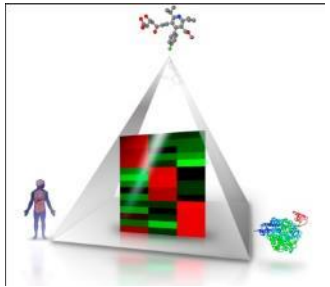
On Chemical Descriptors, Models, Model Validation, and Merging Structural and –Omics Information

Andreas Bender, PhD

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

Chief Technology & Information Officer
PangeAI, part of Pangea Botanica, London/Berlin

UNIVERSITY OF CAMBRIDGE



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

Slides available from http://www.andreasbender.de

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

On data, endpoints, models, and predictions

- Data and endpoints
    - Coverage, conditionality, error, and predictivity

- Descriptors and models
    - Descriptors, machine learning, supervised vs. unsupervised methods

- Validation and application
    - Problems, “Questions to ask your friend, the modeller”

- Merging information from structure and -omics

On data, endpoints, models, and predictions

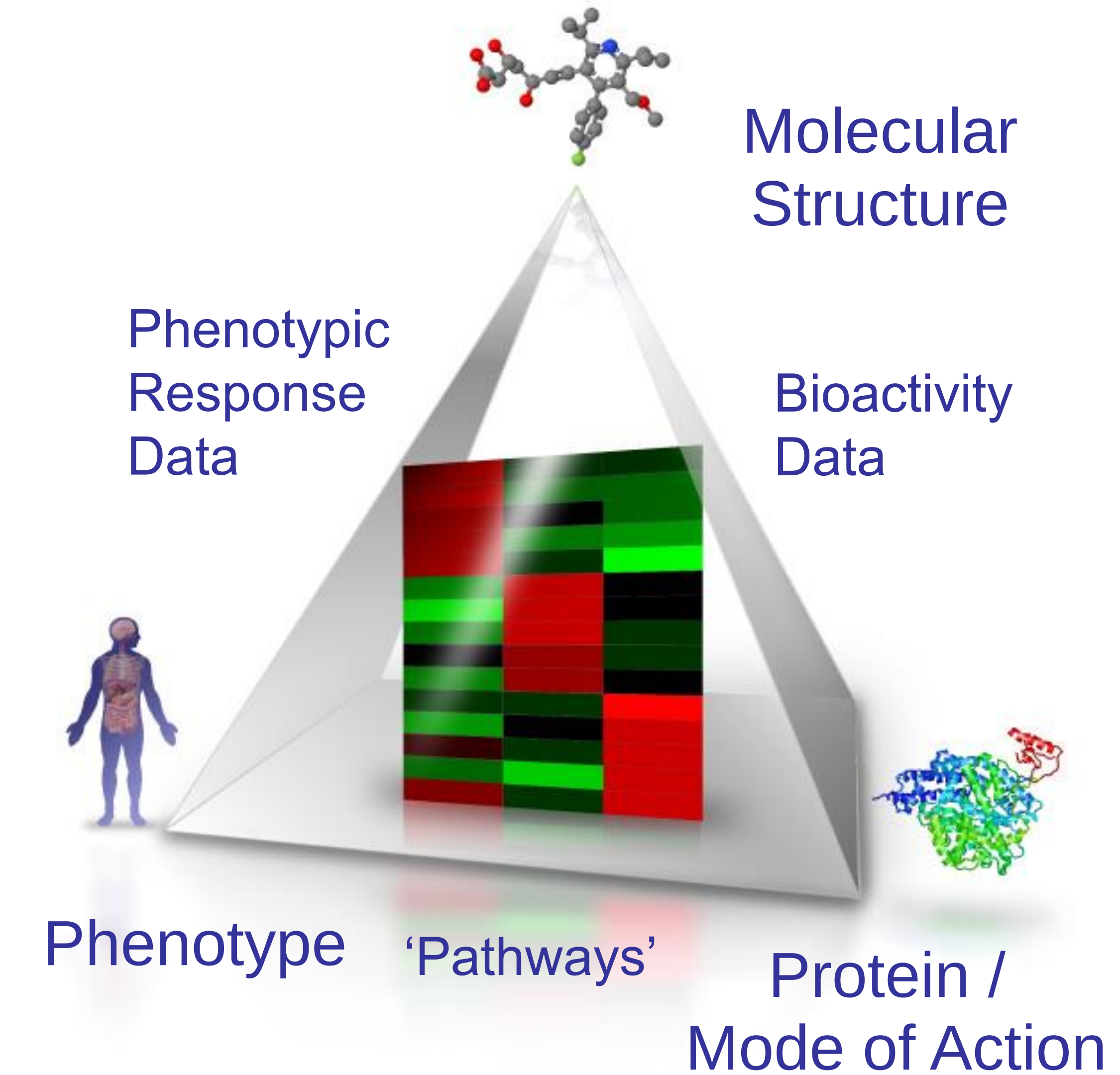
- Data and endpoints
  - Coverage, conditionality, error, and predictivity

- Descriptors and models
  - Descriptors, machine learning, supervised vs. unsupervised methods

- Validation and application
  - Problems, “Questions to ask your friend, the modeller”

- Merging information from structure and -omics

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



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

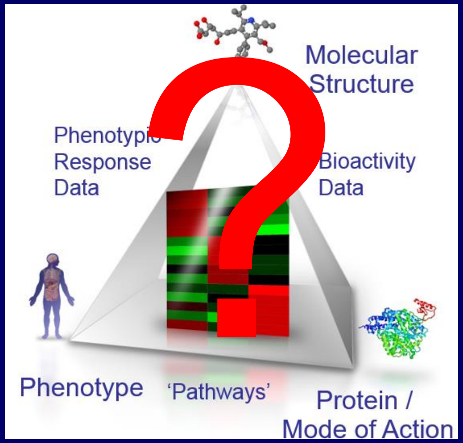
= "We believe our labels"

"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.*

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?**

Key point: We often cannot label our data properly in the life sciences

- Machine learning/AI knows *unsupervised* or *supervised* methods
- Predictive methods are (usually) supervised, and need data points with *labels* (active/not active; or quantitative labels, etc.)
- Those labels need to come from *experiments*

- Experiments (and hence labels) often either fall into the ‘large-scale, but little *in vivo* relevance’ or ‘*in vivo* relevant, but small scale and conditional’ category
- **This is a problem** for AI/ML in drug discovery and safety

- So should we use and analyze our data? Absolutely!
- But we need to work towards *in vivo* relevance of data, jointly

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)

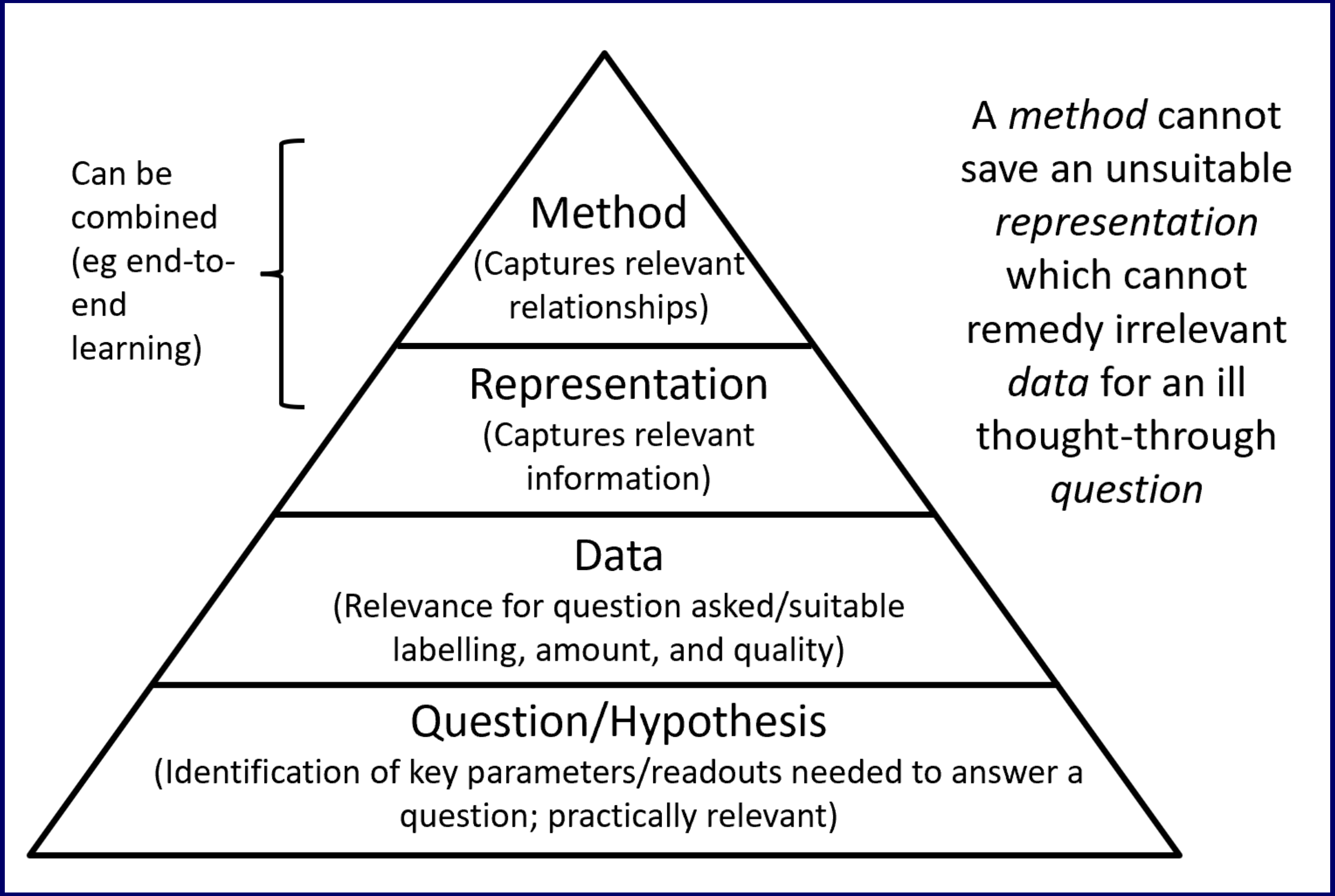
- *Predictive* models (more tricky than generative!)

Much of the data we 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

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

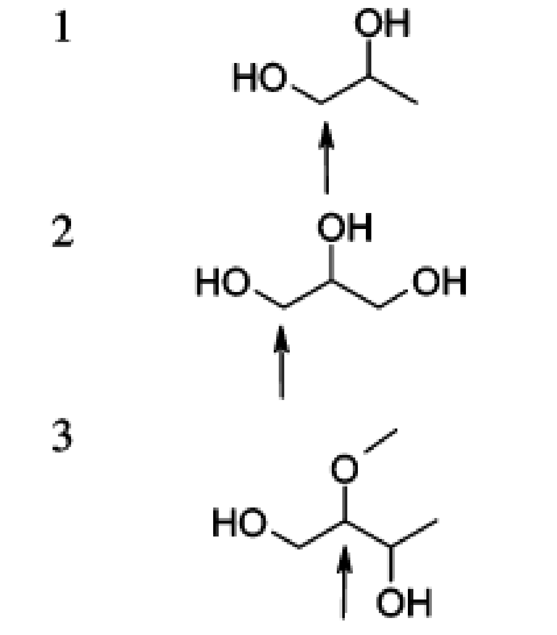
But we need to care more about this

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.

On data, endpoints, models, and predictions

- Data and endpoints
  - Coverage, conditionality, error, and predictivity

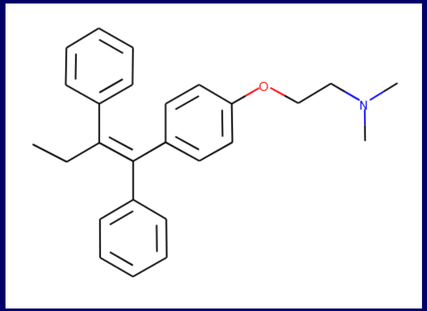
- Descriptors and models
  - Descriptors, machine learning, supervised vs. unsupervised methods

- Validation and application
  - Problems, “Questions to ask your friend, the modeller”

- Merging information from structure and -omics

What is a computational model?

