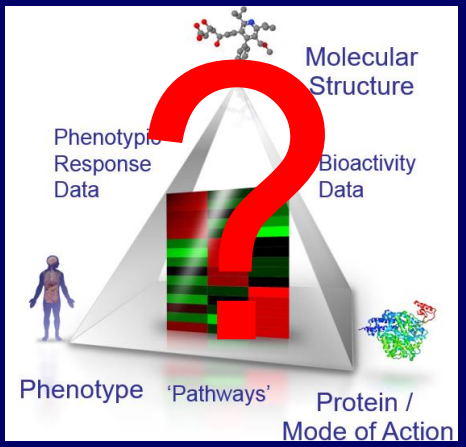
So: Using bioactivity data for ligand-protein activity modelling '*is relatively possible*'

- 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'

Keep in mind:

- 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?



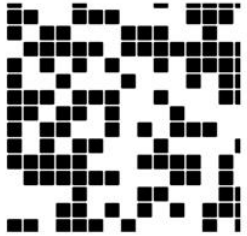
- Links between drugs/targets/diseases are quantitative, incompletely characterized
- Subtle differences in eg compound effects (partial vs full agonists, off-targets, residence times, biased signalling, etc.)
- 'Pathways' from very heterogenous underlying information; dynamic elements not captured etc.
- 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?**

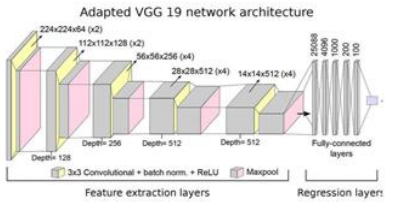
Example of labelling problems: adverse reactions

- ***"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?**

Object | Representation | Model | Object Label

Image Domain





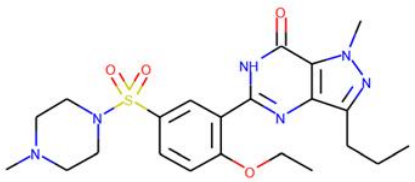
ResNet?
AlexNet?
CapsuleNet?

→ Cat

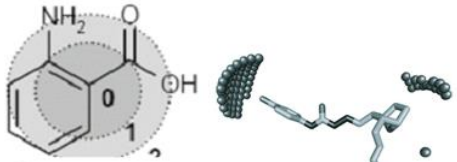
Largely *Unconditional* labels

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

Drug Discovery: Chemical Domain



?



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

?

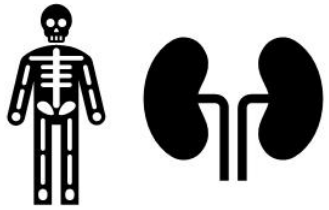
Artificial Neural Network/DNN? Support Vector Machine? Random Forest? Bayesian Classifier?...

→ 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

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!)