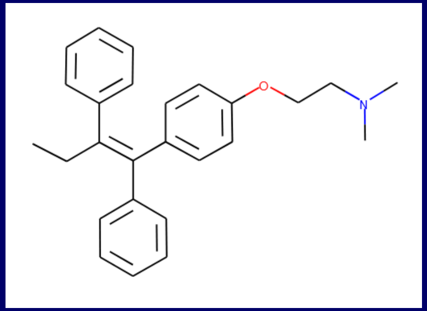
We have (from experiments): Molecule -> Endpoint



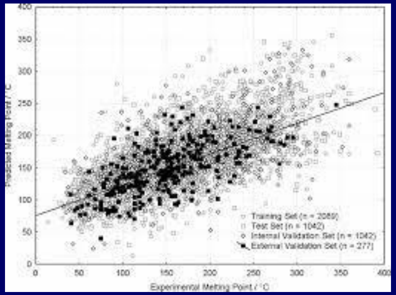
Measured (experiment)

IC50= ..nM

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



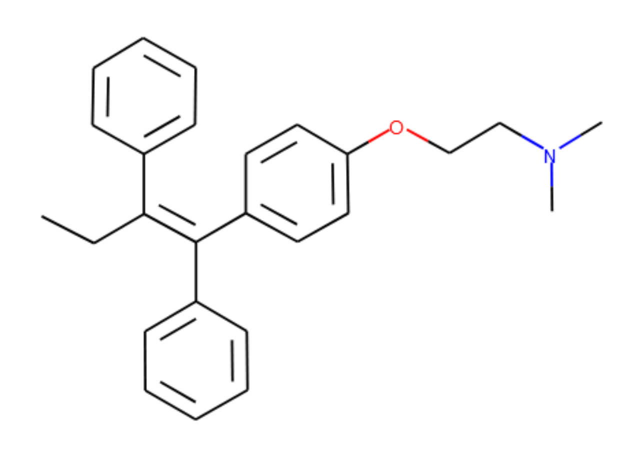
NH2
O
La
C
OH
1
2

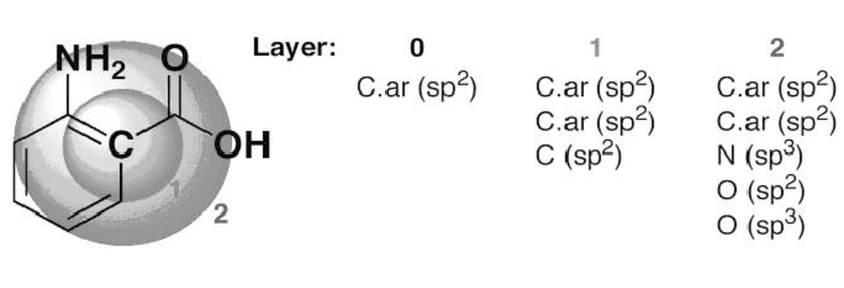


IC50= ..nM

Descriptors

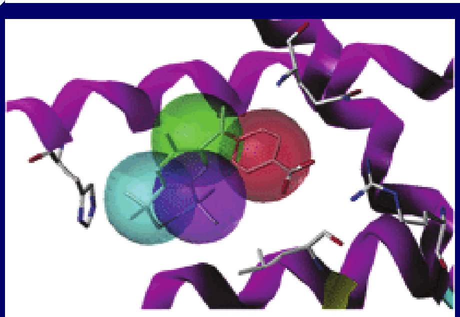
- Provide an *information-preserving* representation of input data (e.g. structures) for the model
- Either knowledge-based (e.g. reactive groups), or (usually) ‘trial and error’
- *Can* be learned from data, but only *if there is enough data, and we can meaningfully label!*

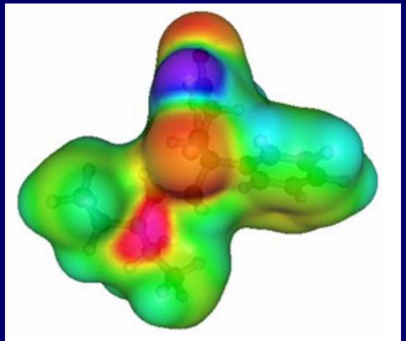




Layer:	0	1	2
	C.ar (sp^2)	C.ar (sp^2) C.ar (sp^2) C (sp^2)	C.ar (sp^2) C.ar (sp^2) N (sp^3) O (sp^2) O (sp^3)

0100101010000…

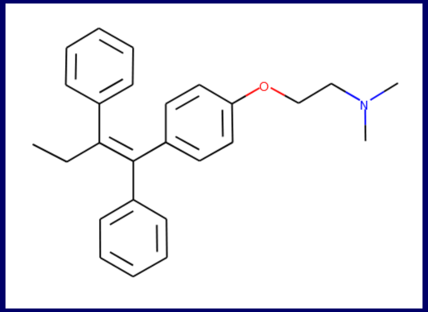


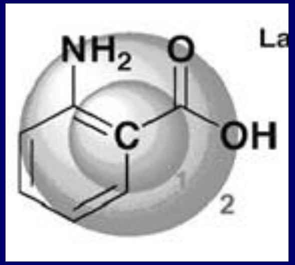


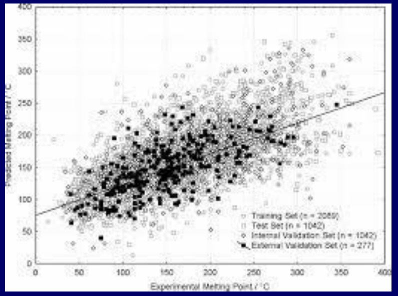
Fingerprints, pharmacophores, surface properties, substructures/ functional groups, shapes, physchem properties *etc.*

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

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





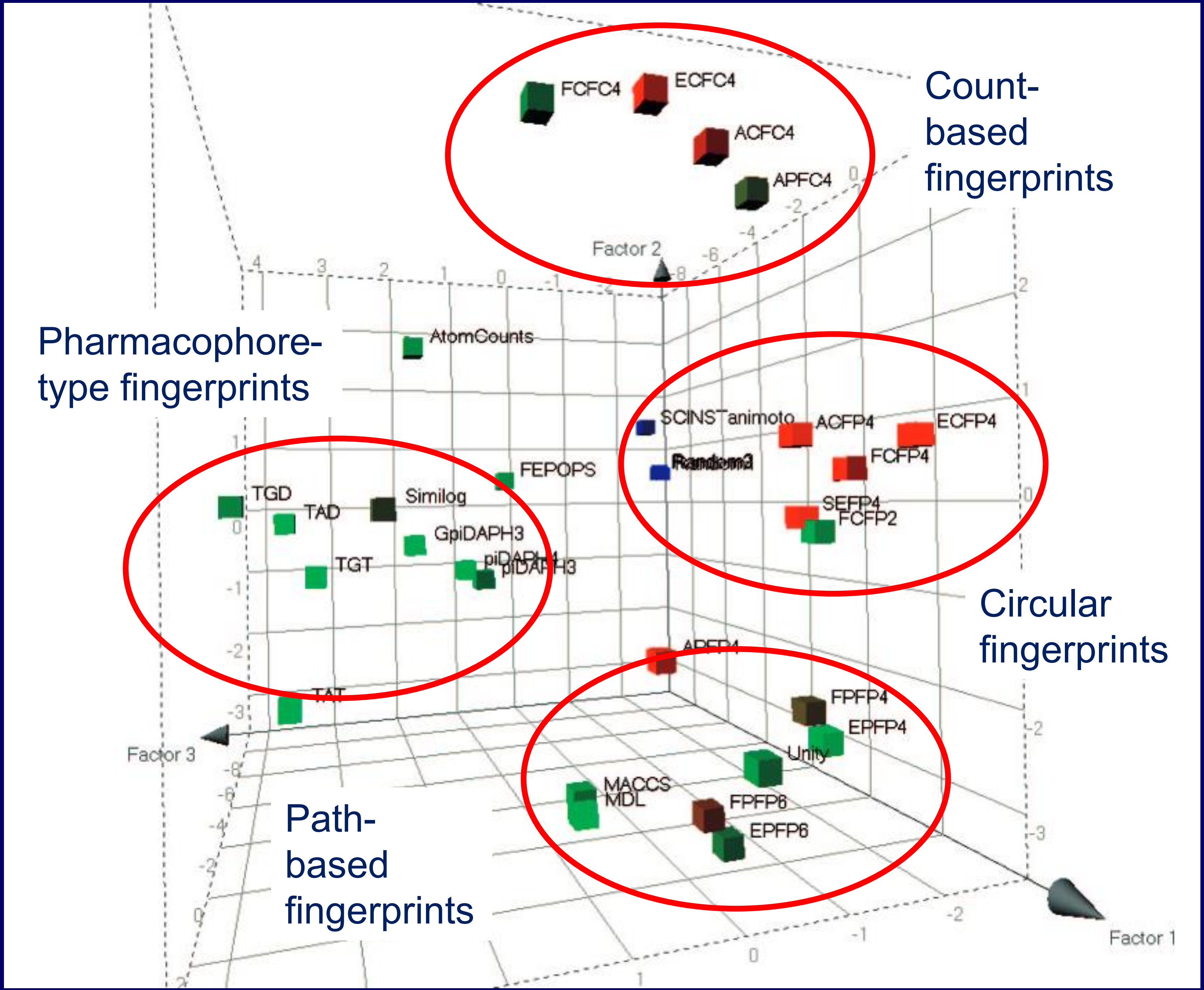


IC50= ..nM

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



Count-based fingerprints

Pharmacophore-type fingerprints

Circular fingerprints

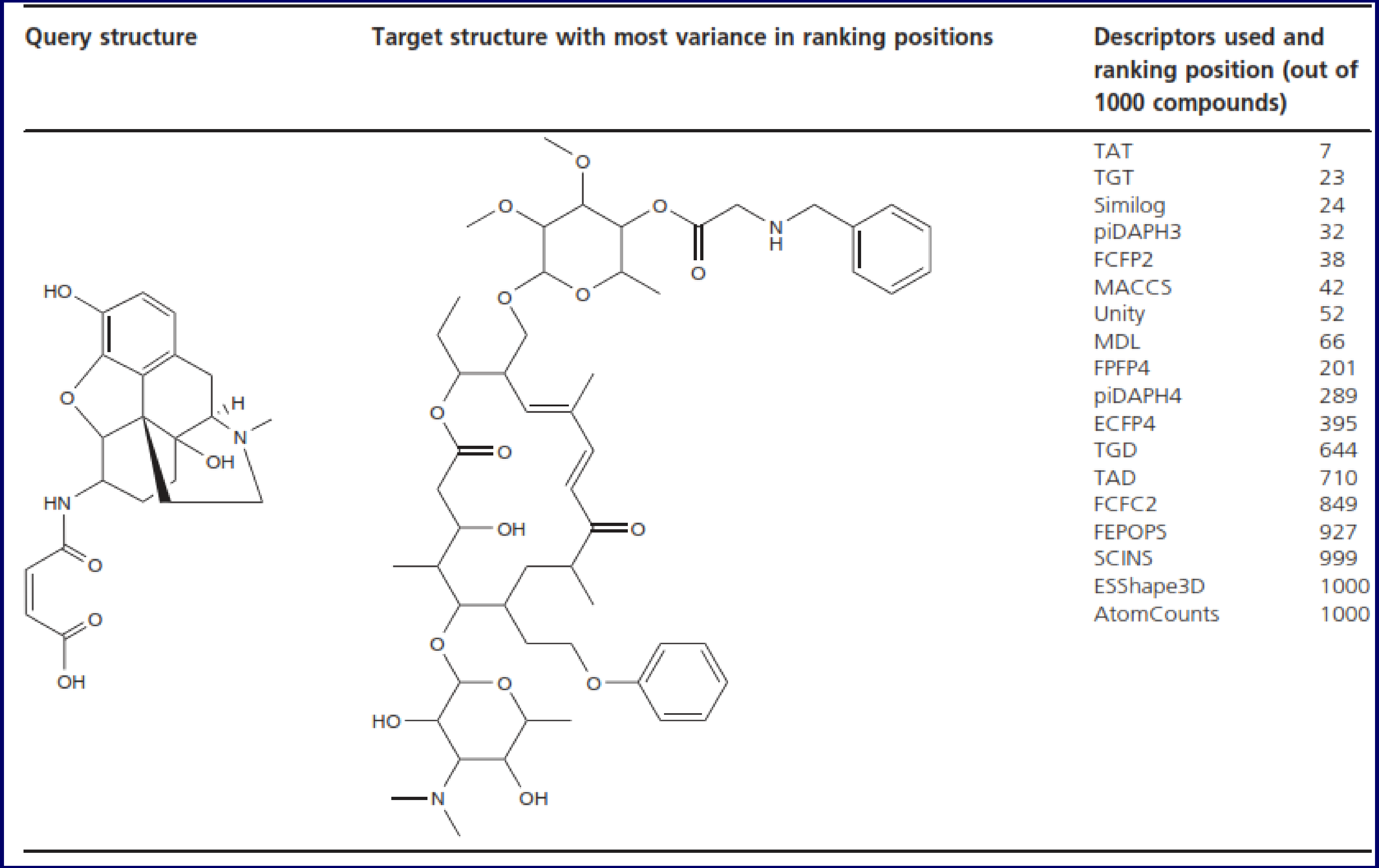
Path-based fingerprints

Generic descriptors behave differently!

So: What do you need?

Bender et al. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space J. Chem. Inf. Model. 2009, 49, 108–119

Generic descriptors – example of differences



Query structure	Target structure with most variance in ranking positions	Descriptors used and ranking position (out of 1000 compounds)	
		TAT	7
		TGT	23
		Similog	24
		piDAPH3	32
		FCFP2	38
		MACCS	42
		Unity	52
		MDL	66
		FPFP4	201
		piDAPH4	289
		ECFP4	395
		TGD	644
		TAD	710
		FCFC2	849
		FEPOPS	927
		SCINS	999
		ESShape3D	1000
		AtomCounts	1000

E.g. some consider size, some are oblivious to it…

A. Bender, Exp. Op Drug Discov. 2010

… some look at feature distributions, some look at local environments (but different typing!)

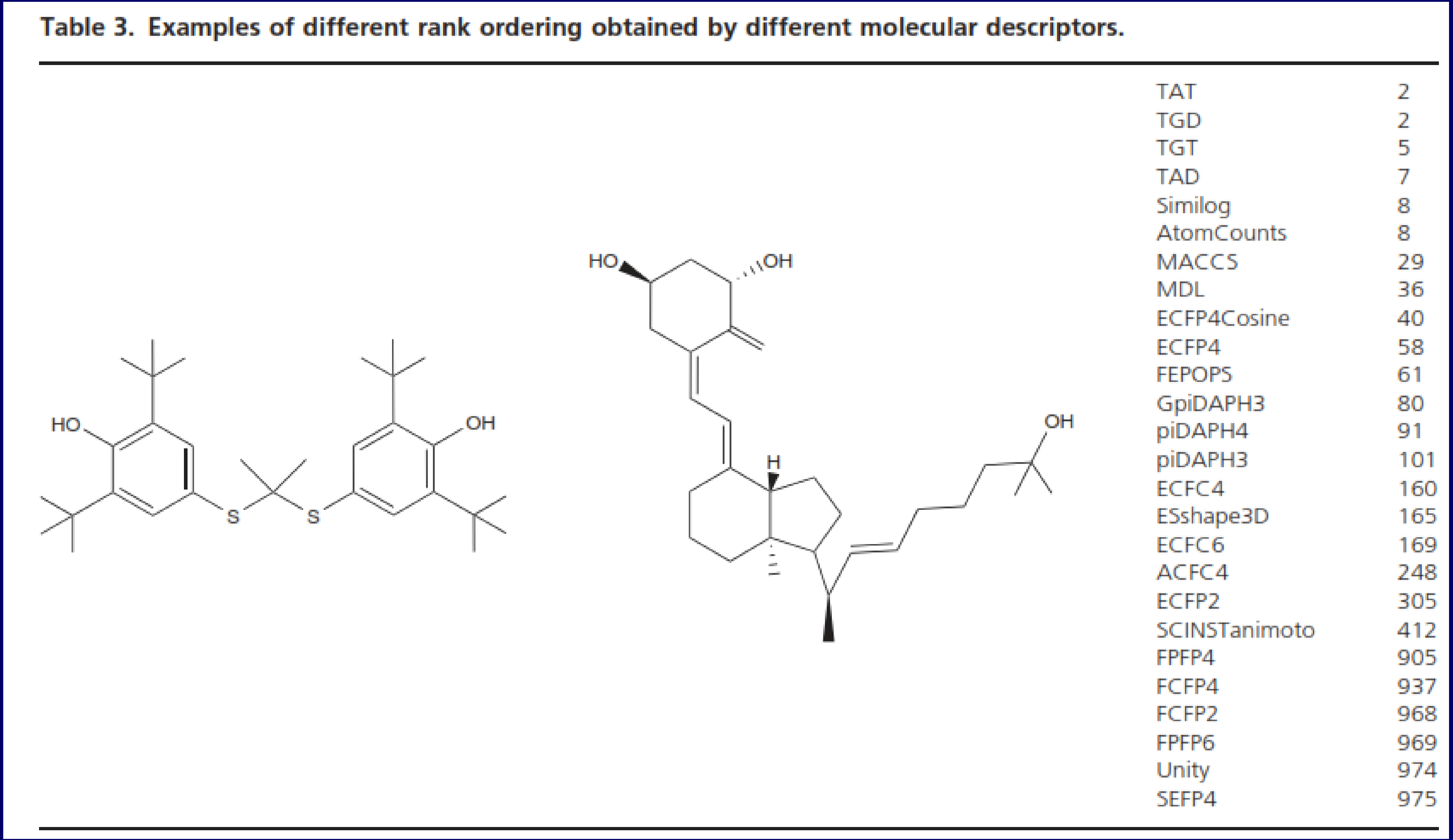
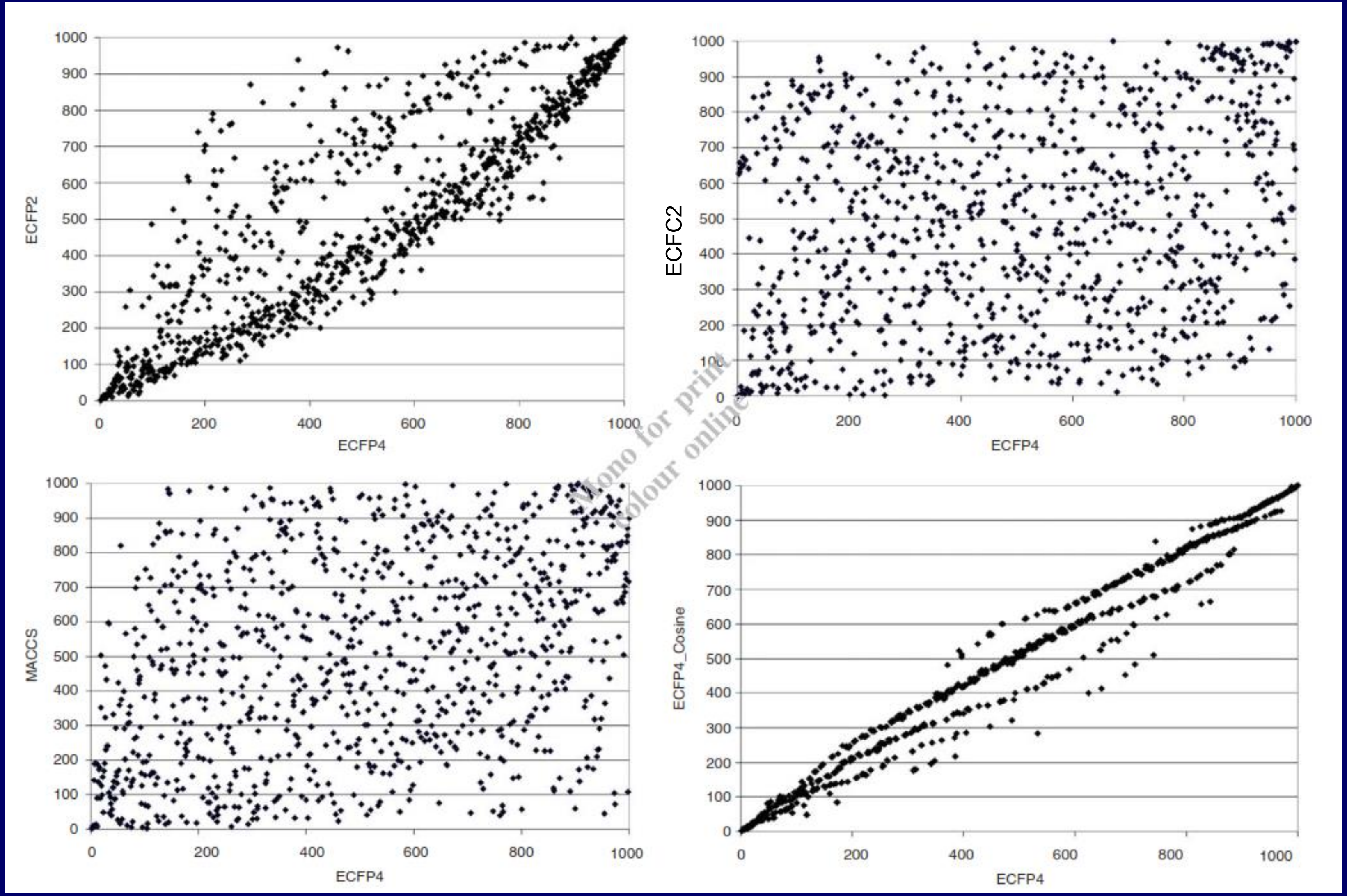


Table 3. Examples of different rank ordering obtained by different molecular descriptors.

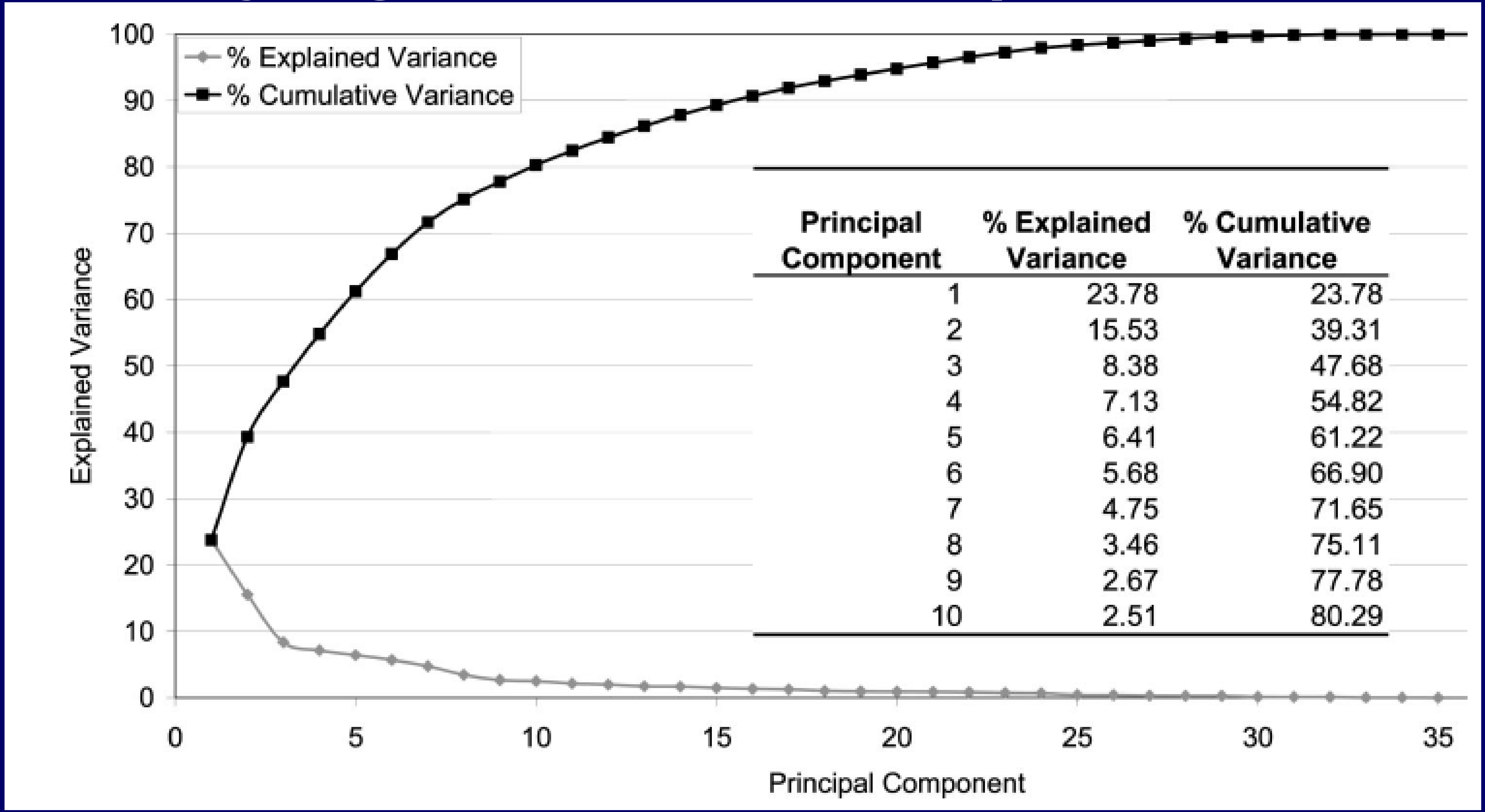
HO
OH
S
S
HO
OH
H
OH

TAT	2
TGD	2
TGT	5
TAD	7
Similog	8
AtomCounts	8
MACCS	29
MDL	36
ECFP4Cosine	40
ECFP4	58
FEPOPS	61
GpiDAPH3	80
piDAPH4	91
piDAPH3	101
ECFC4	160
ESshape3D	165
ECFC6	169
ACFC4	248
ECFP2	305
SCINSTanimoto	412
FPFP4	905
FCFP4	937
FCFP2	968
FPFP6	969
Unity	974
SEFP4	975