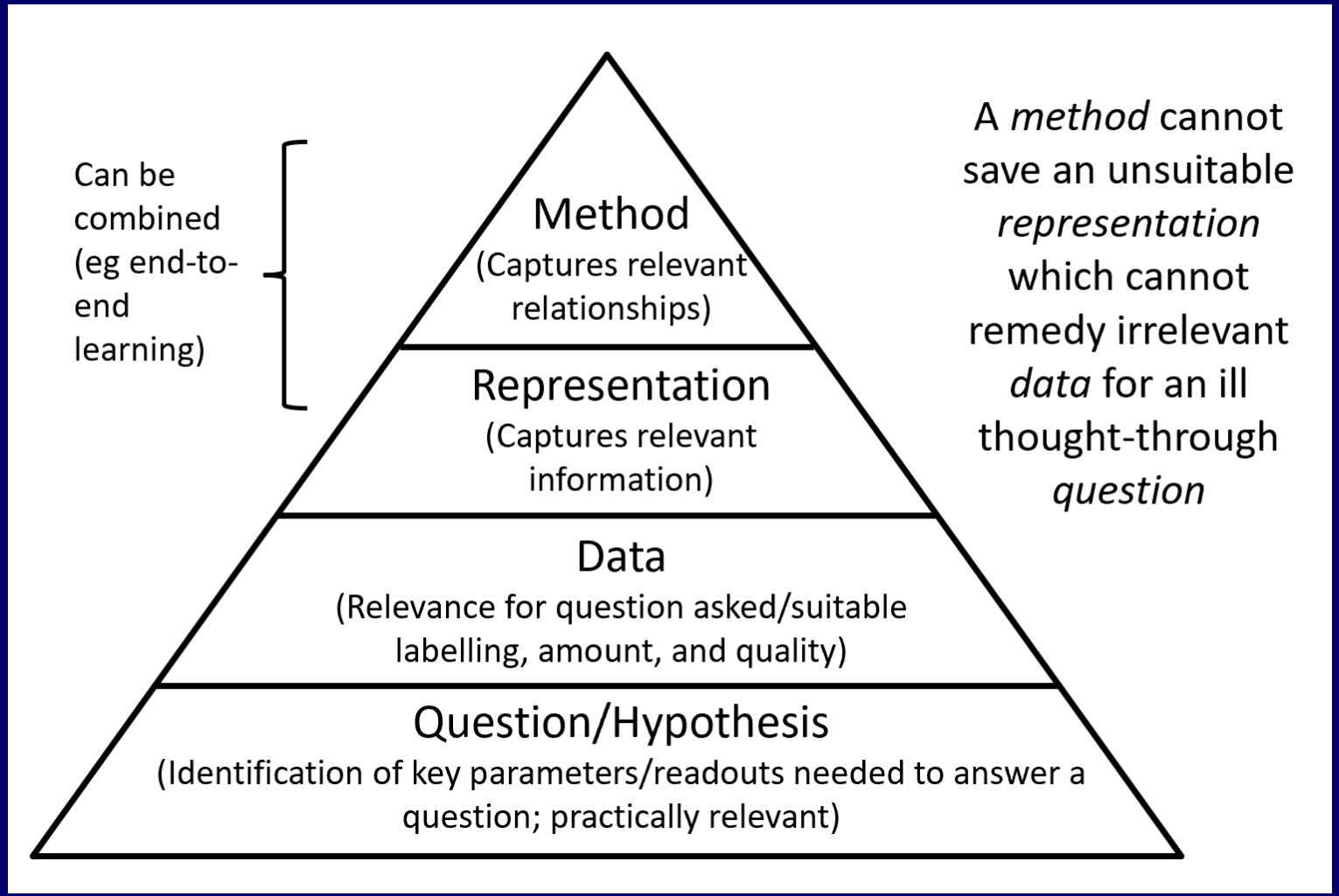
Much of the data we generate is generated for the wrong reasons (or in wrong ways)

- We generate data in *proxy space*, but wish to predict for *in vivo safety/efficacy space*
- Historical data gets now repurposed ‘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
- 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*

-> We need to care more about the data we generate for modelling!

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



Lots of attention currently here…

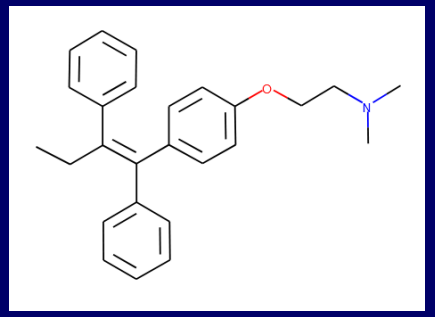
But we need to care more about this

On data, endpoints, models, and predictions

- Data and Endpoints
  - Coverage, conditionality, error, and predictivity
- Models
  - Descriptors, machine learning, validation
- Predictions and Applications
  - “Questions to ask your friend, the modeller”

What is a computational model?

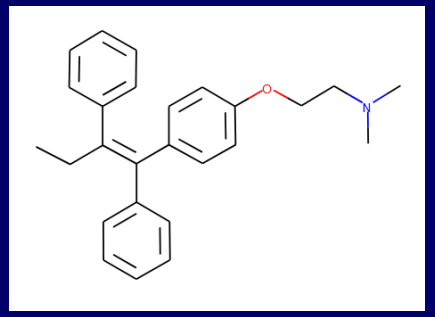
We have (from experiments): Molecule -> Endpoint

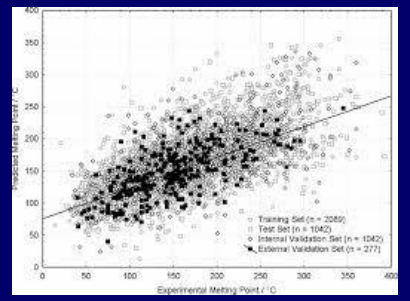


Measured (experiment)