Parameters (e.g. ECFP4/ECFP2) and similarity coefficients (e.g. Cosine vs Tanimoto) matter less; nature of descriptor does (MACCS vs ECFP4; ECFP4 vs ECFC2)



‘Chemical space is ca 15-dimensional (to explain 90% of variance)’ – *but variance is not necessarily signal related to output variable!*



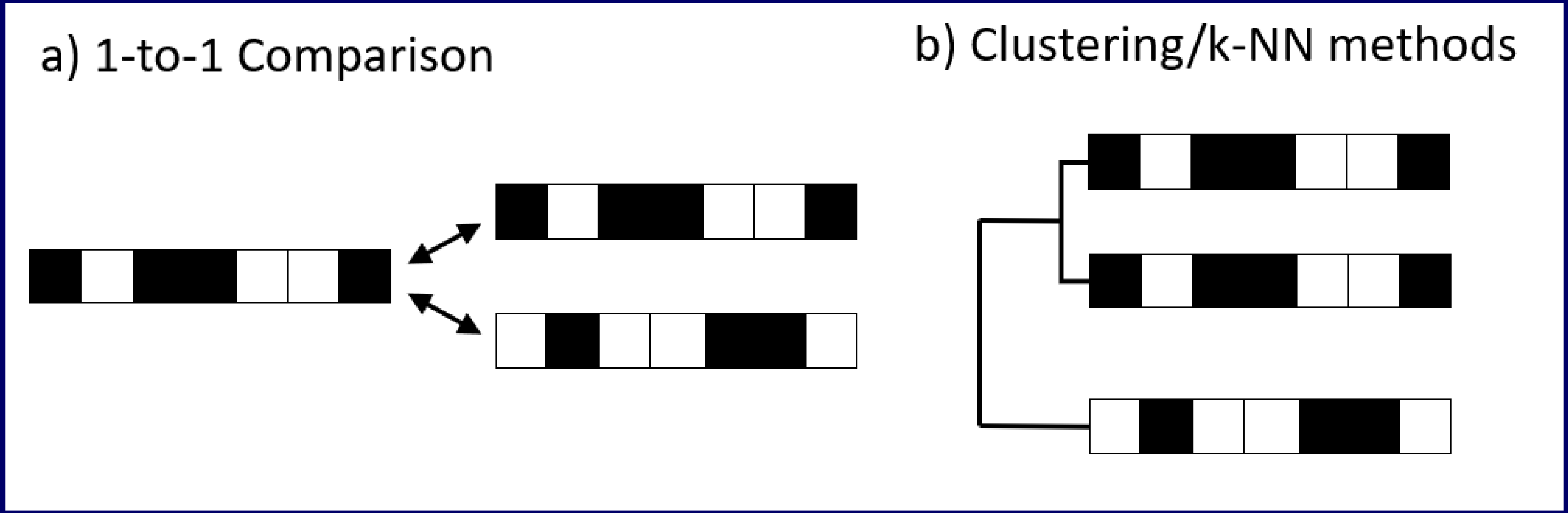
Principal Component	% Explained Variance	% Cumulative Variance
1	23.78	23.78
2	15.53	39.31
3	8.38	47.68
4	7.13	54.82
5	6.41	61.22
6	5.68	66.90
7	4.75	71.65
8	3.46	75.11
9	2.67	77.78
10	2.51	80.29

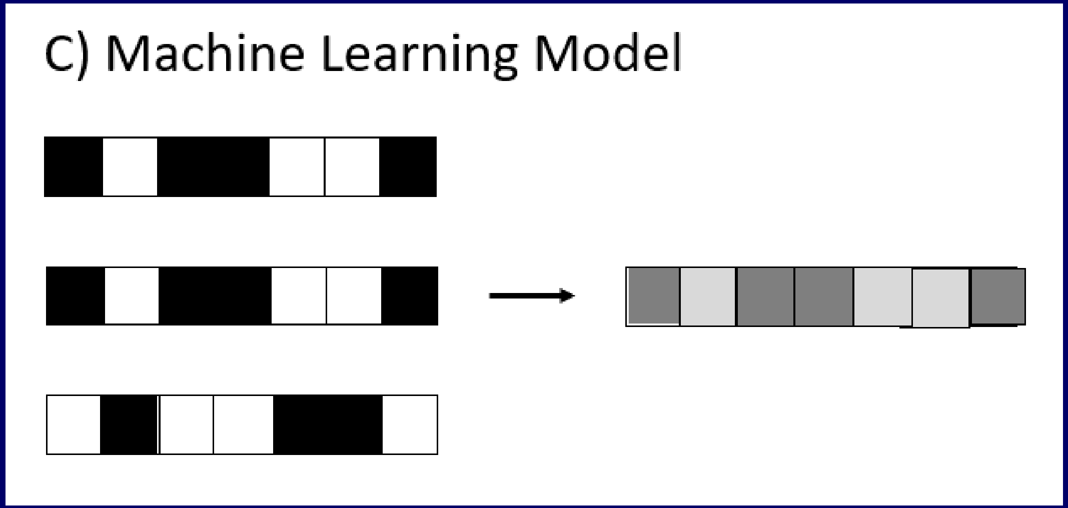
Types of models (all of which can involve feature selection)

- Similarity-based (single neighbour, 1-NN)

- Clustering-based (multiple neighbour, k-NN)

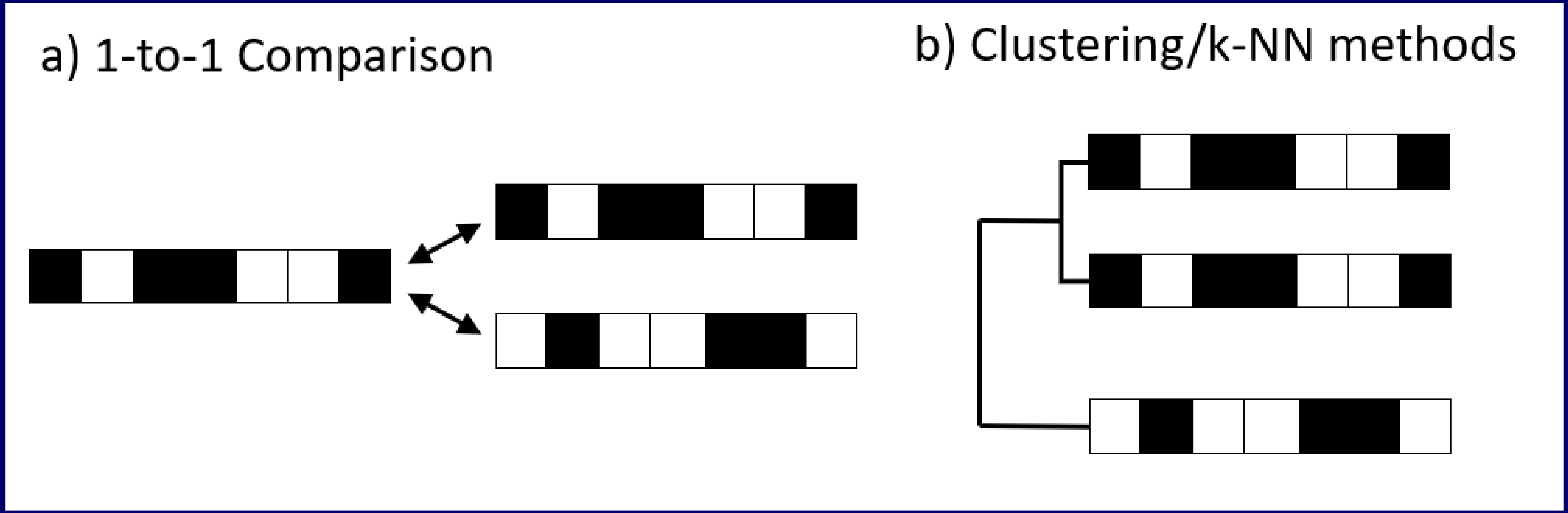
- Machine learning models



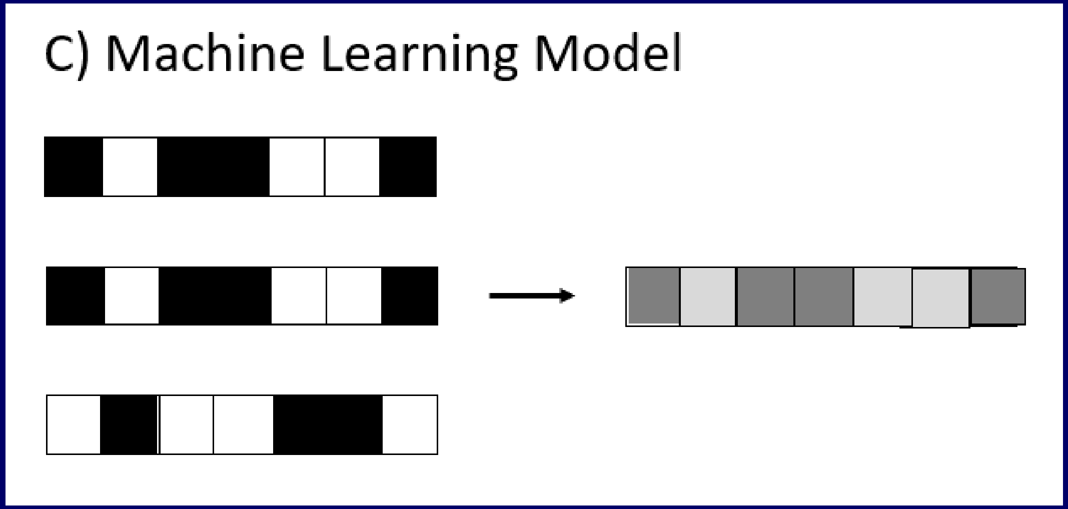


Types of models (all of which can involve feature selection)

- Unsupervised models take features as they are



- Supervised methods fit relative feature importance and conditionalities, based on data



Similarity depends on context (!)

- This is what a model *can* give you

But:

- Usually not enough data available (e.g. for learned representations)
- *Conditionality* of feature changes hence not captured
- Data is often not in vivo relevant

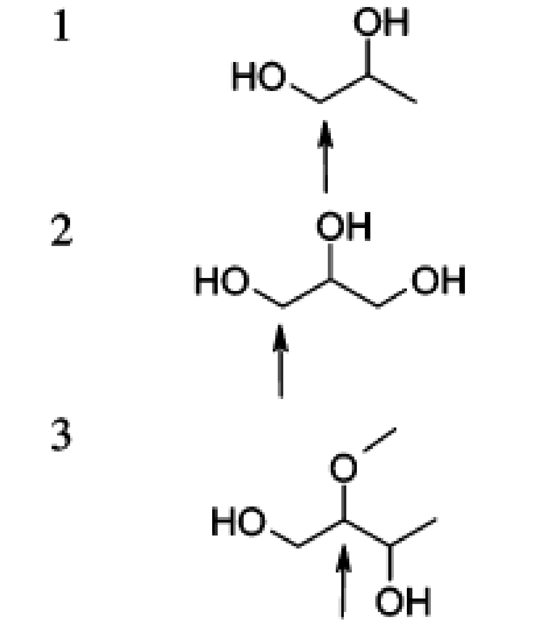
Possible solution: Merging prior information (mechanistic understanding) and data (e.g. Bayesian methods)

Remember: Don’t always trust the features that supervised models select – it all depends on the data you use!

- 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.

On data, endpoints, models, and predictions

- Data and endpoints
  - Coverage, conditionality, error, and predictivity

- Descriptors and models
  - Descriptors, machine learning, supervised vs. unsupervised methods

- Validation and application
  - Problems, “Questions to ask your friend, the modeller”

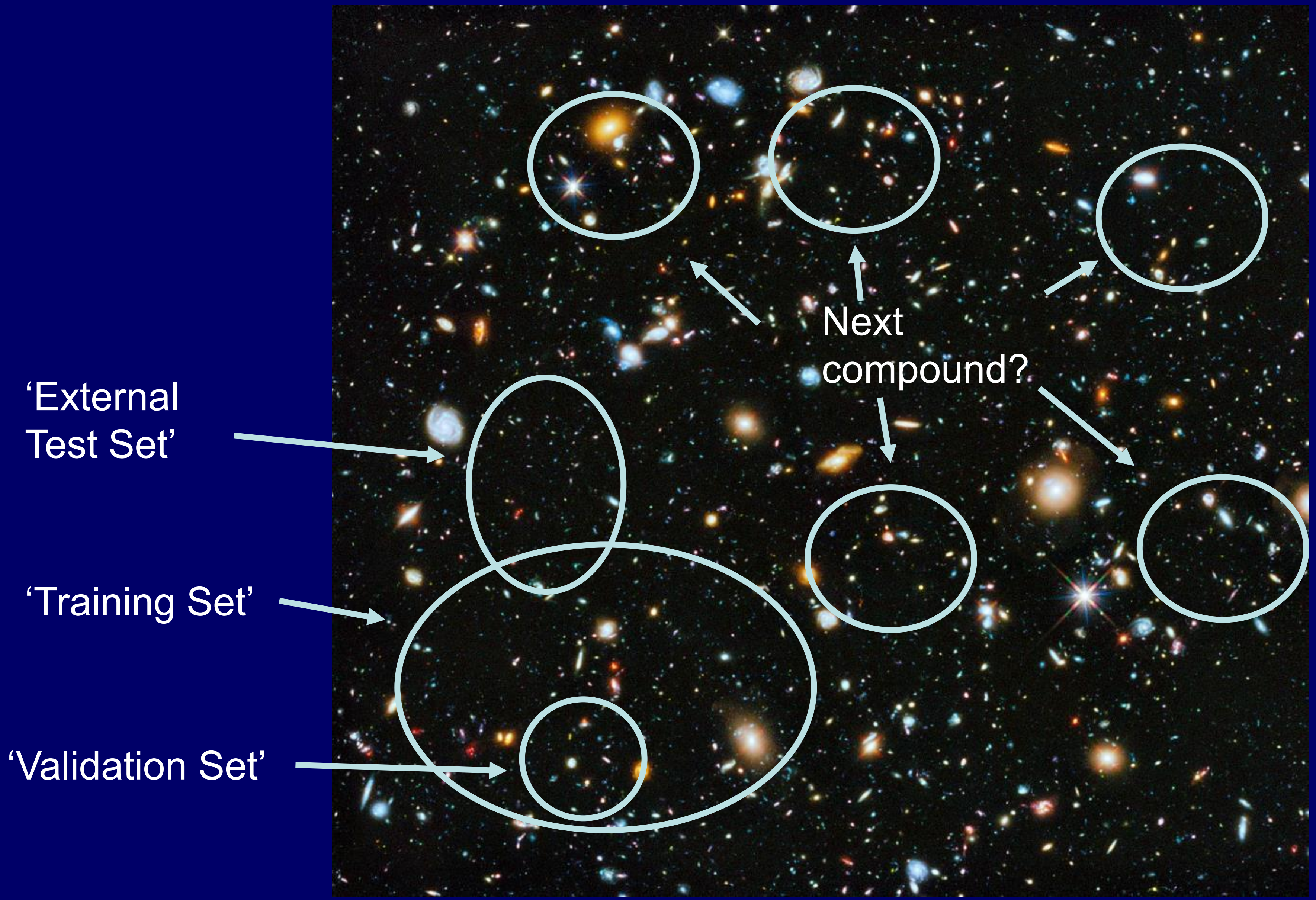
- Merging information from structure and -omics

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 with *in vivo* data –
    - Labels are either mostly only *in vitro*-relevant, or conditional (‘depend’ on dose, *etc*)
    - Validation is costly (e.g. phase II studies for efficacy; *plus controls*), *little prospective data*
    - **Difficult to sample distribution in chemistry/’project’ space well (diversity, number), so performance *depends heavily on test set***

- Retrospective validation is all we can do (but no prospective discovery, predictivity for future projects unknown, all behave differently)

Why ‘validation’ of a model is tricky: You get the numbers you want (depending on the question you ask/data set you use!)



‘External Test Set’

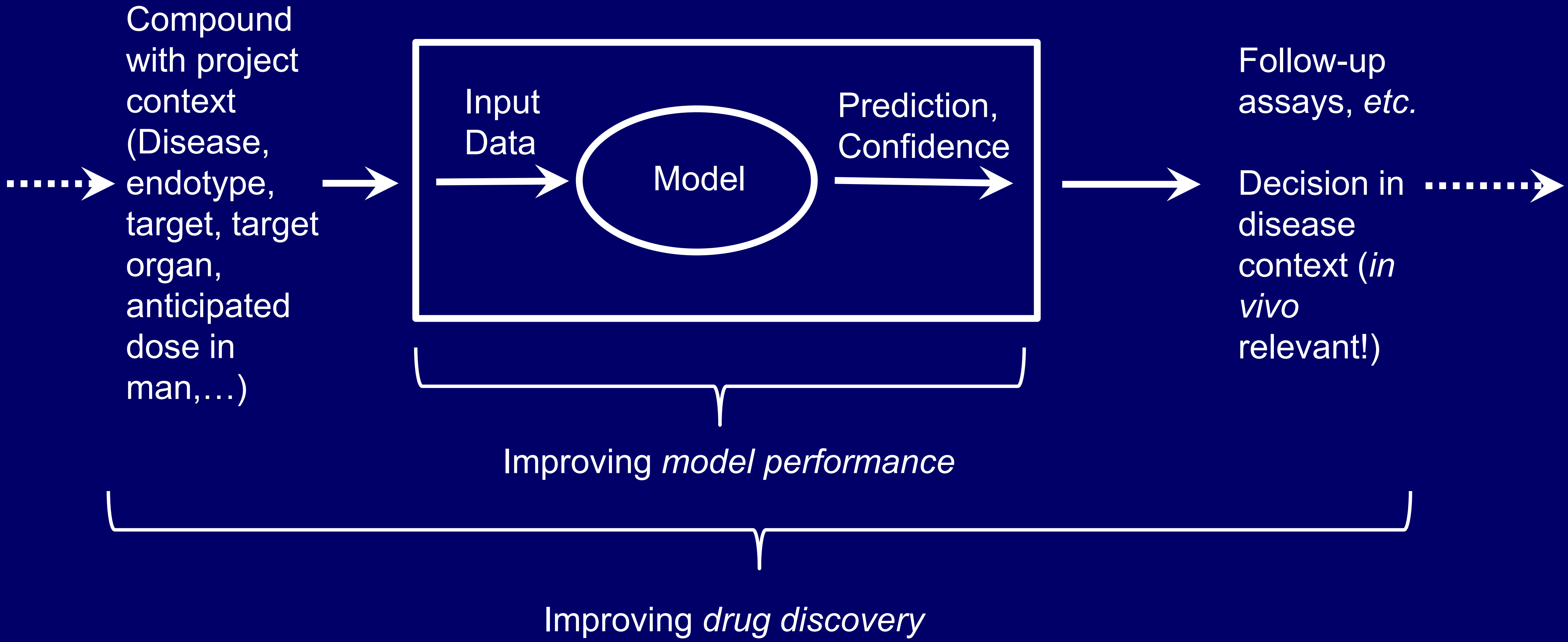
‘Training Set’

‘Validation Set’

Next compound?

- Chemical space is large; data sets are small
- Model is unable to generalize to unseen spaces
- Effect of changes is conditional on scaffold/context
- Sampling of data is generally insufficient
- *“Every model is a local model”*

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



Compound with project context (Disease, endotype, target, target organ, anticipated dose in man,…)

Input Data

Model

Prediction, Confidence

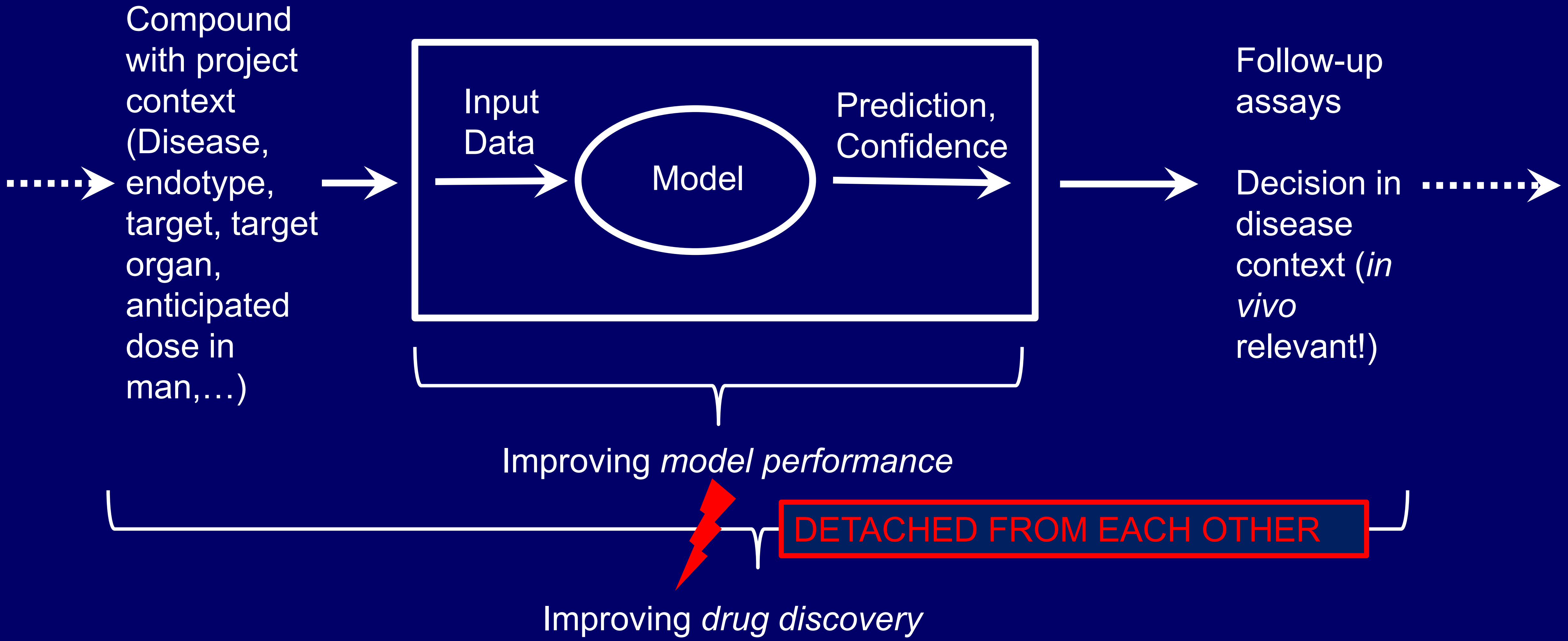
Follow-up assays, *etc.*

Decision in disease context (*in vivo* relevant!)

Improving *model performance*

Improving *drug discovery*

Using computational models for decision making often disappoints since (a) model validation is decoupled from process validation, and (b) many (most!) models use only proxy data (‘model of models’)



Compound with project context (Disease, endotype, target, target organ, anticipated dose in man,…)

Input Data

Model

Prediction, Confidence

Follow-up assays

Decision in disease context (*in vivo* relevant!)