IC50= ..nM

We model: Molecule -> Descriptor -> Model -> Endpoint

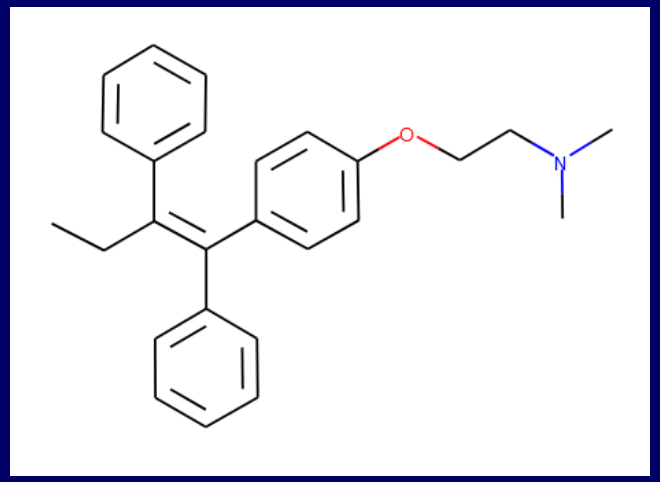


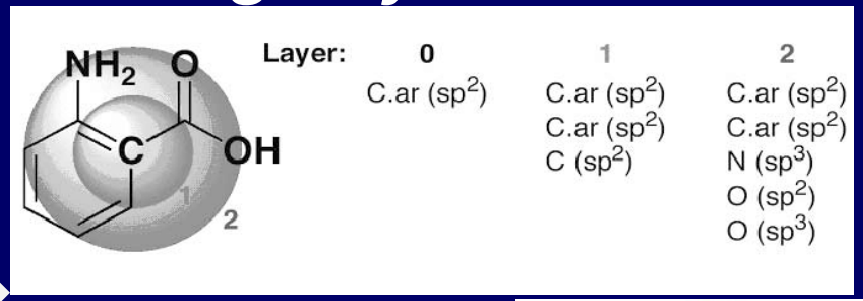


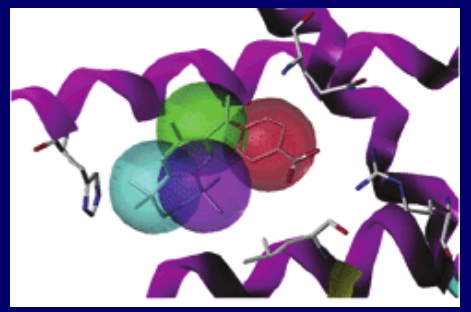
IC50= ..nM

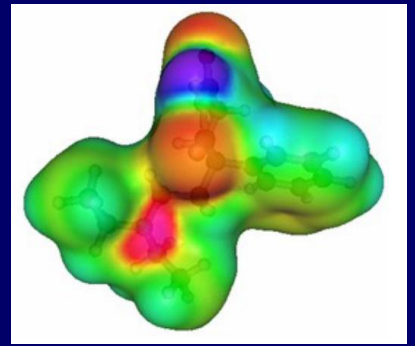
Descriptors

- Provide an *information-preserving* representation of input data (e.g. structures) for the model
- Either knowledge-based (e.g. reactive groups), or (usually) ‘trial and error’
- *Can* be learned from data, but only *if there is enough data, and we can meaningfully label!*







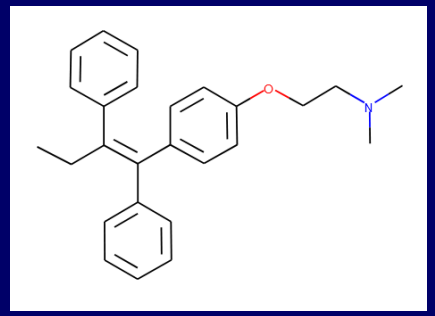


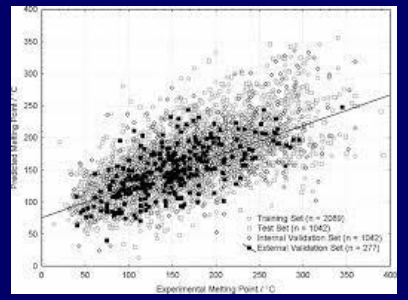
0100101010000…

Fingerprints, pharmacophores, surface properties, substructures/ functional groups, shapes, physchem properties *etc.*

Model: Fit of free model parameters (functional model form can be based on knowledge!) to data