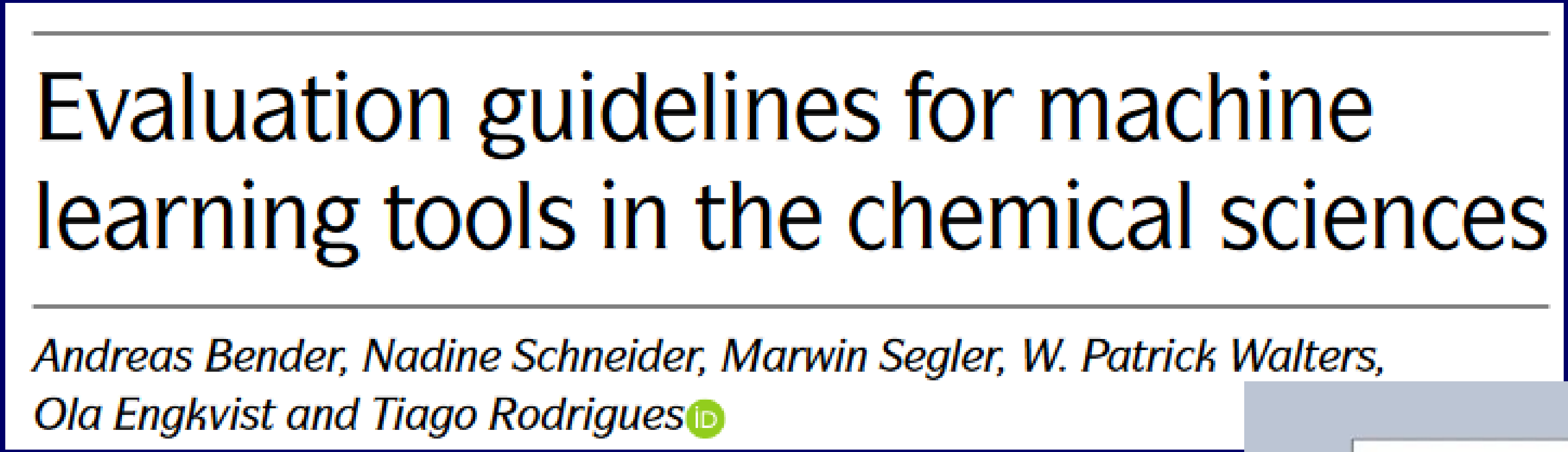
Improving *model performance*

DETACHED FROM EACH OTHER

Improving *drug discovery*

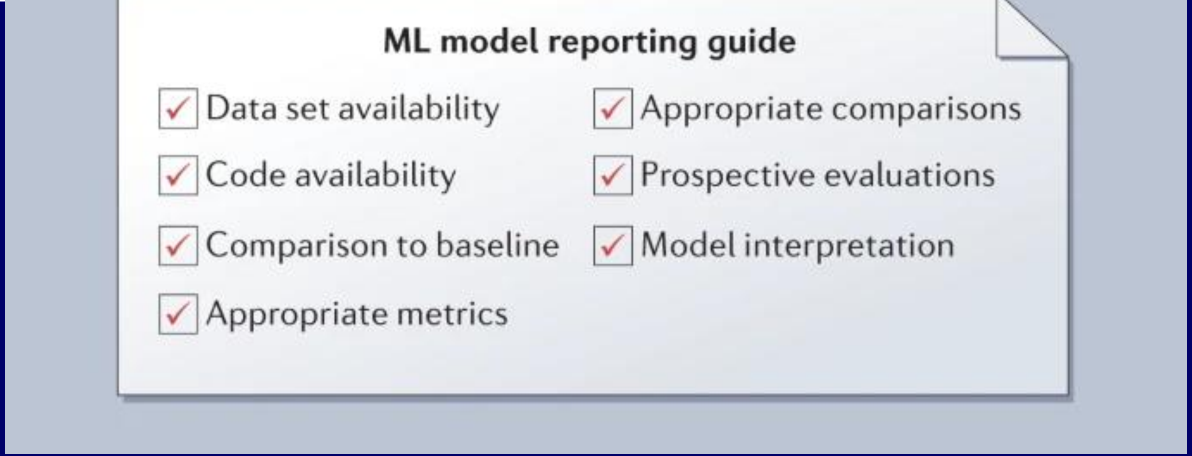
Model validation – two resources

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

On data, endpoints, models, and predictions

- Data and endpoints
  - Coverage, conditionality, error, and predictivity

- Descriptors and models
  - Descriptors, machine learning, supervised vs. unsupervised methods

- Validation and application
  - Problems, “Questions to ask your friend, the modeller”

- Merging information from structure and -omics

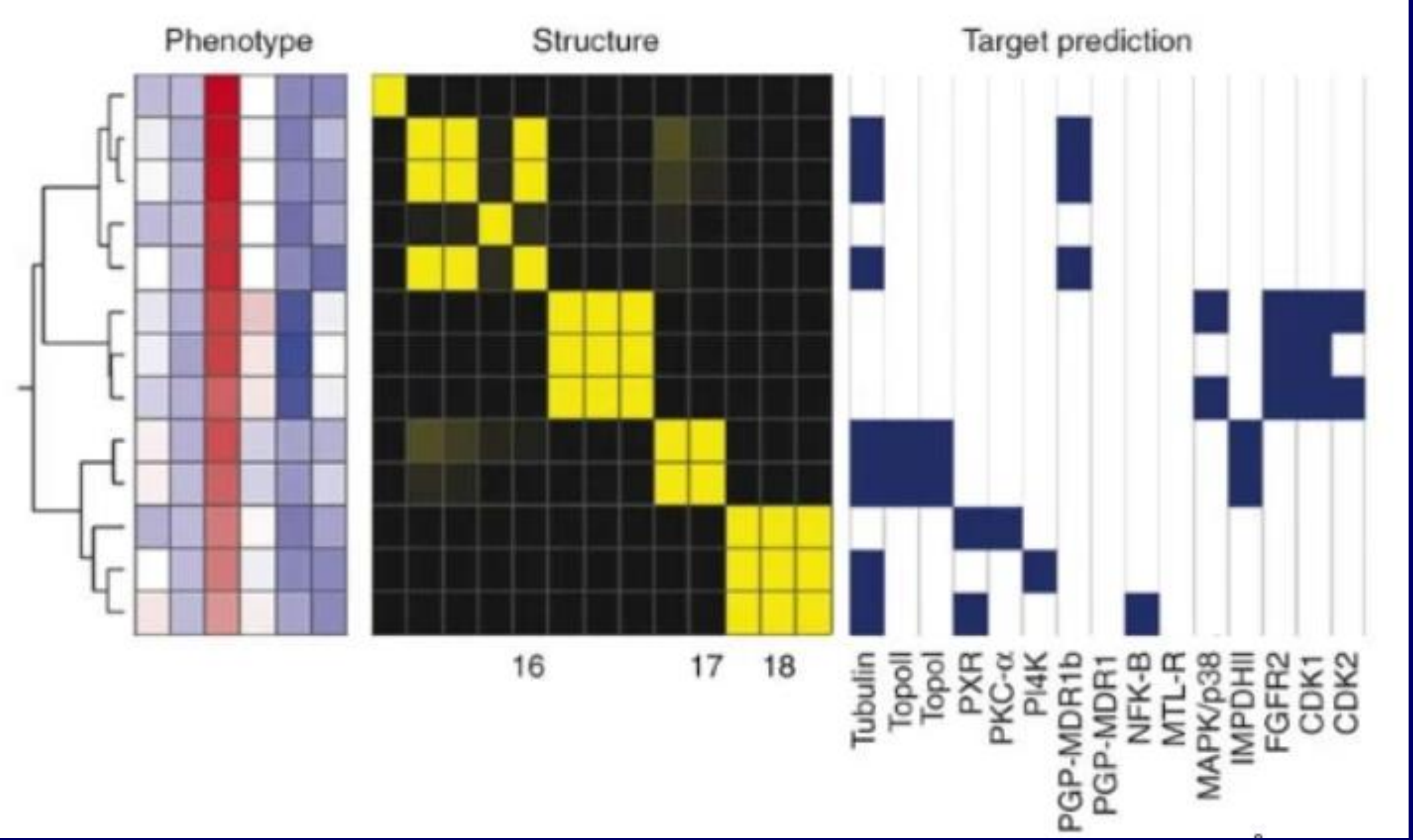
Why merging structure and omics data?

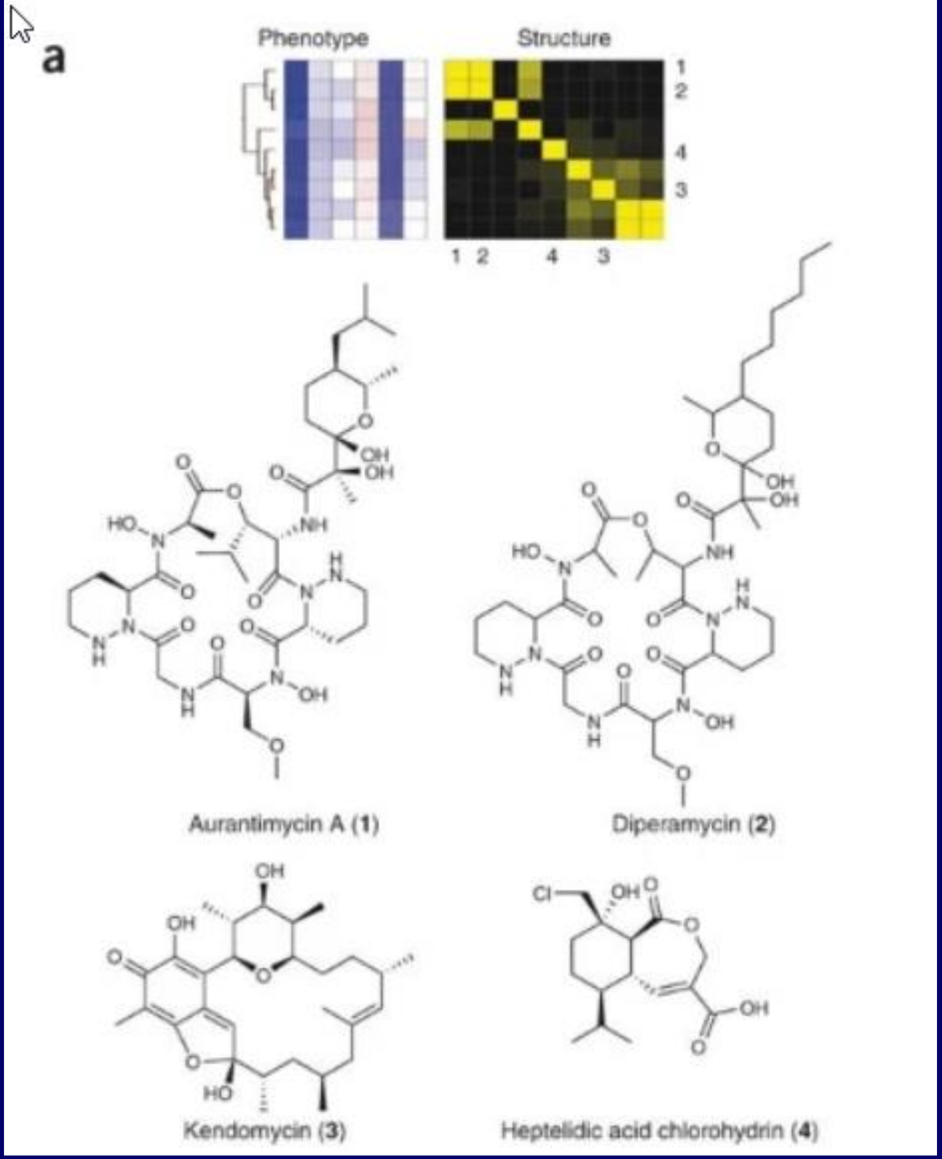
- In many (most?) cases we don’t understand how something works (i.e., biology)

- If we understand how something works we can do *hypothesis-driven, science-pull* driven data generation

- If we *don’t* understand how something works we need to revert to *hypothesis-free, technology-push* driven data generation and describe *variance*
- In this case we need *independent* pieces of information, and we need to *retro-fit* to what is *relevant*

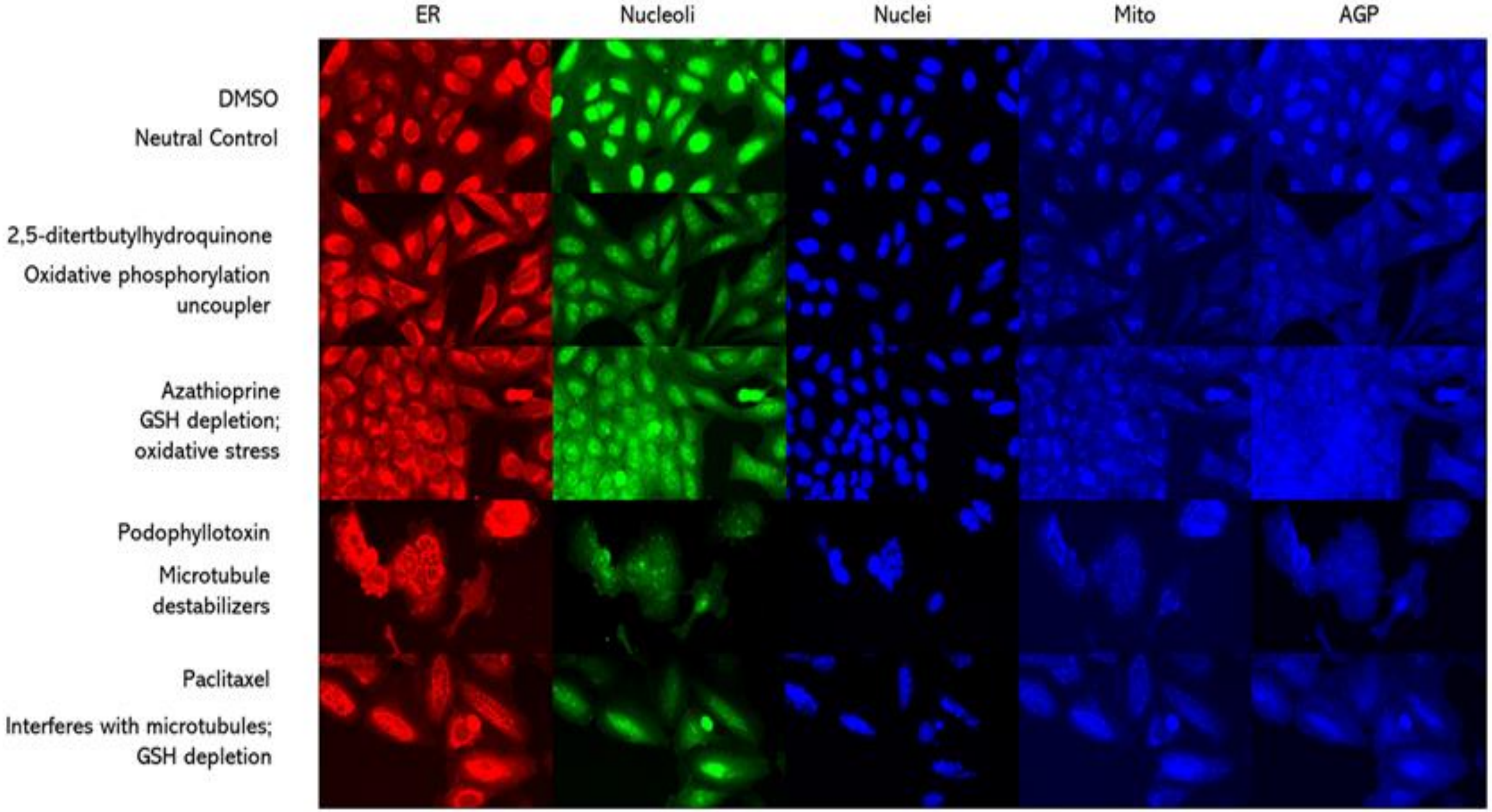
Why –omics, why cell morphology, … if we have the structure? They behave differently!





D. W. Young et al., Integrating high-content screening and ligand-target prediction to identify mechanism of action, Nature Chem. Biol. 2008

Cell Painting cell morphology assays:
Six stains/five channels/eight compartments


ER
Nucleoli
Nuclei
Mito
AGP
DMSO
Neutral Control
2,5-ditertbutylhydroquinone
Oxidative phosphorylation uncoupler
Azathioprine
GSH depletion; oxidative stress
Podophyllotoxin
Microtubule destabilizers
Paclitaxel
Interferes with microtubules; GSH depletion

Anika Liu and Srijit Seal et al

Dataset

Work by Srijit Seal

Training Dataset:

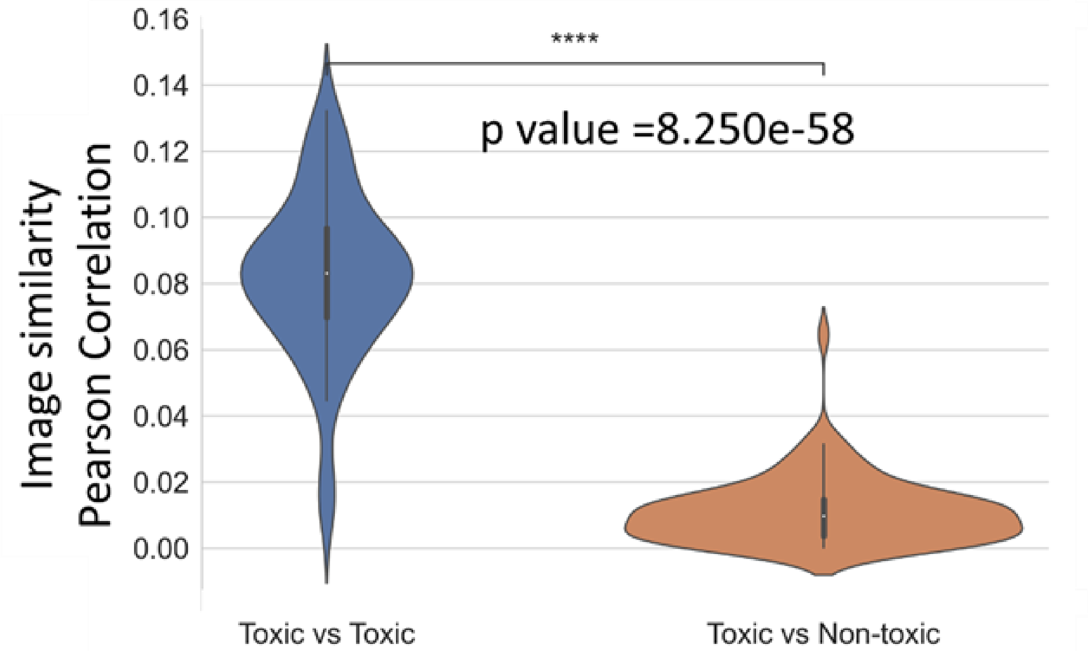
- Tox21 Mitochondrial membrane potential disruption assay hit calls (summary assay)
- 382 compounds
- 62 Mitotoxic

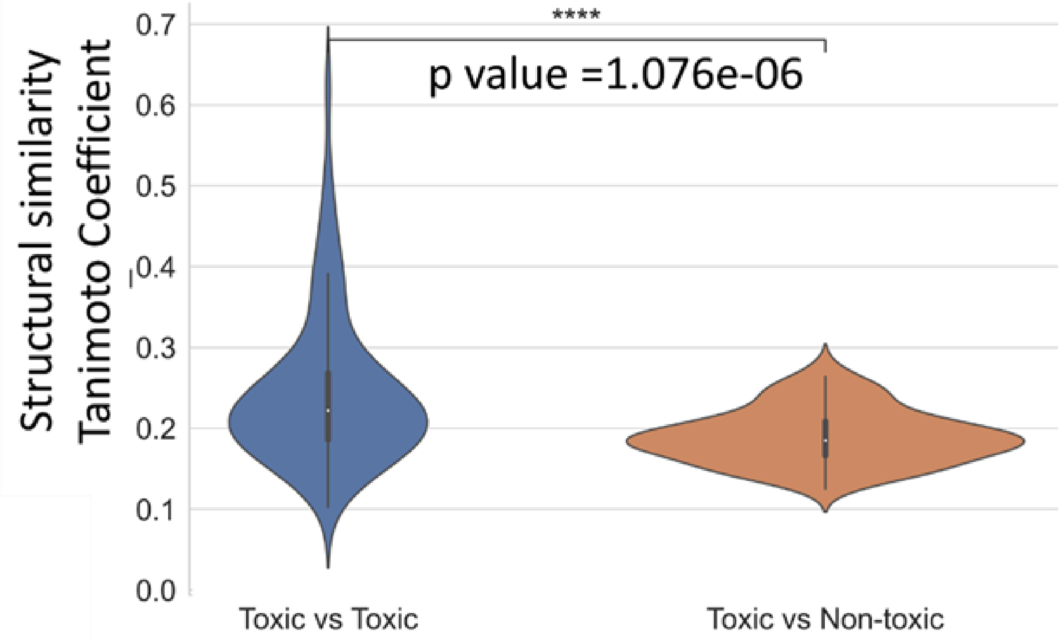
External Test:

- Additional mitotox assays from CHEMBL, PubChem, Mitotox Database relevant to mitochondrial potential
- 244 compounds
- 47 Mitotoxic

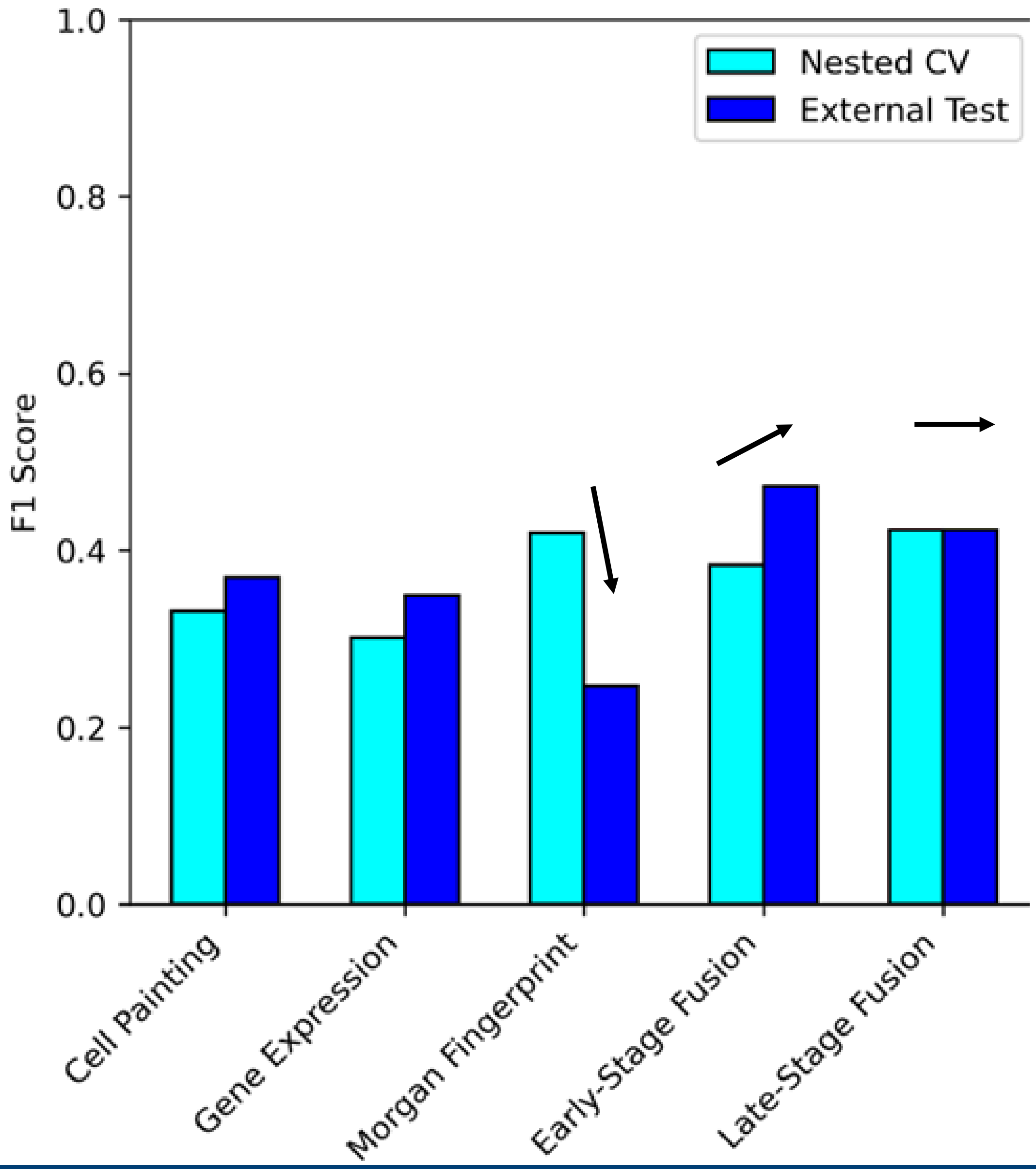
Hemmerich, J., Troger, F., Füzi, B. & F.Ecker, G. Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. Mol. Inform. 39, (2020)

Mitochondrially toxic compounds are more similar in morphology space than fingerprint space





S. Seal et al., Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. Comm Biol. 2022



Fusion models perform better on external test set

- External test set: F1 Score increases by 60% (0.25 to 0.42 in absolute terms) when using fusion models compared to Morgan fingerprints.