We model: Molecule -> Descriptor -> Model -> Endpoint





IC50= ..nM

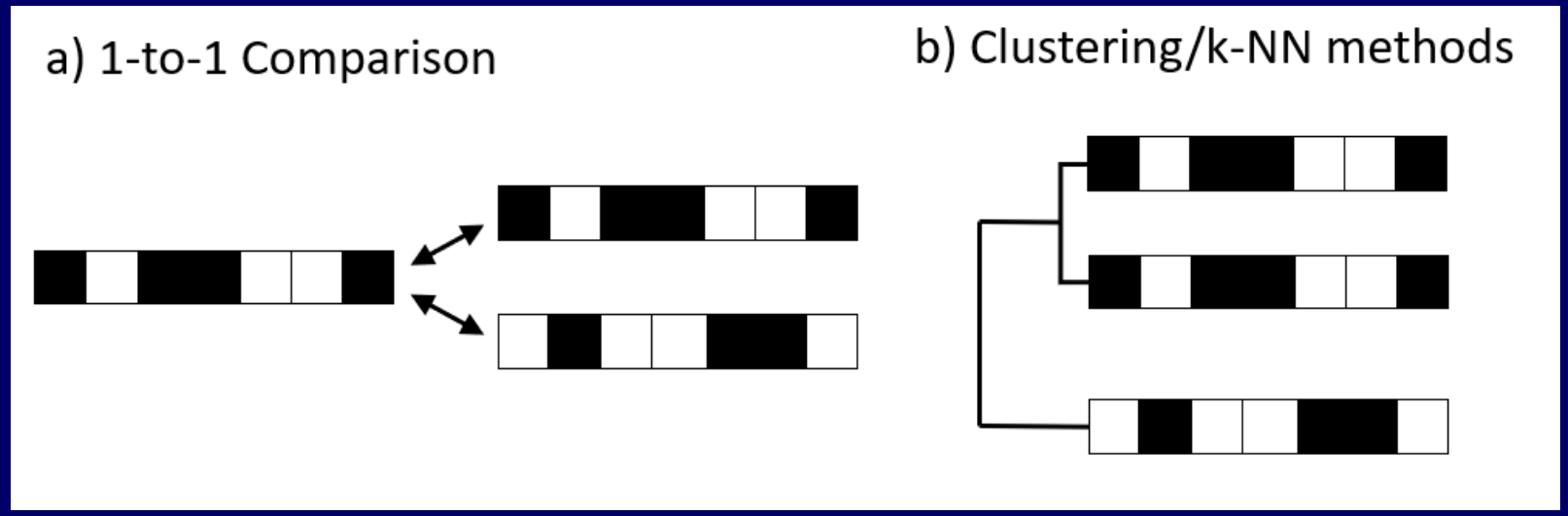
Two things can be done with a model

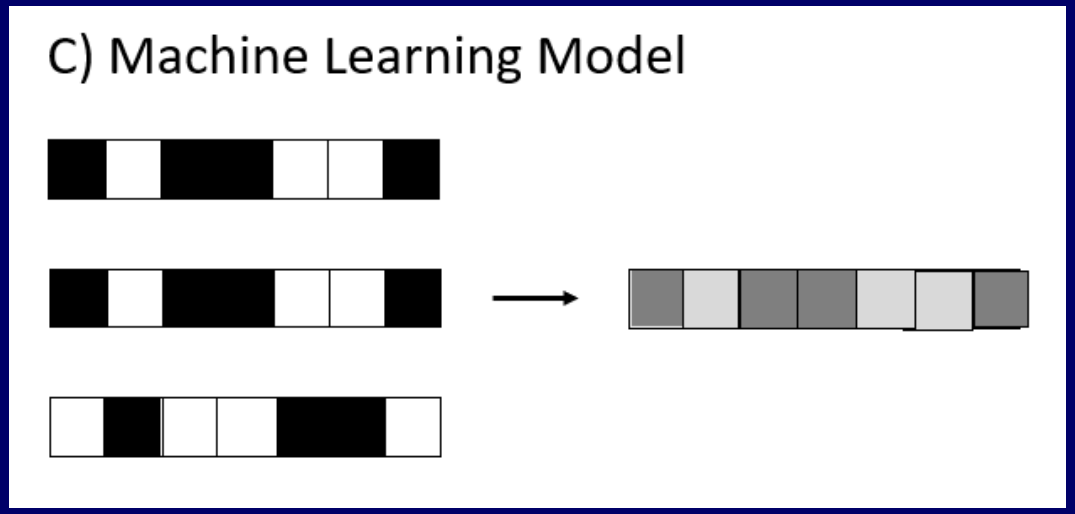
- Training: Fit model to represent experimental endpoints (involves *choice of loss function*, eg RMSE, accuracy, …)
- Application/Test: Predict for any/novel molecules