- Our method achieve higher sensitivity (0.79 in our study vs 0.37 in Apredica MitoMass) with comparable balanced accuracies (0.69 in our study vs 0.65 in Apredica MitoMass).

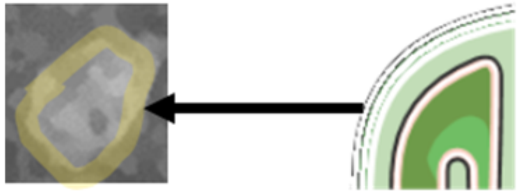
Hallinger, D. R., Lindsay, H. B., Friedman, K. P., Suarez, D. A. & Simmons, S. O. Respirometric screening and characterization of mitochondrial toxicants within the toxcast phase i and II chemical libraries. Toxicol. Sci. 176, 175–192 (2020)

Cell Painting features related to Mitotoxicity are generally interpretable (… but it’s high-dimensional, so not trivial in practice!)

Biological significance of Cell Painting features with respect to Mitochondrial Toxicity :

MITOCHONDRIAL FEATURES

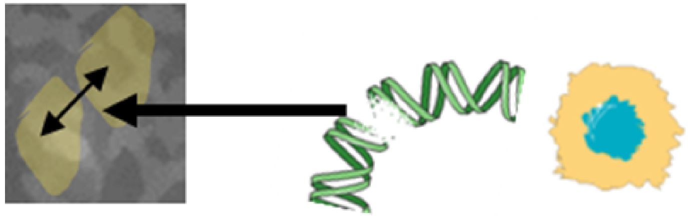
Cells Intensity MaxIntensityEdge Mito (PPV 0.83)



Edge of segmented object potentially indicates loss of membrane integrity

FEATURES FROM OTHER IMAGE CHANNELS

Cells Correlation Costes DNA AGP (PPV 0.52)



Potentially indicates DNA fragmentation and entering apoptosis or cell death

How to move beyond selecting and interpreting individual features in –omics data? Current research on Cell Painting readouts

- Cell morphological readouts contain information on several bioactivity endpoints

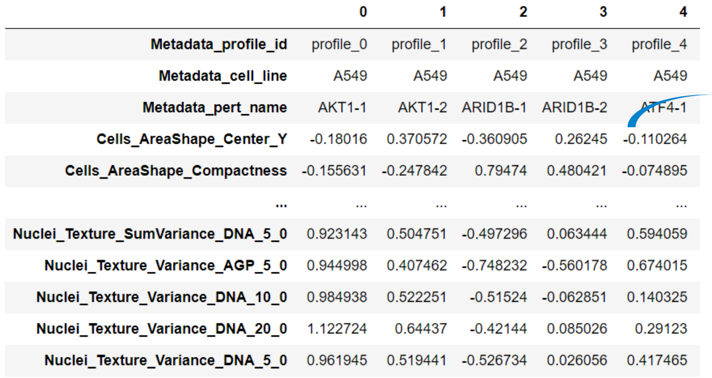
- Features are highly correlated – we *can* remove some of them, but then we lose biologically meaningful information

- We obtain here feature maps which group correlated features, which have importance for a particular endpoint

- We can obtain per-endpoint and per-compound importance heatmaps using Grad-CAM.

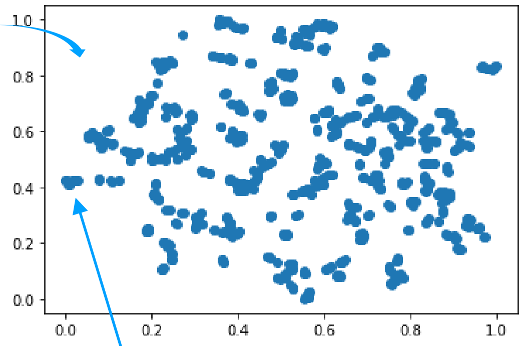
Method

1. Prepare Feature Map



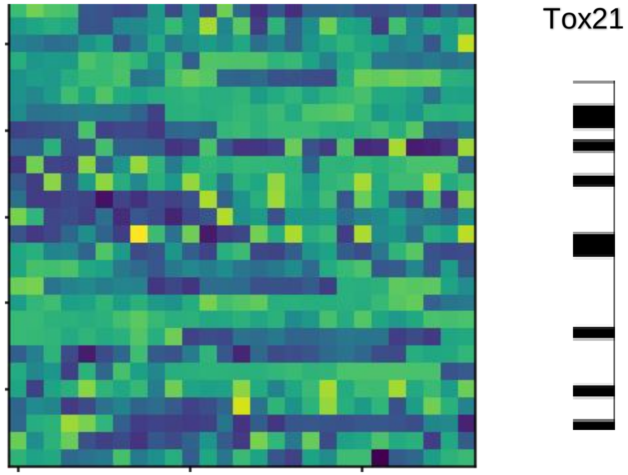
	0	1	2	3	4
Metadata_profile_id	profile_0	profile_1	profile_2	profile_3	profile_4
Metadata_cell_line	A549	A549	A549	A549	A549
Metadata_pert_name	AKT1-1	AKT1-2	ARID1B-1	ARID1B-2	ATF4-1
Cells_AreaShape_Center_Y	-0.18016	0.370572	-0.360905	0.26245	-0.110264
Cells_AreaShape_Compactness	-0.155631	-0.247842	0.79474	0.480421	-0.074895
...	...	...	...	...	...
Nuclei_Texture_SumVariance_DNA_5_0	0.923143	0.504751	-0.497296	0.063444	0.594059
Nuclei_Texture_Variance_AGP_5_0	0.944998	0.407462	-0.748232	-0.560178	0.674015
Nuclei_Texture_Variance_DNA_10_0	0.984938	0.522251	-0.51524	-0.062851	0.140325
Nuclei_Texture_Variance_DNA_20_0	1.122724	0.64437	-0.42144	0.085026	0.29123
Nuclei_Texture_Variance_DNA_5_0	0.961945	0.519441	-0.526734	0.026056	0.417465

t-SNE of Feature Map



Cell_Texture_SumAvergage_AGP_10_0

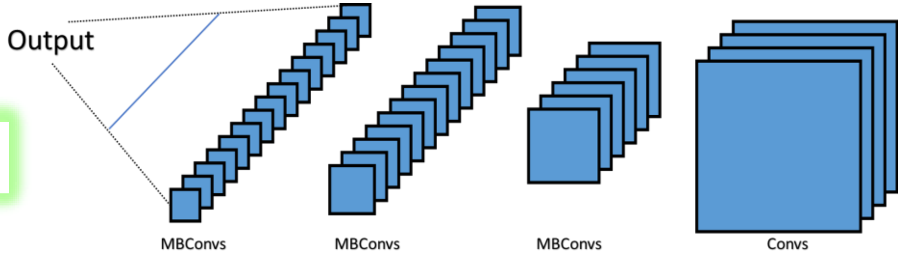
Jonker-Volgenant algorithm



Tox21 Assays

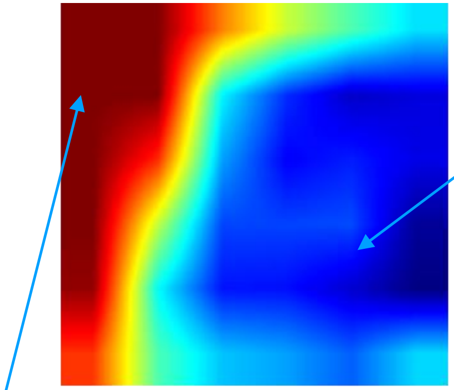
e.g. ER Stress

2. Predict Endpoint of test set



Model: EfficientNet B0

3. Interpretation using Grad-CAM



More important to model

Less contributing to model

UNIVERSITY OF CAMBRIDGE

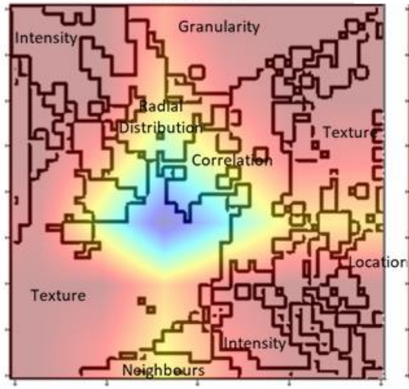
@srijitseal

“The universe of toxic endpoints in cell painting feature space”

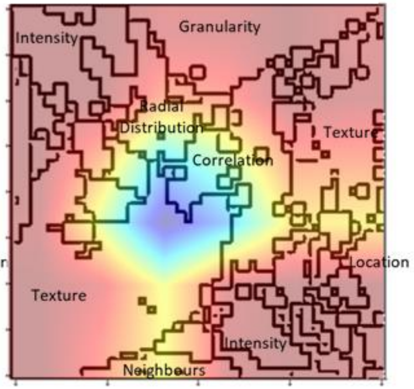
For models predicting proliferation decrease endpoint:

mitochondrial toxicity

Niclosamide

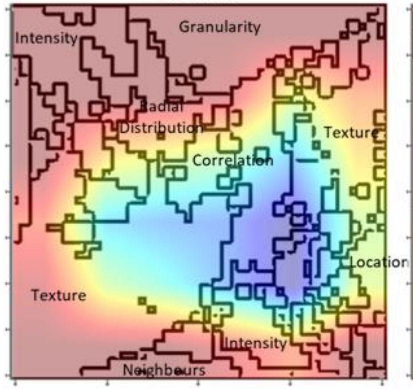


Colchicine

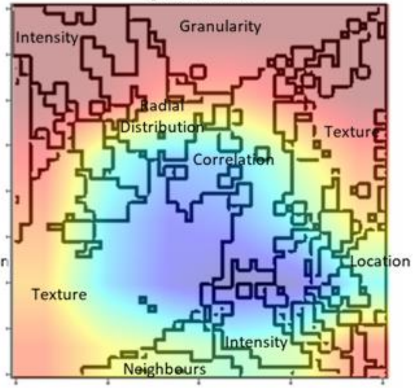


inhibits cell proliferation in the G(1) phase of the cell cycle

Lovastatin

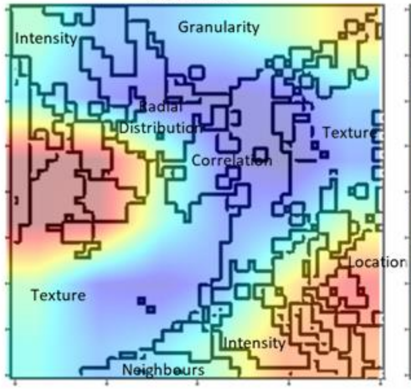


Cycloheximide

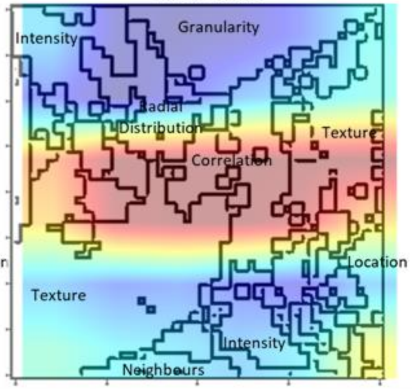


antiandrogens

Nilutamide



Mifepristone



Conclusions

- Life science data is difficult to label, and hence to model
- ‘Big data’ is good, but heterogeneous data makes quantitative decisions often difficult

- *Descriptors all behave differently!*
- We need to either have a reason to select one (unsupervised methods), or retro-fit/learn which features are important, during model generation (supervised methods)

- Merging different types of descriptors generally gives you different pieces of information – but you need to know/learn from the data what matters, more is not always better (!)

‘*In Silico* modelling for dummies’ session organized by the British Toxicology Society

- In November 2022
- 2 Hour session – Background, and seminar on ‘how to build your own models’

- Mail me if you are interested and I will keep you posted: ab454@cam.ac.uk

Thank you for listening!
Any questions?

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