Validation: Repeat training/test on different data

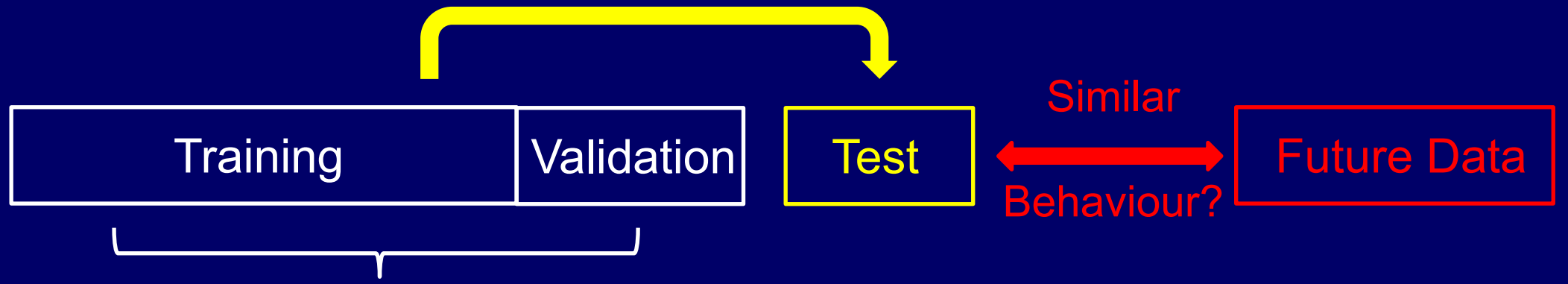
Types of models (all of which can involve feature selection)

- Similarity-based (single neighbour, 1-NN)
- Similarity-based (multiple neighbour, k-NN)
- Machine learning models





Validation … and why it is tricky



- Model validation aims, *based on existing data*, to both optimize the model, *and* to provide estimate of future performance

- Problems:
  - Aims *collide* (we aim for high numbers in one case, and realistic ones in another)!
  - We have only a *given* dataset available for us, and we generally do not know if it resembles *truly prospective* applications of the model

Pitfalls in model validation

- Training/test set split – too small, coverage irrelevant
- Prospective validation – too small, and biased (process!)
- Baseline model not well-chosen/optimized
- Data quality not assessed
- Relevance of model endpoint not assessed
- Result of *process* of validation ascribed to *model*

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

On data, endpoints, models, and predictions

- Data and Endpoints
  - Coverage, conditionality, error, and predictivity
- Models
  - Descriptors, machine learning, validation
- Predictions
  - “Questions to ask your friend, the modeller”

Predictions and what to watch out for

- Computational models take data, in a representation, to (a) parametrize a model, and (b) make a prediction

- *The* Assumption: Rules learned during training apply during model application!

- How can we gain confidence that this is the case?

Questions to ask your friend, the modeler (1/2)

- Key goal: How good is the prediction for *my new compound?*
- Data
  - What is the *number* of data points in the model, and is chemical space coverage *relevant* for my application?
  - What is the closest neighbour (according to mechanistically interpretable space; model space; similarity space), and *is it relevant, given the particular question being asked*?
- Descriptors
  - How was the descriptor chosen, and is there a mechanistic rationale for its choice? (depends on understanding of system; e.g. reactive substructures, bioactivity-based, generic similarity, …)

Questions to ask your friend, the modeler (2/2)

- Models
  - Was there an external test set used in model validation (and was it large, diverse, *relevant to new compound predictions*)?
  - Does model performance change, depending on parameter choices (indicates model instability), and training/test set splits (indicates overfitting)?
  - Is there an applicability domain/confidence that the model assigns – and *does it actually work on the external test set (rather often it does not!)*?
- If all of this is answered satisfactorily, then (a) data in the model covers my new molecule, with (b) a suitable descriptor, and provides (c) a confidence with the prediction

Summary

- Computational models are probably *generally helpful* to find patterns in datasets, also in computational safety
- Limitations of *data*, the *descriptors chosen*, and the *validation process* needs to be understood to put model predictions and performance into context
- Limitations of data when generating models come from at least two sources:
  - Input space (e.g. chemical space coverage)
  - Output space (relevance of endpoint, relevance of experimental assay-setup, assay error, *confounding factors*, …)
- Some general guidelines of questions to ask a modeller regarding the confidence one should put into a model have been outlined here

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